



38189

MILE

D3.3 Metadata Search & Retrieval Seminar 3 & Report

www.mileproject.eu

Deliverable number/name	<i>D3.3 Metadata Search & Retrieval Seminar 3 and Report</i>
Dissemination level	<i>Public</i>
Delivery date	<i>April 2009</i>
Status	<i>Final</i>
Author(s)	<i>Jessica Tier, Project Manager Lucy Geering, Project Administrator Celestine Bramley, Project Assistant</i>



eContentplus

This project is funded under the *eContentplus* programme¹, a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

¹ OJ L 79, 24.3.2005, p. 1.



Contents

Section title	Page No.
Conference Attendance	3
Introduction	5
<i>Towards a Multilingual Future: Report</i>	6
Conclusions	14



MILE Metadata Search & Retrieval and Translation 3 Conference Attendance: Towards a Multilingual Future

9.30am – 5.00pm, Friday 3rd April 2009
Science Gallery, Trinity College, Pearse Street, Dublin 2, Ireland

1	BAPLA	Sarah Saunders	Chair of BAPLA Metadata Group
2	BAPLA	Susanne Kittlinger	Marketing Manager
3	Bildombudsmannen	Staffan Teste	Legal Advisor
4	Bridgeman Berlin	Annette Godefroid	General Manager
5	CACAO / Xerox Research Centre Europe	Frederique Segond	
6	Dublin Institute of Technology	Yvonne Desmond	Delegate
7	Fotofinder	Agnes Folaji	Account Manager
8	Heritage Council	Lesley-Ann Hayden	The Museums Standards Programme Coordinator
9	Imprezzeo	Kurt Dressel	Director
10	Irish Copyright Licensing Agency	Samantha Holman	Executive Director
11	IVARO	Alex Davis	Delegate
12	MILE	Celestine Bramley	MILE Project Assistant
15	MILE	Jessica Tier	MILE Project Manager
13	MILE	Lucy Geering	MILE Project Administrator
14	National Gallery of Ireland	Louise Morgan	Delegate
17	National Gallery of Ireland	Marie McFeely	Delegate
16	National Maritime Museum	Douglas McCarthy	Picture Library Manager
18	plainpicture	Uta Kaack	Knowledge Organisation
19	RTÉ Stills Library	Emma Keogh	Delegate
20	SILVER	Brian Kavanagh	Systems Operator
21	SILVER	Stella Dextre Clarke	Information Consultant
22	Swiss Federal Office of Culture	Genevieve Clavel-Merrin	National and International Cooperation
23	System Simulation	Graham Howard	Design Director
25	The Bridgeman Art Library	Harriet Bridgeman	Chairman
26	The Bridgeman Art Library	Pandora Mather-Lees	SILVER Project Manager
27	The Chester Beatty Library	Sinéad Ward	Rights and Reproductions

D3.3 Metadata Search & Retrieval Seminar 3 & Report



28	The Kenny Gallery	Mick O' Dea	Delegate
29	Trinity College Dublin	Amelie Dorn	Student
37	Trinity College Dublin	Ben Steichen	Delegate
38	Trinity College Dublin	Bo Fu	G.31 O' Reilly Institute
43	Trinity College Dublin	Catherine Giltrap	Curator of the College Art Collections
46	Trinity College Dublin	Daniel King	Delegate
41	Trinity College Dublin	Dr. Rachel Moss	Irish Art Research Centre
47	Trinity College Dublin	Frances Pities	Delegate
30	Trinity College Dublin	Garret McMahon	Institutional Repository Content Manager
35	Trinity College Dublin	Gary Baugh	Delegate
40	Trinity College Dublin	Kangyu Pan	PhD Image Processing
42	Trinity College Dublin	Killian Levacher	Knowledge & Data Engineering
45	Trinity College Dublin	Mercedes Blanco	Delegate
48	Trinity College Dublin	Mohammed Ahmed	Delegate
31	Trinity College Dublin	Niamh Brennan	Programme Manager
33	Trinity College Dublin	Niamh Harte	Assistant Librarian
34	Trinity College Dublin	Pat Carty	IT Coordinator
32	Trinity College Dublin	Professor Anil Kokoram	Associate Professor
39	Trinity College Dublin	Professor Frank Boland	
36	Trinity College Dublin	Rami Ghorab	PhD Research Student
44	Trinity College Dublin	Raya Kakaldehy	Student
50	Trinity College Dublin	Soha Maad	Delegate
51	Trinity College Library, Dublin	Cliona Nishuilleabhaim	Librarian
49	Trinity College Library, Dublin	Esther Walsh	Assistant Librarian
54	Trinity College Library, Dublin	Jane Lawson	Metadata Librarian
53	University College Dublin	Craig Berry	Student
52	University College Dublin	Damian Kelly	Student
55	University College Dublin	David Corrigan	Student
56	University College Dublin	Vincent Wade	Delegate
57	University of California	Gregory Reser	Delegate
58	Unworkable	Dan Ring	Delegate
59		Emma O' Donoghue	Delegate
60		James Saylor	Director



Introduction

The final seminar for Work Package 3, Metadata Search and Retrieval, *Towards a Multilingual Future*, took place at Trinity College, Dublin, Ireland, on the 3rd April, 2009.

Towards a Multilingual Future is the final in the search and retrieval work package (D3.3), focusing on search systems and metadata translation systems. In a networked world, increasing pressure for access to digital image archives information without language or cultural barriers means that there is a strong demand to be able to find information in foreign languages, to be able to read and interpret that information and merge it with information in other languages. This seminar invited participation from experts involved in the development of metadata standards and multilingual information retrieval systems including experts from European and EC-funded projects such as Dr. Gareth Jones of MultiMatch, Dr. Frédérique Segond of CACAO and Genevieve Clavel-Merrin of MACS, in order to extend MILE's reach across other metadata projects to find out how these have tackled the issue of multilingual access, and to facilitate improved European and international access. Agnes Folaji of Fotofinder, a commercial image database, presented a commercial argument for multilingual translation. Kurt Dressel of Imprezeeo presented image recognition technology as an alternative way to deal with image search and retrieval, and one which would partially obviate the need for multilingual translation. This seminar was also used as an opportunity to formally present the findings of MILE's previous Metadata Classification conference workshop – *A Picture's Worth a Thousand Euros* – by seeking industry agreement for a proposed set of core fields in metadata standards.

MILE's first search and retrieval seminar, *Successful Searching? Lost in Translation*, focused in the discussion and evaluation of inhibiting factors to metadata search systems and translation systems, such as poorly defined and inconsistent metadata and complex ontologies. This seminar gave an overview of various in-house solutions, such as metadata hierarchies, ontologies, translation methodologies and in-house Unique Identifiers (UID's). Tools available for translation schemes, including technical formats such as XLIFF (XML Localisation Interchange File Format), manual translation, automated translation systems such as Wordnet, and thesauri were given a broad introduction. Standards were discussed in the context of translation, and questions were asked as to which standard/s would best support translation systems.

The second search and retrieval seminar, *Speaking in Tongues*, concentrated on multilingual facilities, focusing on methods to overcome the problems impeding multilingual search and retrieval of images. A variety of presentations ranged from academic research projects, ISO standard developments, the provision of completed thesauri translation projects and new technological developments in the form of image recognition software.

60 Attendees attended *Towards a Multilingual Future*, with strong international representation from industry professionals, students and tutors. There was also a significant number of attendees who did not have English as their first language; an improvement on the previous S&R and Translation seminar, and appropriate for a seminar focusing on multilingual search and translation systems.



Towards a Multilingual Future: Report

Harriet Bridgeman, Chairman of The Bridgeman Art Library, introduced the day welcoming attendees, followed by Jessica Tier, The MILE Project Manager, who gave an overview of MILE's aims and a brief description of multilingual access.

Stella Dextre Clarke opened the presentations with an introduction to thesaurus standards and issues in the context of image retrieval. Multilingual information retrieval languages exist in different forms, e.g. subject headings lists, thesauri, enumerated classifications, analytico-synthetic classifications, etc. In information technology, a thesaurus represents a database, or list of semantically orthogonal topical search keys – a unique vocabulary defining different kinds of terms and relationships. Dextre Clarke used her extensive knowledge and experience in the development and use of metadata standards and vocabularies to support the argument for using thesauri, stressing that

“...once you have a thesauri it is there for good, and you can use it over and over again”.

Thesauri, such as AAT (Art and Architecture Thesaurus), do exist, but it is unclear whether they are generally suitable for multilingual image retrieval. Furthermore, tangible economic arguments for thesauri were not discussed. Future recommendations should include further investigation into specific uses for the development of thesauri, and consideration of cost and editorial control. Dextre Clarke spoke of the importance of adhering to a set of standards to guide thesauri development and advised the use of existing ISO standards. These are currently *ISO 2788:1986, Monolingual Thesauri* and *ISO 5964:1985, Multilingual Thesauri*. The working group BS 8723:2005/2007, of which Dextre Clarke is a member, is developing *ISO 25964: Monolingual & Multilingual Thesauri*, which is expected 2010+. Dextre Clarke also warned against developing new standards due to the considerable length of time needed to achieve a workable result, and in addition suggested that the necessity for developing new standards is obviated by the flexibility that can be achieved within an existing standard.

She went on to give a brief history of standards and thesauri, pointing out that thesauri were originally developed for text – even the AAT was developed with text in mind. As text-based searching gives way to images, Dextre Clarke illustrated a number of potential issues, suggesting that “recall” might be more relevant to image searching than a precision approach – a problem in respect of understanding needs; that unambiguous terms facilitate better image retrieval – a problem for standards; that standards advice with regards to synonyms, adjectives and verbs can result in confusion when applied to image searching; and that finding equivalence across languages is a semantic problem pertaining to equivalence relations between terms used as preferred and non-preferred terms in information retrieval languages. Equivalence relations exist not only within each separate language involved, but also between the languages. Additional problems pertaining to semantics involve the scope, form and choice of thesaurus terms. Dextre Clarke questioned whether there is a need for an extensive hierarchical system, suggesting that reducing the facets to related terms only might be more appropriate. These points need further discussion and decision if they are to usefully inform thesaurus construction for image search and retrieval.



D3.3 Metadata Search & Retrieval Seminar 3 & Report

Questions from the audience included Sarah Saunders asking “whether there should be a thesaurus dedicated to image retrieval?” to which Dextre Clarke said no “because in a networked world the industry would want to bring images and text together - furthermore, de-facto standards for vocabularies and construction exist aiding a networked environment”.

Genevieve Clavel-Merrin, from the Swiss National Library, followed Dextre Clarke with a presentation on multilingual subject mapping and search in printed documents. Key problems for multilingual access in networked environments are cost, maintenance, the many different types of controlled subject vocabularies used for access to resources in various networks (subject headings, thesauri, classification schemes and ontologies) and in different languages, subject queries across databases or networks limited by a heterogeneous language environment, and a lack of interoperability between subject indexing tools limiting access and use of libraries’ catalogues and databases. Clavel-Merrin described a number of projects and initiatives exploring access through various types of subject indexing tools and thesauri and gave an overview of the development of MACS, a system that permits users to subject search library catalogues in the language of their choice (English, French, German), via widespread, Linked Subject Heading languages (SHLs).

Libraries and image archives tend to index or tag their content through using carefully selected lists of words or phrases such controlled vocabularies, collectively identified as thesauri. Within thesauri, subject headings are used to facilitate access to individual items pertaining to similar subject matter. Linking of subject facilitates interoperability, allowing both indexers and searchers to continue to use the same subject heading language as before, and is done by mapping (manual) headings of subject heading languages. Prior requirements include all documents to be properly indexed, using a preferred vocabulary. As MILE has reported previously, consistency and accuracy are major issues for indexing and for metadata per-se. Linking is a de-centralised way of working whereby each partner works from their own SHL, used as a source language. MACS used existing lists to bring subjects together; the approach was to add value to existing metadata instead of creating new data, i.e. a solution not based on translation but “equivalences”. Equivalences are understood as terms of equal value, corresponding or as having the same meaning or result. The result can be displayed as a table with as many columns as subject heading languages involved. In each row of the table the “equivalent” terms of the different subject heading languages are given. A cell of the table can contain zero, one or more terms.

Politics in the development of MACS dictated parity for linking SHL’s, i.e. no single language source. Time cost and size are significant issues for this type of approach in multilingual development (see D3.2 for further comment).

MACS conformed to linking rules and standards *BS 8723-4:2007* and *ANS/NISO Z39.19-2005*, usefully demonstrating that some standards allow retrospective application. ISO protocols currently assert that in a multilingual thesaurus, to truly conform to ISO standards, the source language acts as a spine on which to hang all other languages – implying prospective use only. Decisions as to whether this criterion is necessary will need to be ascertained if unnecessary limitations are to be avoided.



D3.3 Metadata Search & Retrieval Seminar 3 & Report

MACS has reached an operational phase of production, but text based multilingual interoperability projects focusing on subject headings have not been perfected. Issues in the development of MACS include methodological constraints in linking subject headings; the availability of similarly structured resources; time, labour, and long-term commitment and search interface. Furthermore, because partial equivalences exist, users do not all get the same result, but MACS hopes the results will be relevant and continues to test, refine and add to the facility.

Next Dr. Frédérique Segond (Xerox European Research Centre) presented CACAO, an EC- funded project providing cross-language access to Online Libraries and Online Public Access Catalogue content, in a single translation step. Segond described metadata search (text) and multilingual search as two distinct issues for multilingual retrieval. She gave an overview of the semantic tools needed for monolingual search and retrieval (analyser, stopWords identifier and query expander e.g. thesauri) and spoke of the importance of indexing metadata to achieve successful document retrieval. It appears, from comments made here and in previous presentations, the indexing is key to multilingual search and retrieval, whether mapping is applied or not. Segond explained that indexing text is also possible, to give more weight to certain metadata (e.g. to boost generic fields), and that metadata often contains free text that can be used for better indexing. CACAO sees one of its strengths as its ability to offer extended indexing - this can include abstract, description or even the whole document when free or full text is available; extended indexing allows CACAO to thoroughly analyse text.

Segond went on to say that in addition to the above, multilingual searches require a query translator such as a thesaurus, to disambiguate translation. Viable projects such as CACAO provide further support for the use of thesauri in multilingual translation. CACAO's architecture facilitates analysis of online catalogue content using natural language processing technologies; indexing, using information contained in unstructured fields; expanding and translating queries using various multilingual sources and technologies. CACAO uses the OAI Harvester (**see D2.3** for further explanation) and is supported by standards. Suppliers must conform to the following standards: Dublin Core, Qualified Dublin Core and CACAO Application Profile. Issues for CACAO included time and cost and as with MACS, CACAO is constantly trying to find solutions to providing a user friendly interface. Audience members suggested that having a set of core fields would improve access and that they might be applicable for branches of culture.

Following on from Frederique Segond was a panel discussion. Moderated by Jessica Tier and with a panel consisting of Sarah Saunders, Annette Godefroid, Agnes Folaji and Niamh Brennan, the general goal of this session was to thrash out the underlying necessity for the core fields arrived at in the previous MILE seminar, with focused input from the experts and contributions from the audience.

The ultimate aim of MILE's Search & Retrieval and Translation work package is to recommend specific action for Research and Development (R&D) in cost-effective S&R and translation systems. At MILE's previous seminar *A Picture's Worth a Thousand Euros*, two agreements were reached. Firstly, the acceptance that each user needs their own internal metadata schema, and to therefore provide for this by producing a mapping between the key schemas. A preliminary mapping was created during a workshop in this seminar. Secondly, all agreed that a set of 'core fields' for cross-industry use was



D3.3 Metadata Search & Retrieval Seminar 3 & Report

also key to establishing achievable, cost and time-effective metadata classification standardisation. These twenty proposed core fields are:

For the digital image:

- 1) Supplier picture number
- 2) Headline
- 3) Photographer/ creator**
- 4) Copyright notice**
- 5) Credit**
- 6) Date created**
- 7) Licensing contract
- 8) Instructions
- 9) Rights usage

And for the work of art depicted in the image:

- 10) Title**
- 11) Artist/ creator**
- 12) Subject/ keywords**
- 13) Description
- 14) Date created**
- 15) Medium/ format
- 16) Rights**
- 17) Nationality
- 18) Location
- 19) ID Number
- 20) Size of the artwork.

This set of twenty fields could be condensed further into nine fields (highlighted in bold above) – four for the digital image, five for the work of art depicted in the image.

At this session a core set was generally agreed to be useful not only for image collections, museums, art galleries and the cultural heritage sector for their cataloguing and classification purposes, but also for end users such as publishers and news media, who don't need the rich information and instead could use a base set such as this as for ensuring copyright compliance with regard to credit line structure. Further examination of the structure of a credit line will be one focus of MILE's final seminar, *Know Your Rights*. PLUS's management system was cited by Godefroid as an example of exclusive metadata – it is not publicly viewable as it includes price structuring. This is something to bear in mind.

MILE's final report will define the community served, build around some use cases and define the role of multilingual vocabularies. In this report, MILE will set out the principles of metadata management to include role of internal vs. external schemas, with guidance on mapping and outlining the available options to achieve multilingual access, e.g. automatic translation. Saunders was keen to point out that the preliminary mapping was taken from IPTC data.

There was one dissident in the audience, who believed that multiple fields were necessary for users to choose which of these they used and that IPTC should be



D3.3 Metadata Search & Retrieval Seminar 3 & Report

recommended as the standard for use throughout the industry. However, one crucial factor in MILE's core set is to simplify the workflow process with a straightforward set of fields which can be used both with existing metadata – to ensure that this is compliant – and to create new metadata. There are already existing standards which include a multiplicity of fields, such as CDWA Lite and IPTC Extended, and as discussed in *A Picture's Worth A Thousand Euros*, an overload of fields is deemed to be a weakness rather than a strength for image retrieval. Brennan noted that since the metadata MILE's stakeholders are concerned with is for purposes of access, a minimal set of fields would be most useful, enabling image archives to easily map between one standard and another to achieve interoperability. This was also echoed by Saunders and Dextre Clarke. It was therefore decided that technical and administrative metadata could be created and updated by individual institutions, alongside the standardised metadata. In answer to the request for more fields, Saunders suggested that the museum sector could decide for itself how much info to carry in the files, and suggested that a downloadable template, as an XML panel for installation in PhotoShop, is a feasible option for adding more fields. MILE's focus has to be the end user who wants their metadata simple and manageable, therefore the core set should be kept small.

More tangible results to come from this workshop session were that Trinity College offered to act as a Use Case as regards the core fields upon which to build the final report. Saunders suggested if there IS a recommendation on core fields to improve monolingual and multilingual access, it should be developed as an application profile on existing schemas, which prefers the closest existing schema but draw on others too. Another key factor is to ensure that each field has an unambiguous field name and a definition which coincides with existing schemas, an element which can be asserted through a defined mapping. The next MILE mapping should include as many relevant schemas as possible, e.g. Dublin Core and IEEE-LOM, to ensure as wide an application and take-up as possible. Finally, it was generally thought sensible that if MILE resources do not stretch, the recommendation of a research & development project to address the rest of the work was an appropriate result.

Dr. Gareth Jones of Dublin City University (DCU) followed Segond giving an overview of Multimatch, a project providing enhanced multilingual access to a multimedia collection of cultural heritage objects (text, images, sound and vision), translating documents to English and storing them in a single English index. MILE's network partner Alinari were partners in MultiMatch, along with DCU. MultiMatch's aim was to specifically develop a system for multilingual multimedia search for the cultural heritage domain's digital content. MultiMatch was formed to tackle the abundance of fragmented cultural heritage information by developing a search engine described as 'intelligent access'. This system provides focused, enriched access to multiple digital cultural heritage objects, which helps cultural heritage organisations by raising their profile and disseminating their content more widely.

Multimatch uses query translation for the search query and machine translation and language indexing to support multilingual access. The project is not limited to trusted sources but provides access to all relevant material on the web, regardless of either the source language or the target language. The project uses three search approaches - metadata search e.g. content source, dates, authors - visual search and textual search.



D3.3 Metadata Search & Retrieval Seminar 3 & Report

Text, speech and video indexing are used as the underlying platform. Multimatch facilitates automatic classification of results; automatic extraction of relevant information which is then used to create cross-links between related material, such as the biography of the artist, exhibitions of his/her work, critical analyses etc; the organization and further analysis of material.

Standard text, speech and video indexing tools formed the underlying platform, and domain specific methods were developed to support improved multilingual studies. It uses a mixture of S&R methods. A low-level visual search is combined with a text search, using a Lucene plug-in, the results of which are constrained by a metadata search of features such as content source, dates and authors. It focused on four languages – English, Dutch, Italian and Spanish – and then uses a combination of S&R tools for the text and metadata searches. This is machine-based translation and hybrid translation, which uses domain-specific dictionaries taken from the web. Dr. Jones demonstrated the process of creating a dictionary from the web, and then how this works to translate the text. Its three-stage process aims to iron out any errors thrown up by the machine translation.

Dr. Jones explained that although machine translation is able to provide reasonable translations for general terms, it is not sufficient for domain specific terms in particular, multiple word phrases (personal names, titles of art works, etc.). To overcome this problem Multimatch has developed an automatic process to build a domain-specific phrase dictionary (sources mined from the web) to support machine translation and to improve translation accuracy of phrases previously un-translated, or inappropriately translated. Comments from the audience included Stella Dextre Clarke's observations that translating a query and a document separately could potentially introduce corruption, but experiment shows that it gives better precision. Concerns were raised with regards to the volume of content and storage. Jones responded, saying that this was not a problem. Agnes Folaji explained that Fotofinder made phrase dictionaries and found that they were noisy and time consuming. Folaji said that ideally content would need manual checking.

The second prototype is now available through the Multimatch website; it requires a simple registration process for access, and then the current system is useable. The results are shown alongside the ranked related terms and the different sectors which the results are drawn from, such as web, creators, archives, image and video. The advanced search for refining the results has four extra fields – All fields, Keyword, Title and Location. Initially these seem a little simplistic for an 'advanced' version of the ubiquitous search tool, but the 'all fields' field is actually very useful and does instead of multiple fields such as 'creator', 'dates' or 'medium' for example. However the location field seems a bit specific, and surely less likely to be known than the creator of a work. On the basis of this presentation Multimatch appears to make light work of the semantic and structural difficulties seen in previous presentations. However, one wonders whether it allows optimum handling in linguistic typology - how much work would realistically be needed to build a phrase-dictionary, specific to images, and with the absence of adherence to any standards, how successful the results would be.

Kurt Dressel gave us a run-through through Imprezzeo's image recognition technology, which is billed as "using images to find images." This is a user-friendly competitor to



D3.3 Metadata Search & Retrieval Seminar 3 & Report

Imense, another image recognition technology provider who have featured in previous MILE conferences.

Imprezeeo define their product as “visual metadata”, a new phrase for the semiotic lexicon and a clear definition for a complicated concept. Dressel was also very clear that Imprezeeo’s technology is designed to work in conjunction with existing text metadata, which is of course imperative for the type of specific rich metadata used by art image collections.

Imprezeeo’s product seems to be further developed than Imense, and their justification for using their product is sound. The steep growth in volume of digital image content has impacted on S&R. Add to this the fact that metadata and keywording is expensive to create and maintain, the option of using a visual search tool such as Imprezeeo’s is increasingly attractive.

Imprezeeo have identified their product has a broad market appeal, for both commercial and individual users;

- Creative stock photo
- News agencies
- Publishers
- Digital Asset Management
- Museums and archives
- Corporate marketing departments
- Social networking
- Image hosting & sharing
- Retail & auction sites
- Search engines
- Desktop photo applications

Imprezeeo uses a combination of CBIR, Face Detection and Face Recognition. For fine art images, Dressel foresees Imprezeeo as working in the following ways;

- Integrate visual search with multi-lingual text search
- Streamlining infrastructure and improving workflow by assisting cataloguing departments
- Find duplicates
- Uncover similarities across collections
- Mobile search and link to metadata about asset
- Ecommerce.

Imprezeeo has been developed to be integrated with existing systems as seamlessly as possible. Dressel then gave some demonstrations of Imprezeeo’s results, some of which are available through their website. Finally, he showed the results of searches using images from the Bridgeman collection. Since these were based purely on the images themselves and no accompanying metadata, the results were fairly impressive. As far as current image recognition technologies go, Imprezeeo seems to be a market leader and is definitely one to watch.



D3.3 Metadata Search & Retrieval Seminar 3 & Report

Agnes Folaji ended the presentations with a valuable use case from the perspective of a commercial image organisation, Fotofinder. Fotofinder is an image agency – a portal to images from collections and individual photographers – specifically designed to give access to the German-speaking market. It has a large international list of clients and suppliers, a large range of subjects and its market is 80% editorial and 20% advertising. The website has three main functions for its clients; a search engine, a market place and a solution provider, and is translated into English, German and French.

Fotofinder's company policy is that cataloguers usually only translate the keywords, in addition to enriching the metadata. If they were to directly translate search requests this would not give enough flexibility. However, the steep scale of image uploading from 2003 – 2008 meant rethinking this system. In 2005, Fotofinder launched their new website minus the controlled vocabulary lists, to give a very simple search and retrieval system. This was to deal with the 6,800,000 images and 2,000,000 unique items now available on the site.

Prior to 2003, translation was done manually by Fotofinder's employees. Now then it's been a combination of open-source dictionaries and automatic translation. This machine translation starts from IPTC keywording, and works through a workflow including library specific dictionaries, cache databases, extracts which identify basic forms, split compounds and find synonyms, and then translate to create a new set of metadata. At the same time, this newly translated metadata is then copied to a Cache Database which acts like a 'Big Brain'.

The advantages of such a system are that such visualisation of a dictionary's content helps to identify inappropriate terms while ensuring and enhancing metadata quality. However, the disadvantages are that it returns too many synonyms, keyword variations, ambiguities, in addition to throwing up inappropriate search results and having an inherent inability to handle multiple word phrases.

This system has so far taken 10 years of development; Folaji demonstrated its weakness with two examples, where the metadata has errors and mis-spellings, or where the image is under key-worded but has a lengthy caption. As the amount of digital content increases, so exponentially does the amount of metadata, which is therefore growing uncontrollable. However, to combat this, search engine optimisation is increasing which works to counter-balance this amount of metadata. So Fotofinder's working solution is to use existing technologies as effectively as possible.



CONCLUSION

Towards a Multilingual Future was the final seminar within Work Package 3, Metadata Search and Retrieval, the title proving a fitting conclusion of the day's aims.

This seminar sought to unite the findings from the previous two seminars within this work package, *Successful Searching? Lost in Translation* and *Speaking in Tongues* and to evaluate what recommendations MILE could make to the image industry.

Almost before the first Metadata Search & Retrieval work package, MILE concluded that image metadata is integral to the successful search and retrieval of images both in a monolingual and a multilingual environment. The MILE seminars were then structured around how best to structure and use metadata to improve monolingual and multilingual search and retrieval of images.

Successful Searching? Lost in Translation demonstrated that the primary inhibiting factors to successful search and retrieval of images were poorly defined and inconsistent metadata and complex unstructured keyword ontologies. MILE also concluded from this seminar that the requirements of image libraries and their users were too disparate to achieve realistic recommendation of a single metadata standard for all. Since a similar conclusion was drawn within the closely aligned Metadata Classification work package, MILE sought to investigate these requirements more closely to evaluate whether there was any common ground between image libraries and their users so that the network could assist them in achieving interoperability and improving their search and retrieval of images.

Image library and user requirements were collected during the second seminar, *Speaking in Tongues*, with a multiple-choice questionnaire and a work group exercise to discuss group preferences. Further evaluation of the questionnaire results showed that metadata requirements differ significantly between users seeking precise results and users seeking discovery results. Many uninformed users prefer to view all possible results before narrowing their search down to a select group of images themselves. Other users prefer to rely on the image database to provide a precise set of results. Many of these preferences are down to a lack of trust in the metadata that has been attached to the images. This lack of trust is sometimes due to translation inaccuracies and sometimes a result of lack of consistent metadata. Those users who were experienced users of image databases generally preferred to see all possible results, and maintain control over the refinement of their image search.

Speaking in Tongues evaluated a variety of methods, tools and projects purporting to provide multilingual access. These included multilingual thesauri, automatic translation tools, automatic keywording tools and language-to-

language translation of existing controlled vocabularies. However, many image library professionals and users advised that none of these methods were reliable enough to trust entirely, and negate the need for personal selection or at least the personal refinement of a set of images.



D3.3 Metadata Search & Retrieval Seminar 3 & Report

MILE considered these conclusions carefully in the preparation for the final Search and Retrieval Seminar, *Towards a Multilingual Future*. The network determined that as multilingual access tools are still in a developmental phase and with increased pressure from projects such as Europeana, the World Digital Library and many other digital content portals to provide access to European cultural heritage, that MILE should focus on recommending best practice guides for image metadata for discovery recall rather than precision recall. Multilingual access tools had proven themselves at least sufficient enough for these purposes.

Towards a Multilingual Future presented the MILE network's conclusions and examined them with reference to ongoing and recently-concluded European projects for the provision of a portal to digital content. The network's conclusions on the need to differentiate between precision and discovery recall were reiterated here, as was the recommendation for consistent metadata and the organisation of that metadata into subject headings. The MILE network was interested to hear the experiences and conclusions drawn from projects focussed on multilingual access to textual documents here.

A further panel discussion took place during this seminar to agree upon the core metadata fields needed for discovery recall that were deliberately unbiased towards a particular metadata standard or controlled vocabulary. This discussion was begun at the previous Metadata Classification seminar, *A Picture's Worth a Thousand Euros* with a mapping exercise. At *Towards a Multilingual Future*, each panellist gave their views on the need for a core set of metadata fields, and the majority conclusion was that a core set of unambiguous metadata fields that could be used across known metadata standards would be a useful addition to the image library industry because it would enable and improve cataloguing and classification purposes, as well as improving interoperability and enabling the discovery of digital content from a wide variety of access points.

Delegates and speakers agreed that MILE needed to continue its work in this area by considering specific use-cases, so that the network can define both the role and the use of the proposed set of unambiguous metadata fields.

The COO has already accepted offers from two institutions who want to participate in the use-case scenarios, as well as listing a variety of metadata mapping requirements. The COO will continue to discuss requirements with MILE's stakeholders so that final recommendations and guidelines can be included in the final report.