Cultural Heritage in CLEF (CHiC) 2013 – Multilingual Task Overview[1]

Vivien Petras[1], Toine Bogers[2], Nicola Ferro[3], Ivano Masiero[3]

1 Berlin School of Library and Information Science, Humboldt-Universitätzu Berlin, Dorotheenstr. 26, 10117 Berlin, Germany
vivien.petras@ibi.hu-berlin.de

2 Royal School of Library and Information Science, Copenhagen University,Birketinget 6, 2300 Copenhagen S, Denmark
mvs872@iva.ku.dk

3 Department of Information Engineering, University of Padova, Via Gradenigo 6/B, 35131Padova, Italy
{ferro,masieroi}@dei.unipd.it

Abstract. The Cultural Heritage in CLEF 2013 multilingual task comprised two sub-tasks: multilingual ad-hoc retrieval and semantic enrichment. The multilingual ad-hoc retrieval sub-task evaluated retrieval experiments in 13 languages (Dutch, English, German, Greek, Finnish, French, Hungarian, Italian; Norwegian, Polish, Slovenian, Spanish, Swedish). More than 140,000 documents were assessed for relevance on a tertiary scale. The ad-hoc task had 7 participants submitting 30 multilingual and 41 monolingual runs. The semantic enrichment task evaluated monolingual and multilingual semantic enrichments (suggestions based on a query) in the same 13 languages. Two participants submitted 10 runs. Results indicated that different languages contribute differently to the overall retrieval effectiveness, probably dependent on collection size. Experiments showed that using more or all of the provided languages usually increases retrieval effectiveness, but not always. For a multilingual task of this scale (13 languages), more participants are necessary in order to provide enough variations in runs to allow for comparative analyses.

Keywords: cultural heritage, Europeana, ad-hoc retrieval, semantic enrichment, multilingual retrieval

1 Introduction

Cultural heritage collections – preserved by archives, libraries, museums and other institutions – consist of "sites and monuments relating to natural history, ethnography, archaeology, historic monuments, as well as collections of fine and applied arts" [3]. Cultural heritage content is often multilingual and multimedia (e.g. text, photographs, images, audio recordings, and videos), usually described with metadata in multiple formats and of different levels of complexity. Cultural heritage institutions have dif-

[1] Parts of this paper were already published in the CHIC 2013 LNCS Overview paper [6].

ferent approaches to managing information and serve diverse user communities, often with specialized needs. The targeted audience of the CHiC lab and its tasks are developers of cultural heritage information systems, information retrieval researchers specializing in domain-specific (cultural heritage) and / or structured information retrieval on sparse text (metadata) and semantic web researchers specializing on semantic enrichment with LOD data. Evaluation approaches (particularly system-oriented evaluation) in this domain have been fragmentary and often non-standardized. CHiC aims at moving towards a systematic and large-scale evaluation of cultural heritage digital libraries and information access systems.

After a pilot lab in 2012, where a standard ad-hoc information retrieval scenario was tested together with two use-case-based scenarios (diversity task and semantic enrichment task), the 2013 lab diversifies and becomes more realistic in its tasks organization. The pilot lab has shown that cultural heritage is a truly multilingual area, where information systems contain objects in many different languages. Cultural heritage information systems also differ from some more specified information systems in that ad-hoc searching might not be the prevalent form of access to this type of content. The 2013 CHiC lab therefore focuses on multilinguality in the retrieval tasks and adds an interactive task, where different usage scenarios for cultural heritage information systems were tested. The multilingual tasks described in this paper required multilingual retrieval in up to 13 languages, making CHiC the most multilingual CLEF lab ever.

CHiC has teamed up with Europeana[2], Europe's largest digital library, museum and archive for cultural heritage objects to provide a realistic environment for experiments. Europeana provided the document collection (digital representations of cultural heritage objects) and queries from their query logs. The interactive task also provided a topic clustering algorithm and a customized browsable portal based on Europeana data.

The paper is structured as follows: Chapter 2 introduces the Europeana document collection. Chapters 3 and 4 describe the sub-tasks multilingual ad-hoc and multilingual semantic enrichment in detail, their requirements, participants and results. The conclusion provides an outlook on the future of CHiC and the potential synergies of combining ad-hoc and interactive information retrieval evaluation.

2 The Europeana Collection

The Europeana information retrieval document collection was prepared for the CHiC pilot lab in 2012 (Petras et al., 2012). It consists of the complete Europeana metadata index as downloaded from the production system in March 2012. It contains 23,300,932 documents with a size of 132 GB. With the move of Europeana to an open data license in the summer of 2012 and the subsequent changes in content, this test document collection represents a snapshot of Europeana data from a particular time. However, the overlap to the current content is about 80%.

[2]http://www.europeana.eu

The collection consists of metadata records describing cultural heritage objects, e.g. the scanned version of a manuscript, an image of a painting of sculpture or an audio or video recording. Roughly, 62% of the metadata records describe images, 35% describe text, 2% describe audio and 1% video recordings.

The collection was divided into 14 sub-collections according to the language of the content provider of the record (which usually indicates the language of the metadata record). A threshold was set: all languages with less than 100,000 documents were grouped together under the name "Others". The 13 language collections included Dutch, English, German, Greek, Finnish, French, Hungarian, Italian; Norwegian, Polish, Slovenian, Spanish, Swedish. For the CHiC 2013 experiments, all sub-collections except the "Others" were used, totaling roughly 20 million documents. The 14 sub-collections are listed in table 1.

Table 1. CHiC Collections by Language and Media Type.

Language	**Sound**	**Text**	**Image**	**Video**	**Total**
German	23,370	664,816	3,169,122	8,372	**3,865,680**
French	13,051	1,080,176	2,439,767	102,394	**3,635,388**
Swedish	1	1,029,834	1,329,593	622	**2,360,050**
Italian	21,056	85,644	1,991,227	22,132	**2,120,059**
Spanish	1,036	1,741,837	208,061	2,190	**1,953,124**
Norwegian	14,576	207,442	1,335,247	555	**1,557,820**
Dutch	324	60,705	1,187,256	2,742	**1,251,027**
English	5,169	45,821	1,049,622	6,564	**1,107,176**
Polish	230	975,818	117,075	582	**1,093,705**
Finnish	473	653,427	145,703	699	**800,302**
Slovenian	112	195,871	50,248	721	**246,952**
Greek	0	127,369	67,546	2,456	**197,371**
Hungarian	34	14,134	107,603	0	**121,771**
Others	375,730	1,488,687	1,106,220	19,870	**2,990,507**
Total	**455,162**	**8,371,581**	**14,304,289**	**169,899**	**23,300,932**

The XML metadata contains title and description data, media type and chronological data as well as provider information. For ca. 30% of the records, content-related enrichment keywords were added automatically by Europeana based on a mapping between metadata terms and terms from controlled lists like DBpedia names. In the Europeana portal, object records commonly also contain thumbnails of the object if it is an image and links to related records. These were not included with the test collection, but relevance assessors were able to look at them at the original source. Figure 1 shows an extract example record from the Europeana CHiC collection.

<ims:metadata ims:identifier="http://www.europeana.eu/resolve/record/10105/5E1618BFAF072B8953B30701A6A6C3BB655ACF9D" ims:namespace="http://www.europeana.eu/" ims:language="eng">
<ims:fields>
<dc:identifier>Orn.0240</dc:identifier>
<dc:subject>Tachymarptis melba</dc:subject>
<dc:title>RundunZaqquBajda (Orn.0240)</dc:title>
<dc:title>Alpine Swift (Orn.0240)</dc:title>
<dc:type>mounted specimen</dc:type>
<europeana:country>malta</europeana:country>
<europeana:dataProvider>Heritage Malta</europeana:dataProvider>
<europeana:isShownAt>http://www.heritagemalta.org/sterna/orn.php?id=0240
</europeana:isShownAt>
<europeana:language>en</europeana:language>
<europeana:provider>STERNA</europeana:provider>
<europeana:type>IMAGE</europeana:type>
<europeana:uri>http://www.europeana.eu/resolve/record/10105/5E1618BFAF072B8953B30701A6A6C3BB655ACF9D</europeana:uri>
</ims:fields>
</ims:metadata>

Fig.1. Europeana CHiC Collection Sample Record

3 The CHiC Multilingual Ad-hoc Task

The sub- tasks are a continuation of the 2012 CHiC lab, using a similar task scenarios, but requiring multilingual retrieval and results. Two sub-tasks were defined: multilingual ad-hoc retrieval and multilingual semantic enrichment.

The traditional multilingual ad-hoc retrieval task measures information retrieval effectiveness with respect to user input in the form of queries. The 13 language subcollections form the multilingual collection (ca. 20 million documents) against which experiments were run. Participants were asked to submit ad-hoc information retrieval runs based on 50 topics (provided in all 13 languages) and including at least 2 and at most all 13 collection languages. For pooling purposes, participants were also asked to submit monolingual runs choosing any of the collection languages. Because the topics were provided in all collection languages, the focus of the task was not on topic translation, but on multilingual retrieval across different collection languages.

3.1 Topic Creation

A new set of 50 topics was created for the 2013 edition of CHiC, where topic selection was determined partially by the potential for retrieving a sufficient number of relevant documents in each of the collection languages. CHiC 2012 used topics from the Europeana query logs alone, which resulted in zero results for some of the 3 languages [13]. The problem of having zero relevant results is aggravated when collec-

tion languages are varied, especially in the cultural heritage area. Many topics are relevant for only a few languages or cultures. For 2013, more focus was put on testing all topics in all languages for retrieving relevant documents, which resulted in fewer zero relevant result topics. The topic creation process started with creating a pool of candidate topics, which derived from four different sources:

- 15 topics that showed promising retrieval performance were re-used from the 2012 topic set (only in 3 languages) to test their performance in 13 languages.
- Another 19 topics that were not specific to only a handful of languages were taken from an annotated snapshot of the Europeana query log (the same procedure was used for the 2012 topics).
- The Polish task also suggested topics, 17 were not considered to be relevant only in Polish and input in the candidate pool.
- Finally, two of the track organizers generated another 21 test queries covering a wide range of topics contained in Europeana's collections that would span all collection languages.

These 73 candidate topics were then translated into all 13 languages by volunteers. The translated candidate topics were run against the 13 language collections using Indri 5.2 with default settings[3]. We retained the 50 topics that returned the highest number of relevant documents for all thirteen languages. Another factor that affected the final selection of the 2013 topics was the abundance of named-entity queries (around 60%) in the 2012 topic set. While named-entity queries are a common type of query for Europeana [9], they are less challenging than non-entity queries that describe a more complex information need. For this we wished to down-sample the proportion of named-entity queries to around 20%.

The final topics set covers a wide range of topics and consisted of 12 topics from the 2012 topic set, 13 log-based topics, 13 topics from the Polish subtask, and 12 intellectually derived queries. In form and type, the different query types are indistinguishable and usually include 1-3 query terms (e.g. "silent film", "ship wrecks", and "last supper"). The underlying information need for a query can be ambiguous if the intention of the query is not clear. In this case, the track organizers discussed the query and agreed on the most likely information need. These were not admissible for information retrieval. Figure 2 shows an example of an English query.

```
<topic lang="en">
  <identifier>CHIC-004</identifier>
  <title>silent film</title>
  <description>documents on the history of silent film, silent film videos, biographies of actors and directors, characteristics of silent film and decline of this genre</description>
</topic>
```

Fig. 2. CHiC Sample Query

[3]Jelinek-Mercer smoothing with λ set to 0.4 and no stemming or stopword filtering.

3.2 Pooling and Relevance Assessments

This year, we produced 13 pools, one for each target language using different depths depending on the language and the available number of documents. The pools were created using all the submitted runs. A 14th pool, for the multilingual task, is the union of the 13 pools described above. Table 2 provides details about the created pools, their size, the number of relevant and not relevant documents, and the pooled runs.

Table 2. CHiC 2013 Multilingual Pools

CHiC 2013 Multilingual - Dutch Pool		
Size	Depth	125
	Total documents	10,548
	Highly Relevant documents	1,583
	Partially Relevant documents	811
	Not relevant documents	8,154
	Topics with relevant documents / Total Topics	48 out of 50
	Assessors	2
CHiC 2013 Multilingual - English Pool		
Size	Depth	50
	Total documents	16,696
	Highly Relevant documents	2,530
	Partially Relevant documents	70
	Not relevant documents	14,096
	Topics with relevant documents / Total Topics	49 out of 50
	Assessors	2
CHiC 2013 Multilingual - Finnish Pool		
Size	Depth	200
	Total documents	2,465
	Highly Relevant documents	276
	Partially Relevant documents	19
	Not relevant documents	2,170
	Topics with relevant documents / Total Topics	16 out of 50
	Assessors	1
CHiC 2013 Multilingual - French Pool		
Size	Depth	50
	Total documents	17,978
	Highly Relevant documents	2,508
	Partially Relevant documents	436
	Not relevant documents	15,034
	Topics with relevant documents / Total Topics	50 out of 50
	Assessors	1
CHiC 2013 Multilingual - German Pool		
Size	Depth	50
	Total documents	18,460

	Highly Relevant documents	3,510
	Partially Relevant documents	50
	Not relevant documents	14,900
	Topics with relevant documents / Total Topics	50 out of 50
	Assessors	2
CHiC 2013 Multilingual - Greek Pool		
Size	Depth	125
	Total documents	10,032
	Highly Relevant documents	265
	Partially Relevant documents	145
	Not relevant documents	9622
	Topics with relevant documents / Total Topics	40 out of 50
	Assessors	1
CHiC 2013 Multilingual - Hungarian Pool		
Size	Depth	200
	Total documents	5,834
	Highly Relevant documents	332
	Partially Relevant documents	491
	Not relevant documents	5,011
	Topics with relevant documents / Total Topics	48 out of 50
	Assessors	1
CHiC 2013 Multilingual - Italian Pool		
Size	Depth	75
	Total documents	13,387
	Highly Relevant documents	2,176
	Partially Relevant documents	721
	Not relevant documents	10,490
	Topics with relevant documents / Total Topics	47 out of 50
	Assessors	1
CHiC 2013 Multilingual - Norwegian Pool		
Size	Depth	125
	Total documents	10,287
	Highly Relevant documents	1,723
	Partially Relevant documents	289
	Not relevant documents	8,275
	Topics with relevant documents / Total Topics	43 out of 50
	Assessors	2
CHiC 2013 Multilingual - Polish Pool		
Size	Depth	125
	Total documents	11,342
	Highly Relevant documents	1,086
	Partially Relevant documents	624
	Not relevant documents	9,632
	Topics with relevant documents / Total Topics	46 out of 50

	Assessors	1
CHiC 2013 Multilingual - Slovenian Pool		
Size	Depth	200
	Total documents	6,718
	Highly Relevant documents	481
	Partially Relevant documents	195
	Not relevant documents	6,042
	Topics with relevant documents / Total Topics	37 out of 50
	Assessors	1
CHiC 2013 Multilingual - Spanish Pool		
Size	Depth	100
	Total documents	11,373
	Highly Relevant documents	1,689
	Partially Relevant documents	446
	Not relevant documents	9,238
	Topics with relevant documents / Total Topics	46 out of 50
	Assessors	1
CHiC 2013 Multilingual - Swedish Pool		
Size	Depth	150
	Total documents	11,640
	Highly Relevant documents	941
	Partially Relevant documents	342
	Not relevant documents	10,357
	Topics with relevant documents / Total Topics	43 out of 50
	Assessors	1

We used graded relevance, i.e. highly relevant, partially relevant, and not relevant. To compute the standard performance measures reported in Section 3.3, we used binary relevance and conflated highly relevant and partially relevant to just relevant. The DIRECT system [1] was used to collect runs, perform relevance assessment, and compute performances. The system's interfaces and processes were also described in last year's CHiC Paper [5]

For all languages except English, native language speakers performed the relevance assessments. Fifteen assessors took 2 weeks to assess the ca. 140,000 documents. The assessors received detailed instructions on how to use the assessor interface and guidelines, how the relevance assessments were to be approached. Constant communication via a common mailing list ensured that assessors across languages treated topics from the same perspective.

Despite our efforts in topic creation, some topics in some languages did not have any relevant documents in the pool. Besides not all queries having relevant documents in the Europeana collection, the problem was exacerbated by receiving very few monolingual runs that could be used for pooling, sometimes resulting in very small pools. While 11 languages have at least 40 topics with relevant documents (5 with 48 or more topics with relevant documents), Finnish (only 16 topics with relevant docu-

ments) and Slovenian (only 37 topics with relevant documents) give raise for concern in comparative analyses.

3.3 Participants and Runs

Seven different teams participated in the 2013 edition of the ad-hoc track (table 3).

Table 3. Participating groups and country.

Group	Country
CEA LIST	France
Department of Computer Science, University of Neuchâtel	Switzerland
MRIM/LIG, University of Grenoble	France
RSLIS, University of Copenhagen & Aalborg University	Denmark
School of Information, UC Berkeley	USA
Technical University of Chemnitz	Germany
University of Westminster	Great Britain

Out of the 71 runs submitted, 30 were multilingual runs using at least 2 collection languages; 10 runs used all available languages for both topics and collections. All languages were also represented in the monolingual or bilingual runs (41 total). English, German, French and Italian were the popular languages for the monolingual runs, all other languages had only 1 or 2 runs. Toine Bogers (RSLIS) provided 2 more baseline runs for each language collection using the Indri information retrieval system using language modelling with either the Dirichlet (no stopword list, no stemming) or the Jelinek-Mercer smoothing algorithm (with stopword list, no stemming), which are used in the comparison. Table 4 shows the submitted runs and their language combinations including the baseline runs.

Table 4. Submitted Runs in the CHiC 2013 Multilingual Ad-hoc Retrieval Task

Topic Language(s)	Collection Language(s)	Runs		Topic Language(s)	Collection Language(s)	Runs
Monolingual runs				**Multilingual runs**		
DE	DE	6		All	All	10
EL	EL	3		DE	All	1
EN	EN	10		EN	All	1
ES	ES	4		FR	All	1
FI	FI	3		All NOT EL	All NOT EL	1
FR	FR	6		All NOT EL, HU, SL	All NOT EL, HU, SL	4
HU	HU	3		All	DE,EN,FR	1
IT	IT	8		DE, EN, ES, FR, IT	DE,EN,FR	1
NL	NL	4		DE,EN,FR	DE,EN,FR	1

Topic Language(s)	Collection Language(s)	Runs		Topic Language(s)	Collection Language(s)	Runs
NO	NO	4		DE	DE,EN,FR	1
PO	PO	4		EN	DE,EN,FR	1
SL	SL	3		ES	DE,EN,FR	1
SV	SV	4		FI	DE,EN,FR	1
				FR	DE,EN,FR	1
Bilingual runs				IT	DE, EN, FR	1
DE	FR	1		NL	DE,EN,FR	1
DE	EN	1		EN	EN, IT	1
EN	DE	1		IT	EN, IT	1
EN	FR	1				
FR	DE	1				
FR	EN	1				

3.4 Results & Participant Approaches

Because of the many variations in topic and collection language configurations, comparisons between runs is difficult. Since language combinations are then varied by different system configurations, the matrix of possible impact factors becomes very big. However, several comparisons can give indications into further research questions that should be analyzed.

3.4.1 Multilingual Runs: All Languages vs. Fewer languages

Table 5 shows the best multilingual run per participating group ordered by MAP showing the topic and collection languages that were used for retrieval. Note that only the best run is selected for each group, even if the group may have more than one top run.

Table 5. Best Multilingual Experiments per Group (in MAP)

Participant	Experiment Identifier	Topic Languages	Collection Languages	MAP
Chemnitz	TUC_ALL_LA	All	All	23.38%
CEA List	MULTILINGUALNOEXPANSION	All NOT EL, HU, SL	All NOT EL, HU, SL	18.78%
Neuchatel	UNINEMULTIRUN5	All	All	15.45%
RSLIS	RSLIS_MULTI_FUSION_COMBSUM	All	All	8.37%
Westminster	R005	EN	EN,IT	6.30%
Berkeley	BERKMLENFRDE19	EN,FR,DE	EN,FR,DE	3.93%

Figure 3 shows the best 5 multilingual runs in an interpolated recall vs. average precision graph.

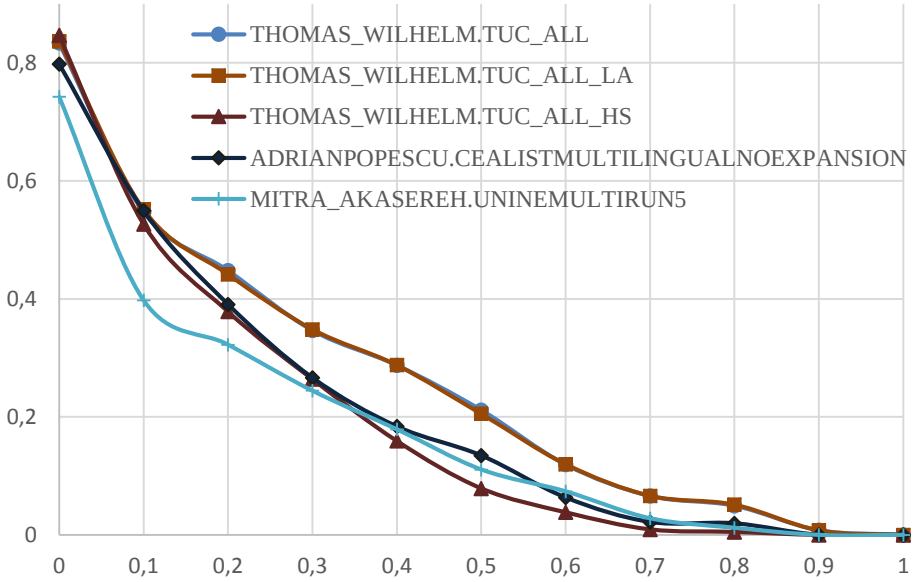

Fig. 3. Best 5 Multilingual Runs – Interpolated Recall / Precision

It is difficult to interpret these figures in terms of which languages have the most input for retrieval success as the applied IR systems play a much bigger role in this cross-system comparison.

UC Berkeley compared experiments with different topic languages against a multilingual collection of English, French and German combined. Results show that using the exact same languages for topics achieves a slightly higher result than using just one of the topic languages or even more languages (table 6). In this experiment, differences between runs are probably not all statistically significant. However it is interesting to note that English and French seem not to contribute to the retrieval effectiveness as much as German, for example, and that a topic language, which is not represented in the collection languages (ES) can still achieve almost as high a MAP as the topic language English.

Table 6. UC Berkeley: Comparing Topic and Collection Languages (in MAP) [4]

Experiment Identifier	Topic Languages	Collection Languages	MAP
BERKMLENFRDE19	EN,FR,DE	EN,FR,DE	3.93%
BERKMLALL17	All	EN,FR,DE	3.57%
BERKMLSPENFRDEIT18	EN,FR,DE, ES, IT	EN,FR,DE	3.53%
BERKMLDE12	DE	EN,FR,DE	3.31%
BERKMLFR11	FR	EN,FR,DE	2.22%
BERKMLEN10	EN	EN,FR,DE	1.66%
BERKMLSP16	ES	EN,FR,DE	1.33%

RSLIS used a similar approach with equivalent results: using one topic language against the whole multilingual index did result in lower retrieval effectiveness than the fusion runs using 3 topic languages (table 7).

Table 7. RSLIS: Comparing Topic and Collection Languages (in MAP)[8]

Experiment Identifier	Topic Languages	Collection Languages	MAP
MULTI_FUSION_COMBSUM	EN,FR,DE	All	8.37%
MULTI_FUSION_COMBMNZ	EN,FR,DE	All	8.36%
MULTI_MONO_GER	DE	All	6.79%
MULTI_MONO_FRE	FR	All	4.30%
MULTI_MONO_ENG	EN	All	3.70%

Both groups found that the German topics seem to have the highest retrieval impact. The Westminster group [11] showed in a similar experiment that English seemed to have a higher impact than Italian. More runs would be necessary to be able to perform a complete analysis.

Unine experimented with removing topic and collection languages equally and different fusion algorithms (merging results from separate language indexes) and showed that leaving out the smaller collection languages can result in an increase in performance, however, the impact of an individual language is unclear (table 8).

Table 8. Unine: Comparing Topic and Collection Languages (in MAP) [2]

Experiment Identifier	Topic Languages	Collection Languages	MAP
UNINEMULTIRUN5	All	All	15.45%
Inofficial Unine Run, Z-score	All NOT EL, HU, SL	All NOT EL, HU, SL	16.22%
Inofficial Unine Run, RR	All	All	13.88%
Inofficial Unine Run, RR	All NOT EL	All NOT EL	13.87%

Finally, TU Chemnitz experimented with different stemming algorithms for all languages and found that using a less aggressive stemmer worked best compared to the standard rule-based stemmers used in Solr or a no-stemming approach (table 9).

Table 9. Chemnitz: Comparing Stemming Approaches (in MAP) [12]

Stemming Approach	MAP
Less aggressive	23.38%
Standard (rule-based)	23.36%
No stemmer	15.34%

3.4.2 Monolingual Runs

For pooling purposes, participants submitted monolingual runs as well. We can compare them using the whole multilingual pool (results are also available in the DIRECT[4] system) or using the monolingual pools. While a multilingual pool is what the real use case prescribes (all languages are potentially relevant), we can also look at monolingual pools to achieve an improved system comparison (less variation because of language). We will concentrate on the 4 languages with the most submitted experiments: English (10), Italian (8), German and French (6). Table 10 shows the best monolingual run for each participant in those languages.

Table 10. Best Monolingual Experiments per Group (in MAP)

Participant	**Experiment Identifier**	**MAP**	**Participant**	**Experiment Identifier**	**MAP**
Monolingual English			**Monolingual Italian**		
MRIM	MRIM_AR_2	40.43%	Westminster	R004	29.41%
Westminster	R001	28.30%	RSLIS	BASELINE.ITA3	24.90%
Berkeley	BERKBIDEEN04	19.42%	CEA List	CEALISTITALIANFILTERED	16.50%
RSLIS	BASELINE.ENG1	18.35%			
CEA List	CEALISTENGLISHFILTERED	16.68%			
Monolingual French			**Monolingual German**		
CEA List	CEALISTFRENCHNOEXPANSION	27.62%	RSLIS	BASELINE.GER2	29.79%
Berkeley	BERKMONOFR02	20.14%	CEA List	CEALISTGERMANNOEXPANSION	28.99%
RSLIS	BASELINE.FRE3		Berkeley	BERKBIENDE09	17.85%

Unfortunately, only 2 groups (RSLIS & CEA List) submitted runs to all 4 languages so that a comparison among even those 4 languages becomes difficult.

3.4.3 Participant Approaches

Table 11 briefly summarizes the participants' approaches to the ad-hoc track.

Table 11. Participating groups and their approaches to the multilingual ad-hoc track.

Group	**Description of approach**
Chemnitz	Apache Solr with special focus on comparing different types of stemmers (generic, rule-based, dictionary-based) [12].
CEA LIST	Query expansion of a Vector Space model with tf-idf weighting by using related concepts extracted from Wikipedia using Explicit Semantic Analysis [7].

[4] http://direct.dei.unipd.it

MRIM	Language modeling approach using Dirichlet smoothing and Wikipedia as external document collection to estimate the word probabilities in case of sparsity of the original term-document matrix [10].
Neuchâtel	Probabilistic IR using Okapi model with stopword filtering and light stemming. Collection fusion on the results lists from 13 different monolingual indexes using z-score normalization merging [2].
RSLIS	Language modeling with Jelinek-Mercer smoothing and no stopword filtering or stemming. One run each for English, French, and German where these topic languages are run against a multilingual index. Two fusion runs using the CombSUM and CombMNZ methods combining these three monolingual runs against the multilingual index [8].
UC Berkeley	Probabilistic text retrieval model based on logistic regression together with pseudo-relevance feedback for all of the runs. Runs with English, French, and German topic sets and sub-collections, as well translations generated by Google Translate [4].
Westminster	Divergence from randomness algorithm using Terrier on the English and Italian collections [11].

4 The CHiC Multilingual Semantic Enrichment Task

The multilingual semantic enrichment task requires systems to present a ranked list of related concepts for query expansion. Related concepts can be extracted from Europeana data or from other resources in the Linked Open Data cloud or other external resources (e.g. Wikipedia). Participants were asked to submit up to 10 query expansion terms or phrases per topic. This task included 25 topics in all 13 languages. Participants could choose to experiment on monolingual or multilingual semantic enrichments. The suggested concepts were assessed with respect to their relatedness to the original query terms or query category.

Only 2 groups participated in the semantic enrichment task, making a comparison more difficult. Almost all experiments contained either only English concepts or concepts from several languages (multilingual). In total, 10 experiments were submitted.

MRIM/LIG (Univ. of Grenoble) used Wikipedia as a knowledge base and the query terms in order to identify related Wikipedia articles for enrichment candidates. Both in-links and out-links to and from these related articles (in particular their titles) were then used to extract terms for enrichment [10].

CEA List used Explicit Semantic Analysis (documents are mapped to a semantic structure) also with Wikipedia as a knowledge base. Whereas MRIM/LIG used the title of Wikipedia articles and their in- and out-links for concept expansion, CEA List concentrated on the categories and the first 150 characters within a Wikipedia article. When Wikipedia category terms overlapped with query terms, these concepts were boosted for expansion. In ad-hoc retrieval, the topic and expanded concepts were matched against the collection and the results were then matched again to a consolidated version of the topics (favoring more frequent concept phrases) before outputting

the result. For multilingual query expansion, the interlingua links to parallel language versions of a Wikipedia article were used in a fusion model. For most expansion experiments, only concepts were considered that appear in at least 3 Wikipedia language versions, allowing for multilingual expansions [7].

The semantic enrichments were evaluated using a tertiary relevance assessment (definitely relevant, maybe relevant, not relevant) and P@1, P@3 and P@10 measurements. Table 12 shows the results for the best 2 runs for each participants using either the strict relevance measurement (just definitely relevant) or the relaxed relevance measurement (definitely relevant and maybe relevant).

Table 12. Semantic Enrichment: Best 2 Runs for each Participant

Run name	**P@1**	**P@3**	**P@10**
	Strict relevance		
ceaListEnglishMonolingual	0.5200	0.5467	0.4680
ceaListEnglishRankMultilingual	0.4800	0.4533	0.3400
MRIM_SE13_EN_WM_1	0.0800	0.0667	0.0522
MRIM_SE13_EN_WM	0.0400	0.0533	0.0422
	Relaxed relevance		
ceaListEnglishRankMultilingual	0.6800	0.7200	0.5600
ceaListEnglishMonolingual	0.6800	0.7067	0.6600
MRIM_SE13_EN_WM_1	0.2800	0.1467	0.1598
MRIM_SE13_EN_WM	0.2800	0.1333	0.1448

Only CEA List experimented with multilingual enrichments. Interestingly, a multilingual enrichment run was the best with a relaxed relevance measurement, while the monolingual run was the best with a strict relevance measurement.

5 Conclusion and Outlook

The results of this year's multilingual CHiC task show that multilingual information retrieval experiments are challenging not only because of the number of languages that need to be processed but also because of the number of participants necessary in order to produce comparable results. As the number of possible language variations increases (CHiC had 13 source languages and 13 target languages), very few experiments across participants can be compared. While this year's results have shown that searching in several languages increases the overall performance (an obvious result), we could not show which languages contributed more to retrieval results. Future research in the multilingual task needs to focus on narrower defined tasks (e.g. particular source languages against the whole collection) or define a GRID experiment where a particular information retrieval system performs all possible run variation to arrive at better answers.

The interactive study collected a rich data set of questionnaire and log data for further use. Because the task was designed for easy entrance (predetermined system and

research protocol, this is somewhat different that the traditional lab and is planned to follow a 2-year cycle (assuming the lab's continuation). In year two, the data gathered this year should be released to the community in aggregate form having been assessed by the user interaction community with the goal of identifying a set of objects that need to be developed. The ad-hoc retrieval tasks can benefit from the interactive task by re-using the real queries in ad-hoc retrieval test scenarios – effectively merging both evaluation methods.

Acknowledgements.
This work was supported by PROMISE (Participative Research Laboratory for Multimedia and Multilingual Information Systems Evaluation, Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191. We would like to thank Europeana for providing the data for collection and topic preparation and providing valuable feedback on task refinement. We would like to thank Maria Gäde, Preben Hansen, Anni Järvelin, Birger Larsen, Simone Peruzzo, Juliane Stiller, Theodora Tsikrika and Ariane Zambiras for their invaluable help in translating the topics. We would also like to thank our relevance assessors Tom Bekers, Veronica Estrada Galinanes, Vanessa Girnth, Ingvild Johansen, Georgios Katsimpras, Michael Kleineberg, Kristoffer Liljedahl, Giuliano Migliori, Christophe Onambélé, Timea Peter, Oliver Pohl, Siri Soberg, Tanja Špec, Emma Ylitalo.

References

1. Agosti, M., Ferro, N.: Towards an Evaluation Infrastructure for DL Performance Evaluation. In Tsakonas, G. and Papatheodorou, C. (eds.), Evaluation of Digital Libraries: An Insight to Useful Applications and Methods, pp 93-120. Chandos Publishing, Oxford, UK (2009).
2. Akasereh M., Naji N., Savoy J. UniNE at CLEF – CHIC 2013. In *Proceedings CLEF 2013, Working Notes* (2013).
3. International Council of Museums (2003). Scope Definition of the CIDOC Conceptual Reference Model. http://www.cidoc-crm.org/scope.html
4. Larson, R. Pseudo-Relevance Feedback for CLEF-CHiC Adhoc. In *Proceedings CLEF 2013, Working Notes* (2013).
5. Petras V., Ferro N., Gäde M., Isaac A., Kleineberg M., Masiero I., Nicchio M., Stiller J. Cultural Heritage in CLEF (CHiC) Overview 2012. In *Proceedings CLEF-2012, Working Paper* (2012).
6. Petras, V., Bogers, T., Toms, E., Hall, M., Savoy, J., Malak, P., Pawłowski, A., Ferro, N., Masiero, I. Cultural Heritage in CLEF (CHiC) 2013. In *Proceedings of CLEF 2013, LNCS, Springer* (forthcoming).
7. Popescu, A. CEA LIST's participation at the CLEF CHiC 2013. In *Proceedings CLEF 2013, Working Notes* (2013).
8. Skov, M., Bogers, T., Lund, H., Jensen, M., Wistrup, E., Larsen, B. RSLIS/AAU at CHiC 2013. In *Proceedings CLEF 2013, Working Notes* (2013).
9. Stiller, J., Gäde, M., & Petras, Vivien (2010). Ambiguity of Queries and the Challenges for Query Language Detection. In *CLEF 2010 LABs and Workshops*. Retrieved from http://clef2010.org/resources/proceedings/clef2010labs_submission_41.pdf

10. Tan, K., Almasri, M., Chevallet, J., Mulhem, P., Berrut, C. Multimedia Information Modeling and Retrieval(MRIM)/Laboratoire d'Informatique de Grenoble (LIG) at CHiC2013. In *Proceedings CLEF 2013, Working Notes* (2013).
11. Tanase, D. Using the Divergence Framework for Randomness: CHiC 2013 Lab Report. In *Proceedings CLEF 2013, Working Notes* (2013).
12. Wilhelm-Stein, T., Schürer, B., Eibl, M. Identifying the most suitable stemmer for the CHiC multilingual ad-hoc task. In *Proceedings CLEF 2013, Working Notes* (2013).