Published in Information Access Evaluation, 192-211, 2013
which should be used for any reference to this work

Cultural Heritage in CLEF (CHiC) 2013

Vivien Petras[1], Toine Bogers[2], Elaine Toms[3], Mark Hall[3], Jacques Savoy[4], Piotr Malak[4], Adam Pawłowski[5], Nicola Ferro[6], and Ivano Masiero[6]

[1] Berlin School of Library and Information Science, Humboldt-Universität zu 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] The Information School, University of Sheffield, Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
{e.toms,m.hall}@sheffield.ac.uk
[4] Department of Computer Science, University of Neuchatel, rue Emile Argand 11, 2000 Neuchatel, Switzerland
{jacques.savoy,piotr.malak}@unine.ch
[5] Institute of Library and Information Science, University of Wrocław, pl. Uniwersytecki 9/13, 50-137 Wrocław, Poland
apawlow@uni.wroc.pl
[6] 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 lab comprised three tasks: multilingual ad-hoc retrieval and semantic enrichment in 13 languages (Dutch, English, German, Greek, Finnish, French, Hungarian, Italian, Norwegian, Polish, Slovenian, Spanish, and Swedish), Polish ad-hoc retrieval and the interactive task, which studied user behavior via log analysis and questionnaires. For the multilingual and Polish sub-tasks, more than 170,000 documents were assessed for relevance on a tertiary scale. The multilingual task had 7 participants submitting 30 multilingual and 41 monolingual runs. The Polish task comprised 3 participating groups submitting manual and automatic runs. The interactive task had 4 participating research groups and 208 user participants in the study. For the multilingual task, results show that more participants are necessary in order to provide comparative analyses. The interactive task created a rich data set comprising of questionnaire of log data. Further analysis of the data is planned in the future.

Keywords: cultural heritage, Europeana, ad-hoc retrieval, semantic enrichment, multilingual retrieval, Polish, interactive, user behavior.

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" [8]. 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 different 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 in 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 task 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 other 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 task required multilingual retrieval in up to 13 languages, making CHiC the most multilingual CLEF lab ever. The Polish task concentrated on a rarely tested language in detail. Combining ad-hoc information retrieval and interactive information retrieval test scenarios in one lab provided an environment where both methodologies could overlap and benefit from each other.

CHiC has teamed up with Europeana[1], 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, which is used in all 3 tasks. Chapters 3-5 describe the tasks 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

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

index as downloaded from the production system in March 2012. It contains 23,300,932 documents. 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%.

The collection consists of metadata records describing cultural heritage objects, e.g. the scanned version of a manuscript, an image of a painting or 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 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.

3 The CHiC Multilingual Task

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

The traditional ad-hoc retrieval task measures information retrieval effectiveness with respect to user input in the form of queries. The 13 language sub-collections 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.

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 other external resources (e.g. Wikipedia or other resources from the Linked Open Data cloud). 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.

3.1 Topic Creation

A 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 collection 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 of which 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[2]. 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 [18], 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"). For later relevance assessment, descriptions of the underlying information needs were added, but were not admissible for information retrieval. The underlying information need for a query can be ambiguous if the intention of the query is

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

not clear. In this case, the track organizers discussed the query and agreed on the most likely information need.

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. 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] has been used to collect runs, perform relevance assessment, and compute performances.

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.

3.3 Participants and Results

Multilingual Ad-hoc

Seven different teams participated in the 2013 edition of the ad-hoc track. Out of the 71 runs submitted, 30 were multilingual runs using at least 2 collection languages; 10 runs used all available languages for topics and documents. All languages were also represented in the monolingual runs (41 total). English (10 runs), German (6), French (6) and Italian (8) were the popular languages for the monolingual runs, all other languages had only 1 or 2 runs. Table 1 shows the best runs by participating group ordered by MAP showing the 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 1. Best Experiments per Group (in MAP)

Participant	Experiment Identifier	Collection Languages	MAP
Chemnitz	TUC_ALL_LA	All	23.38%
CEA List	MULTILINGUALNOEXPANSION	All except EL, HU, SL	18.78%
Neuchatel	UNINEMULTIRUN5	All	15.45%
RSLIS	RSLIS_MULTI_FUSION_COMBSUM	All	8.37%
MRIM	MRIM_AR_2	EN	6.43%
Westminster	R005	EN,IT	6.30%
UC Berkeley	BERKMONODE03	DE	4.14%

It is difficult to interpret these figures as all runs regardless of the language sub-collections used were measured against the multilingual pool. Monolingual runs or runs using fewer languages could not have reached better numbers. The working notes paper includes a more detailed analysis for the different run types [14]. Table 2 below lists the participating groups and briefly summarizes their approaches to the ad-hoc track.

Table 2. Participating groups and their approaches to the multilingual ad-hoc track

Group	Description of approach
RSLIS, University of Copenhagen & Aalborg University (Denmark)	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 [17].
University of Neuchâtel (Switzerland)	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].
MRIM/LIG, University of Grenoble (France)	Language modeling approach using Dirichlet smoothing that uses Wikipedia as an external document collection to estimate the word probabilities in case of sparsity of the original term-document matrix [20].
CEA LIST (France)	Query expansion of a Vector Space model with tf-idf weighting by using related concepts extracted from Wikipedia using Explicit Semantic Analysis [15].
Technical University of Chemnitz (Germany)	Apache Solr with special focus on comparing different types of stemmers (generic, rule-based, dictionary-based) [22].
School of Information, UC Berkeley (USA)	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 [9].
University of Westminster (Great Britain)	Divergence from randomness algorithm using Terrier on the English and Italian collections [21].

Multilingual Semantic Enrichment

Only 2 groups participated in the semantic enrichment task, making a comparison more difficult. Participants could choose between monolingual and multilingual runs. Almost all experiments contained only English concepts.

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 (particularly their titles) were then used to extract terms for enrichment.

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 interlingual 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.

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 3 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 3. Semantic enrichment results

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

4 The CHiC Polish Task

The main objective of the Polish task was to obtain a better understanding of information retrieval problems for complex languages such as Polish [19] when facing short text descriptions. We know that the complex morphology of the Polish language may have an impact on both retrieval effectiveness and its relevance. Can this aspect be ignored under the assumption that the morphological complexity will not or have only a small impact on the retrieval performance? If not, can we evaluate the extent of the retrieval effectiveness variations when having a poorer or a better understanding of the Polish morphology? With a related language like Czech, previous studies indicate

that the stemming phase might improve the overall retrieval effectiveness of around 44% over an approach ignoring this word normalization procedure [4]. Can we achieve similar findings with relatively short description of CH objects?

To answer these questions we have organized a Polish task as a standard ad-hoc retrieval task, measuring the information retrieval effectiveness with respect to user input in the form of queries. The resulting ranked list of retrieved items is produced without any prior knowledge about either the user needs or the context.

The Polish collection is a part of the CHiC 2013 multilingual collection and each descriptor contains on average 35 terms. For this task, we have offered both an automatic and manual submission mode. In both cases, the participants are free to use the logical tags they want for indexing the various CH objects. Regarding those titles or the CH objects descriptions, participants are free to manually or automatically enrich the corresponding queries and/or document surrogates (e.g., using specific thesauri, dedicated ontologies or the web in general). Moreover, automatic blind feedback or query expansion mechanisms are allowed to hopefully improve the proposed ranking.

4.1 Topic Creation

Based on the Europeana query logs, we have generated a set of 50 topics consisting of a mixture of topical and named-entity queries. The 50 short topics in title-format only (e.g., "królowie polscy w 18 wieku" – "Polish kings in 18 century") tend to reflect information needs as expressed by real Europeana users. To provide an overview of the topic meaning, we manually translated them into the English language. For each topic, an additional description was provided to give the relevance assessor an idea of what subjects were intended to be retrieved. This last field cannot be used during the search process. When inspecting the number of search keywords in the title section only, we can count 10 titles composed only by a single word, and 11 titles with two terms. On average, the topic contains 2.82 search keywords.

As this year Poland has celebrated the 150th anniversary of the January uprising, we have added topics related to Polish territories and history within the 18th and 19th centuries. There are also 8 topics on certain historical periods (e.g., "chłopi w 18 lub 19 wieku" – "peasants in 18 or 19 century") as well as 8 on temporary issues concerning Poland. 12 topics contain also personal names (e.g., "obrazy Jana Matejki" – "Jan Matejko's paintings"), but we also have 6 topics with geographical names (e.g., "kościoły w Toruniu" – "churches in Torun") or five with historical names (e.g., "Powstanie Styczniowe" – "January Uprising"). Finally, we can find 5 topics about religion or beliefs (e.g., "Matka Boża w sztuce" – "Our Lady in art"), and 7 on social groups or functions (e.g., "ruch robotniczy" – "workers movement").

4.2 Pooling and Relevance Assessments

Relevance assessments were done manually first by collaboratively generating an assumed information need for the topic and then describing it. The pooled documents (with a pool depth = 100, resulting in 32,144 judged documents) were then assessed

for their relevance according to the topic and the information need. This assumption is built around the perspective of an average user. We assumed that the majority of users typing that particular query would like to obtain that particular piece of information. Two experts have done the relevance assessments.

For this task, we have selected a three graded relevance value, with "fully relevant," "partially relevant," and "irrelevant". By default, we will opt for a strict interpretation assuming that only items judged "fully relevant" are judged relevant. The assessors have found 8,530 fully relevant CH objects. On the other hand, 4,758 CH objects have been judged as partially relevant to the corresponding query.

Fully relevant items can be found for every topic, with a minimum of 5 relevant CH objects for Topic #17 ("Czesław Miłosz"), and a maximum of 562 pertinent items for Topic#20 ("PRL" People's Republic of Poland). On average, we can find 170.6 relevant objects per topic (median: 125; stdev: 139.6).

Under the lenient option, we will consider as pertinent items judged fully or partially relevant. Under this condition, all topics have at least 22 relevant CH objects. This minimum value of 22 can be found for Topic#43 ("II Wojna Światowa" – "2nd World War") and the maximum of 562 pertinent items for Topic#3 ("medycyna w 19 wieku" – "medicine in 19 century"). On average, we can find 265.8 relevant objects per topic (median: 263; stdev: 132.2).

4.3 Participants and Results

From the 7 teams having expressed an interest in this task, we only obtained runs from 3 groups, namely 1 in the automatic mode, and 2 in the manual mode. We have also received request for information from 2 other teams in Poland but they were not able to send their runs in time. Table 4 shows the list of active participants.

Table 4. Polish Task 2013 Participating Groups and Country

Institute of Information Science and Book Studies, Nicolaus Copernicus University	Poland
Institute of Library and Information Science Institute, University of Wrocław	Poland
Computer Science Dept., University of Neuchâtel	Switzerland

When analyzing their results, we have considered mainly mean average precision (MAP), an evaluation measure corresponding to a user who wants to retrieve all pertinent CH objects. As a second measure, we have also reported P@10, a measure reflecting the result given by the Europeana search engine in its first result screen.

Automatic Runs

In this mode, our intent was to explore the best search strategy to automatically search within a morphologically rich language. As a general overview of the automatic runs, Table 5 depicts the main results together with their descriptions, ordered by MAP. The third row (PLWR0Base) corresponds to an automatic run submitted by the Torun's team [10] and used as a baseline for comparison for their manually enrichment query modifications. The University of Neuchatel (UniNE) sent the other runs [2]. To test for significant improvements, we applied a paired *t*-test. In our analysis,

statistically significant differences were detected by a two-sided test (α=5%) and are denoted by "†". There is no statistically significant difference between the first three runs.

Table 5. Strict Evaluation of Official Runs of the Automatic Mode

Rank	Name	Parameter Setting	MAP	P@10
1	UniNEFusion	Data fusion (Okapi: no stem, light stem, truc-5)	0.3433	0.614
2	UniNEDFR	DFR-I(n_e)B2, light stemming, with stopword	0.3308	0.568
3	PLWR0Base	Okapi, no stemming, with stopword	0.3140	0.552
4	UniNEPRF	Data fusion, PRF (Rocchio, 5 docs, 10 terms)	0.2578 †	0.494
5	UniNEBaseline	*tf idf* (cosinus), no stemming, with stopword	0.2566 †	0.492
6	UniNE-	Data fusion, 5-gram, PRF	0.2203 †	0.472

From the runs depicted in Rank#2, #3, and #5, we can see the performance differences achieved mainly when using the classical *tf*idf* IR model [11], the Okapi model [16] and 1 implementation of the DFR probabilistic paradigm [3]. The MAP of the DFR-I(n_e)B2 without stemming is 0.3028. Comparing the Okapi with the classical *tf*idf* model, we notice a relative improvement of +22.4% (from 0.2566 to 0.3140).

Additional runs presented by UniNE [2] indicate that indexing the CH objects with isolated words tends to perform better than either the *n*-gram of trunc-*n* indexing approaches. For example, the DFR-I(n_e)B2 based on the trunc-6 indexing scheme achieves a MAP of 0.3078 (or a MAP of 0.2641 for the 6-gram scheme). Using the same IR model with a light stemming (word-based), we can obtain a MAP of 0.3308 (see UniNEDFR in Table 5). Of course, in the CH domain where names can be an important source of evidence to discriminate between relevant and irrelevant objects, taking into account the short sequences of terms (e.g., "Jaroslaw city" instead of only "Jaroslaw" because this might also be a personal name) may hopefully improve these retrieval performances. The use of a stopword list also seems a good practice. Based on additional runs described in [2], the Okapi model with stemming and without a stopword list achieves a MAP of 0.3258. When applying a stopword list (composed of 304 terms), the MAP increases to 0.3433 (a relative improvement of +5.3%). Indexing the CH objects with the Europeana automatically enrichment tags (indicated by the prefix *europeana:*) does not have any impact of the retrieval effectiveness because only a few enrichment tags have been added in the Polish corpus.

An interesting question is to analyze the retrieval performance comparing the performance difference between different stemming strategies as well as the use of a lemmatizer. Based on the submitted runs, only a partial answer can be provided. The UniNE group has compared the use of a light stemmer (removing only the inflectional suffixes related to the gender, number and grammatical cases) with approaches ignoring this word normalization procedure. Based on the *tf*idf*, Okapi and DFR-I(n_e)B2 models, the mean relative improvement of applying a light stemmer is 5.3%.

The run "UniNEFusion" indicates the retrieval effectiveness when combining 2 word-based Okapi models (with and without a light stemming procedure) with an Okapi model based on trunc-5 indexing scheme (only the first 5 letters of each word are considered). This data fusion strategy does not seem to be really effective because

we have another run based only on the Okapi model that already obtains a MPA of 0.3433. The runs "UniNEPRF" and "UniNEGramPRF" were also based on a data fusion between runs using pseudo-relevant feedback. According to unofficial runs described in [2], this automatic query expansion does not result in better retrieval effectiveness. For example, adding 5 terms extracted from the first 5 top-ranked retrieved items (Rocchio's approach [11]) with the DFR-I(n_e)B2 changes the MAP from 0.3028 before the query expansion to 0.2189 (after a relative decrease of -27.7%).

Manual Runs

Within the manual mode, the participants are free to use any source of knowledge, tools, or strategies to modify and enrich the topics. No further user-system interaction is assumed after the first set of results is retrieved (but automatic blind feedback or query expansion mechanisms are allowed, although not used by the participants).

In Table 6, we have regrouped the evaluation of the official runs submitted in the manual mode, ordered by MAP. The run prefixed by the string "PLWR" comes from the Wroclaw University group [12] while those with the prefix "PLTO" are from the Torun group [10]. In both cases, the searchers have added a text description to semantically enrich the topic title. These additional terms were added under an "<enrich>" tag in the topic formulation. As depicted in Table 6, there is no statistically significant difference between the runs submitted by the Torun group. However, the retrieval performance differences are statistically significant between the best run (PLTO1EduLS) and all runs provided by the Wroclaw's group.

Table 6. Strict Evaluation of Official Runs of the Manual Mode

Rank	Name	Enrichment (Parameter Setting)	MAP	P@10
1	PLTO1EduLS	Educated, light stemmer	**0.2774**	0.454
2	PLTO1EduNO	Educated, no stemmer	0.2724	**0.460**
3	PLTO2HighLS	High, light stemmer	0.2709	0.528
4	PLTO2HighNO	High, no stemmer	0.2690	0.528
5	PLWR2Exp	Experts (Okapi, no stemming)	0.1795 †	0.378
6	PLWR1Edu	Educated (Okapi, no stemming)	0.1529 †	0.350
7	PLWR3Stu	Students (Okapi, no stemming)	0.1279 †	0.268
	PLWR0Base	Basic (Okapi, no stemming)	0.3140	0.552

The Torun group wants to compare the difference in retrieval performance that can be achieved when comparing "educated" users vs. "specialists". In the first case, the educated users have considered spelling variations, added other spellings for the same location or name or enriched the title by considering alternative formulations. With the specialists, the enrichment was based mainly on encyclopedias and a deeper elaboration of the main topic by including narrower terms (e.g., a list of writer names for a topic about "stories"). The educated users have added, on average, 3.3 terms, letting the mean length of the queries increase from 2.8 terms to 6.1 search keywords. With the specialists, this manual enrichment increases the mean topic length from 2.8 to 9.8 search terms.

As depicted in Table 5, these different forms of manual query enrichments do not improve the MAP over a simple search strategy using the title of the topic (run PLWR0Base). A first overview shows that mainly broad terms were added by the different user types and therefore the search system was not able to improve the ranking of the pertinent items. A query-by-query analysis reveals that the manual enrichment (PLTO1EduLS) improves the average precision (AP) for 22 queries over 50 compared to the automatic run (PLWR0Base). The largest improvement was obtained with the Topic #29 ("Warszawa w 19 wieku w sztuce" – "Warsaw in 19 century in art"). In this case, the AP increases from 0.001 (automatic run) to 0.3463, mainly by adding the terms "*architektura*" (architecture) and "*dzielnica*" (district). The specialists have also obtained a better retrieval performance for 20 topics over 50. The largest improvement was achieved with the Topic #32 ("kobiety w powstaniachi w wojsku" – "uprising or military and women") for which the MAP increases from 0.004 to 0.2825.

Moreover, the retrieval effectiveness of the various runs presented in Table 5 seems to indicate that applying a light stemming approach produces mixed results (see the performance difference between runs "*nnn*LS" and "*nnn*NO").

When analyzing Wroclaw's run, we can use the same search strategy (Okapi in this case) and baseline performance as with Torun. The manual query enrichment done by experts (run "PLWR2Exp") produces the best overall performance within this group. The performance difference with run "PLWR1Edu" is however not statistically significant (based on a paired *t*-test, two-sided, α=5%). With the students' run (run "PLWR3Stu"), the performance difference is larger (0.1795 vs. 0.1279, a relative difference of -28.7%), close to a statistically significant one (*p-value* = 0.0706).

As unofficial runs, the Torun team suggests that we can apply a Boolean search model [10]. In this approach, all keywords appearing in the title of the topic must be present in the retrieved items. With this model, they can achieve an MAP of 0.3484, the highest retrieval performance for this task. Of course this search strategy will not provide the best answer for all queries. An interesting example is Topic#24 ("Fryderyk Szopen" – "Fryderyk Chopin") that achieves an AP of 0.113 when using the Okapi search engine (PLWR0Base) but an AP of 0.996 (+881%) when using a Boolean search model. Clearly having both terms in the retrieved documents implies higher chance to be pertinent. However, such a Boolean strategy does not perform well in all cases. For example, with Topic #41 ("barok") the ranking provided by the Okapi model was better (AP: 6162) than that proposed by the Boolean model (AP: 0.004) based though on a single search keyword.

5 The CHiC Interactive Task

The intent of the CHiC*i* task was to collect a large enough data set that represented user interactivity with the Europeana collection so as to a) model user search/browse behaviour initially, and b) build a collection of user-centred data that might be augmented and used in future for testing various types of hypotheses about the process, the context and the nature of the interactivity. With that broad objective, the research

task focused on one user task: one with an implicit goal that reflects the exploratory nature of the interaction with culture and heritage information objects, particularly when the user is not an expert in the topic. As such it was designed to encourage interactivity and immersion in a culture and heritage environment, and the research design enabled multiple questions: what do people do when exposed to such an environment? How does the search process change over the course of that immersion? How do people interact with the images and their associated metadata? What can we learn from a user "session"? For this task, one common experimental system, one set of content and one interface was deployed and used by all teams [6].

5.1 Research Protocol, i.e., the Lab Task

The 'task' thus was a multi-part protocol that extracted multiple types of data from participants and observed participants virtually in their interactivity with the system. The protocol followed the pattern outlined in Fig. 1. All teams used the same protocol, which could be accessed remotely over the internet.

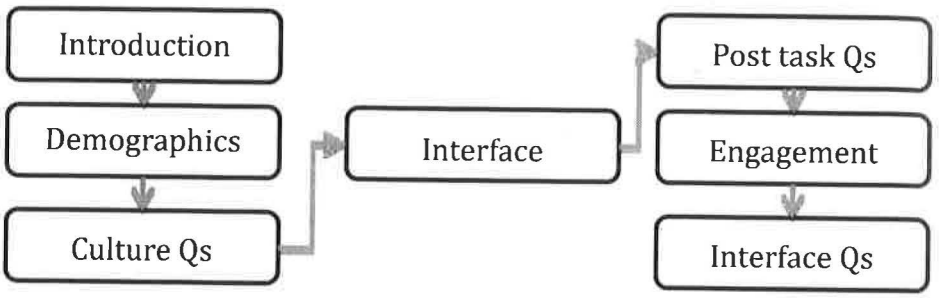

Fig. 1. CHiCi Research Protocol

An information sheet and informed consent (required by the University of Sheffield's research ethics review process) was first presented to participants, followed by sets of questions about:

- basic demographic questions to create a profile of participant group;
- country of birth and residence, mother tongue, and language used to speak at home or search the web, to understand the potential impact of an individual's culture;
- museum visits, familiarity and interest in European culture and heritage and experience with the European Digital Library, to address whether the participant was 'of convenience' or interested in the topic matter.

All of these may have influenced the level and intensity of their interaction with this resource. While participants were engaged in the assigned experimental task, the system logged and time stamped the entire set of user actions and events including:

queries, category selection, items examined, added to the bookbag, and so on. After the assigned task (see section 5.3), participants:

- responded to a 31-item User Engagement Scale to assess the overall experience;
- provided a narrative explanation of why they included the objects in the bookbag, and their level of satisfaction with what they found;
- assessed the usefulness of each object on the interface;
- assessed the usefulness of each piece of metadata in assisting with assessing an item.

5.2 IR System and Interface

The content contained 1,107,176 million records from the English-language collections of the Europeana Digital Library. The IR system was based on Apache Solr[3], which provides the text search, spelling checker, and the "more like this" suggestions. The default settings were used for all components and all fields specified in the source records were loaded without any pre-processing.

Access to the IR system was provided using a novel Cultural and Heritage Explorer (see Fig. 2); it offered three key ways of accessing the content and additional features intended to support the assigned task.

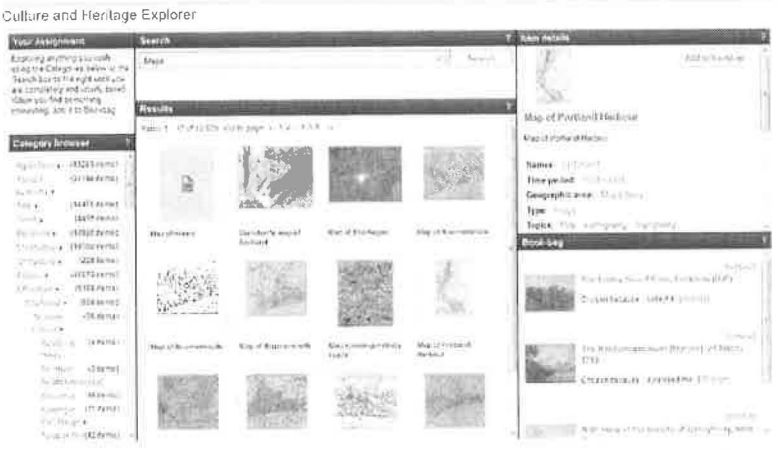

Fig. 2. CHiCi Cultural and Heritage Explorer

In addition, a hierarchical category browser was added, based on the work of [5]. This process resulted in a set of 24 top-level categories, with between 3 and 14 sub-levels (median 5). The individual levels in the category hierarchy had between 1 and 384 sub-categories (median 3). A total of 267,768 items were automatically mapped into the category hierarchy. When the item – category mappings were loaded into

[3] http://lucene.apache.org/solr/

Solr, each item was linked both to the category the pre-processing had linked it to, and to all of that category's ancestors. When the user selected a category from the category browser, the Solr index was searched for all items that were mapped to this category. Because of the way the items were linked not only to their category, but also to the category's ancestors, this query would also return all items that were linked to the selected category's descendants.

In addition to the task assignment in the upper left corner, the interface contained:

1) Category hierarchy: The hierarchy was navigated using the right arrow located to the right of each category, which expanded the level within the space.
2) Search box: a conventional implementation of query entry that accepted keywords. After submitting a query, the results display below the box was updated.
3) Results display: displayed 16 thumbnails and titles of the thumbnails in a 3x4 grid layout that also enabled navigation within the list. When an item in this display was clicked, it appeared in the item display to the extreme upper right.
4) Item display: contained the thumbnail and metadata fields associated with the item; unfortunately, only the thumbnail is present in the data collection. The metadata use the Dublin core standard, but some Dublin core labels used expert jargon and were modified for a naïve participant. At this point, an item could be added to the bookbag using the button in the upper right corner. At the bottom of each item, the "more like this" was displayed using thumbnail images.
5) Bookbag: used for collected images that were deemed useful. Items in the bookbag could be redisplayed or removed. The display included the item and the rationale for including the item as well.

On startup, no query was inserted, but the results grid was populated with randomly selected images to serve as a stimulus for starting the task. At that point, a participant could enter a query, scan the categories, examine the results or an individual item, or select from "more like this." At the item display, a participant could search by any of the metadata contents, or add an item to the bookbag. Once the "add to Bookbag" was selected, a popup box asked why the object was selected with the following options:

- I wanted to show someone
- I wanted to use the image in something
- I wanted to collect for a future purpose
- It surprised me!
- I simply liked it! No particular reason.

5.3 Experimental Task

The implicit task (which remained stationary in the upper left corner of the Explorer) was: "Your Assignment: exploring anything you wish using the Categories below or the Search box to the right until you are completely and utterly bored. When you find something interesting, add it to the Bookbag." Prior to being assigned the task,

participants were presented with a situation to set the stage for the task: "Imagine you are waiting to meet a friend in a coffee shop or pub or the airport or your office. While waiting, you come across this website and explore it looking at anything that you find interesting, or engaging, or relevant..." No further guidance was given, and participants were free to explore the resource; a mouse click on a 'Next Page' button disengaged the participant from the activity.

5.4 Research Teams

Four teams participated in this task, which required each team to process 30 participants via the web and 10 in a fixed observable lab-based location; not all participants met this objective as illustrated in Table 7. The language of operation was English, and all protocols and systems were expressed only in that language.

Table 7. Participating Research Teams

	Web	**Lab**	**Total**
Humboldt Universität	18	8	26
Royal School of Library and Information Science	12	19	31
Stockholm University	9	0	9
University of Sheffield	117	20	137
Other	4	1	5
Total	160	48	208

5.5 Participants

The participant group (208) contained a well-educated group of about 1/3 male (f=136, m=72), about 2/3 were under 35, and about half had undergraduate degrees, and all were currently enrolled in a programme of study. Participants came from 16 countries but more than half are residents in the UK, but originated, i.e., by birth, in 35 countries. 20 languages are spoken today, but they speak 26 languages at home. However, the predominant language is English, both as a mother tongue and as the current language spoken.

On a scale of 1 to 5 (from not familiar to very familiar), participants rated familiarity with European culture and heritage at 2.2, and their interest in the topic in the middle of the scale at 2.5. Of the participants, 78% indicated that they have never visited Europeana and 81% visited museums and galleries on the web or in person less than monthly. Thus, participants were dominated by well-educated, English-speaking and origin, females under 35 who were relatively non-expert in European culture and heritage and neither particularly interested or uninterested in the topic, and who primarily had never visited Europeana.

5.6 Results

From both user responses and the log files, we aggregated selected measures by participant. See Table 8 for that summary. Because data was collected in two types of locations: via the Web and in the Lab, we present data by location as it became apparent in preliminary analyses that there may be differences. But, because of the variation in size of the two location groups we are hesitant to say that these differences are statistically significant, and thus report the result and identify what looks suggestive (identified with an asterisk *).

Table 8. Summary Results across all participants

		Web		**Lab**		**Mean**	
Measure	**Definition**	#	*SD*	#	*SD*	#	*SD*
Queries	# of queries	3.5	*8.6*	5.3	*6.6*	3.9	*8.2*
Categories*	# of categories selected (hierarchy)	9.3	*11.3*	19.6	*22.8*	11.7	*15.3*
Metadata facets*	# of metadata facets examined	0.7	*2.1*	2.4	*6.4*	1.1	*3.6*
Query Time	Time (sec) spent querying	187.5	*600.4*	234.3	*253.1*	198.1	*541.2*
Category time*	Time (sec) spent using categories	239.2	*299.8*	493.0	*362.1*	296.8	*331.7*
Metadata time*	Time (secs) spent using metadata	22.8	*78.1*	65.7	*179.4*	32.5	*110.5*
Objects*	# of objects viewed	12.9	*16.7*	22.9	*18.4*	15.1	*17.6*
Objects (query)	# of objects viewed from query	5.4	*11.0*	7.7	*9.78*	5.92	*10.8*
Objects (categories)*	# of objects viewed from categories	5.7	*9.1*	13.2	*12.8*	7.4	*10.5*
Objects (metadata)	# of objects viewed from metadata	1.1	*5.4*	1.8	*6.0*	1.2	*5.5*
Interaction*	# of events/actions with system	57.1	*63.4*	97.1	*67.6*	66.2	*66.4*
Results page used*	# of results pages viewed	24.7	*36.2*	42.4	*41.8*	28.7	*38.2*
Bookbag	# of objects	6.0	*8.3*	4.5	*4.2*	5.7	*7.6*
Bookbag (category)	# of objects saved after category	2.9	*4.7*	2.5	*3.1*	2.8	*4.4*
Bookbag (metadata)	# of objects saved after metadata	0.3	*1.3*	0.3	*0.7*	0.3	*1.2*
Bookbag (query)	# of items in Bookbag after query	2.5	*5.8*	1.6	*2.5*	2.3	*5.2*

Table 8. (*Continued*)

Expected	Scale of 1-5, degree to which objects were as expected	1.54	0.977	1.94	1.099	1.63	1.017
Satisfied	Scale of 1-5, degree to which objects were as expected	1.74	1.119	1.92	1.145	1.78	1.125

As illustrated, participants issued on average approximately 4 queries, examined almost 12 categories, and about one of the metadata items associated with each object. They examined on average about 15 of the objects, with about 6 of those resulting from queries to the system and seven emerging from using the category explorer. Of these objects approximately 6 (50%) were deemed interesting enough to add to the Bookbag. On average they clicked on something on the interface 66 times, and clicked through the results pages 28 times. Overall, they were dissatisfied with what they found, and found the objects they examined not to be what they would have expected of Europeana.

In addition to understanding the effect of the interface, we also asked about the usefulness of each of the objects in the Explorer, but all were rated on the negative side on a five-point scale. Similarly, each object had a set of metadata associated with it, and of the set the Title, Description, and Thumbnail were considered to be useful in helping to assess the object with the title rated 2.8. Thus, in general neither the interface nor the details associated with each object were considered useful in exploring the content. There may be many reasons for this including the limited amount of information associated with an object and the very limited thumbnail associated with the original object.

Of all of the potential differences between their use in the Lab versus on the Web, most notable is no difference in terms of interesting objects saved. The differences appear at the level of interactivity – both in aggregate and in use of the Category Explorer, suggesting that being overseen in the lab may have changed their behavior, or doing the test off the web similarly gave them the anonymity that ensured participation without commitment. The individual lab studies in which people came into the lab should illuminate this issue.

The results presented here are descriptive and summary. What resulted from the work is a rich data set that contains both user response and log data. Unlike other tracks and/or tasks in which each lab uses the same data set to test multiple algorithms, this track *jointly* collected a data set using a common procedure and system which has resulted in a large data set that may now be used for multiple types of studies.

6 Conclusion and Outlook

The results of this year's CHiC lab 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 more narrowly 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 variations 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 from 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 intention of this second cycle is that the interactive experiment results of year one should inform system designers about which features are desirable for cultural heritage access and thus make it easier to focus development efforts into systems and interfaces. In a second year, any such developed system and interface features could be evaluated in more controlled interactive experiments. The ad-hoc retrieval tasks can benefit from the interactive task as well 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. This research was supported in part by the Sciex-NMS under Grant POL 11.219. 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. Last but not least, we would like to thank all participants (either in the lab or online) in the interactive study.

References

1. Agosti, M., Ferro, N.: Towards an Evaluation Infrastructure for DL Performance Evaluation. In: Tsakonas, G., Papatheodorou, C. (eds.) Evaluation of Digital Libraries: An Insight to Useful Applications and Methods, pp. 93–120. Chandos Publishing, Oxford (2009)
2. Akasereh, M., Naji, N., Savoy, J.: UniNE at CLEF – CHIC 2013. In: Proceedings CLEF 2013, Working Notes (2013)

3. Amati, G., van Rijsbergen, C.J.: Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. ACM Transactions on Information Systems 20, 357–389 (2002)
4. Dolamic, L., Savoy, J.: Indexing and Stemming Approaches for the Czech Language. Information Processing & Management 45, 714–720 (2009)
5. Fernando, S., Hall, M.M., Agirre, E., Soroa, A., Clough, P., Stevenson, M.: Comparing taxonomies for organising collections of documents. In: Proceedings of COLING 2012: Technical Papers, pp. 879–894 (2012)
6. Hall, M.M., Toms, E.: Building a common framework for IIR evaluation. In: CLEF 2013. LNCS, vol. 8138, pp. 17–28. Springer, Heidelberg (2013)
7. Hall, M., Villa, R., Rutter, S., Bell, D., Clough, P., Toms, E.: Sheffield Submission to the CHiC Interactive Task: Exploring Digital Cultural Heritage. In: Proceedings CLEF 2013, Working Notes (2013)
8. International Council of Museums, Scope Definition of the CIDOC Conceptual Reference Model (2003), http://www.cidoc-crm.org/scope.html
9. Larson, R.: Pseudo-Relevance Feedback for CLEF-CHiC Adhoc. In: Proceedings CLEF 2013, Working Notes (2013)
10. Malak, P.: The Polish Task within Cultural Heritage in CLEF (CHiC) 2013. Torun runs. In: Proceedings CLEF 2013, Working Notes (2013)
11. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
12. Pawlowski, A.: Polish Monolingual Task within Cultural Heritage in CLEF (CHiC) 2013. Wroclaw Runs. In: Proceedings CLEF 2013, Working Notes (2013)
13. 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)
14. Petras, V., Bogers, T., Ferro, N., Masiero, I.: CHiC Multilingual Task Overview and Analysis. In: Proceedings CLEF 2013, Working Notes (2013)
15. Popescu, A.: CEA LIST's participation at the CLEF CHiC 2013. In: Proceedings CLEF 2013, Working Notes (2013)
16. Robertson, S.E., Walker, S., Beaulieu, M.: Experimentation as a Way of Life: Okapi at TREC. Information Processing & Management 36, 95–108 (2000)
17. Skov, M., Bogers, T., Lund, H., Jensen, M., Wistrup, E., Larsen, B.: RSLIS/AAU at CHiC 2013. In: Proceedings CLEF 2013, Working Notes (2013)
18. Stiller, J., Gäde, M., Petras, V.: Ambiguity of Queries and the Challenges for Query Language Detection. In: CLEF 2010 LABs and Workshops (2010), http://clef2010.org/resources/proceedings/clef2010labs_submission_41.pdf (retrieved)
19. Swan Oscar, E.: Polish Grammar in a Nutshell, http://polish.slavic.pitt.edu/firstyear/nutshell.pdf
20. Tan, K., Almasri, M., Chevallet, J., Mulhem, P., Berrut, C.: Multimedia Information Modeling and Retrieval(MRIM)/Laboratoire d'Informatique de Grenoble (LIG) at CHiC 2013. In: Proceedings CLEF 2013, Working Notes (2013)
21. Tanase, D.: Using the Divergence Framework for Randomness: CHiC 2013 Lab Report. In: Proceedings CLEF 2013, Working Notes (2013)
22. 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)