Cultural Heritage in CLEF (CHiC) Overview 2012

Vivien Petras[1], Nicola Ferro[2], Maria Gäde[1], Antoine Isaac[3], Michael Kleineberg[1],
Ivano Masiero[2], Mattia Nicchio[2] and Juliane Stiller[1]

1Berlin School of Library and Information Science, Humboldt-Universität zu Berlin,
Dorotheenstr. 26, 10117 Berlin, Germany
{vivien.petras,maria.gaede,michael.kleineberg,
juliane.stiller}@ibi.hu-berlin,
2 Department of Information Engineering, University of Padova, Via Gradenigo 6/B, 35131
Padova, Italy
{ferro,masieroi,nicchio}@dei.unipd.it
3 Europeana, The Europeana Office, Koninklijke Bibliotheek, Prins Willem-Alexanderhof 5,
2595 BE Den Haag, Netherlands
aisaac@few.fu.nl

Abstract. The paper for the CHiC pilot lab describes the motivation, tasks, Europeana collections and topics, evaluation measures as well as the submitted and analyzed information retrieval runs. In its first year, CHiC offered three tasks: ad-hoc, which measured retrieval effectiveness according to relevance of the ranked retrieval results (standard 1000 document TREC output), variability, which required participants to present a list of 12 records that represent diverse information contexts and semantic enrichment, which asked participants to provide a list of 10 semantically related concepts to the one in the query to be used in query expansion experiments. All tasks were offered in monolingual, bilingual and multilingual modes. 126 different experiments from 6 participants were evaluated using the DIRECT system.

Keywords: cultural heritage, Europeana, variability, diversity, semantic enrichment

1 Introduction

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. Institutions in this domain have different approaches to managing information and serve diverse user communities, often with specialized needs and information contexts (native language, search environment, etc.).

Evaluation approaches (particularly system-oriented evaluation) in this domain have been fragmentary and often non-standardized. The CHiC 2012 pilot evaluation lab aimed at moving towards a systematic and large-scale evaluation of cultural heritage digital libraries and information access systems. The lab's goal is to increase our understanding on how to integrate examples from the cultural heritage community into a CLEF-style evaluation framework and how results can be fed back into the CH community.

The CHiC lab researches information retrieval systems for the cultural heritage environment by using real data, real user queries and real tasks. 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.

At the CLEF 2011 conference, a first workshop on information retrieval evaluation was put on by the organizers of the lab to discuss information needs, search practices and appropriate information retrieval tasks for this domain. The outcome of this workshop was a pilot lab proposal for the CLEF conference series suggesting three tasks relevant for cultural heritage information systems. Even as a pilot lab, CHiC was able to use real data and real search topics gathered from Europeana.

The paper is structured as follows: sections 2-4 explain the data collection, the preparation of topics and the CHiC tasks as well as the used evaluation measures. Sections 5 and 6 provide an overview of the participants and submitted experiments and describe the relevance assessment process. Section 7 discusses the experimental results, whereas section 8 provides an outlook for the next lab.

2 Collection

In March 2012, the complete Europeana data index was downloaded for collection preparation. The Europeana index as used in Europeana's Solr search portal contained 23,300,932 documents with a size of 132 GB.

Europeana data consists of metadata records describing digital representations of 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 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.

The original Europeana index contained fields from different schemas: Simple Dublin Core, e.g. dc:title, dc:description, Qualified Dublin Core, e.g. dcterms:provenance, dcterms:spatial and Europeana Semantic Elements, e.g. europena:type, europeana:isShownAt. On top of these schema-related fields, there were additional fields used internally in the Lucene index to improve search performance or to support specific application functionalities.

These fields were removed from the data collection and the index data was wrapped in a special XML format. The whole collection was then divided into 14 subcollections according to the language of the content provider of the record (which usually indicates the language of the metadata record). If all the provider languages had been used, the number of subcollections would have reached 30. Thus, in order to

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

reduce this amount, a threshold was set: all the languages with less than 100,000 documents were grouped together under the name “Other”.

The resultant 14 subcollections are listed in table 1. For the CHiC 2012 experiments, only the English, French and German subcollections as well as the entire collection were used.

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 data for all collections were made available and released to participants. Figure 1 shows an extract example record from the Europeana CHiC collection.

```
<ims:metadata
ims:identifier="http://www.europeana.eu/resolve/record/10105/5E1618BFAF072B8953B3070
1A6A6C3BB655ACF9D"
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>Rundun Zaqqu Bajda (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:isS
hownAt>
<europeana:language>en</europeana:language>
<europeana:provider>STERNA</europeana:provider>
<europeana:type>IMAGE</europeana:type>
```

```
<europeana:uri>http://www.europeana.eu/resolve/record/10105/5E1618BFAF072B8953B3070
1A6A6C3BB655ACF9D</europeana:uri>
</ims:fields>
</ims:metadata>
```

Fig. 1. Europeana CHiC Collection Sample Record

In the Europeana portal, object records commonly also contain thumbnails of the object if it is an image and links to related records. The thumbnails were not contained in the collection given to CHiC participants, but relevance assessors were able to look at them at the original source.

Finally, each file in the collection contained specific copyright information about the metadata record themselves and their providers. The XML code shown in Figure 2 was used for this purpose.

```
<dc:rights>The metadata contained in this file is made available by Europeana
(http://europeana.eu) only to the members of the Europeana Network
(http://pro.europeana.eu/about/network) that have agreed to use it for the research purposes of
the CLEF initiative (http://www.clef-initiative.eu). This usage falls within the more general
conditions of the Europeana Terms for Re-use of Europeana Metadata
(http://pro.europeana.eu/terms-of-use).</dc:rights>
```

Fig. 2. Copyright declaration XML code

3 Topics

For all experiments, original user queries were extracted from Europeana query logs. From all user search sessions in August 2010, those queries were extracted that resulted in a user viewing at least one complete object (in order to ensure that the session contained more than one user-system interaction). The queries were then further filtered to not include wildcards or automatically generated queries (for example by Europeana features).

Over 500 queries were then annotated according to their query category, i.e. topical, personal name, geographical name, work title or other. Queries could be either in the English language or ambiguous in language but would also appear in English. Ambiguous queries could include personal or location names that do not change across languages, e.g. William Shakespeare.

For CHiC, 50 queries were selected that covered a wide range of topics and represented a distribution of query categories that was found in a previous study [9]. For later relevance assessments, descriptions of the underlying information need 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 not clear. In this case, the research group discussed the query and agreed on the most likely information need. Figure 3 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. 3. CHiC English Example Query

All 50 queries were then translated into French and German. For the variability and semantic enrichment tasks, only the first 25 topics were used for the experiments.

4 CHiC Tasks

For the pilot lab of CHiC, three experimental tasks were selected that represented realistic use cases for cultural heritage information systems like Europeana but were also relatively simple in their set-up and to evaluate. The goal for this year's lab was to create baselines for topic and task development but also generate ground-truth in relevance assessments for experimental results.

All tasks were offered with the same set of topics and in three language modes: (i) monolingual (query and document language are the same), (ii) bilingual (query and document languages are different), (iii) multilingual (documents in multiple languages, i.e. the whole Europeana collection will be searched). This allowed the participants to experiment with a number of language variations (table 2).

Participants were asked to submit at least one monolingual experiment in any language per chosen task and were allowed to submit up to 4 experiments in the same language mode and combination.

Table 2. Language Modes for CHiC Experiments

Possible monolingual runs	DE → DE, EN → EN, FR → FR
Possible bilingual runs	X → DE, X → EN, X → FR, whereas X is a topic language the document language is not in
Possible multilingual runs	X → MUL, whereas X is either DE, FR or EN

4.1 Ad-hoc Information Retrieval

This task is a standard ad-hoc retrieval task, which measures information retrieval effectiveness with respect to user input in the form of queries. No further user-system interaction is assumed although automatic blind feedback or query expansion mechanisms are allowed to improve the system ranking. The ad-hoc setting is the standard setting for an information retrieval system - without prior knowledge about the user need or context, the system is required to produce a relevance-ranked list of documents based entirely on the query and the features of the collection documents.

Participants were allowed to use all collection fields and had to submit 1000 ranked documents (TREC-style) for relevance assessment.

4.2 Variability

A particular user type - the casual user or “information tourist” - does not follow the conventional pattern of a targeted information need being expressed in a targeted query but poses particular challenges for access or entry points and result presentation.

The variability task required systems to present a list of 12 objects (represents the first Europeana results page), which are relevant to the query and should present a particular good overview over the different object types and categories targeted towards a casual user, who might like the "best" documents possibly sorted into "must sees" and "other possibilities." This task is about returning diverse objects and resembles the diversity tasks of the Interactive TREC track or the CLEF Image photo tracks and other research [1], [5], [7-8], [11].

For CHIC, this task resembles a typical user of a cultural heritage information system, who would like to get an overview over what the system has with respect to a certain concept or what the best alternatives are. It is also a pilot task for this type of data collection using different assumptions about diversity or variability. Documents returned should be relevant but also as diverse as possible with respect to:

- media type of object (text, image, audio, video)
- content provider
- query category
- field match (which metadata field contains a query term)

Several approaches or measures have been suggested to measure diversity in an information retrieval result set [2-4], [6], [10], [12]. For the pilot variability task, we decided to measure cluster recall, i.e. the number of retrieved diverse categories (media type, content providers, query categories etc.) divided by the number of possible diverse categories per query. The evaluation of the results of this task was therefore two-fold. First, all returned documents were assessed for their relevance and then the cluster recall for relevant documents in the 4 categories above was determined.

4.3 Semantic Enrichment

Semantic enrichment is an important task in cultural heritage information systems with short and ambiguous queries like Europeana, which will support the information retrieval process either interactively (the user is asked for clarification, e.g. "Did you mean?") or automatically (the query is automatically expanded with semantically related concepts to increase the likely search success).

The semantic enrichment task required systems to present a ranked list of at most 10 related concepts for a query to semantically enrich the query and / or guess the user's information need or original query intent. For CHiC, this task resembles a typical user interaction, where the system should react to an ambiguous query with a clarification request (or a result output from an expanded query).

Related concepts could be extracted from Europeana data (internal information) or from other resources in the LOD cloud or other external resources (e.g. Wikipedia). Europeana already enriches about 30% of its metadata objects with *concepts, names* and *places* (included in the test collection). It uses the vocabularies GeoNames, GEMET and DBPedia for its included semantic enrichments, which could be explored further as well.

For the semantic enrichment task, participants could also use the Europeana Linked Open Data collections. Europeana released metadata on 2.5 million objects as linked open data in a pilot project[2]. The data is represented in the Europeana Data Model (RDF) and encompasses collections from ca. 300 content providers. Other external resources are allowed but need to be specified in the description from participants. The objects described in the LOD dataset are included in the Europeana test collection, but the RDF format might be convenient for accessing object enrichments.

System effectiveness was assessed in two phases. First all submitted enrichments were assessed manually for use in an interactive query expansion environment (e.g. "does this suggestion make sense with respect to the original query?").

During the second phase, the submitted terms and phrases were used in a query expansion experiment, i.e. the enrichments were added to the query and submitted as new experimental runs. All new topics were searched against the same standard Lucene indexes of the Europeana collections (according to the language of the enrichments). The results of those runs were then assessed according to ad-hoc retrieval standards.

5 CHiC Participation and Experiments

Although 21 groups registered for participation in CHiC, only 6 research groups submitted experimental results for evaluation. Table 3 shows the experiment participants for CHiC.

Table 3. CHiC 2012 Participating Groups and Country

Chemnitz University of Technology, Dept. of Computer Science	Germany
GESIS – Leibniz Institute for the Social Sciences	Germany
Unit for Natural Language Processing, Digital Enterprise Research Institute, National University of Ireland	Ireland
University of the Basque Country, UPV/EHU & University of Sheffield	Spain / UK
School of Information at the University of California, Berkeley.	USA
Computer Science Department, University of Neuchatel	Switzerland

Humboldt Universität (one of the organizers) also submitted experiments for assessment, which can be seen as baselines, because these multilingual ad-hoc runs used

[2] http://pro.europeana.eu/web/guest/linked-open-data

Europeana's Solr index to retrieve results. Two multilingual Europeana experiments were submitted, one using Solr's standard vector space ranking model, the other an adapted version of the BM-25 ranking model. Table 4 the number of experiments per task and language.

Table 4. CHiC Runs per Task and Language

	Language	Runs			Language	Runs
Ad-hoc				**Variability**		
Monolingual	English	17		Monolingual	English	8
	French	9			French	4
	German	8			German	4
Bilingual	X→English	8		Bilingual	X → English	4
	X→French	4			X → French	4
	X→ German	4			X → German	4
Multilingual		6		Multilingual		4
Semantic Enrichment						
Monolingual	English	17				
	French	4				
	German	8				
Bilingual	X→English	4				
	X→French	4				
	X→ German	4				
Multilingual		4				

5.1 The DIRECT System

DIRECT[3] (Distributed Information Retrieval Evaluation Campaign Tool) has supported the different stages of the CHiC evaluation activity, from the experiment submission phase to the relevance assessment and metrics computation. DIRECT manages different types of users, i.e. participants, assessors, organizers, and visitors, who need to have access to different kinds of features and capabilities. A personal username and password has been assigned to each participant/assessor [13].

6 Relevance Assessments

6.1 Pooling

The number of documents in large test collections such as CLEF makes it impractical to judge every document for relevance. Instead approximate recall values are calculated using pooling techniques. The results submitted by the groups participating in

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

the tasks are used to form a pool of documents for each topic and language by collecting the highly ranked documents from selected runs according to a set of predefined criteria. One important limitation when forming the pools is the number of documents to be assessed. Traditionally, the top 100 ranked documents from each of the runs selected are included in the pool; in such a case we say that the pool is of depth 100. This pool is then used for subsequent relevance judgments. After calculating the effectiveness measures, the results are analyzed and run statistics produced and distributed. The main criteria used when constructing the pools in CLEF are:

- favor diversity among approaches adopted by participants, according to the descriptions that they provide of their experiments;
- for each task, include at least one experiment from every participant, selected from the experiments indicated by the participants as having highest priority;
- ensure that, for each participant, at least one mandatory title+description experiment is included, even if not indicated as having high priority;
- add manual experiments, when provided;
- for bilingual tasks, ensure that each source topic language is represented.

This year, we produced three pools, one for each target language (English, French, and German) using a depth of 100. The pools have been created using all the runs in the ad-hoc monolingual and variability tasks, two runs per participant in the ad-hoc bilingual tasks, and all the runs in the bilingual variability task. A fourth pool, for the multilingual task, is the union of the three pools described above.

Table 5 provides details about the created pools, their size, the number of relevant and not relevant documents, and the pooled runs. You can note that English and French pools one run was not pooled from the monolingual tasks: this is a late arriving run, submitted after the closure of the submission phase.

Table 5. CHiC 2012 Pools

CHiC 2012 English Pool		
Size	Total documents	35,161
	Relevant documents	1,566
	Not relevant documents	33,595
	Topics with relevant documents / Total Topics	36 out of 50
	Assessors	8
Experiments	Pooled experiments / Submitted Experiments	32 out of 37
	ad-hoc monolingual	16 out of 17
	ad-hoc bilingual	4 out of 8
	variability monolingual	8 out of 8
	variability bilingual	4 out of 4
CHiC 2012 French Pool		
Size	Total documents	22,378
	Relevant documents	1,623
	Not relevant documents	20,755
	Topics with relevant documents / Total Topics	39 out of 50

	Assessors	2
Experiments	Pooled experiments / Submitted Experiments	18 out of 21
	ad-hoc monolingual	8 out of 9
	ad-hoc bilingual	2 out of 4
	variability monolingual	4 out of 4
	variability bilingual	4 out of 4
CHiC 2012 German Pool		
Size	Total documents	22,828
	Relevant documents	2,272
	Not relevant documents	20,556
	Topics with relevant documents / Total Topics	48 out of 50
	Assessors	2
Experiments	Pooled experiments / Submitted Experiments	18 out of 20
	ad-hoc monolingual	8 out of 8
	ad-hoc bilingual	2 out of 4
	variability monolingual	4 out of 4
	variability bilingual	4 out of 4

The box plot of Fig. **4** compares the distributions of the relevant documents across the topics of each pool for the different CHiC pools; the boxes are ordered by decreasing mean number of relevant documents per topic.

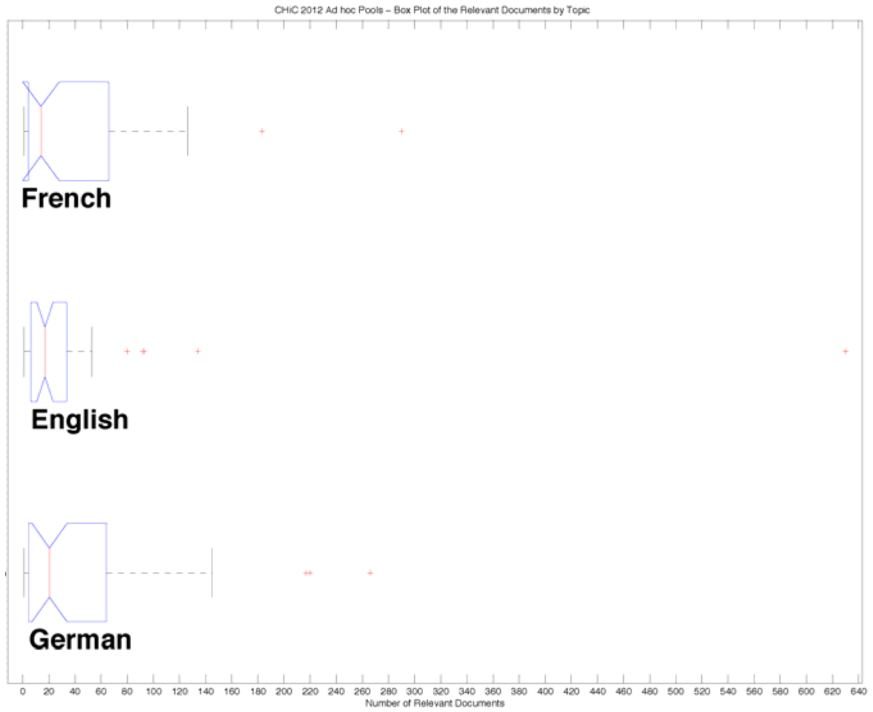


Fig. 4. Distribution of the relevant Documents across the CHiC Pools.

We see that the French and German distributions appear similar and are slightly asymmetric towards topics with a greater number of relevant documents whereas the English distribution is almost balanced. All the distributions show some upper outliers, i.e. topics with a greater number of relevant documents with respect to the behavior of the other topics in the distribution. These outliers are probably due to the fact that CHiC topics have to be able to retrieve relevant documents in all the collections; therefore, they may be considerably broader than typical monolingual topics.

6.2 Assessment Rules

During the relevance assessment phase, all eight assessors followed the same guidelines for relevance. Unclear or ambiguous cases were discussed within the group. A final validation by one of the organizers went through all relevant documents to check for consistency among the assessments.

The following general assumption guided the decision process: a record is relevant, when it fulfills the information need represented by the original query (in title) and by the suggested information need description (in description). Three relevance criteria were defined:

- Not relevant – the record does not fulfill the information need, the information is not relevant,
- Relevant – the record as represented in the DIRECT system fulfills the information need,
- Europeana relevant – the record only as represented in the Europeana portal fulfills the information need (only the whole Europeana record, i.e. the thumbnail and other related documents, contains enough information to make this object relevant, not just the record in the DIRECT system).

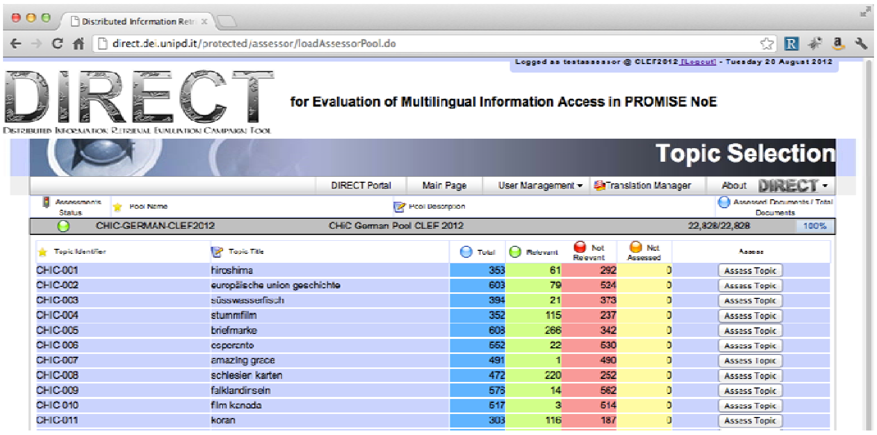
For the analysis, Europeana relevant and not relevant were counted as not relevant, the remaining documents as relevant.

6.3 The Assessment Interface

Figure 5 shows the main assessment interface of the DIRECT framework. It provides the assessor with an overview on the status of each pool. In particular, it displays the current number of relevance judgments for each topic in a specific pool.

The assessment stage is supported by the interface shown in Figure 6. The assessor can easily navigate through the list of document for a given topic. The interface includes a set of buttons to select relevance criteria for each document (yellow color for the not assessed documents, red for *not relevant* documents, green for *relevant* documents, grey for *Europeana relevant* documents). The document preview displays two direct links to:

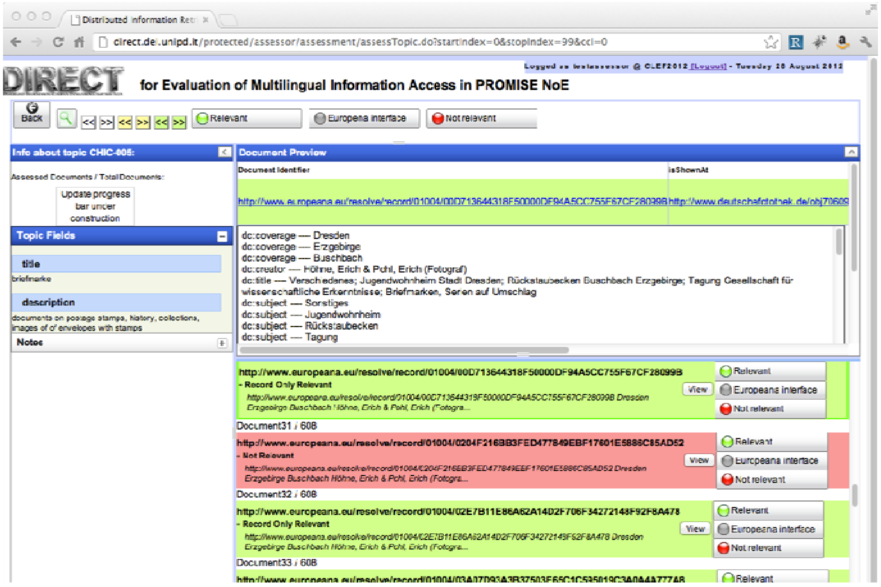
1. the original record in the Europeana website;
2. the content of the original europena_isShownAt field.


Distributed Information Retr
direct.dei.unipd.it/protected/assessor/loadAssessorPool.do
Logged as testassessor @ CLEF2012 [Logout] - Tuesday 20 August 2012
DIRECT
Distributed Information Retrieval Evaluation Campaign Tool
for Evaluation of Multilingual Information Access in PROMISE NoE
Topic Selection
DIRECT Portal
Main Page
User Management
Translation Manager
About DIRECT

Assessments Status	Pool Name	Pool Description	Assessed Documents / Total Documents	
●	CHIC-GERMAN-CLEF2012	CHIC German Pool CLEF 2012	22,828/22,828	100%

Topic Identifier	Topic Title	Total	Relevant	Not Relevant	Not Assessed	Assess
CHIC-001	hiroshima	353	61	292	0	Assess Topic
CHIC-002	europäische union geschichte	603	79	524	0	Assess Topic
CHIC-003	süsswasserfisch	394	21	373	0	Assess Topic
CHIC-004	stummfilm	352	115	237	0	Assess Topic
CHIC-005	briefmarke	608	266	342	0	Assess Topic
CHIC-006	esperanto	652	22	630	0	Assess Topic
CHIC-007	amazing grace	491	1	490	0	Assess Topic
CHIC-008	schlesier karten	472	220	252	0	Assess Topic
CHIC-009	falklandinseln	576	14	562	0	Assess Topic
CHIC-010	film kanada	517	3	514	0	Assess Topic
CHIC-011	koran	303	116	187	0	Assess Topic

Fig. 5. Main Assessment Interface in DIRECT


Distributed Information Retr
direct.dei.unipd.it/protected/assessor/assessment/assessTopic.do?startIndex=0&stopIndex=99&ccl=0
Logged as testassessor @ CLEF2012 [Logout] - Tuesday 28 August 2012
DIRECT
for Evaluation of Multilingual Information Access in PROMISE NoE
Back
<< >> << >> << >>
Relevant
Europena interface
Not relevant
Info about topic CHIC-005:
Assessed Documents / Total Documents:
Update progress bar under construction
Topic Fields
Notes
Document Preview
View
Relevant
Europeana interface
Not relevant

title
briefmarke

description
documents on postage stamps, history, collections, images of of envelopes with stamps

Document Identifier	isShownAt
http://www.europeana.eu/resolve/record/01004/00D713644318F50000DF94A5CC755F67CF28099B	http://www.deutschefotothek.de/obj70609

dc:coverage — Dresden
dc:coverage — Erzgebirge
dc:coverage — Buschbach
dc:creator — Höhne, Erich & Pohl, Erich (Fotograf)
dc:title — Verschiedenes; Jugendwohnheim Stadt Dresden; Rückstaubecken Buschbach Erzgebirge; Tagung Gesellschaft für wissenschaftliche Erkenntnisse; Briefmarken, Serien auf Umschlag
dc:subject — Sonstiges
dc:subject — Jugendwohnheim
dc:subject — Rückstaubecken
dc:subject — Tagung

http://www.europeana.eu/resolve/record/01004/00D713644318F50000DF94A5CC755F67CF28099B
- Record Only Relevant
http://www.europeana.eu/resolve/record/01004/00D713644318F50000DF94A5CC755F67CF28099B Dresden Erzgebirge Buschbach Höhne, Erich & Pohl, Erich (Fotogra...

Document31 / 608

http://www.europeana.eu/resolve/record/01004/0204F216BB3FED477849EBF17601E5886C85AD52
- Not Relevant
http://www.europeana.eu/resolve/record/01004/0204F216BB3FED477849EBF17601E5886C85AD52 Dresden Erzgebirge Buschbach Höhne, Erich & Pohl, Erich (Fotogra...

Document32 / 608

http://www.europeana.eu/resolve/record/01004/02E7B11E86A62A14D2F706F34272148F92F8A478
- Record Only Relevant
http://www.europeana.eu/resolve/record/01004/02E7B11E86A62A14D2F706F34272148F92F8A478 Dresden Erzgebirge Buschbach Höhne, Erich & Pohl, Erich (Fotogra...

Document33 / 608

http://www.europeana.eu/resolve/record/01004/03A07D93A3B37503F65C1C535019C3A0A4A777A8

Fig. 6. Assessment Interface in DIRECT

6.4 Semantic Enrichment Task

The semantic enrichment task results were first evaluated for the relevance or "semantic appropriateness" of the individual suggested terms or phrases. All enrichments for a query were looked at by the same assessor.

All submitted enrichments were assessed on a 3-point scale: definitely relevant as enrichment to the query, maybe relevant, and not relevant. If more than 10 suggestions were submitted, they were not included. If less than 10 suggestions were submitted, all suggestions were counted.

7 Results Analysis

7.1 Ad-hoc Information Retrieval

Monolingual Experiments.

Monolingual retrieval was offered for the following target collections: English, German, and French.

Table 6 shows the top five groups for each target collection, ordered by mean average precision. Note that only the best run is selected for each group, even if the group may have more than one top run. The table reports: the short name of the participating group; the experiment identifier; the mean average precision achieved by the experiment; and the performance difference between the first and the last participant.

Table 6. Best monolingual Experiments and Performance Difference between best and last (up to 5) Experiment (in MAP)

Track	Rank	Part.	Experiment Identifier	MAP
Monolingual English	**1st**	UPV	EXP_UKB_WN100	51.61%
	2nd	Chemnitz	QE0X20NO	48.60%
	3rd	Neuchatel	UNINEENEN1	44.87%
	4th	Gesis	GESIS_WIKI_ENTITY_EN_EN	43.96%
	5th	Berkeley	MONO_EN_TD_T2FB	36.40%
	Diff.			**41.78%**
Monolingual German	**1st**	Chemnitz	QE_NO	60.39%
	2nd	Gesis	GESIS_WIKI_ENTITY_DE_DE	54.80%
	Diff.			**10.20%**
Monolingual French	**1st**	Neuchatel	UNINEFRFR3	37.92%
	2nd	Chemnitz	QE_BO2_3D_10T	35.90%
	3rd	Berkeley	MONO_FR_TD_T2FB	20.85%
	Diff.			**81.87%**

Figures 7 to 9 show the interpolated recall vs. average precision for the top groups of the monolingual tasks.

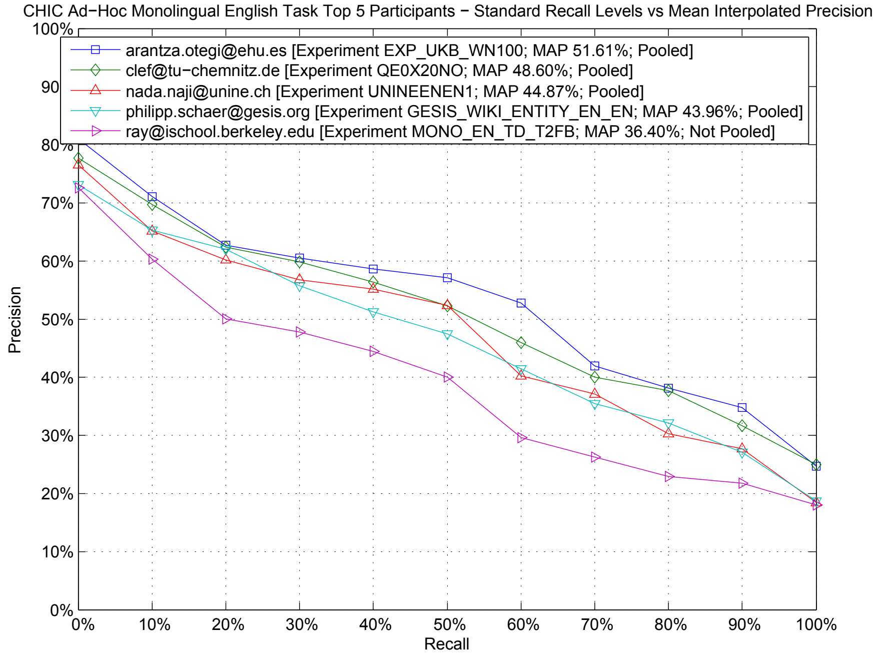


Fig. 7. Monolingual English Top Groups. Interpolated Recall vs. Average Precision

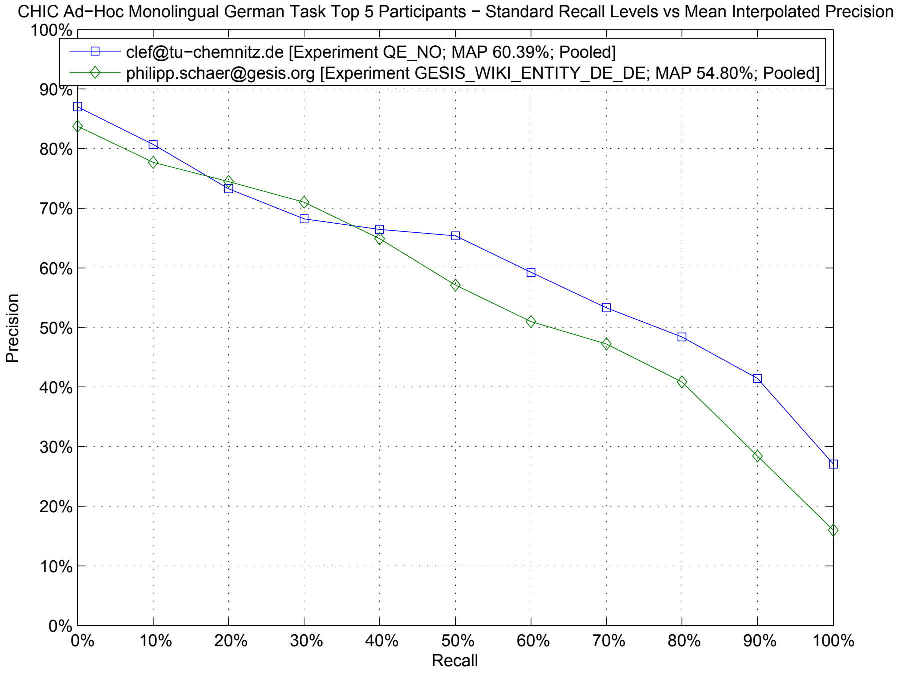


Fig. 8. Monolingual German Top Groups. Interpolated Recall vs. Average Precision

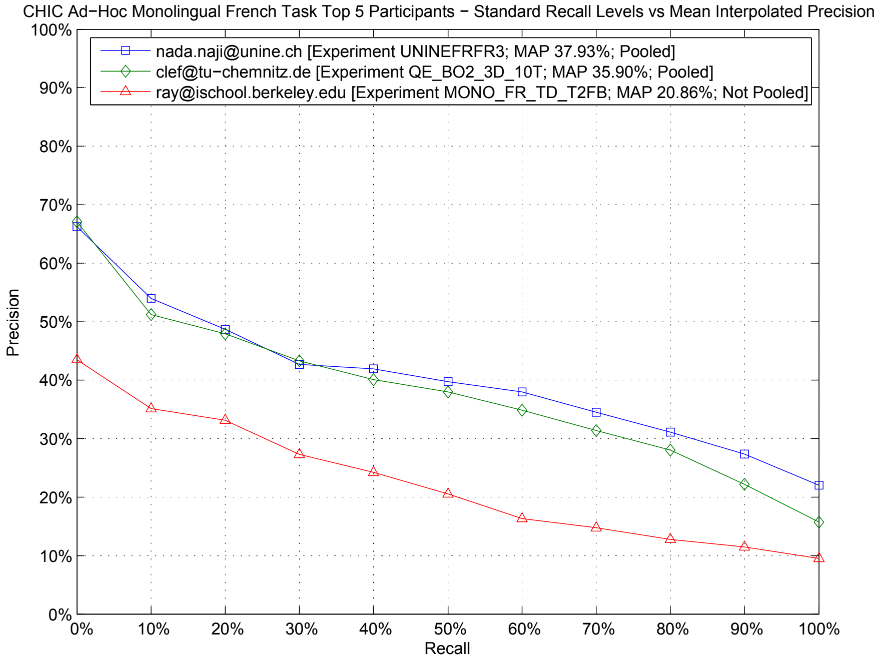


Fig. 9. Monolingual French Top Groups. Interpolated Recall vs. Average Precision

Bilingual Experiments.
The bilingual task was structured in three subtasks (X → DE, EN, or FR target collection). Table 7 shows the best results for this task. For bilingual retrieval evaluation, a common method is to compare results against monolingual baselines:

- X → EN: 86.40% of best monolingual English IR system
- X → DE: 63.52% of best monolingual German IR system
- X → FR: 81.32% of best monolingual French IR system

Table 7. Best bilingual Experiments and Performance Difference between best and last (up to 5) Experiment (in MAP)

Track	Rank	Part.	Experiment Identifier	MAP
Bilingual English	**1st**	Neuchatel	`UNINEDEEN3`	44.59%
	2nd	Chemnitz	`FR2EN_QE_DBPEDIA_SUBJECTS_MICROSOFT`	35.49%
	Diff.			**25.67%**
Bilingual German	**1st**	Chemnitz	`FR2DE_QE_DBPEDIA_SUBJECTS_MICROSOFT`	38.36%
	Diff.			-
Bilingual French	**1st**	Chemnitz	`DE2FR_QE_DBPEDIA_SUBJECTS_MICROSOFT`	30.84%
	Diff.			-

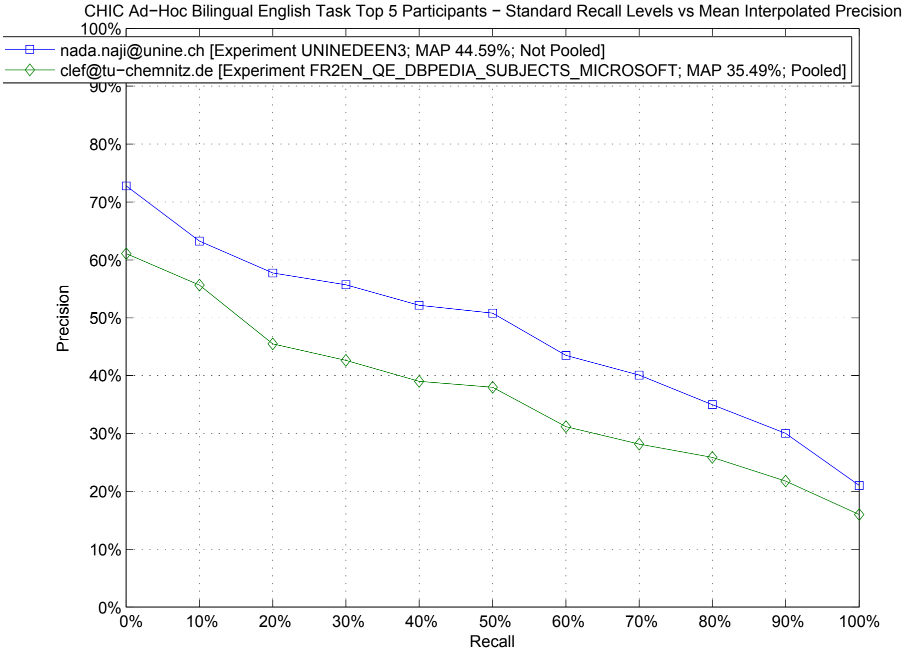


Fig. 10. Bilingual English Top Groups. Interpolated Recall vs. Average Precision

Figure 10 shows the interpolated recall vs. average precision graph for the top groups of the English bilingual tasks. Bilingual German and French had only one participant and are not shown here.

Multilingual Experiments.

Table 8 shows the best results for this task with the same logic of Table 6 and 7.

Table 8. Best Multilingual Experiments and Performance Difference between best and last (up to 5) Experiment (in MAP)

Track	Rank	Part.	Experiment Identifier	MAP
Multilingual	1st	Humboldt	`HUBEUNEW`	23.02%
	2nd	Chemnitz	`FR2X_DBPEDIA_SUBJECTS_MS`	13.33%
	Diff.			**72.61%**

Figure 11 shows the interpolated recall vs. average precision graph for the top participants of the multilingual task.

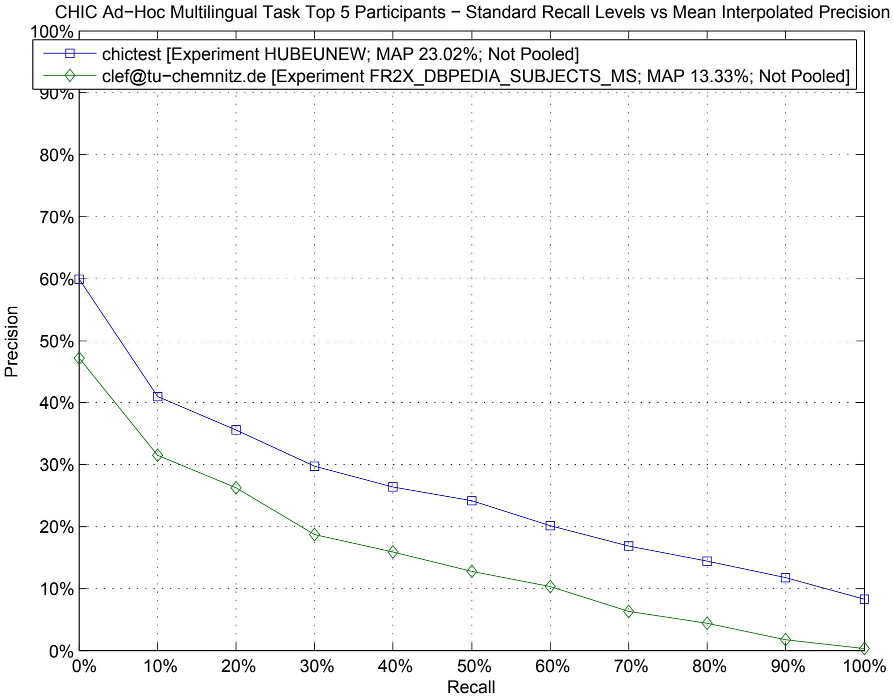


Fig. 11. Multilingual Top Groups. Interpolated Recall vs. Average Precision

7.2 Variability

Unfortunately, at the time of writing, the cluster recall analysis was not completed so that only the first phase evaluation results (retrieval effectiveness in finding relevant documents) can be shown.

For now, we report precision@5 and precision@15 values. Recall that participants were asked to submit 12 results for each query, representing a Europeana result page. The calculated p@15 measure comes closes to evaluating how many relevant documents were found even though it overdraws the boundaries of the precision@k. The corrected evaluation measures will be published on the CHiC website[4].

Monolingual Experiments.

Monolingual retrieval was offered for the following target collections: English, German, and French. Table 9 shows the best results for this task.

[4] http://www.culturalheritageevaluation.org

Table 9. Best monolingual Experiments and Performance Difference between best and last (up to 5) Experiment (mean of P@5 and P@15)

Track	Rank	Part.	Experiment Identifier	P@5	P@15
Monolingual English	**1st**	UPV	`SIMFACETS`	45.26%	28.42%
	2nd	Chemnitz	`QE_BO2_3D_10T`	27.36%	10.87%
	Diff.			**65.42%**	**161.45%**
Monolingual German	**1st**	Chemnitz	`QE_NO`	48.00%	22.93%
	Diff.			-	-
Monolingual French	**1st**	Chemnitz	`QE_DBPEDIA_SUBJECTS_QE`	27.82%	11.88%
	Diff.			-	-

Bilingual and Multilingual Experiments.
Only one group (Chemnitz) submitted results for these tasks, so Table 10 shows the best runs without the difference to other tasks. For bilingual retrieval evaluation, a common method is to compare results against monolingual baselines:

- Mean of P@5
  - X → EN: 17.83% of best monolingual English IR system
  - X → DE: 70.00% of best monolingual German IR system
  - X → FR: 87.49% of best monolingual French IR system
- Mean of P@15
  - X → EN: 74.06% of best monolingual English IR system
  - X → DE: 70.91% of best monolingual German IR system
  - X → FR: 73.14% of best monolingual French IR system

Table 10. Best bilingual and Multilingual Experiments (mean of P@5 and P@15)

Track	Rank	Part.	Experiment Identifier	P@5	P@15
Bilingual English	**1st**	Chemnitz	`FR2EN_QE_DBPEDIA_SUBJECTS_MICROSOFT`	21.05%	8.07%
Bilingual German	**1st**	Chemnitz	`FR2DE_QE_DBPEDIA_SUBJECTS_MICROSOFT`	33.60%	16.26%
Bilingual French	**1st**	Chemnitz	`EN2FR_QE_DBPEDIA_SUBJECTS_MICROSOFT`	24.34%	8.69%
Multilingual	**1st**	Chemnitz	`FR2X_DBPEDIA_SUBJECTS_MS`	19.20%	13.60%

7.3 Semantic Enrichment

We first report the overall results of the first phase evaluation of the semantic relevance (appropriateness) of the enrichments, then the overall results of the query expansion runs using the semantic enrichments.

Semantic Relevance.

For the evaluation of the “semantic appropriateness” of the suggested enrichments, two relevance measures were used - definitely relevant and maybe relevant – to be able to distinguish a strict and a relaxed evaluation. Precision (strong) is the average precision (over 25 queries) of "relevant" suggestions over all suggestions. Precision (weak) is the average precision (over 25 queries) of "relevant" and "maybe relevant" over all suggestions.

Table 11 shows average precision numbers (over all topics and all runs) for each language mode in this task. The weaker precision measure is, as should be expected, higher than the strict precision measure, by an average of 10 percentage points. The strict precision measure shows that on average about half of the suggested terms or phrases can be considered a good fit for the query.

German monolingual suggestions seem to have a lower precision than other experiments. The reason for this is that two experiments were submitted containing errors that would assign enrichments to the wrong queries after about half of the topics. We kept the experiments in the analysis for completeness, however.

Bilingual and multilingual experiments also seem to perform better than the monolingual experiments on average. This is probably due to averaging as most of the bilingual and monolingual runs were submitted by one group (Chemnitz Univ. of Techn.), which achieved higher results.

The detailed results for every run can be found on the CHiC website.

Table 11. Average Precision (over all 25 topics and all runs) for semantic Relevance of Enrichments

Run Mode	Language	Avg. Precision (weak)	Avg. Precision (strong)
Monolingual	English	0.6834	0.5470
	French	0.6120	0.5600
	German	0.6045	0.4721
Bilingual	X→English	0.7260	0.6390
	X→French	0.7010	0.6050
	X→ German	0.6970	0.6290
Multilingual		0.6960	0.5970

Monolingual Experiments.

Table 12 shows the best results for each group in this task.

Table 12. Best monolingual Experiments and Performance Difference between best and last (up to 5) Experiment (in Precision (weak and strong))

Track	Rank	Part.	Experiment Identifier	Precision (weak)	Precision (strong)
Monolingual English	1st	UPV	EHU.ES.UKBWIKI	0.8520	0.7520
	2nd	Gesis	WIKI_ENTITY_EN_EN	0.9240	0.7000
	3rd	Deri	DERI_SE1_CLEF-se	0.8000	0.6800
	4th	Chemnitz	CUT_T3_EN_EN_R4	0.7880	0.6520
	Diff.			**117.25%**	**115.34%**
Monolingual German	1st	Gesis	WIKI_ENTITY_DE_DE	0.8794	0.7448
	2nd	Chemnitz	CUT_T3_DE_DE_R1	0.7720	0.6080
	Diff.			**113.92%**	**122.49%**
Monolingual French	1st	Chemnitz	CUT_T3_FR_FR_R2	0.6240	0.5720
	Diff.				-

Bilingual and Multilingual Experiments.

Only one group (Chemnitz) submitted results for these tasks, so Table 13 shows the best runs without the difference to other runs.

Table 13. Best monolingual Experiments (in Precision (weak and strong))

Track	Rank	Part.	Experiment Identifier	Precision (weak)	Precision (strong)
Bilingual English	1st	Chemnitz	CUT_T3_DE_EN_R2	0.7680	0.6760
Bilingual German	1st	Chemnitz	CUT_T3_EN_DE_R1	0.8400	0.7600
Bilingual French	1st	Chemnitz	CUT_T3_EN_FR_R1	0.7920	0.6800
Multilingual	1st	Chemnitz	CUT_T3_FR_EN_DE_R2	0.7360	0.6440

Query Expansion.

Monolingual Experiments.

Monolingual retrieval was offered for the following target collections: English, German, and French. Table 14 shows the best results for this task. As can be seen, the original topic runs (without expansion) as denoted by the ORIGINALQUERIES identifier outperforms all other runs.

Table 14. Best monolingual Experiments and Performance Difference between best and last (up to 5) Experiment (in MAP)

Track	**Rank**	**Part.**	**Experiment Identifier**	**MAP**
Monolingual English	**1st**	Humboldt	`ORIGINALQUERIESEN-se`	34.11%
	2nd	Deri	`DERI_SE_CLEF_R1-se`	30.23%
	3rd	UPV	`UKBWIKI-se`	29.05%
	4th	Gesis	`GESIS_WIKI_ENTITY_EN_EN-se`	23.38%
	5th	Chemnitz	`CUT_T3_EN_EN_R1-se`	10.92%
	Diff.			**212.36%**
Monolingual German	**1st**	Humboldt	`ORIGINALQUERIESDE-se`	57.01%
	2nd	Gesis	`GESIS_WIKI_ENTITY_DE_DE-se`	31.92%
	3rd	Chemnitz	`CUT_T3_DE_DE_R3-se`	26.00%
	Diff.			**119.26%**
Monolingual French	**1st**	Humboldt	`ORIGINALQUERIESFR-se`	32.29%
	2nd	Chemnitz	`CUT_T3_FR_FR_R1-se`	14.67%
	Diff.			**120.10%**

Figures 12 to 14 show the interpolated recall vs. average precision for the top groups of the monolingual tasks.

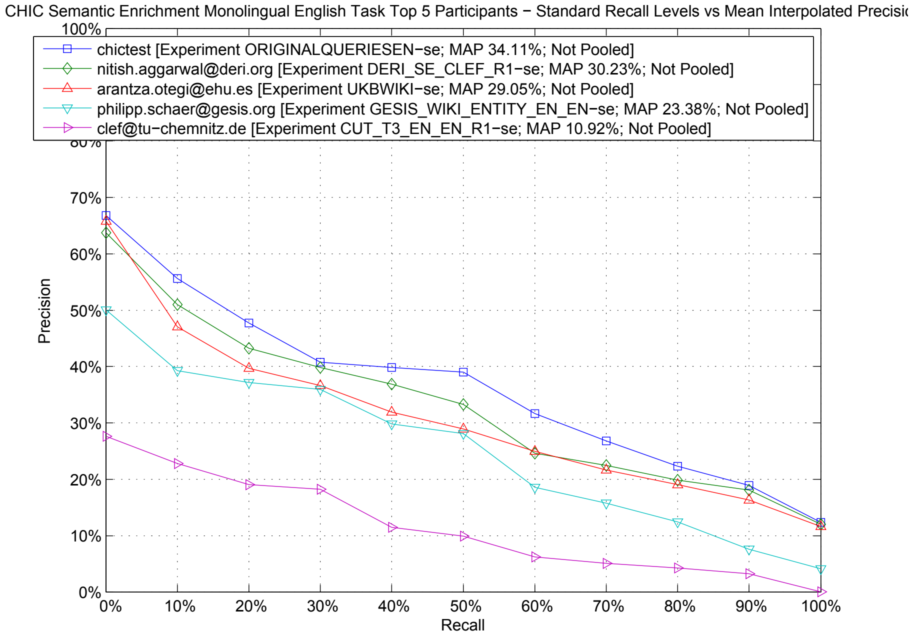


Fig. 12. Monolingual English Top Groups. Interpolated Recall vs. Average Precision

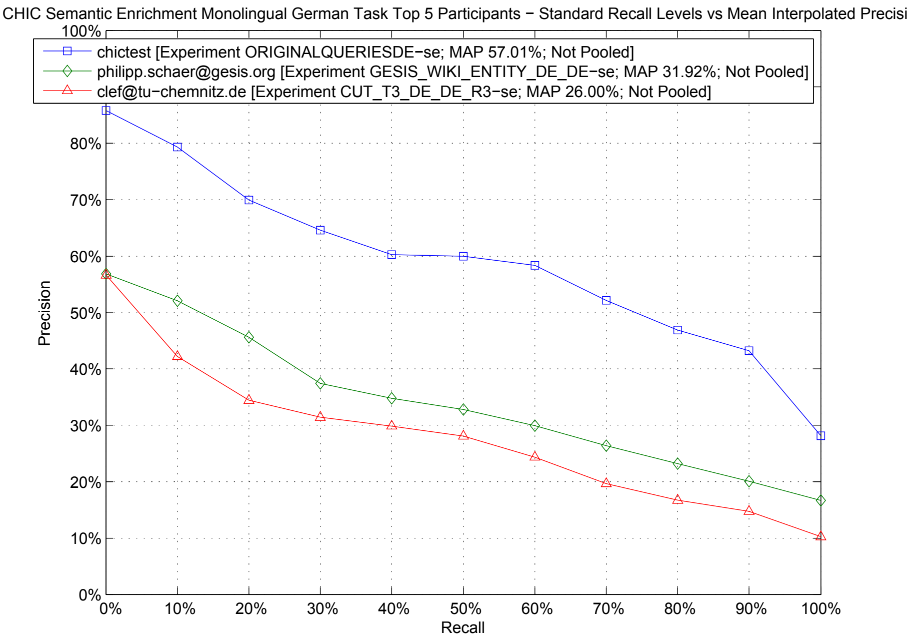


Fig. 13. Monolingual German Top Groups. Interpolated Recall vs. Average Precision

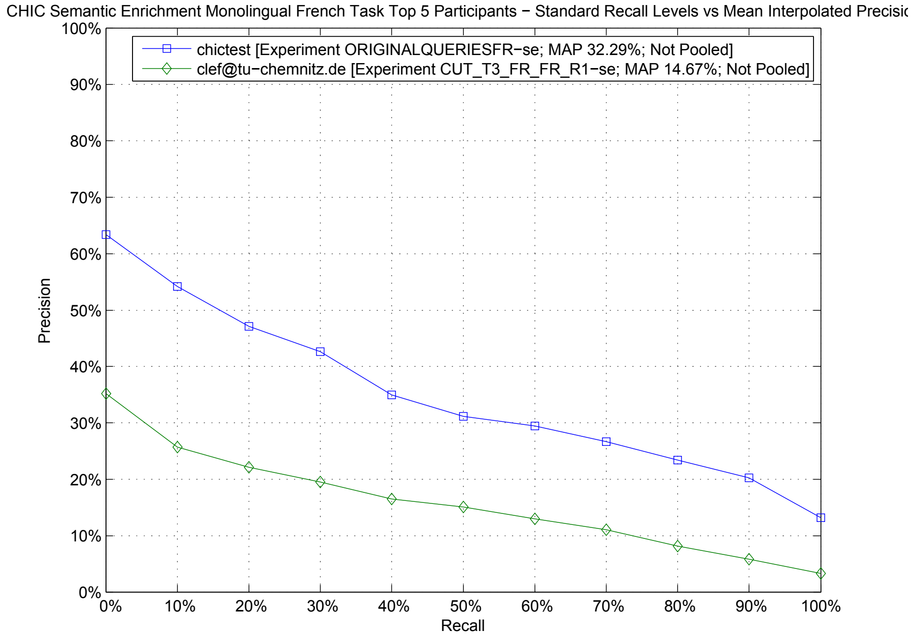


Fig. 14. Monolingual French Top Groups. Interpolated Recall vs. Average Precision

Bilingual and Multilingual Experiments.
Only one group (Chemnitz) submitted results for these tasks, so Table 15 shows the best runs without the difference to other runs.

Table 15. Best Bilingual and Multilingual Experiments (in MAP)

Track	Rank	Part.	Experiment Identifier	MAP
Bilingual English	1st	Chemnitz	`CUT_T3_DE_EN_R1-se`	13.12%
Bilingual German	1st	Chemnitz	`CUT_T3_FR_DE_R4-se`	00.00%
Bilingual French	1st	Chemnitz	`CUT_T3_EN_FR_R1-se`	19.13%
Multilingual	1st	Chemnitz	`CUT_T3_FR_EN_DE_R2-se`	6.14%

7.4 Approaches

Five groups submitted experimental results for the ad-hoc experiments, two groups for the variability task, and five groups submitted experiments for the semantic enrichment task. Most groups concentrated on the monolingual tasks (mostly English), only Chemnitz participated in all monolingual, bilingual and multilingual tasks.

For the ad-hoc task, most groups used open information retrieval systems like Cheshire, Indri, Lucene (in its Chemnitz Xtrieval implementation) and Solr. Many ranking algorithms were tested: vector space, language modeling, DFR and Okapi.

For translations in the bilingual and multilingual tasks, Google Translate, Wikipedia entries (with associated translations) and Microsoft's translation service were used.

For the variability task, Chemnitz used its ad-hoc retrieval implementation to retrieve results and then used the least recently used (LRU) algorithm to prioritize documents describing different media types from different providers. UPV used different document collection fields and two approaches for retrieving diverse results: using maximal-marginal relevance (MMR) to cluster results and then use cosine similarity to select the most dissimilar documents.

For the semantic enrichment task, the most often used external source for terms was Wikipedia at different levels of detail (article titles, first paragraph, full text). Wordnet and DBpedia (two groups) were also used. Gesis also used co-occurrence analysis to add related terms from the Europeana collection itself.

More details on methodologies and approaches can be found in the working papers of the individual groups.

8 Conclusion and Outlook

The results of this year's pilot CHiC lab have shown that working with data from the cultural heritage domain is possible but also poses many challenges due to the ambiguity of the users' information needs and the sparseness of the retrievable data. The preparation of new collections, the extraction of real queries and the organization of three realistic tasks with their respective evaluation measures was a challenge for organizers and participants, but it provided a lot of insight and more experience to continue this work in the next year.

After reviewing the tasks, their descriptions and the results, we believe that we can work on improving the current tasks by fine-tuning both the requirements and the evaluation measures (especially in the variability and semantic enrichment tasks). For 2012, we have only used three of the 14 language subcollections that were prepared and didn't put a lot of focus on the entire collection. Using the other collections to introduce more languages into the evaluation as well as putting more focus on the entire dataset (the actual use case for the Europeana portal) are both viable directions for additional instances of this lab.

Europeana is moving towards a linked data model for its objects[5] and one direction for this lab would be to combine experts from the information retrieval and linked data domains to research new retrieval approaches for this kind of data.

Finally, cultural heritage information systems are looking to incorporate more user interactions into their systems. The information retrieval evaluation field has often been criticized for viewing the viewer as outside of the scope of study. This domain and the available system (Europeana) enable us to combine and collaborate on information retrieval and information interaction research. CHiC is attempting to move towards this direction.

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 and assessment. Elaine Toms (University of Sheffield, UK) and Birger Larsen (Royal School of Library and Information Science, Copenhagen, Denmark) have shaped the lab's organization from the beginning and are instrumental in integrating more interactive features into the lab's tasks in further instances. We would like to thank our external relevance assessors Anthi Agoropoulou, Christophe Onambélé and Astrid Winkelmann.

References

1. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining, pp. 5–14. ACM, New York (2009)

[5] http://pro.europeana.eu/edm-documentation

2. Carbonell, J.G., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: SIGIR ’98, pp. 335-336. ACM, New York (1998)
3. Chen, H., Karger, D. R.: Less is more: probabilistic models for retrieving fewer relevant documents. In: SIGIR ’06, pp. 429-436. ACM, New York (2006)
4. Clarke, C. L.A., Craswell, N., Soboro, I.: Overview of the TREC 2009 Web Track. In: Voorhees, E. M., Buckland, L.P. (eds.) TREC 2009. NIST (2009)
5. Clarke, C. L. A., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: SIGIR ‘08, pp.659–666. ACM, NewYork (2008)
6. Over, P.: TREC-6 interactive track report. In: Voorhees, E. M., Harman, D.K. (eds.) TREC 1998, p.73. NIST (1998)
7. Sanderson, M.: Ambiguous queries: test collections need more sense. In: SIGIR ’08, pp. 499-506. ACM, New York (2008)
8. Sanderson, M., Tang, J., Arni, T., Clough, P.: What Else Is There? Search Diversity Examined. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR '09, pp. 562-569. Springer, Heidelberg (2009)
9. Stiller, J., Gäde, M., Petras, V.: Ambiguity of Queries and the Challenges for Query Language Detection. CLEF 2010 LogCLEF Workshop. In: Braschler, M., Harman, D., Pianta, E. (eds) CLEF 2010 Labs and Workshops Notebook Papers. Padua, Italy, 22-23 September 2010. (2010)
10. Voorhees, E. M.: Overview of the TREC 2004 robust retrieval track. In: Voorhees, E. M., Buckland, L.P. (eds.) TREC 2004. NIST (2004)
11. Xu, Y., Yin, H.: Novelty and topicality in interactive information retrieval. J. Am. Soc. Inf. Sci. Technol. 59(2), 201-215 (2008)
12. Zhai, C., Cohen, W., Lafferty, J: Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In: SIGIR ’03, pp. 10-17. ACM, New York (2003)
13. 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)