



# Report on Evaluation of Multilingual Information Access to Europeana

---

Task 2.5 – Sandbox integration, testing and evaluation of translation modules

co-funded by the European Union



The project is co-funded by the European Union, through the **eContentplus** programme  
<http://ec.europa.eu/econtentplus>



EuropeanaConnect is coordinated by the Austrian National Library



ECP-2008-DILI-528001

## EuropeanaConnect

### Report on Evaluation of Multilingual Information Access to Europeana

<b>Dissemination level</b>	<i>Public</i>
<b>Delivery date</b>	<i>XXX</i>
<b>Status</b>	<i>Draft</i>
<b>Author(s)</b>	<i>Nicola Ferro</i> <i>University of Padua (UNIPD)</i>



**eContentplus**

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



EuropeanaConnect is coordinated by the Austrian National Library

### Distribution

Version	Date of sending	Name	Role in project
0.1	02/12/2010	Nicola Ferro	Task Leader

### Approval

Version	Date approval	of	Name	Role in project
1.0				

### Revisions

Version	Status	Author	Date	Changes
0.1	D	University of Padua	02/12/2010	First draft circulated to Task 2.5 participants for comments
0.2	D	HUB	08/12/2010	workplan paragraph added

### Abstract

This report describes the evaluation activities that will be carried out for assessing multilingual information access components of Europeana.

It provides details about the evaluation paradigm adopted (laboratory evaluation according to the Cranfield methodology) and why this evaluation paradigm has been chosen.

Then, it will present the document collections that will be used taken from the catalogues of the British National Library, French National Library, and German National Library, as well as the topics that will be used to model the actual user information needs.

Afterwards, it discusses the evaluation tasks that will be carried out: monolingual tasks will analyse language resources when using queries against collections in the same language; bilingual tasks will investigate language resources and translation modules when querying in a language different from the one of the target collection. These evaluation tasks will give the possibility of getting a complete and exhaustive picture of the performances of the multilingual information access components of Europeana.

Finally, the report presents the infrastructure that will be used to manage and give access to all the experimental data, performance measures, and statistical analyses that will be produced.



## Table of Contents

<b>1. Introduction</b> .....	<b>5</b>
<b>2. Experimental Setup</b> .....	<b>6</b>
2.1. Document Collections .....	6
2.2. Topics.....	9
2.3. Relevance Assessments.....	11
<b>3. Evaluation Tasks</b> .....	<b>12</b>
<b>4. Experimental Data Management</b> .....	<b>13</b>
<b>5. Workplan</b> .....	<b>14</b>
<b>References</b> .....	<b>15</b>

## 1. Introduction

Large-scale evaluation campaigns provide qualitative and quantitative evidence over the years as to which methods give the best results in certain key areas, such as indexing techniques, relevance feedback, multilingual querying, and results merging, and contribute to the overall problem of evaluating a digital library system [Fuhr et al., 2007].

All the major evaluation initiatives at international level, such as the Text REtrieval Conference (TREC)<sup>1</sup> in the United States, the Cross-Language Evaluation Forum (CLEF)<sup>2</sup> in Europe, and the NII-NACSIS Test Collection for IR Systems (NTCIR)<sup>3</sup> adopt the Cranfield paradigm [Cleverdon, 1997], which makes use of *experimental collections*. An experimental collection is a triple  $C = (D, T, J)$ , where:  $D$  is a set of documents, called also collection of documents;  $T$  is a set of topics, which expresses the user's information needs from which the actual queries are derived;  $J$  is a set of relevance judgements, i.e. for each topic  $t \in T$  and for each document  $d \in D$  it is determined whether  $d$  is relevant to  $t$  or not.

An experimental collection  $C$  allows the comparison of information access systems according to some measurements, which quantify their performances. The main goal of an experimental collection is both to provide a common test-bed to be indexed and searched by information access systems and to guarantee the possibility of replicating the experiments.

The Cranfield paradigm is usually referred to as *laboratory evaluation* since it uses the experimental collection mechanism to abstract from the reality and allows for repeatable, reproducible, systematic, and comparable experiments. The other option is to conduct a *user-centered evaluation*, where real users are involved, they have to carry out some task, their actions may be recorded (e.g. system logs, videos, and so on) measured (e.g. completion time for a task) and they may have to fill in post-task questionnaires and interviews to describe their experience with a system.

The evaluation of multilingual information access components to Europeana will adopt a laboratory approach since there is no definitive user interface for accessing these functionalities yet and any intermediate interface may lead to not meaningful results<sup>4</sup>. Moreover, in this moment, we need a deep and systematic evaluation of each component, e.g. to compare its contribution with respect to other similar components, and this is possible only by means of a laboratory evaluation, due to the high number of possibilities and combinations to be tested. In a sense, we could think at laboratory evaluation as a kind of indispensable step to determine the "best configuration" of a system; afterwards, after removing all the not optimal configurations, we could think at a user-centered evaluation to assess the impact of the chosen solution.

In order to ensure comparability with existing literature and existing systems whose performances are known, we will make use of the CLEF collections developed for the Ad-hoc TEL Tasks in CLEF 2008 and CLEF 2009 [Agirre et al., 2009; Ferro and Peters, 2010]. This task offered monolingual and cross-language search on library catalogs. It was organized in collaboration with *The European Library* and used three collections derived from the catalogs of the British Library, the Bibliothèque Nationale de France, and the Austrian National Library. These collections contain catalog records expressed in an embryonal version of what then has become the Europeana Semantic Elements (ESE) and, thus, they are representative for what is currently used, before a full deployment of the newest Europeana Data Model (EDM).

---

<sup>1</sup> <http://trec.nist.gov/>

<sup>2</sup> <http://www.clef-campaign.org/>

<sup>3</sup> <http://research.nii.ac.jp/ntcir/>

<sup>4</sup> Note that an evaluation of the "European Portal" from the Web usability point of view has been proposed in the past and it has been decided to not carry it out for the same reason.

The report is organized as follows: Section 2 describes the experimental setup (documents collections, topics, relevance assessments); Section 3 describes the evaluation tasks that will be conducted to assess the multilingual information access components of Europeana; Section 4 explains the infrastructure that will be used for managing the experimental data and their analyses.

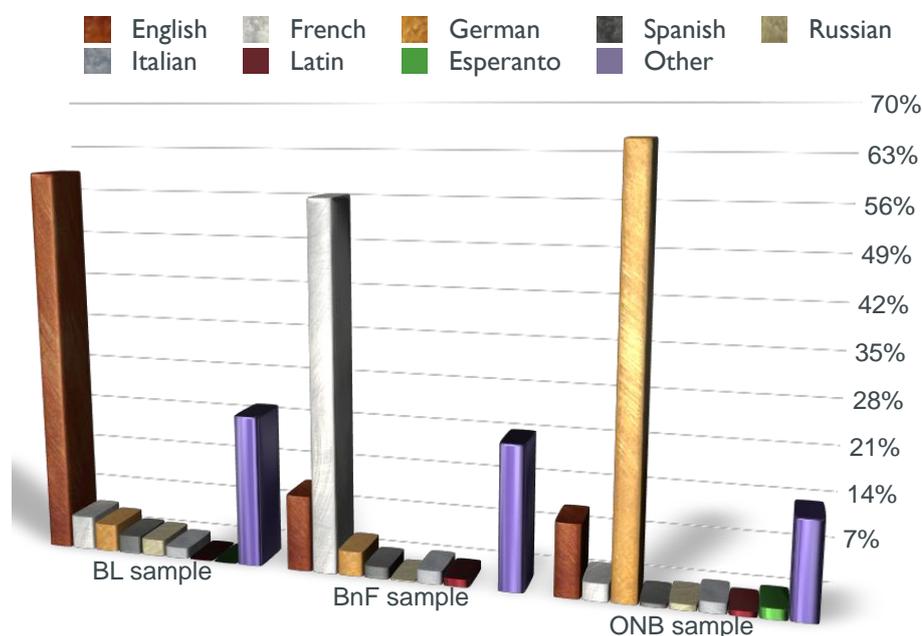
## 2. Experimental Setup

### 2.1. Document Collections

We use three collections:

- **British Library (BL):** 1,000,100 catalog records, 1.2 GByte of uncompressed XML;
- **Bibliothèque Nationale de France (BNF):** 1,000,100 catalog records, 1.3 GByte of uncompressed XML;
- **Austrian National Library (ONB):** 869,353 catalog records, 1.3 GByte of uncompressed XML.

We refer to the three collections (BL, BNF, ONB) as English, French and German because, in each case, this is the main and expected language of the collection. However, each of these collections is to some extent multilingual and contains documents in many additional languages; roughly speaking, about 60%-70% of the collections is in the “main language” and the remaining 30%-40% is in other languages as shows in Figure 1.

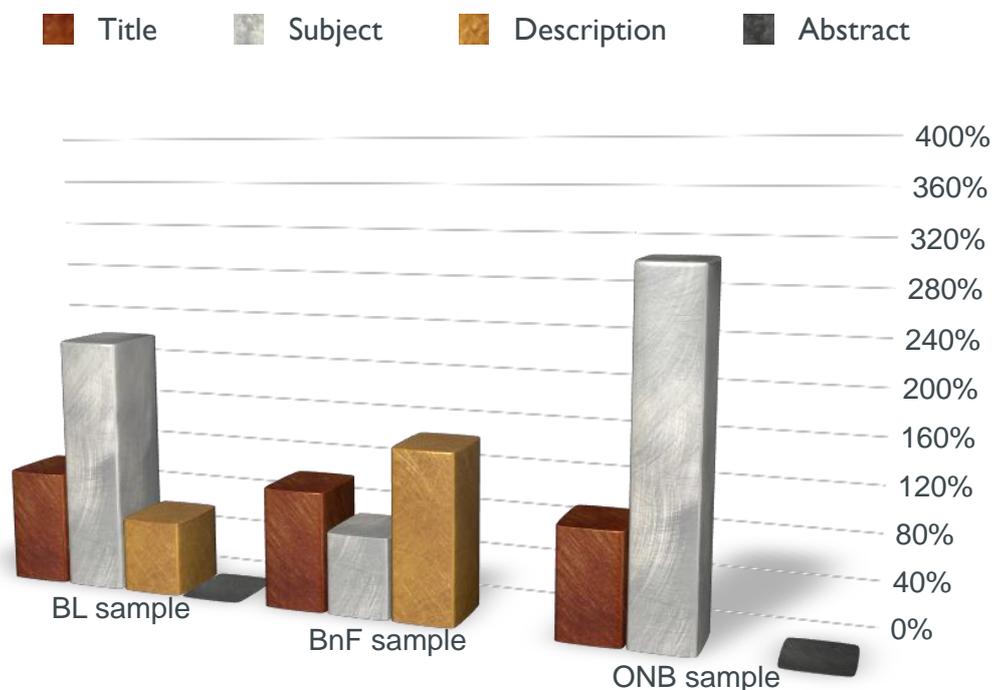


**Figure 1: Distribution of the languages in the BL, BnF, and ONB collections.**

Many records contain only title, author and subject heading information; other records provide more detail. The title and (if existing) an abstract or description may be in a different language to that understood as the language of the collection. The subject heading information is normally in the main language of the collection. About 66% of the documents in the English and German collection have textual subject headings, while only 37% in the French collection. Dewey Classification (DDC) is not



available in the French collection; negligible (<0.3%) in the German collection; but occurs in about half of the English documents (456,408 docs to be exact). Figure 2 shows the distributions of the records fields in the three collections.



**Figure 2: Distribution of the record fields in the BL, BnF, and ONB collections (percentages greater than 100% means that the field is repeated more than one in a record, on average).**

The following figures (Figure 3, Figure 4, and Figure 5) provide an example of the BL, BnF, and ONB records.

```

<record>
  <set>TEL_BL_opac</set>
  <header>
    <id>010734316</id>
  </header>
  <document format="index">
    <index>
      <topic>BL_opac</topic>
    </index>
  </document>
  <document format="dcx">
    <oai_dc:dc>
      <dc:title>Country wives </dc:title>
      <dc:contributor> Shaw, Rebecca, 1931- </dc:contributor>
      <dc:publisher>London : Orion, c2001</dc:publisher>
      <dc:terms:issued>c2001</dc:terms:issued>
      <dc:terms:extent>263 p. : 1 map ; 20 cm.</dc:terms:extent>
      <dc:language xsi:type="IS0639-2">eng</dc:language>
      <dc:description>Originally published: 2001.</dc:description>
      <dc:abstract>Country Wives is the second in a series of novels set in a veterinary practice on the outskirts of a Dorset town, by the bestselling author of the Turnham Malpas novels.</dc:abstract>
      <dc:subject xsi:type="dcterms:DDC">823.914</dc:subject>
      <dc:subject>Veterinarians</dc:subject>
      <dc:identifier xsi:type="lib:ISBN">0752844733 (pbk.) </dc:identifier>
      <dc:type>text</dc:type>
      <dc:identifier>
        <xlink:href="http://catalogue.bl.uk/F/?func=direct-doc-set& ; l_base=BLL01& ; from=TELgateway& ; doc_number=010734316">010734316</dc:identifier>
      <dc:identifier xsi:type="dcterms:URI">
        >http://catalogue.bl.uk/F/?func=direct-doc-set& ; l_base=BLL01& ; from=TELgateway& ; doc_number=010734316</dc:identifier>
      <mods:location>British Library HMNTS H.2002/3337</mods:location>
    </oai_dc:dc>
  </document>
</record>

```

Figure 3: Example of BL record.

```

<record>
  <set>TEL_BnF_opac</set>
  <id>oai:bnf.fr:catalogue/ark:/12148/cb346737691/description</id>
  <document format="index">
    <index>
      <topic>BnF_opac</topic>
    </index>
  </document>
  <document format="dcx">
    <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:dc="http://purl.org/dc/elements/1.1/">
      <dc:identifier>http://catalogue.bnf.fr/ark:/12148/cb346737691/description</dc:identifier>
      <dc:title>Chine : la terre, l'eau et les hommes / écrit par Han Suyin ; photographié par Claude Sauvageot</dc:title>
      <dc:creator>Han, Suyin (1917-....)</dc:creator>
      <dc:contributor>Sauvageot, Claude (1935-....). Illustrateur</dc:contributor>
      <dc:publisher>Édition J.A. (Paris)</dc:publisher>
      <dc:date>1980</dc:date>
      <dc:description>Collection : Grands livres</dc:description>
      <dc:subject xml:lang="fre"> Chine -- Histoire</dc:subject>
      <dc:subject xml:lang="fre"> Chine -- Ouvrages illustrés</dc:subject>
      <dc:description>ISBN 2852581973</dc:description>
      <dc:language>fre</dc:language>
      <dc:type xml:lang="fre">texte imprimé</dc:type>
      <dc:type xml:lang="eng">printed text</dc:type>
      <dc:type xml:lang="eng">text</dc:type>
      <dc:rights xml:lang="fre">Catalogue en ligne de la Bibliothèque nationale de France</dc:rights>
      <dc:rights xml:lang="eng">French National Library online Catalog</dc:rights>
    </oai_dc:dc>
  </document>
</record>

```

Figure 4: Example of BnF record.

```

<record>
  <set>TEL_ONB_onb01</set>
  <id>oai:aleph.onb.ac.at:ONB01-000628794</id>
  <document format="index">
    <index>
      <topic>ONB_onb01</topic>
    </index>
  </document>
  <document format="dcx">
    <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:dc="http://purl.org/dc/elements/1.1/"
      xmlns:dcterms="http://purl.org/dc/dc/terms/"
      xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
      <dc:identifier xsi:type="onb:ACCRecordId">AC03970200</dc:identifier>
      <dc:language xsi:type="dcterms:IS0639-2">ger</dc:language>
      <dc:creator>Salema, Teresa</dc:creator>
      <dc:title>&lt;&lt;Des&gt;&gt; Widerspenstigen Zähmung in der
        Gesellschaft Wilhelm Meisters: Ordnung der Natur oder Ironie der Kultur</dc:title>
      <dcterms:issued>1986</dcterms:issued>
      <dc:identifier xsi:type="onb:CallNumber">834301-B.NEU-Per.C,31</dc:identifier>
      <dcterms:bibliographicCitation>&lt;&lt;Der&gt;&gt; Widerspenstigen
        Zähmung</dcterms:bibliographicCitation>
      <dc:publisher xsi:type="onb:PlaceofPublisher">Innsbruck</dc:publisher>
      <dcterms:issued>1986</dcterms:issued>
      <dcterms:bibliographicCitation>(1986), S.143 - 156</dcterms:bibliographicCitation>
      <dc:subject>Romantik</dc:subject>
      <dc:subject>Feministische Literaturwissenschaft</dc:subject>
      <dc:subject>Klassik</dc:subject>
      <dc:subject>Frauenbild</dc:subject>
      <dc:subject>Weiblichkeitsideal</dc:subject>
      <dc:subject>Rollenbild</dc:subject>
      <dc:subject>Goethe, Johann Wolfgang &lt;&lt;von&gt;&gt;; Wilhelm
        Meisters Lehr- und Wanderjahre</dc:subject>
      <dcterms:abstract>Die feministische Literaturwissenschaft der letzten Jahre hat sich
        erfolgreich bemüht, die patriarchalisch orientierten Denk- und Handlungsstrategien
        aufzudecken, die seit der Aufklärung der Befreiung des Subjekts als Citoyen und
        Bourgeois zugrundeliegen. Daß eine solche "Bfreiung" sowohl mit der Negierung der
        Frau als Subjekt als auch mit der Unterdrückung von für weiblichgehaltenen Werten im
        Mann zusammenhängt, ist bisher und nicht nur aus feministischer Perspektive
        weitgehend bestätigt worden. Daher erscheint es sinnvoll, nach der Notwendigkeit
        einer solchen "Zähmung" grundsätzlich zu fragen, indem Alternativmodelle aus jener
        Zeit analysiert werden, die zwar einen Zivilisationsprozeß mit dazugehöriger
        Triebbändigung postulieren, dies jedoch nicht aus eindeutig sexistischer Sicht tun.
        Goethes "Wilhelm Meister" ist hier in doppelter Hinsicht ein positives
        Gegenbeispiel: seiner frauenfreundlichen Einstellung wegen, und auch weil die beiden
        Romane im Gegensatz zur Literatur der Romantik mit der Wirklichkeit in positiver
        Konfrontation stehen</dcterms:abstract>
    </oai_dc:dc>
  </document>
</record>

```

Figure 5: Example of ONB record.

## 2.2. Topics

Topics are structured statements representing information needs. Each topic typically consists of three parts: a brief **title** statement; a one-sentence **description**; a more complex **narrative** specifying the relevance assessment criteria.



For the evaluation, we use a common set of 100 topics in each of the 3 main collection languages (English, French and German). These topics are translated to all the 10 EuropeanaConnect languages, namely: English, French, German, Italian, Polish, Spanish, Portuguese, Swedish, Dutch and Hungarian.

Only the Title and Description fields are used in the evaluation because the narrative was prepared to provide information for the assessors on how the topics should be judged during CLEF. The topic sets were prepared on the basis of the contents of the collections, i.e. by interactively searching the collections to ensure the existence of relevant documents for each topic.

More in detail, when a task uses data collections in more than one language, we consider it important to be able to use versions of the same core topic set to query all collections. This makes it easier to compare results over different collections and also facilitates the preparation of extra topic sets in additional languages. However, it is never easy to find topics that are effective for several different collections and the topic preparation stage requires considerable discussion between the coordinators for each collection in order to identify suitable common candidates. The sparseness of the data makes this particularly difficult for the TEL task and leads to the formulation of topics that are quite broad in scope so that at least some relevant documents could be found in each collection. Figure 6 shows an example of a topic.

```

<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<topic>
  <identifier>10.2452/711-AH</identifier>

  <title lang="zh">深海生物</title>
  <title lang="en">Deep Sea Creatures</title>
  <title lang="fr">Créatures des fonds océaniques</title>
  <title lang="de">Kreaturen der Tiefsee</title>
  <title lang="el">Πλάσματα στα βάθη των ωκεανών</title>
  <title lang="it">Creature delle profondità oceaniche</title>

  <description lang="zh">
    查找有关世界上任何深海生物的出版物。
  </description>
  <description lang="en">
    Find publications about any kind of life in the depths
    of any of the world's oceans.
  </description>
  <description lang="fr">
    Trouver des ouvrages sur toute forme de vie dans les
    profondeurs des mers et des océans.
  </description>
  <description lang="de">
    Finden Sie Veröffentlichungen über Leben und
    Lebensformen in den Tiefen der Ozeane der Welt.
  </description>
  <description lang="el">
    Αναζήτηση δημοσιεύσεων για κάθε είδος ζώης στα
    βάθη των ωκεανών
  </description>
  <description lang="it">
    Trova pubblicazioni su qualsiasi forma di vita nelle
    profondità degli oceani del mondo.
  </description>
</topic>

```

Figure 6: Example of topic.

### 2.3. Relevance Assessments

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 the ad hoc 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. 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 stability of pools constructed in this way and their reliability for post-campaign experiments is discussed in [Braschler, 2003].

The main criteria used when constructing the pools in CLEF are:

- favour 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;
- or bilingual tasks, ensure that each source topic language is represented.

For the CLEF 2008 ad hoc test collections, [Tomlinson, 2009] reported some sampling experiments aimed at estimating the judging coverage. He found that this tended to be lower than the estimates he produced for the CLEF 2007 ad hoc collections. With respect to the TEL collections, he estimated that at best 50% to 70% of the relevant documents were included in the pools – and that most of the unjudged relevant documents were for the 10 or more queries that had the most known answers.

These discussions show how complex the creation of relevance judgements is and how much care is devoted to ensure that they are reliable and robust. The net results, from our evaluation point of view, is that the CLEF relevance judgements are fair also for systems that have not participated in the CLEF campaigns, as it is the case for Europeana, and can provide third-party assessment.

### 3. Evaluation Tasks

For Europeana we will evaluate the following:

- **monolingual tasks** where the language of the source query is the same as that of the target collection, for example an English query against an BL collection;
- **bilingual tasks** where the language of the source query is different from that of the target collection, for example a Dutch query against an BL collection.

Monolingual tasks will offer the possibility of assessing the performances provided by different language resources. For each of the three target collections, we will provide a baseline run and then evaluate the different language resources available in the Europeana language resources repository.

In order to do that, CELI will provide a standard information retrieval system, where all the components will be kept fixed except for the language resource under testing.

In the end, it will be possible to create a matrix, as that reported below, where performance scores are associated to each language resource. If it is a language-independent resource, such as a statistical stemmer, it will have performance scores associated for each of the three target languages; if it is a language-dependent resource, it will have performance scores associated only for its language.

Resource	English (BL)	French (BnF)	German (ONB)
Resource 1	xx.xx%	yy.yy%	zz.zz%
Resource 2	xx.xx%	–	–
Resource 3	–	yy.yy%	–
...	...	...	...

Bilingual tasks will offer the possibility of evaluating the translation modules together with their interaction with language resources (both monolingual ones, such as stemmers, and bilingual ones, such as dictionaries).

For each of the three target collections, we will provide a baseline run and then evaluate the different (translation module, language resource) pairs with respect to the ten EuropeanaConnect languages. In order to do that, CELI will provide a standard information retrieval system, where all the components will be kept fixed except for the (translation module, language resource) under testing.

In the end, it will be possible to create a multi-level matrix, as that reported below, where for each target language all the ten source languages will be evaluated. Then, for each (source language, target language) pairs, different translation modules will be evaluated. Finally, for each (source



language, target language, translation module) triple, different languages resources will be evaluated. This will give the possibility of an accurate comprehension of the interaction of the different translation modules, language resources, and languages.

Source	Target English (BL)						Target French (BnF)						Target German (ONB)					
	it	de	fr	es	po	...	it	de	fr	es	po	...	it	de	fr	es	po	...
<b>Module 1</b>																		
Resource 1																		
Resource 2																		
<b>Module 2</b>																		
Resource 2																		
Resource 3																		
<b>Module 3</b>																		
Resource 4																		
Resource 5																		
....																		

Moreover, exploiting the monolingual runs, for each of the combinations in the above matrix, we will be able to compute the best bilingual to best monolingual ratio, which gives an idea of the relative performances of a system when passing from one language to another.

## 4. Experimental Data Management

The experimental data and all the performance measures will be managed by means of the DIRECT system<sup>5</sup> [Agosti and Ferro, 2009] which has been developed and adopted since CLEF 2005. DIRECT manages all the aspects of an evaluation campaigns, provides access and curates all the experimental data, performance measure, and statistical analyses produced during evaluation. Figure 7 provides an overview of some of the functionalities offered by DIRECT.

---

<sup>5</sup> <http://direct.dei.unipd.it/>

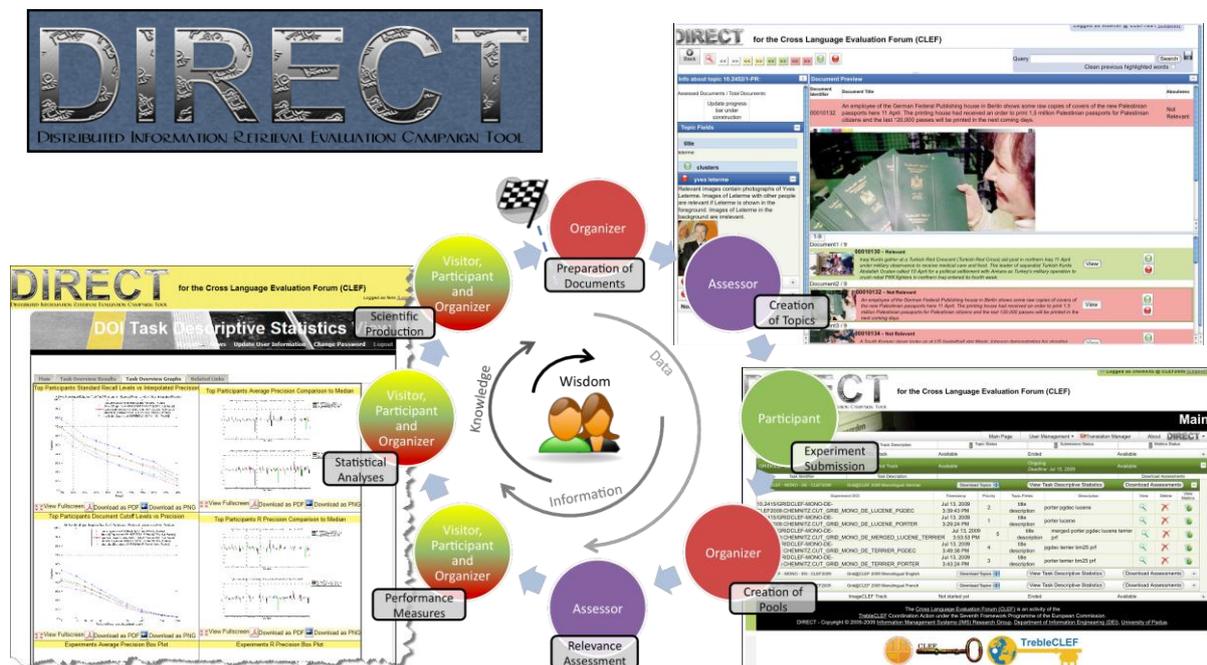


Figure 7: Overview of some of the DIRECT functionalities.

Moreover, DIRECT keeps the data produced in the last ten years of CLEF [Agosti et al., 2010]. This will give us the opportunity to compare the evaluation results for the multilingual information access components of Europeana with the research systems evaluated in CLEF, thus comparing them with the current state of the art.

## 5. Workplan

### 1. Table containing languages and components

- Available at LifeRay:

Partners: UBER

### 2. Make TEL data (queries, collections) available

- Provided 100 topics in French, German and English
- 50 of these were already translated into Dutch, Italian and Spanish

Partner: UNIPD

Deadline: October 2010

### 3. Translation of TEL queries

- Translation of the remaining 50 topics into Dutch, Italian, Spanish
- Translation of the whole set of topics (100) into Hungarian, Portuguese, Swedish

Partners: UBER

Deadline: January 2010

#### 4. Concept for evaluation (paper)

Partner: UNIPD

Deadline: December 2010

#### 5. Test set up (workflow)

- See description above

Partner: UNIPD

Deadline: January 2011

#### 6. Finalization of Prototype

Partner: CELI

Deadline: December 2010

#### 7. Report on Evaluation activities

Partner: all

Deadline: June 2011

## References

[Agirre et al., 2009]	Agirre, E., Di Nunzio, G. M., Ferro, N., Mandl, T., and Peters, C. (2009). CLEF 2008: Ad Hoc Track Overview. In Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G. J. F., Kurimo, M., Mandl, T., and Peñas, A., editors, <i>Evaluating Systems for Multilingual and Multimodal Information Access: Ninth Workshop of the Cross-Language Evaluation Forum (CLEF 2008). Revised Selected Papers</i> , pages 15–37. Lecture Notes in Computer Science (LNCS) 5706, Springer, Heidelberg, Germany.
[Agosti et al., 2010]	Agosti, M., Di Nunzio, G. M., Dussin, M., and Ferro, N. (2010). 10 Years of CLEF Data in DIRECT: Where We Are and Where We Can Go. In Sakay, T., Sanderson, M., and Webber, W., editors, <i>Proc. 3rd International Workshop on Evaluating Information Access (EVIA 2010)</i> , pages 16–24. National Institute of Informatics, Tokyo, Japan.
[Agosti and Ferro, 2009]	Agosti, M. and Ferro, N. (2009). Towards an Evaluation Infrastructure for DL Performance Evaluation. In Tsakonias, G. and Papatheodorou, C., editors, <i>Evaluation of Digital Libraries: An insight into useful applications and methods</i> , pages 93–120. Chandos Publishing, Oxford, UK.
[Braschler, 2003]	Braschler, M. (2003). CLEF 2002 – Overview of Results. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, <i>Advances in Cross-Language Information Retrieval: Third Workshop of the Cross-Language Evaluation Forum (CLEF 2002) Revised Papers</i> , pages 9–27. Lecture Notes in Computer Science (LNCS) 2785, Springer, Heidelberg, Germany.
[Cleverdon,	Cleverdon, C. W. (1997). The Cranfield Tests on Index Languages Devices. In

1997]	Spärck Jones, K. and Willett, P., editors, <i>Readings in Information Retrieval</i> , pages 47–60. Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA.
[Ferro and Peters, 2010]	Ferro, N. and Peters, C. (2010). CLEF 2009 Ad Hoc Track Overview: TEL & Persian Tasks. In Peters, C., Di Nunzio, G. M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., and Roda, G., editors, <i>Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments – Tenth Workshop of the Cross–Language Evaluation Forum (CLEF 2009). Revised Selected Papers</i> , pages 13–35. Lecture Notes in Computer Science (LNCS) 6241, Springer, Heidelberg, Germany.
[Fuhr et al., 2007]	Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., Klas, C.-P., Kovács, L., Landoni, M., Micsik, A., Papatheodorou, C., Peters, C., and Sølvberg, I. (2007). Evaluation of Digital Libraries. <i>International Journal on Digital Libraries</i> , 8(1):21–38.
[Tomlinson, 2009]	Tomlinson, S. (2009). Sampling Precision to Depth 10000 at CLEF 2008. In Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G. J. F., Kurimo, M., Mandl, T., and Peñas, A., editors, <i>Evaluating Systems for Multilingual and Multimodal Information Access: Ninth Workshop of the Cross–Language Evaluation Forum (CLEF 2008). Revised Selected Papers</i> , pages 163–169. Lecture Notes in Computer Science (LNCS) 5706, Springer, Heidelberg, Germany.