europeana
connect

# 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 **eContent*plus***programme
http://ec.europa.eu/econtentplus

Österreichische Nationalbibliothek

EuropeanaConnect is coordinated by the Austrian National Library


**ECP-2008-DILI-528001**

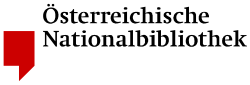
# EuropeanaConnect

## Report on Evaluation of Multilingual Information Access to Europeana

| | |
|---|---|
| **Dissemination level** | *Public* |
| **Delivery date** | *09-09-2011* |
| **Status** | *Final* |
| **Author(s)** | *MaristellaAgosti (UNIPD), AlessioBosca (CELI), Franco Crivellari (UNIPD), GrazianoDeambrosis (UNIPD), Giorgio Maria Di Nunzio (UNIPD), Marco Dussin (UNIPD), Nicola Ferro (UNIPD), Maria Gäde (HUB), Vivien Petras (HUB)* |

***e*Content*plus***

This project is funded under the *e*Content*plus* 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 |
| 0.3 | 09/09/2011 | Nicola Ferro | Task Leader |
| | | | |

**Approval**

| Version | Date of approval | Name | Role in project |
|---|---|---|---|
| 1.0 | | | |
| | | | |
| | | | |

**Revisions**

| Version | Status | Author | Date | Changes |
|---|---|---|---|---|
| 0.1 | D | UNIPD | 02/12/2010 | First draft circulated to Task 2.5 participants for comments |
| 0.2 | D | HUB | 08/12/2010 | workplan paragraph added |
| 0.3 | D | UNIPD | 09/09/2011 | Analysis of the experimental results |


**Abstract**

This report describes the evaluation activities 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 presents the used document collections, taken from the catalogues of the British National Library, French National Library, and German National Library, as well as the topics that have been used to model the actual user information needs.

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

Then, the report presents the infrastructure that has been used to manage and give access to all the experimental data, performance measures, and statistical analyses that have been produced.

Finally, the report presents the results of the evaluation activities for both the monolingual and the bilingual task according to the selected performance measures.

# Table of Contents

# 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)[1] in the United States, the Cross-Language Evaluation Forum (CLEF)[2] in Europe, and the NII-NACSIS Test Collection for IR Systems (NTCIR)[3]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 needsfrom 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[4]. 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 made 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 BibliothéqueNationale de France, and the Austrian National Library. These collections containcatalog 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).

The report is organized as follows: Section 2 describes the experimental setup (documents collections, topics, relevance assessments); Section 3 describes the evaluation tasks that have been conducted to assess the multilingual information access components of Europeana; Section 4

[1] http://trec.nist.gov/

[2] http://www.clef-campaign.org/

[3] http://research.nii.ac.jp/ntcir/

[4] Note that an evaluation of the "European Portal" from the Web usabilitypoint of viewhasbeenproposed in the past and ithasbeendecidedto notcarryit out for thesamereason.

explains the infrastructure that have been used for managing the experimental data and their analyses; Section 5 presents the experimental results.

# 2. Experimental Setup

## 2.1. DocumentCollections

We use three collections:

- **British Library (BL)**: 1,000,100 catalog records, 1.2 GByte of uncompressed XML;
- **BibliothéqueNationale 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.

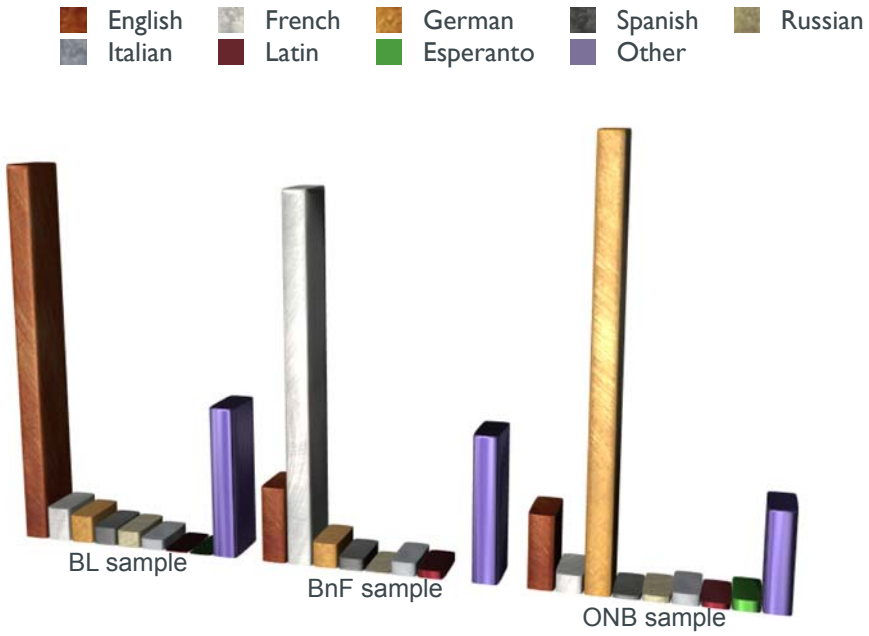



**Figure 1: Distribution of the languages in the BL, BnF, and ONB collections.**

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.

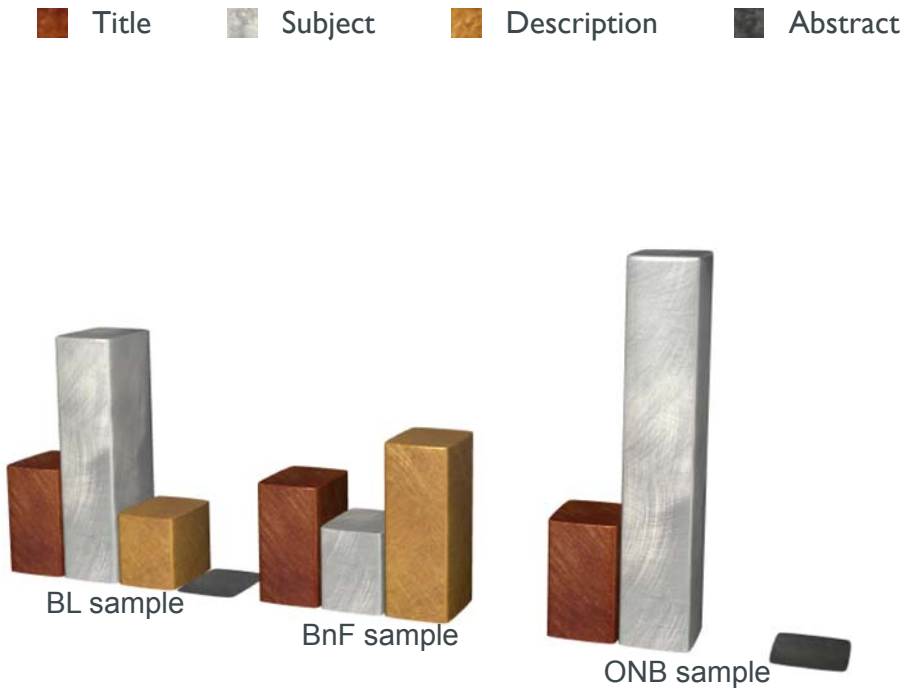



**Figure 2: Distribution of the record fields in the BL, BnF, and ONB collections (percentages greater than 100% means that the field is repeated more than one in a record, on average).**

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>
            <dcterms:issued>c2001</dcterms:issued>
            <dcterms:extent>263 p. : 1 map ; 20 cm.</dcterms:extent>
            <dc:language xsi:type="ISO639-2">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&amp;amp;l_base=BLL01&amp;from=TELgateway&amp;doc_number=010734316"
                >010734316</dc:identifier>
            <dc:identifier xsi:type="dcterms:URI"
                >http://catalogue.bl.uk/F/-?func=direct-doc-set&amp;amp;l_base=BLL01&amp;from=TELgateway&amp;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:ISO639-2">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 feministishcer Perspektive
                weitgehend bestätigt worden. Daher erscheint es sinnvoll, nach der Notwendigkeit
                einer solchen "Zähmung" grundsätzlicher 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 Ltieratur 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 ζωής 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 un-judged relevant documents were for the 10 or more queries that had the most known answers.

These discussions show how complex the creationof 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 weevaluate 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 offer the possibility of assessing the performances provided by different language resources. For each of the three target collections, we provide a baseline run and then evaluate the different language resources available in the Europeana language resources repository.

In order to do that, CELI 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 is 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 has performance scores associated for each of the three target languages; if it is a language-dependent resource, it has 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 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 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 provides 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 is 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 are evaluated. Finally, for each (source language, target

language, translation module) triple, different languages resources are evaluated. This gives the possibility of an accurate comprehension of the interaction of the different translation modules, language resources, and languages.

| | Target English (BL) | | | | | | Target French (BnF) | | | | | | Target German (ONB) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Source** | **it** | **de** | **fr** | **es** | **po** | **…** | **it** | **de** | **fr** | **es** | **po** | **…** | **it** | **de** | **fr** | **es** | **po** | **…** |
| **Module 1** | | | | | | | | | | | | | | | | | | |
| Resource 1 | | | | | | | | | | | | | | | | | | |
| Resource 2 | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | |
| **Module 2** | | | | | | | | | | | | | | | | | | |
| Resource 2 | | | | | | | | | | | | | | | | | | |
| Resource 3 | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | |
| **Module 3** | | | | | | | | | | | | | | | | | | |
| Resource 4 | | | | | | | | | | | | | | | | | | |
| Resource 5 | | | | | | | | | | | | | | | | | | |
| …. | | | | | | | | | | | | | | | | | | |

Moreover, exploiting the monolingual runs, for each of the combinations in the above matrix, we are 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 have managed by means of the DIRECT system[5][Agosti and Ferro, 2009] which has been developed and adopted since CLEF 2005 and it is now supported by the PROMISE[6] network of excellence (grant agreement no. 258191) on multilingual and multimodal information access evaluation . 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.

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

[6] http://www.promise-noe.eu/

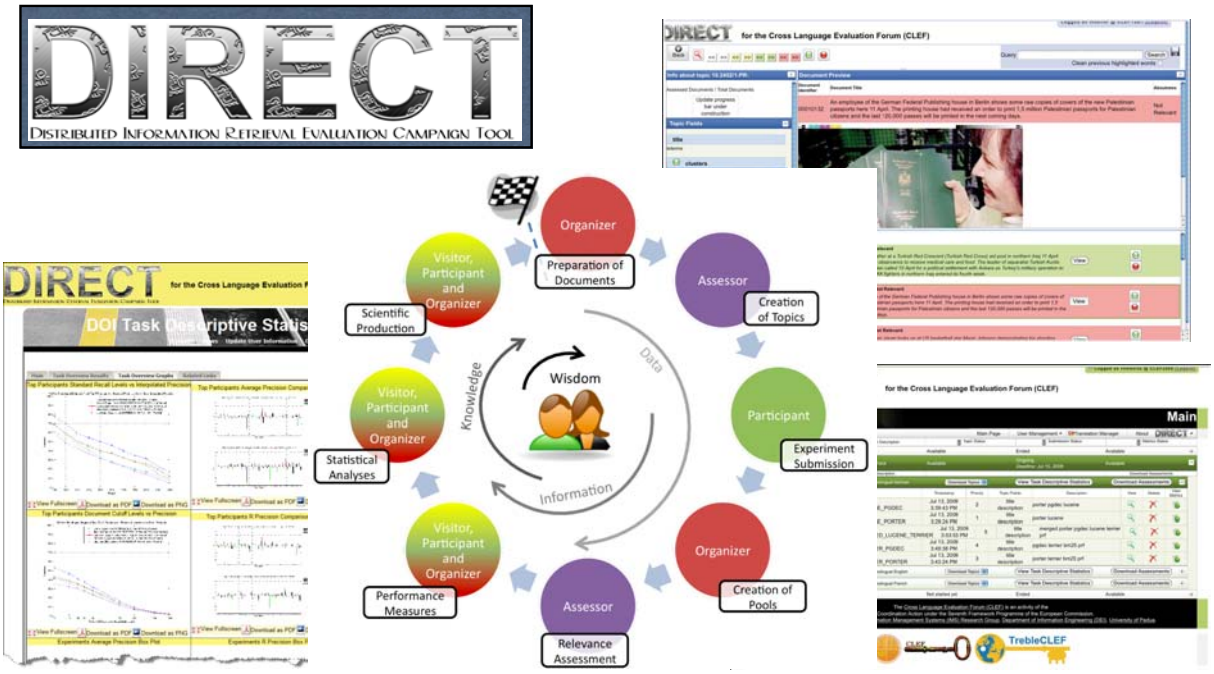



**Figure 7: Overview of some of the DIRECT functionalities.**

Figure 7summarizes the main functionalities of the DIRECT systems, which are reported below and related to the traditional DIKW (Data – Information – Knowledge – Wisdom) hierarchy [Ackoff, 1989; Zeleny, 1987]:

- *acquisition and preparation of documents*: the organizers are responsible for acquiring, formatting, and preparing the set of documents that are released to the participants. These documents are part of the data on which the experiments are built.
- *creation of topics*: the organizers and the assessors cooperate to create the topics for the test collection. For each topic, this step usually requires preparing a first draft of the topics and searching the set of document to verify that there are relevant documents for that topic; then the topics are refined by discussing their content and facets until a final version is reached. These topics are part of the data on which the experiments are built.
- *experiment submission*: the participants submit their experiments, which are built using the documents and the topics created in the previous steps. The result of each experiment is a list of retrieved documents in decreasing order of relevance for each topic and represents the output of the execution of the Information Retrieval System (IRS) developed by the participant. The experiments are part of the data that are produced during an evaluation campaign.
- *creation of pools*: the organizers collect all the experiments submitted by the participants and, using some appropriate sampling technique, select a subset of the retrieved documents to be manually assessed in the next step to determine their actual relevance. The pools are midway between data and information, since they are still raw elements but represent a first form of processing of the experiments.
- *relevance assessment*: the organizers and the assessors cooperate to assess each document in the pool with respect to the topic, i.e. for determining whether the document is relevant or not for the given topic. As in the case of the pools, the relevance judgments are midway between data and information, since they are raw elements which constitute an experimental collection but represent human-added information about the relationship between the topics and documents of an experiment.
- *measures and statistics*: the organizers exploit the relevance assessments to compute the performance measures and plots about each experiment submitted by a participant; then, these measurements are used for computing descriptive statistics about the overall behaviour of both an

experiment and all the experiments in a given task; furthermore, these measurements are also employed for conducting statistical analyses and tests on the submitted experiments. As discussed above, performance measures are information, since they are the results of data processing; descriptive statistics and hypothesis tests are knowledge, since they provide some more insights into the meaning of the obtained performance.

- *scientific production*: both organizers and participants prepare reports where the former describe the overall trends and provide an overview for the evaluation campaign and the latter explain their experiments, the techniques that have been adopted, and the findings. This work usually continues even after the conclusion of the campaign, since the investigation and understanding of the experimental results require deep analysis and reasoning, which usually takes the form of conference papers, journal articles, talks, and discussion among researchers. Furthermore, not only the organizers and the participants but also external visitors may exploit the information resources produced during the evaluation campaign to carry out their research activity. As explained above, the outcomes of this process are wisdom.

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.

A specific instance of DIRECT has been setup and installed for EuropeanaConnect and it is available at the following address:

http://svrims.dei.unipd.it:8080/europeana-mlia-evaluation/

Figure 8 shows the login page of the running DIRECT instances available at the above address.

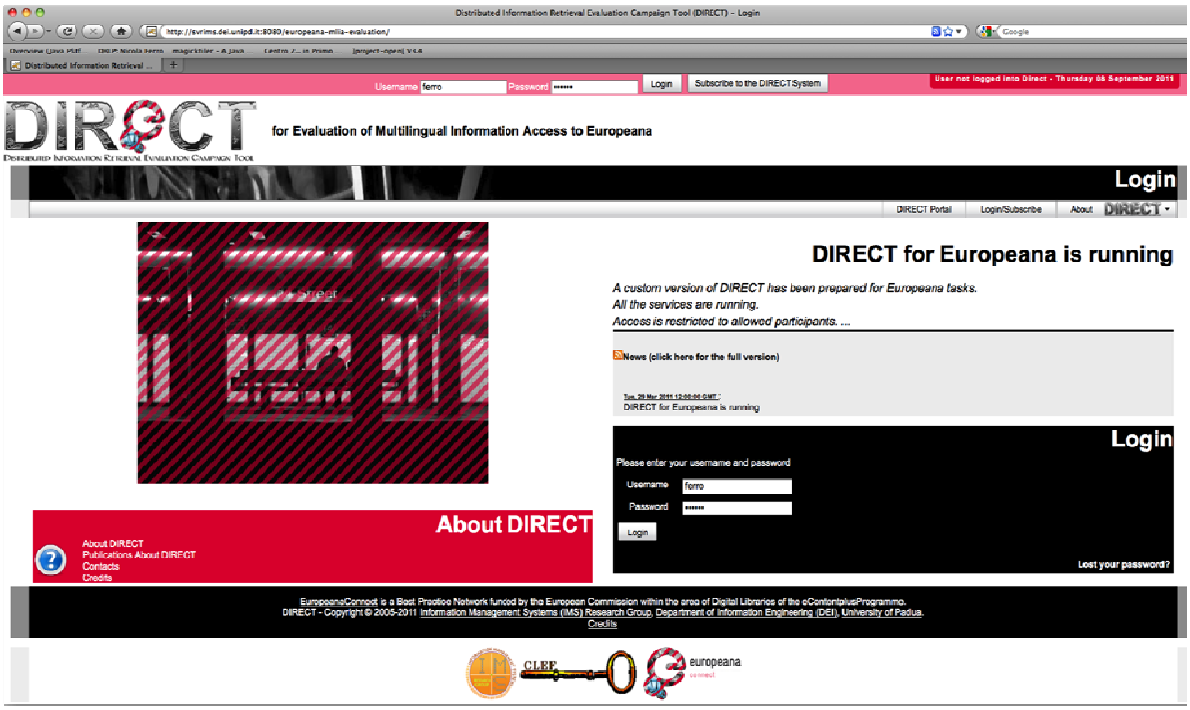



**Figure 8: Screenshot of the login page of the DIRECT instance for Europeana MLIA evaluation.**

# 5. Experimental Results

This sections discusses the results of the conducted experiments. As discussed in the previous section, all the experimental data reported here are available also online through the DIRECT instance set up for Europeana.

In the following, we describe the adopted metrics, the results of the monolingual tasks, and the results of the bilingual tasks.

## 5.1. AdoptedMetrics

The most common metrics used to assess performances of information access systems are based on the concepts of:

- **recall**: a measure of the ability of a system to present all relevant items

$$\text{recall} = \frac{\text{number of relevant items retrieved}}{\text{number of relevant items in collection}}$$

- **precision**: a measure of the ability of a system to present only relevant items

$$\text{precision} = \frac{\text{number of relevant items retrieved}}{\text{total number of items retrieved}}$$

Precision and recall are set-based measures. That is, they evaluate the quality of an unordered set of retrieved documents. To evaluate ranked lists, variations of these basics measures are developed in order to take into account at which rank positions relevant documents are retrieved.

In particular, the following metrics have been adopted to assess the Europeana multilingual information access components:

- **average precision**: is a single-valued measure that reflects the performance over all relevant documents. It rewards systems that retrieve relevant documents quickly (highly ranked).
- **precision@5**: is the precision after 5 documents have been retrieved. If you think at it in terms of a results list of a search engine with 10 results displayed per page, it gives you an idea of the performances at the mid of the first page.
- **precision@10**: is the precision after 10 documents have been retrieved. If you think at it in terms of a results list of a search engine with 10 results displayed per page, it gives you an idea of the performances at the end of the first page.
- **precision@20**: is the precision after 5 documents have been retrieved. If you think at it in terms of a results list of a search engine with 10 results displayed per page, it gives you an idea of the performances at the end of the second page.
- **R_precision**: is the precision after R documents have been retrieved, where R is the total number of relevant document that can be retrieved. It de-emphasizes the exact ranking of the retrieved relevant documents.

## 5.2. Monolingual Results

Table 1 reports the best results achieved in the three monolingual tasks (English, French, and German) for the above descripted metrics.

These results provide also the baseline against which bilingual performances will be compared to understand which is the performance penalty due to the passing from one language to another.

| Task | Best Mean Average Precision | Best Mean Precision@5 | Best Mean Precision@10 | Best Mean Precision@20 | Best Mean R_Precision |
|---|---|---|---|---|---|
| **Monolingual English** | 27.46% | 51.20% | 45.30% | 36.35% | 29.90% |
| **Monolingual French** | 23.35% | 39.20% | 34.50% | 27.05% | 25.68% |
| **Monolingual German** | 13.48% | 33.00% | 27.20% | 20.85% | 16.03% |

**Table 1: Best results for the monolingual tasks.**

## 5.3. Bilingual Results

Table 2 reports the best results achieved in the three bilingual tasks (X → English, X → French, and X → German) for the above descripted metrics.

For each target language (English, French, and German), the results achieved with different source languages (Dutch, English, French, German, Hungarian, Italian, Polish, Portuguese, Spanish, and Swedish) are reported as well as the comparison with respect to the corresponding monolingual baseline.

For example, the best mean average precision for the bilingual Dutch to English experiments is 19.49% while the best mean average precision for the monolingual English experiments (see Table 1) is 27.46%; therefore, the bilingual Dutch to English achieves 70.98% of the performances of the monolingual English.

| Task | Best Mean Average Precision | Best Mean Precision@5 | Best Mean Precision@10 | Best Mean Precision@20 | Best Mean R_Precision |
|---|---|---|---|---|---|
| **Bilingual To English** | | | | | |
| *Dutch* | 19.49% | 40.80% | 33.70% | 22.30% | 21.73% |
| wrt monolingual baseline | 70.98% | 79.69% | 74.39% | 61.35% | 72.68% |
| *French* | 19.18% | 39.40% | 33.40% | 26.40% | 21.73% |
| wrt monolingual baseline | 69.85% | 76.95% | 73.73% | 72.63% | 72.68% |
| *German* | 17.65% | 34.40% | 29.30% | 24.60% | 20.10% |
| wrt monolingual baseline | 64.28% | 67.19% | 64.68% | 67.68% | 67.22% |
| *Hungarian* | 14.02% | 29.40% | 25.90% | 20.25% | 15.91% |
| wrt monolingual baseline | 51.06% | 57.42% | 57.17% | 55.71% | 53.21% |

| Task | Best Mean Average Precision | Best Mean Precision@5 | Best Mean Precision@10 | Best Mean Precision@20 | Best Mean R_Precision |
|---|---|---|---|---|---|
| *Italian* | 18.97% | 38.60% | 31.80% | 25.80% | 21.16% |
| wrt monolingual baseline | 69.08% | 75.39% | 70.20% | 70.98% | 70.77% |
| *Polish* | 18.24% | 35.20% | 29.40% | 24.20% | 20.52% |
| wrt monolingual baseline | 66.42% | 68.75% | 64.90% | 66.57% | 68.63% |
| *Portuguese* | 21.05% | 39.80% | 33.90% | 28.85% | 23.80% |
| wrt monolingual baseline | 76.66% | 77.73% | 74.83% | 79.37% | 79.60% |
| *Spanish* | 16.74% | 36.60% | 28.60% | 23.85% | 19.70% |
| wrt monolingual baseline | 60.96% | 71.48% | 63.13% | 65.61% | 65.89% |
| *Swedish* | 16.77% | 33.00% | 29.50% | 25.00% | 19.46% |
| wrt monolingual baseline | 61.07% | 64.45% | 65.12% | 68.78% | 65.08% |
| **Bilingual To French** | | | | | |
| *Dutch* | 11.76% | 20.60% | 18.10% | 13.35% | 12.53% |
| wrt monolingual baseline | 50.36% | 52.55% | 52.46% | 49.35% | 48.79% |
| *English* | 15.77% | 27.40% | 23.90% | 18.40% | 16.55% |
| wrt monolingual baseline | 67.54% | 69.90% | 69.28% | 68.02% | 64.45% |
| *German* | 12.77% | 22.80% | 18.80% | 15.25% | 13.78% |
| wrt monolingual baseline | 54.69% | 58.16% | 54.49% | 56.38% | 53.66% |
| *Hungarian* | 9.29% | 17.00% | 14.10% | 11.20% | 10.50% |
| wrt monolingual baseline | 39.79% | 43.37% | 40.87% | 41.40% | 40.89% |
| *Italian* | 15.73% | 28.00% | 22.60% | 17.04% | 16.50% |
| wrt monolingual baseline | 67.37% | 71.43% | 65.51% | 62.99% | 64.25% |
| *Polish* | 12.42% | 23.80% | 18.90% | 14.10% | 13.62% |
| wrt monolingual baseline | 53.19% | 60.71% | 54.78% | 52.13% | 53.04% |
| *Portuguese* | 15.74% | 30.60% | 25.30% | 19.00% | 16.90% |
| wrt monolingual baseline | 67.41% | 78.06% | 73.33% | 70.24% | 65.81% |
| *Spanish* | 10.17% | 21.00% | 17.20% | 13.00% | 11.76% |
| wrt monolingual baseline | 43.55% | 53.57% | 49.86% | 48.06% | 45.79% |
| *Swedish* | 11.29% | 20.80% | 18.10% | 14.00% | 12.97% |
| wrtmonolingual baseline | 48.35% | 53.06% | 52.46% | 51.76% | 50.51% |
| **Bilingual To German** | | | | | |
| *Dutch* | 11.08% | 22.60% | 18.40% | 15.10% | 12.30% |
| wrt monolingual baseline | 82.20% | 68.48% | 67.65% | 72.42% | 76.73% |
| *English* | 9.12% | 22.00% | 16.80% | 12.60% | 10.14% |
| wrt monolingual baseline | 67.66% | 66.67% | 61.76% | 60.43% | 63.26% |
| *French* | 10.08% | 23.00% | 17.70% | 13.55% | 11.34% |
| wrt monolingual baseline | 74.78% | 69.70% | 65.07% | 64.99% | 70.74% |
| *Hungarian* | 9.02% | 15.20% | 14.00% | 11.55% | 9.92% |
| wrt monolingual baseline | 66.91% | 46.06% | 51.47% | 55.40% | 61.88% |
| *Italian* | 10.00% | 23.40% | 18.60% | 13.50% | 11.05% |

| Task | Best Mean Average Precision | Best Mean Precision@5 | Best Mean Precision@10 | Best Mean Precision@20 | Best Mean R_Precision |
|---|---|---|---|---|---|
| wrt monolingual baseline | 74.18% | 70.91% | 68.38% | 64.75% | 68.93% |
| *Polish* | 10.10% | 19.60% | 17.20% | 13.30% | 11.65% |
| wrt monolingual baseline | 74.93% | 59.39% | 63.24% | 63.79% | 72.68% |
| *Portuguese* | 12.77% | 27.20% | 23.40% | 18.25% | 14.55% |
| wrt monolingual baseline | 94.73% | 82.42% | 86.03% | 87.53% | 90.77% |
| *Spanish* | 9.52% | 21.00% | 16.00% | 12.95% | 10.58% |
| wrt monolingual baseline | 70.62% | 63.64% | 58.82% | 62.11% | 66.00% |
| *Swedish* | 11.12% | 22.40% | 19.40% | 15.50% | 11.80% |
| wrt monolingual baseline | 82.49% | 67.88% | 71.32% | 74.34% | 73.61% |

**Table 2: Best results for the bilingual tasks.**

# References


| | |
|---|---|
| [Ackoff, 1989] | Ackoff, R. L. (1989). From Data to Wisdom. *Journal of Applied Systems Analysis*, 16, pp. 3-9. |
| [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, *Evaluating Systems for Multilingual and Multimodal Information Access: Ninth Workshop of the Cross–Language Evaluation Forum (CLEF 2008). Revised Selected Papers*, 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, *Proc. 3rd International Workshop on Evaluating Information Access (EVIA 2010)*, 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 Tsakonas, G. and Papatheodorou, C., editors, *Evaluation of Digital Libraries: An insight into useful applications and methods*, 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, *Advances in Cross-Language Information Retrieval: Third Workshop of the Cross–Language Evaluation Forum (CLEF 2002) Revised Papers*, pages 9–27. Lecture Notes in Computer Science (LNCS) 2785, Springer, Heidelberg, Germany. |
| [Cleverdon, 1997] | Cleverdon, C. W. (1997). The Cranfield Tests on Index Languages Devices. In Spärck Jones, K. and Willett, P., editors, *Readings in Information Retrieval*, 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., Pen˜as, A., and Roda, G., editors, *Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments – Tenth Workshop of the Cross–Language Evaluation Forum (CLEF* |

| | |
|---|---|
| | *2009). Revised Selected Papers*, 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. *International Journal on Digital Libraries*, 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, *Evaluating Systems for Multilingual and Multimodal Information Access: Ninth Workshop of the Cross–Language Evaluation Forum (CLEF 2008). Revised Selected Papers*, pages 163–169. Lecture Notes in Computer Science (LNCS) 5706, Springer, Heidelberg, Germany. |
| [Zeleny, 1987] | Zeleny, M. (1987). Management Support Systems: Towards Integrated Knowledge Management. *Human Systems Management*, 7 (1), 59-70. |
| | |