ECP-2008-DILI-528001

EuropeanaConnect

Report on User Preferences and Information Retrieval Scenarios for Multilingual Access in Europeana

Deliverable number/name	*D2.1.1 – Report on User Preferences for Multilingual Access in Europeana*
Dissemination level	*Public*
Delivery date	*7th December 2009*
Status	*Final*
Author(s)	*Maristella Agosti[2], Franco Crivellari[2], Graziano Deambrosis[2], Nicola Ferro[2], Maria Gäde[1], Vivien Petras[1], Juliane Stiller[1]* *[1]Humboldt-Universität zu Berlin (HUB)* *[2]University of Padua (UNIPD)*

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

D2.1.1 – Multilingual Information Access in Digital Libraries

Report on User Preferences and Information Retrieval Scenarios for Multilingual Access in Europeana

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

Abstract

While the need for multilingual access to multinational and multicultural information systems like Europeana is undoubted, few truly operational systems exist and can serve as examples. Several projects have done extensive studies on user needs and requirements for features for information system access, but few have paid specific attention to multilingual issues. This report attempts to give an overview over a number of projects within Europe that have dealt – expressively or not – with multilingual access issues to their content representing their results. We will then briefly describe different perspectives on what multilingual access to an information system like Europeana could mean, organizing different approaches by their impact on the system overall. The main focus of this report is on user needs and desired features for multilingual access learned in part from a thorough screening of associated Europeana user studies and results from other projects as well as from a survey targeted specifically toward multilingual access issues within Europeana. The outcome of these studies is a description of user requirements and suggested usage scenarios for two multilingual access features (query translation and result representation), which will hopefully serve as a starting point for a discussion on multilingual access options in Europeana. We conclude with open questions and challenges for the way forward.

Table of Contents

1 Introduction

"EuropeanaConnect will support the creation of a diverse and inclusive Europeana facilitating access to culture by all communities and individuals and representative of various cultures and language-groups." (EuropeanaConnect, 2009)

Europeana's portal now attempts to provide access to currently over 4.6 million digitized cultural heritage objects. Some of them are famous or appear in many versions, others are unique, rare or at-risk objects, which would not be seen by the public otherwise. Access to these objects is hampered by several obstacles:

- Digital objects are provided through the metadata description efforts of the organizations and agencies curating the objects, usually with specified technical vocabularies suited for their particular domains, which need to be mapped in order to provide homogenous access.
- Different heterogeneous media types (images, texts, sound, videos) have to be searched and presented simultaneously and similarly.
- Users from different cultures with different needs and requirements need to be presented with suitable search and access options.
- Both the users and the content of Europeana appear in many different languages and need to be matched to each other in order to provide seamless access across language barriers.

Within the EuropeanaConnect project (www.europeanaconnect.eu), a separate work package (WP2) and a substantial number of resources are devoted to develop solutions to cope with multilingual access issues for users and objects alike within Europeana. Through the provision of multilingual access capabilities, that is the translation and mapping of the portal interfaces, object descriptions, browsing categories and user queries, all content should be leveraged by all Europeana users equally, regardless of their native language or the available native language resources.

EuropeanaConnect will develop solutions for the multilingual access to the Europeana portal and resources. In order to facilitate the creation of viable solutions for multilingual access, a better understanding of user needs and user requirements within a multilingual framework is imperative. A number of questions need to be answered:

- *What do we know about multilingual access to digital libraries?*

  Section 2 and Appendices A&B review the state-of-the-art in current multilingual information systems and describe projects and initiatives within a European context

- *Which lessons and best practices have been learned from existing information systems dealing with multilingual content and users?*

  Section 3 and Appendix C report on user needs and desired features for multilingual access learned from previous user studies and best practices.

- *What do users really want with respect to multilingual access within Europeana?*

  Section 4 and Appendix D report on the survey that has been conducted to investigate multilingual access issues within Europeana.

- *Which steps should be taken on the way to a truly multilingual system and which scenarios for multilingual access can we implement for a scalable, operational system?*

Section 5 suggests usage scenarios for three multilingual access features (query translation, result representation & multilingual subject mapping for document enrichment) and conclude with open questions and challenges for the way forward.

The analysis and suggestions detailed in this report come from a variety of sources. A thorough review of on-going and completed initiatives dealing with multilingual access issues in information systems preceded our own discussions within the Europeana community. A workshop was organized in September 2009 (MLIA4DL Workshop, 2009) inviting researchers, stakeholders, and representatives of pan-European digital library projects to discuss multilingual user needs, assessment methods for requirements, and usage logging of multilingual resources use as well as practical implementation issues when incorporating multilingual capabilities into a digital library. The discussion was extended to Europeana specific issues by rich discussions with stakeholders and core experts at the Europeana plenary conference in September 2009. Finally, a survey on user preferences for multilingual access was developed and conducted during the summer of 2009 and is analyzed here.

The goal of this report is thus to serve as a starting point for the discussion on multilingual access features with the Europeana developers and Europeana stakeholders. To that end, the survey results, usage scenarios and open questions can be looked at initial suggestions for directions to pursue.

2 Aspects of Multilingual Information Access in Digital Libraries

By MultiLingual Information Access (MLIA) we usually denote procedures for search on collections of information items that are potentially stored in multiple languages. True multilingual access is more than just being able to search in more than one language. It means that the intended result is retrieved in each target collection regardless of language, character-encoding, metadata-schema, or normalisation rules (Agosti et al. 2007; Braschler and Ferro 2007). Usually the term is used for situations in which the user is allowed to query the collection across languages, i.e. retrieving information items in a language that is different from the language used by the user to formulate his/her information need. In this narrower sense, the term Cross-Language Information Retrieval (CLIR) is often used.

Approaches for CLIR can be classified according to different schemes. (Oard 1997) proposes a taxonomy for CLIR approaches in terms of what he calls types (free-text vs. controlled vocabulary) and aspects (knowledge-based vs. content-based). We follow the definition in (Braschler et al. 1998; Peters and Sheridan 2001; Oard 2006) which uses a first-level classification according to how the query and information items (documents in the cited paper) are matched across languages – be it by translating the query, the information item, or both. These three basic options can in some situations be extended to include a fourth, which does not use translation on either query or information items, but instead uses matching at sub-word level (see e.g. (McNamee and Mayfield 2004)). Since this option typically relies on lengthy textual representations of queries and information items, it seems to be less suitable for the present problem of matching short metadata records, and is not pursued further in the following.

Today, the mainstream research on cross-language information retrieval in Europe is carried out in the confines of the Cross-Language Evaluation Forum (CLEF) campaign in Europe (Peters et al. 2009). The campaign gives researchers the possibility to compare different approaches to CLIR in a common setting and provides tools for both in-depth analysis and curation of the experimental results. Most of the experiments in CLEF concentrate on retrieval on lengthy, unstructured full-text documents using a general vocabulary. In such a setting, evaluations have shown that query translation is a good compromise between effectiveness in terms of retrieval quality and efficiency, and query translation is therefore the prevailing method used by participants in the CLEF campaign. An overview of the recent

achievements in CLIR with a special focus on metadata and bibliographic records from The European Library can be found in (Agirre et al. 2008; Ferro and Peters 2009).

Being “multilingual” can mean different things for different information systems. Different components of an information system can deal with multilingual issues such as content, user queries or interface issues. This section gives a brief overview of different levels of multilinguality the user will encounter during a retrieval process in a system focusing in more details on query translation as the most common approach to multilingual information retrieval.

Finally, in chapter 5 of the outline functional specification for Europeana (Europeana 2009), a first attempt to face multilinguality in Europeana has been made by outlining different aspects, which we will draw on. We also take into consideration the report on best practices in system-oriented and user-oriented multilingual information access produced by the TrebleCLEF coordination action (TrebleCLEF 2009).

2.1 Localization and Internationalization of the Interface

As explained by the World Wide Web (W3C) consortium in its Internationalization Activity (W3C 2009; Ishida and Miller 2006), localization, often abbreviated in l10n, and internationalization, often abbreviated in i18n, are two distinct activities.

Localization refers to the *adaptation* of a product, application or document content to meet the language, cultural and other requirements of a specific target market (a "locale"). Often thought of only as a synonym for translation of the user interface and documentation, localization is often a substantially more complex issue. It can entail customization related to:

- Numeric, date and time formats;
- Use of currency;
- Keyboard usage;
- Collation and sorting;
- Symbols, icons and colours;
- Text and graphics containing references to objects, actions or ideas which, in a given culture, may be subject to misinterpretation or viewed as insensitive;
- Varying legal requirements.

Localization may even necessitate a comprehensive rethinking of logic, visual design, or presentation if the way of doing business (e.g., accounting) or the accepted paradigm for learning (egg., focus on individual vs. group) in a given locale differs substantially from the originating culture.

Internationalization is the design and development of a product, application or document content that *enables* easy *localization* for target audiences that vary in culture, region, or language. Internationalization typically entails:

- Designing and developing in a way that removes barriers to localization or international deployment. This includes such things as enabling the use of Unicode, or ensuring the proper handling of legacy character encodings where appropriate, taking care over the concatenation of strings, avoiding dependence in code of user-interface string values, etc.;
- Providing support for features that may not be used until localization occurs. For example, adding mark-up in your DTD to support bidirectional text, or for identifying language. Or adding to CSS support for vertical text or other non-Latin typographic features.

- Enabling code to support local, regional, language, or culturally related preferences. Typically this involves incorporating predefined localization data and features derived from existing libraries or user preferences. Examples include date and time formats, local calendars, number formats and numeral systems, sorting and presentation of lists, handling of personal names and forms of address, etc.;
- Separating localizable elements from source code or content, such that localized alternatives can be loaded or selected based on the user's international preferences as needed.

Notice that these items do not necessarily include the localization of the content, application, or product into another language; they are design and development practices which allow such a migration to take place easily in the future but which may have significant utility even if no localization ever takes place.

Localization, achieved by means of proper internationalization, is often perceived as an elementary level of multilinguality, since it offers users with the interface in their own language. In this sense, the Europeana public prototype launched in November 2008 was already based on such architecture.

The choice of the interface language is usually one of many personalization options but could give some useful indications about other choices of the user regarding multilingual access functions. In general, it needs to be identified whether the user wants to select the interface language or whether he is satisfied with a default interface language like English.

Within this level of multilinguality it is also essential to be aware of the different options that could be used to determine the language interface, as for example:

- The user selects the language interface (drop-down-menu);
- The language interface is selected automatically, e.g. based on the language settings of the user agent or the geographic location of the user determined via IP-address.

2.2 Multilingual Search

The core component of a truly multilingual information system is the multilingual search capability, that is, for a query in a particular language to retrieve documents in other languages as well. Two methods have been identified to provide multilingual search functionality:

- *Query translation*: the original query is translated into additional languages that the document collection may contain;
- *Document translation*: the documents in the collection are translated into different languages that are supported at query time.

Both approaches require having some mechanisms or linguistic resources – bilingual dictionaries, bilingual lexica, machine translation systems, and so on – to cross the language boundaries for each language pair that needs to be supported. In general, given a set of n languages that have to be supported, you would have n(n – 1) possible language pairs – if, for example, you consider German to Dutch different from Dutch to German – or n(n – 1)/2 possible language pairs – if, for example, you consider German to Dutch the same as Dutch to German. Thus, if you consider the 23 official languages of the European union, you would have to support, at least, 253 language pairs.

A possible solution to avoid this quadratic grow in the number of language pairs to be supported is to adopt an *interlingua* approach (Ballesteros 2000), which applies to both query translation and document translation. In such cases, instead of having support and multilingual resources for all the possible pairs of source and target languages, one language is selected as *pivot* and all the translations are made to and from this pivot language. For example, if we need to translate from

Portuguese to Bulgarian, instead of performing a direct translation, we may choose German as the interlingua, and perform a translation from Portuguese to German and from German to Bulgarian. Therefore, given a set of n languages that have to be supported, you would have 2(n – 1) possible language pairs to/from the pivot language – if, for example, you consider German to Dutch different from Dutch to German – or (n – 1) possible language pairs to/from the pivot language – if, for example, you consider German to Dutch the same as Dutch to German. Thus, if you consider the above case of the 23 official languages of the European union, you would have to support, at least, 22 language pairs. Obviously, the interlingua approach deteriorates a little bit the overall performances since it require to cross the language boundaries twice.

Query translation is the most commonly adopted method for multilingual information systems today since it is usually less demanding from a computational and storage point of view than document translation and it may be more flexible in incorporating new languages or new documents within a collection, not requiring to compute new translations.

Therefore, in the following, we will focus on one possible variation of the query translation approach, which is the most readily applicable to Europeana, and leave a discussion of a possible exploitation of the document translation approach to Section 5.3 “Open Questions and Challenges”.

2.2.1 Query Translation

For the envisioned query translation, the following phases can be identified. The steps in brackets are optional and depend on the design of the system and the query type.

- Query formulation
- (Language identification)
- (Named entity identification)
- Term translation
- (Disambiguation of candidate translations)
- Query processing in the target language(s)

Note that the overall query translation can be conducted either in a completely automatic way or involving the user by letting him/her the possibility of modify the proposed translations. The outcomes of the questionnaire about user expectations with respect to multilinguality reported in Section 4 will provide some hints about these two alternatives.

Whenever the language of the query is not explicitly known, the language of the query needs to be determined in order to apply the correct language resources for processing and translation. With respect to the very short length of most queries, language detection can be highly ambiguous and might introduce errors into the translation process. Therefore, it would be very useful to ask the user to select the language of the query term but will this would add an additional step before a search can be performed and could be perceived as annoying.

If the document collection contains documents in more than one language, it could also be useful to know if all them are to be considered as target languages or if the user is interested only in a subset of them. This could be controlled by user input – usually through an advanced search interface where the target languages can be indicated.

For the translation phase, particular care should be paid to the problem of translation disambiguation, should it happen by means of machine translation, parallel corpora, bilingual dictionaries, and so on. One example for this problem is the polysemous French query “avocat”, which could either mean lawyer or avocado (fruit). Translating the term into other languages will most likely lead to different

translations, which need to be disambiguated according to context. This can happen in an additional automated step or - in an interactive system - the user can select the preferred translation from the candidates. Other forms of user interaction could be the possibility to add, change or deselect translation candidates, as well as creating personal vocabularies.

2.3 Multilingual Result Representation

Once the search has performed, two options arise: either present the result list containing the retrieved documents in their original language or translate the retrieved documents into a language selected by the user.

For textual documents within a collection, it also needs to be determined whether result translation happens at the metadata level (as would be the case for other media types) or the original document level. For Europeana only the metadata level is relevant as the original documents are hosted at the partner institutions. The questions are whether users prefer snippets and to which extend metadata should be translated.

2.3.1 Multilingual Result Filtering

Another elementary option implemented for almost any kind of information system is the capability of filtering a result set by determining the desired languages of the documents a user wants to view as the result of a search. Commonly, this feature can be implemented in two ways:

- Filter option in advanced search interface: when inputting a query, a user can determine the desired language of the documents in the result set by choosing from a list of available languages.
- Refinement filter for result set: the user is presented with the option to filter a result set by language after the first search has been processed.

This second option is already available in the Europeana public prototype today. Whereas commonly both options allow the user to only choose one language, thereby restricting the result set to a monolingual set, the Europeana prototype refinement filter allows to select as many languages as are available in the result set.

2.4 Multilingual Browsing

Browsing within an information system is usually provided through a hierarchical classification or subject ontology for content descriptions. Other browsing modalities are people and organizations (who), places (where) and time periods (when). It needs to be determined what browsing functionalities are preferred by users.

3 Findings from Previous User Studies on MLIA

This section reports the outcomes of previous user studies on the user preferences and needs in multilingual digital libraries.

Project name	Methodology	Participants	Reference
Multimatch	Questionnaire Interviews workshops	Educational users Tourist users Cultural heritage users	D 1.2, 2006
Clarity	Questionnaire Search tasks	Monolingual users Bilingual users	Petrelli et al., 2002
Eurovision	Questionnaire Search tasks	Multilingual users	Clough & Sanderson, 2006
Tate Online	Online questionnaire Evaluation task	635 responses	Marlow et al., 2007
TrebleCLEF	workshop	Workshop participants	D 3.3, 2009
iCLEF 2008	Search tasks (Flickr) Log file analysis Questionnaire	307 users in search logs	Srinivasarao, 2008
CACAO	Usability test with MuSiL interface Unsupervised test Moderated test Survey Log file analysis Expert review	Unsupervised test: 6 responses Moderated test: 4 participants	D 4.1, 2008
Google Translate	Search tasks	Non-native English speakers	Aula & Keller, 2009 Marlow et al., 2008
Gabriel & EDLproject	Questionnaire Log file analysis	560 responses	Janssen, 2003 EDL M 1.4, 2007
TEL & Telplus	Log files	Data from 7 months	EDL M 1.4, 2007
Europeana	Online survey	3,204 participants	IRN Research ,2009

The list above is a compilation of user studies which dealt with or at least briefly touched on user needs in multilingual digital libraries. A more detailed description of the different user studies can be found in Appendix C. The following discussion focuses on previous findings and its relevance for projects like Europeana. Most of the mentioned initiatives collected results through a combination of methods such as: Questionnaires, Log file analysis and search tasks. Although similar methods were used the comparison or generalization of these findings seems to be difficult because of the very different user groups involved. As the table shows, some studies selected their users in regard to their information needs, others in regard to their language skills. Also the number of participants varies considerably between the multiple studies. Following some statements are highlighted and categorized in regard to the different aspects of multilingual information access as mentioned in Section 2.

Multilingual user interface:

- Users are more likely to visit a collection if the site was translated in their preferred language. (*all projects*)
- The most frequency used interface language is English (*Tel & Telplus*)

Multilingual search:

- Search/query language does not inevitably result from the interface language. (*TrebleCLEF*)
- Users would like to start searching in their native language. (*Clarity, Google Translate*)
- Users would like to choose their own query language. (*Clarity*)
- Users routinely search and browse in languages other than their native language, especially if the native language is not English. (*Clarity, Eurovision*)
- It is common to repeat a query in another language (usually English) if the first language version was not successful. (*Google Translate*)
- Users search in English when they want a broader result set. (*CACAO, Google Translate*)
- Users do not necessarily trust in automatic query translation and disambiguation. (*Eurovision, Tel & Telplus*)
- Users would like to modify suggested translations prior to searching. (*Clarity, Eurovision, TrebleCLEF*)
- User-created dictionaries should be created harvesting user input when harvesting translation options. (*Clarity*)
- Successful searchers reformulate queries frequently (*iCLEF 2008*).
- Users often search for place names and subject keywords. (*Tate Online, Tel & Telplus*)
- The disambiguation of translation candidates is necessary because otherwise extraneous terms will be added to the query and make the search more imprecise. (*Eurovision*)
- Named entities need special consideration as they commonly do not appear in bilingual dictionaries. (*Eurovision*)

Result representation / translation:

- Users routinely refine results by language. (*Europeana)*
- Users are willing to accept imperfect translations of texts. (*Tate Online*)
- Users with a passive knowledge of a language do not require full document translation in order to decide whether a document could be relevant for them. (*TrebleCLEF, Google Translate*)
- Translation candidates should be displayed in a wider context. (*TrebleCLEF*)
- The translation of whole documents seems not be desired or sensible. (*TrebleCLEF, Google Translate, Tel & Telplus*)
- Subject translation seems to be useful. (*Multimatch, Tel & Telplus*)

TrebleClef also compiled a list of multilingual user requirements at a very general level (TrebleCLEF D3.2, 2008), some of which are repeated here:

Integration:

- Systems must be transparent regarding cross-language search.
- Multilingualism is a feature of information access systems, which must be seamlessly integrated.

Multilingual Interface:

- The interface should adapt to a profile specifying language skills and translation preferences.

Multilingual search:

- There should be an advanced search mode that gives user full control over multilingual features (target languages, query translations), but this might not be helpful in very complex systems.
- Default translation should only be provided for languages unknown to the user.

Results presentation / translation:

- Interfaces should organize results by at least two choices: separated by target language or merged.
- There should be a choice of observing the original document or a translation.
- If translation for a certain language pair is not available, one option is to show metadata: named entities, categories, etc.
- When few monolingual results are available, the system should alert the user whenever there is more information available in other target languages.
- The system should warn about the quality of machine translation to avoid wrong expectations from the user.

4 Survey on Multilingual Information Access to Europeana

To specify needs assessment for the particular Europeana environment, we prepared a survey that was available for online use by University of Padua Library Centre (CAB) in June 2009. For testing, two groups were selected: librarians and researchers within the field of multilingual information retrieval. Additionally, we also presented the survey to the participants of the workshop “Multilinguality in Information Access to Digital Libraries - User Needs and Evaluation of Multilingual Resources Use” in Trento (MLIA4DL Workshop, 2009).

The survey consisted of 7 themes, including 23 questions:

- User profile including native & other languages and DL use
- Multilingual content interaction
- Multilingual user interface
- Information access and retrieval
- Multilingual information retrieval
- Multilingual query formulation & expansion
- Multilingual results presentation

Most of them are multiple-choice questions with a restriction to one selection; only with some of them multiple selections are possible. Other questions ask for specifications or give the opportunity for comments. See appendix B for an overview of the complete survey.

In the following, we report the analysis of the results of the survey presented to the attendees (students, researchers, librarians) of the TrebleCLEF Summer School on Multilingual Information Access (http://www.trebleclef.eu/summerschool.php), where we gathered 25 questionnaires. The surveyed users represent a sample with a strong interest in multilinguality and a good knowledge of what can be expected from MLIA systems. All the elaborations have been performed by using the R language (http://www.r-project.org) for statistical computing while Excel has been used for the plots.

Note that due to the size of the sample (25 questionnaires) and to the specific interests of the surveyed users, the indications gathered from the questionnaires have to be considered preliminary and cannot be taken as representative of all the possible users and stakeholders of Europeana.

However, it should considered that this is valuable sample since it is often difficult to gather in the same survey users with such a "multilingual profile" and with a so wide and different cultural and linguistic background that give the possibility of studying and appreciating how the different facets of multilinguality are perceived; individual profiles of the participants in the survey can be found at http://www.trebleclef.eu/ss09_participants.php. Indeed, it is much more common to have more homogenous samples, e.g. students from the same nation and faculty, practitioners in the same discipline, and so on.

Finally, the approach adopted in preparing the questionnaire, the kind of analyses that have been conducted, and what we have learnt from it provide us with a robust basis which could be adopted to investigate the requirements of other user communities by conducting additional surveys in the future.

4.1 User Profile

4.1.1 Age

The following table reports the answers for the age ranges.

Age Range	Absolute Frequency	Relative Frequency
15-24	3	0.12
25-34	19	0.76
35-44	2	0.08
45-54	1	0.04

As it can be noted from the above table and from Figure **4.1**, most of the surveyed users are "young" users, since 88% of them fall in the 15-34 range. This represents an important target for Europeana, since the young generation will be one of the main user categories of the Europeana system.

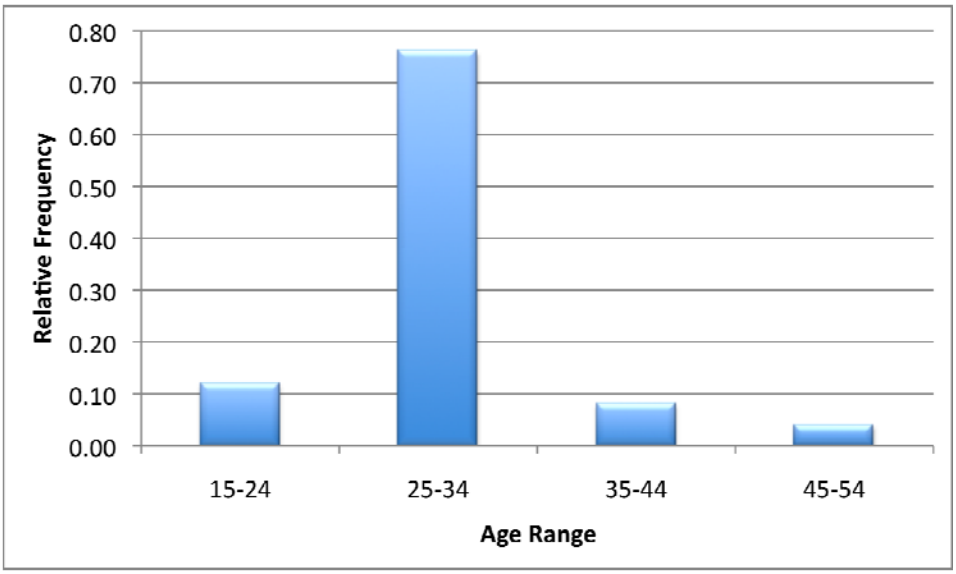

Figure 4.1: Distribution of the ages.

4.1.2 Country

Country	Absolute Frequency	Relative Frequency
Austria	1	0.04
Egypt	1	0.04
Finland	2	0.08
France	3	0.12
Greece	1	0.04
Ireland	1	0.04
Italy	6	0.24
Malaysia	1	0.04
Romania	1	0.04
Spain	5	0.20
The Netherlands	1	0.04
United Kingdom	1	0.04
Vietnam	1	0.04

The table above and the histogram of Figure **4.2** report the nationalities of the participants in the survey. As it can be noted the participants mostly come from European countries, even if there is an interesting participation (12%) also from other countries (Egypt, Malaysia, and Vietnam) with a quite different cultural and linguistic background.

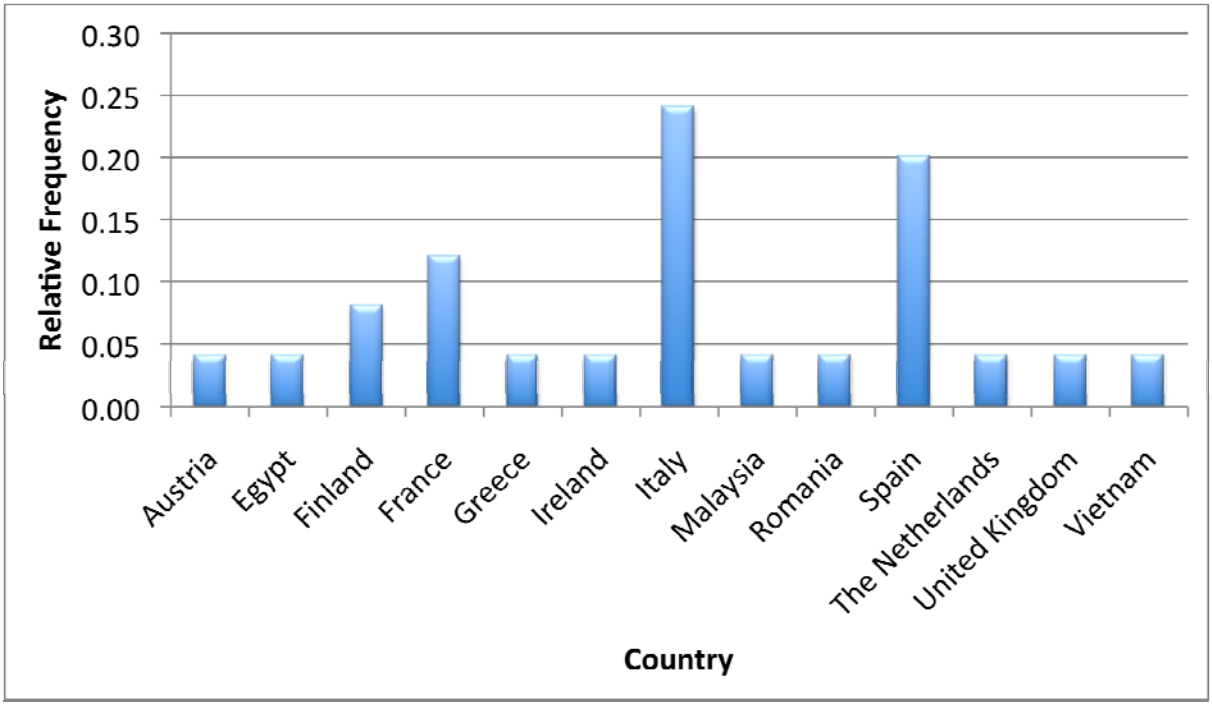

Figure 4.2: Distribution of the countries.

4.1.3 Occupation

Occupation	Absolute Frequency	Relative Frequency
Head of IT	1	0.04
Librarian	1	0.04
Researcher	14	0.56
Student	7	0.28
Teacher	2	0.08

The table above and the histogram of Figure **4.3** report the occupations of the participants in the survey.

Note that many of the PhD students attending the school have defined themselves as “Researchers” while others have chosen “Student”. On the other hand, all the undergraduate students fall under “Student”.

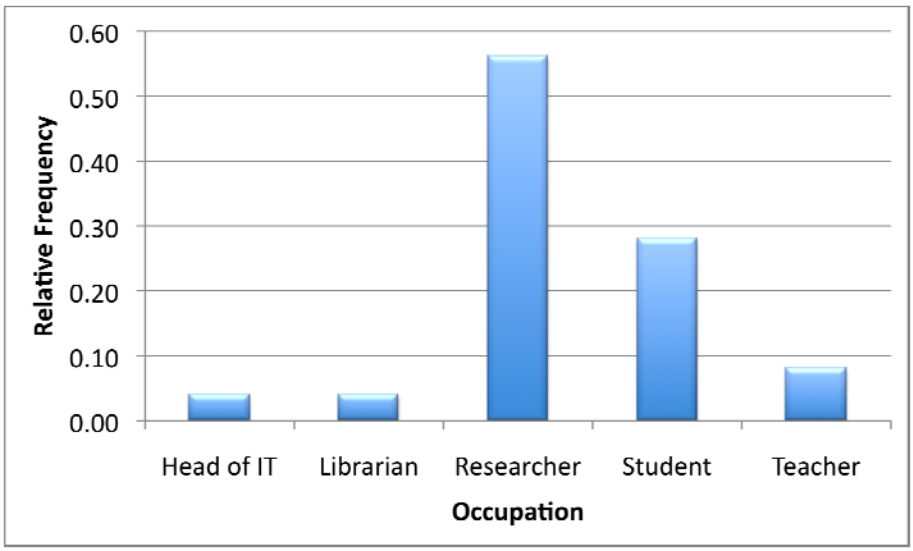

Figure 4.3: Distribution of the occupations.

4.1.4 Native Language

Native Language	Absolute Frequency	Relative Frequency
Abkhazian	1	0.04
Arabic	1	0.04
Dutch	1	0.04
English	1	0.04
Farsi	1	0.04
Finnish	2	0.08
French	3	0.12
German	2	0.08
Greek	1	0.04
Italian	4	0.16
Malay	1	0.04
Romanian	2	0.08
Russian	1	0.04
Spanish	3	0.12
Vietnamese	1	0.04
TOTAL	25	1.00

The table above and the histogram of Figure **4.4** report the native languages of the participants in the survey. The choice of the “Abkhazian” language might be due to an error since it is the first item in the drop down list for selecting the native language.

It can be noted as the distribution of the native languages is more widespread than the one of the countries: this is due to the fact that many participants have moved and now live in a country different from their (or their parents’) original one. Moreover, many different languages are represented in the considered sample – giving an account of the great cultural diversity that characterize Europe – and English is not the prominent one, as it may happen in other contexts such as the Web.

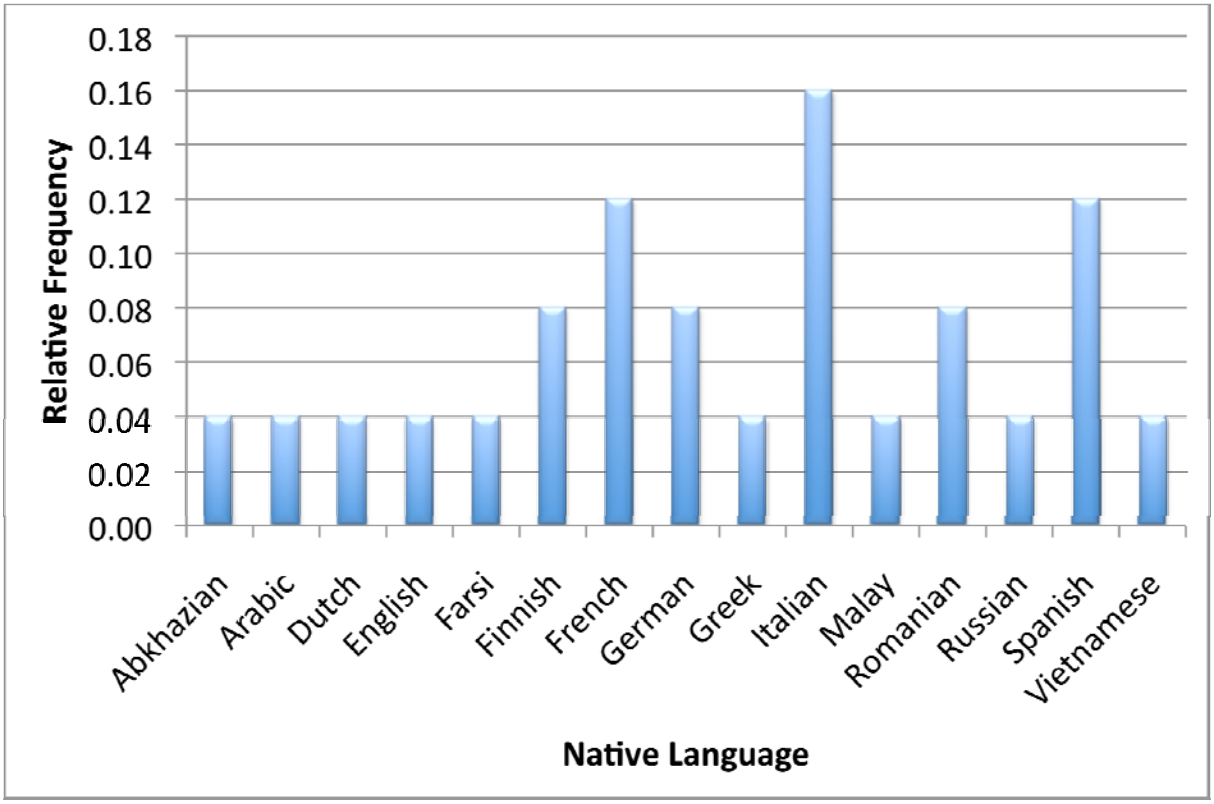

Figure 4.4: Distribution of the native languages.

4.1.5 Known Languages

Language	Absolute Frequency	Relative Frequency
Arabic	1	0.018
English	24	0.428
Estonian	1	0.018
French	11	0.196
German	6	0.107
Italian	4	0.071
Malay	1	0.018
Romanian	1	0.018
Russian	1	0.018
Spanish	2	0.036
Swedish	3	0.054
Turkish	1	0.018

The table above and the histogram of Figure **4.5** report the languages known by the participants in the survey. 56 answers have been provided to this question indicating that, on average, participants know 2.2 languages other than their native one, i.e. that they can deal with about 3 languages.

Almost everybody knows English, even if with different skills. The difference between the number of participants in the survey (25) and the frequency for English (24) is due to the fact that the participant who speaks English as native language has correctly declared to not know it as an additional language.

It is interesting to note that overall among native and known languages, the 25 participants involve 18 different languages. This is quite important from the Europeana point of view because even a very focused user community – the MLIA researchers and students in this case – could require to target a

very high number of languages to deliver them the expected service. This underlines again as multilinguality should be a key concern in the Europeana agenda.

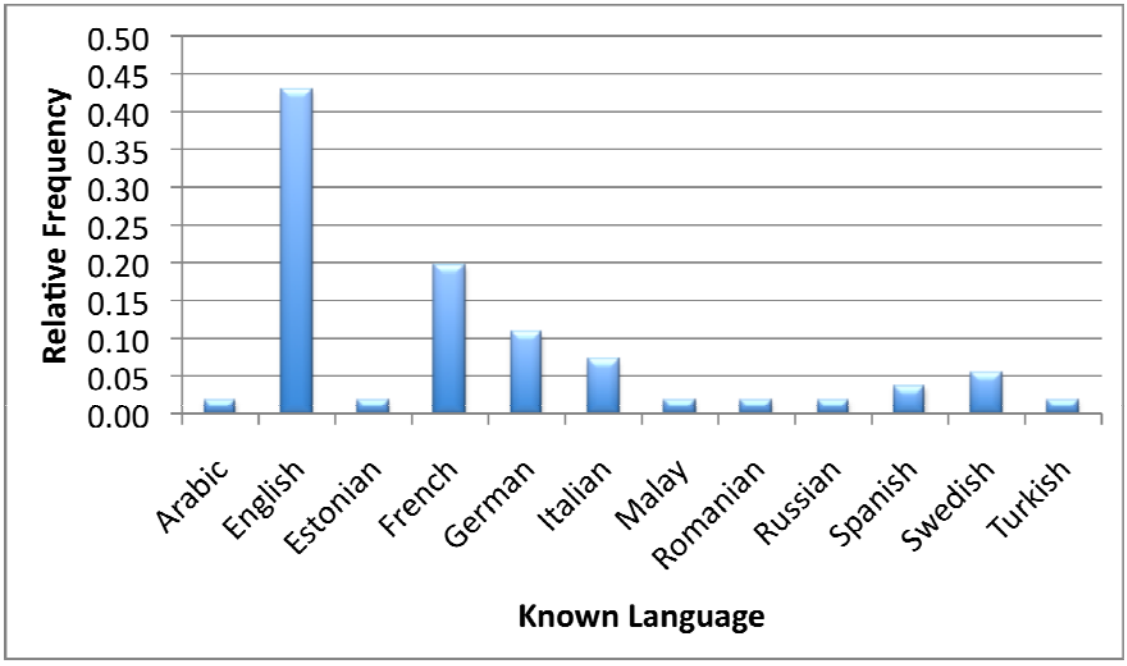

Figure 4.5: Distribution of the known languages.

The table below reports the different language skills for each one of the known languages.

	Very Good		Good		Basic	
	Absolute Frequency	Relative Frequency	Absolute Frequency	Relative Frequency	Absolute Frequency	Relative Frequency
Arabic	0	0.000	1	0.018	0	0.000
English	13	0.232	9	0.161	2	0.036
Estonian	0	0.000	0	0.000	1	0.018
French	2	0.036	3	0.054	6	0.107
German	1	0.018	2	0.036	3	0.054
Italian	1	0.018	1	0.018	2	0.036
Malay	0	0.000	1	0.018	0	0.000
Romanian	1	0.018	0	0.000	0	0.000
Russian	1	0.018	0	0.000	0	0.000
Spanish	1	0.018	0	0.000	1	0.018
Swedish	0	0.000	2	0.036	1	0.018
Turkish	0	0.000	0	0.000	1	0.018

The tables above and the histogram of Figure **4.6** report the different language skills for each one of the known languages.

The three most known languages are: English (quite obvious), French, and German. Nevertheless, while the language skills for English are mostly “very good” and “good”, participants who know French and German mainly have “basic” skills in these languages.

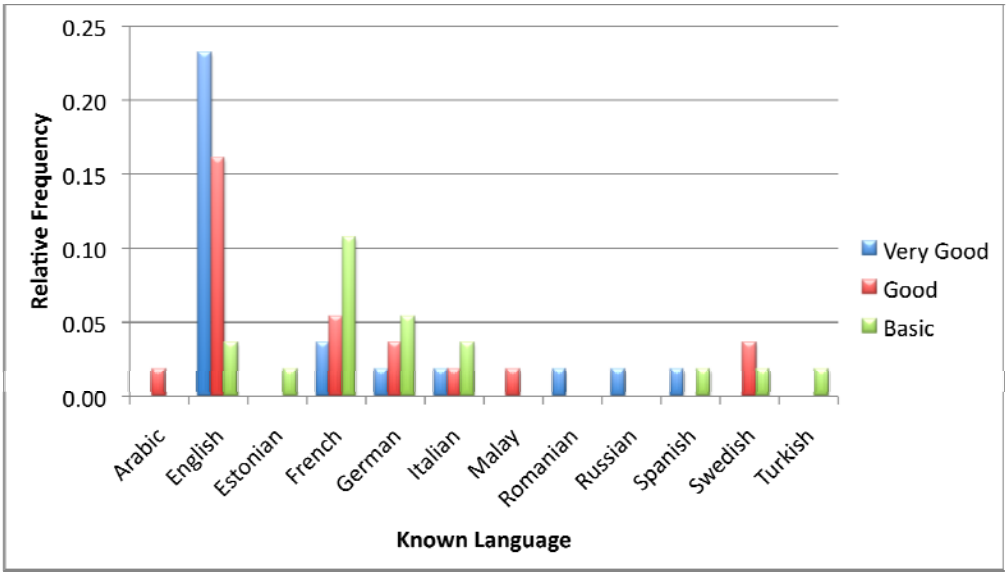

Figure 4.6: Distribution of the language skills.

	Relative frequencies with respect to the row total		
	Very Good	Good	Basic
Arabic	0.000	1.000	0.000
English	0.542	0.375	0.083
Estonian	0.000	0.000	1.000
French	0.182	0.273	0.545
German	0.167	0.333	0.500
Italian	0.250	0.250	0.500
Malay	0.000	1.000	0.000
Romanian	1.000	0.000	0.000
Russian	1.000	0.000	0.000
Spanish	0.500	0.000	0.500
Swedish	0.000	0.667	0.333
Turkish	0.000	0.000	1.000

From the table above, it emerges that 54.2% of the participants who know English know it very well, 37.5% know it well and only 8.3% has a basic knowledge of English. These ratios tend to be reversed when it comes to the other known languages such as French and German, where 54.5% and 50.0%, respectively, has only a basic knowledge.

Relative frequencies with respect to the column total			
	Very Good	Good	Basic
Arabic	0.000	0.053	0.000
English	0.650	0.474	0.118
Estonian	0.000	0.000	0.059
French	0.100	0.158	0.353
German	0.050	0.105	0.176
Italian	0.050	0.053	0.118
Malay	0.000	0.053	0.000
Romanian	0.050	0.000	0.000
Russian	0.050	0.000	0.000
Spanish	0.050	0.000	0.059
Swedish	0.000	0.105	0.059
Turkish	0.000	0.000	0.059

From the table above, it emerges that 65% of the participants who declare to know a language very well know English while only 10% holds for French. Also in the case of a basic knowledge, English is still in the first place. This confirms the role of English as “lingua franca” in the present scenario.

Figure **4.7** provides a view of the participants’ skills by language; it shows the cumulative percentages computed over each group of language skills; for example, if you consider the “Very good” group you have 20 answers out of the 56 total answers, which means that users that have “Very good” knowledge of at least a non-native language amount to 36% of the total. Therefore, from Figure **4.7** it emerges that about 70% of the participants have “very good” or “good” skills in a non-native language even if the abilities vary dramatically from language to language, as it emerges from the previous discussion. However, this overall tendency indicates that for Europeana it could be worth to invest on multilingual information access functionalities since a good portion of the users could be able to actually exploit them.

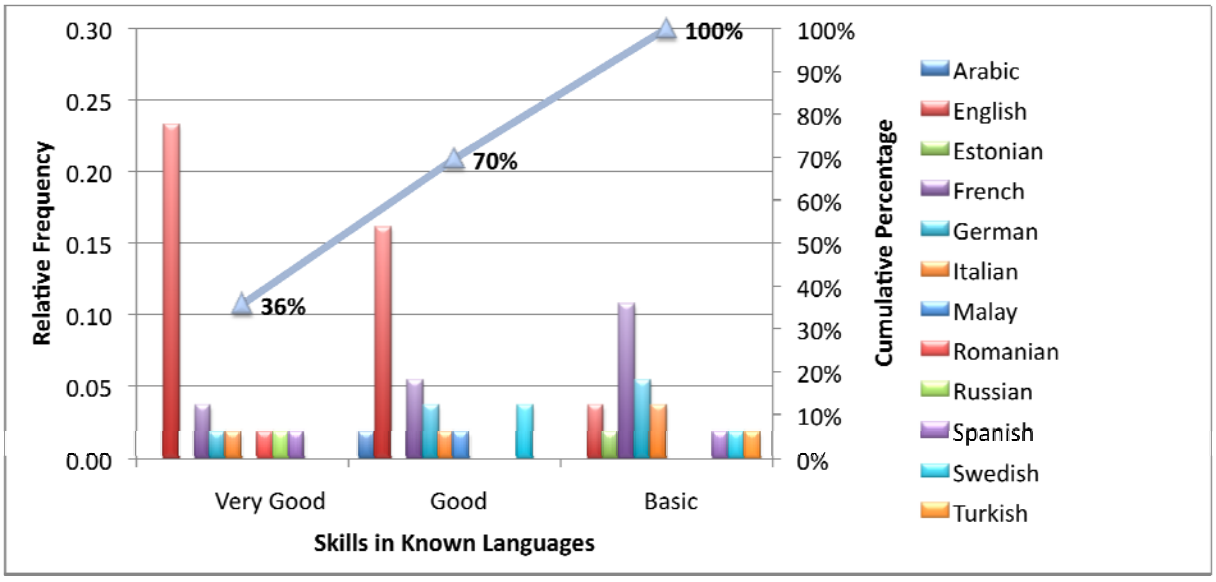

Figure 4.7: Skills in Non Native Languages.

4.1.6 Activities Performed in Other Languages

	Never		Rarely		Sometimes		Often		Always	
	Abs Freq	Rel Freq	Abs Freq	Rel Freq	Abs Freq	Rel Freq	Abs Freq	Rel Freq	Abs Freq	Rel Freq
Reading	1	0.04	0	0.00	0	0.00	12	0.48	12	0.48
Writing	1	0.04	0	0.00	5	0.20	13	0.52	6	0.24
Speaking	0	0.00	2	0.08	6	0.24	11	0.44	6	0.24
Thinking	2	0.08	2	0.08	8	0.32	10	0.40	3	0.12

As it emerges from the table above, 24 out of 25 participants declare to use a language different from the native one *always* or *often* when they read. Only the participant from United Kingdom says to never use a language different from English, which is his native one. Moreover, even if not explicitly mentioned in the questionnaire, the "language other than the native one" to which participants refer to answer the question is English, since it is the language that almost all the participants know very well or well (see previous question).

More than 50% of the participants say to *often* use a non-native language when they write and about 25% say that they *always* use a non-native language when they write. We can find a similar situation when it comes to speaking in a non-native language: 44% of participants *often* speak in a non-native language and 24% *always* speak in a non-native language.

The situation is slightly different when it comes to thinking in a non-native language: 40% of the participants *often* think in another language; 32% *sometimes*; and only 12% *always*. You should note that many of the participants come from northern Europe and, therefore, they are either Anglophone or culturally close to English; moreover, as you can note from the difference between countries and native languages, some of the participants are living in a different country from their original one and this makes it easier for them to think in a non-native language.

Overall, the language skills of the participants are interesting from an Europeana point of view, since they suggest that this kind of users will able to benefit from multilingual information access functionalities and to actively exploit contents provided in multiple languages. Figure **4.8** shows the cumulative percentages computed over each group of the table above; for example, if you consider the "Always" group you have 27 answers out of the 100 total answers, which means that users that "Always" perform such activities in non-native language amount to 27% of the total. From Figure **4.8** you can note that 73% of the participants *often* or *always* perform these activities in a non-native language and, overall, 92% of the participants is somehow used to deal with complex intellectual activities in a non-native language.

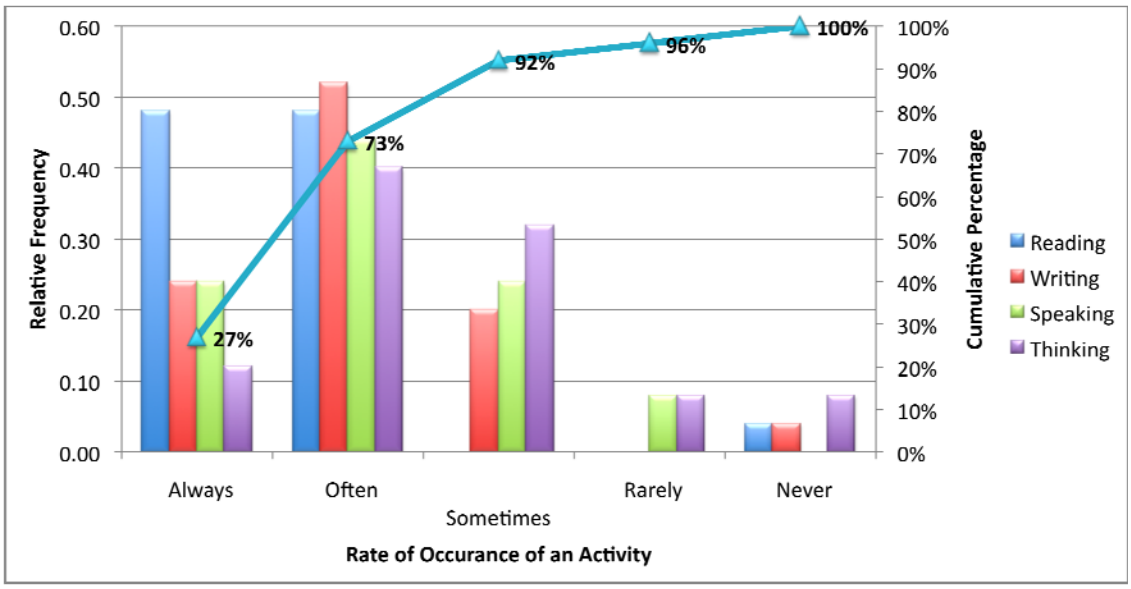

Figure 4.8: How often reading, writing, speaking, and thinking activities are performed.

5.1.7 Used DL Services

Service	Absolute Frequency	Relative Frequency
Online Library Catalogues	18	0.154
Online Journals	22	0.188
Literature Databases	14	0.120
Digital Repositories	15	0.128
Image Archives	17	0.145
Audio Archives	15	0.128
Video Archives	16	0.137

The table above and the histogram of Figure **4.9** report the different DL services that are used by the participants in the survey and show how users know a wide array of the services offered by a DL.

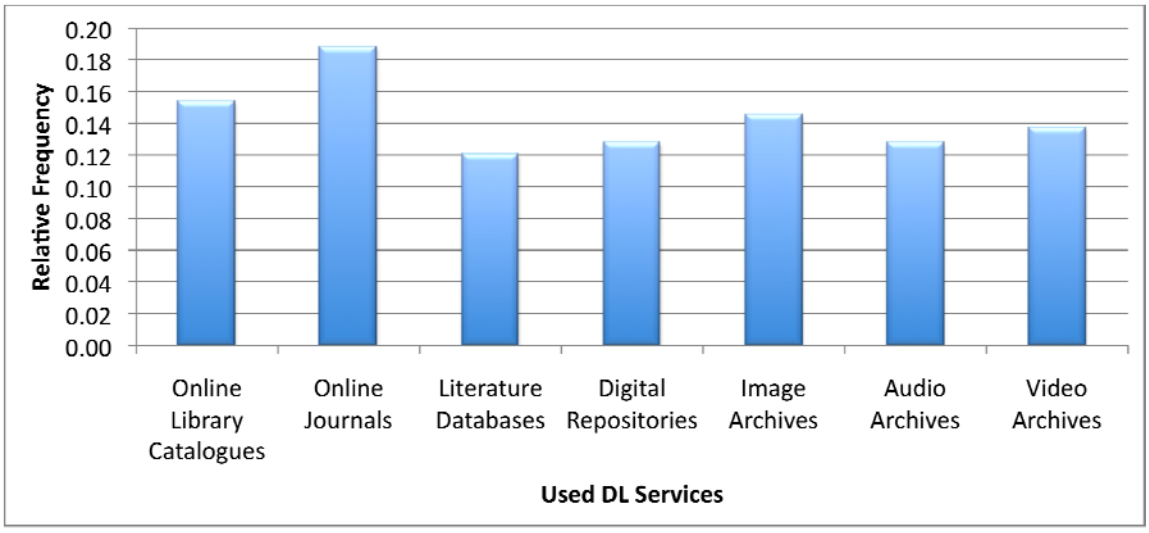

Figure 4.9: Distribution of the used DL services.

Number of Known DL Services	Absolute Frequency	Relative Frequency
2	2	0.08
3	6	0.24
4	5	0.20
5	3	0.12
6	3	0.12
7	6	0.24

The participants in the questionnaire declare to be familiar with one or more DL services, as summarized in the table above and in the histogram of Figure **4.10**.

About a quarter (24%) of the participants is familiar with all the services usually offered by a DL and this confirms the idea the questionnaire has been filled in by subjects quite expert in the field.

However, we still need to get a better understanding of their actual abilities with respect to DL services since this is relevant for Europeana in general, and for the multilingual functionalities, in particular, which can impact each one of the services listed above.

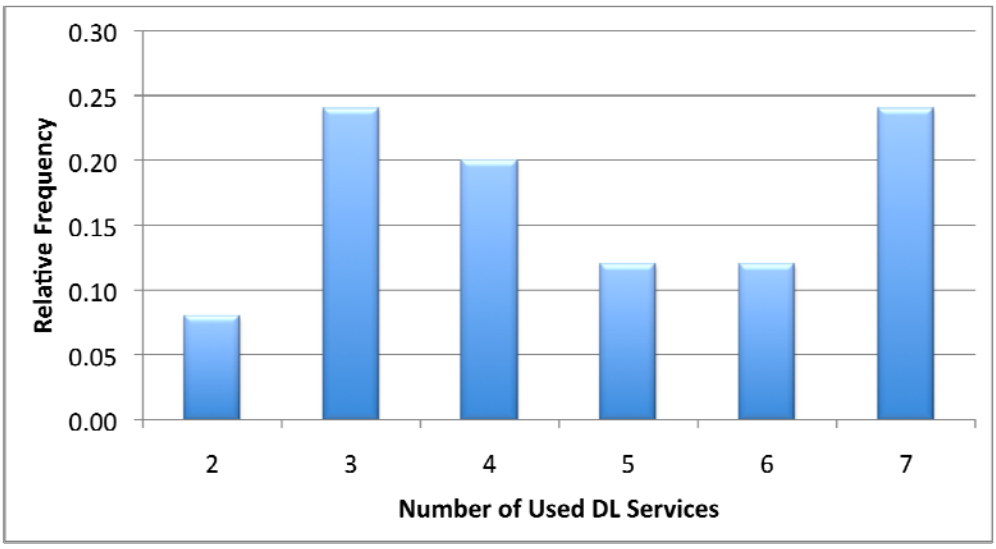

Figure 4.10: Distribution of the number of used DL services.

Let us label the different DL services as follows: A = Online library catalogues, B = Online journals; C = Literature databases; D = Digital repositories; E = Image archives; F = Audio archives; G = Video archives. The following table reports combinations of obtained answers.

Combination of Used DL Services	Absolute Frequency	Relative Frequency
ABC	1	0.04
ABCD	1	0.04
ABCDEFG	6	0.24
ABCDEG	2	0.08
ABCEFG	1	0.04
ABD	1	0.04
ABDF	1	0.04
ABEFG	2	0.08
ADEFG	1	0.04
AEF	1	0.04
AEG	1	0.04
BCD	1	0.04
BCE	1	0.04
BCEF	1	0.04
BD	1	0.04
BDFG	1	0.04
BEFG	1	0.04
BG	1	0.04

Let us group the answers in the following categories: *Limited Use* for those who use 2 or 3 DL services; *Average Use* for those who use 4 o 5 DL service; and, *Advanced Use* for those who use more than 5 DL services.

Kind of Usage of DL Services	Absolute Frequency	Relative Frequency
Limited Use	8	0.32
Average Use	8	0.32
Advanced Use	9	0.36

The table above seems to not provide any interesting indication apart from that Europeana should expect to serve users with varied expertise and that one category seems to be not more prominent than another. This represents a challenge per se, since it would require to put the same effort both in developing basic services as well as advanced ones.

However, we can combine the above information with the one about the occupation of the participants in the questionnaire, as shown below and in Figure **4.11**.

	Head of IT		Librarian		Researcher		Student		Teacher	
	Abs Freq	Rel Freq	Abs Freq	Rel Freq	Abs Freq	Rel Freq	Abs Freq	Rel Freq	Abs Freq	Rel Freq
Limited Use	0	0	0	0	7	0.28	0	0	1	0.04
Average Use	1	0.04	1	0.04	4	0.16	1	0.04	1	0.04
Advanced Use	0	0	0	0	3	0.12	6	0.24	0	0.00

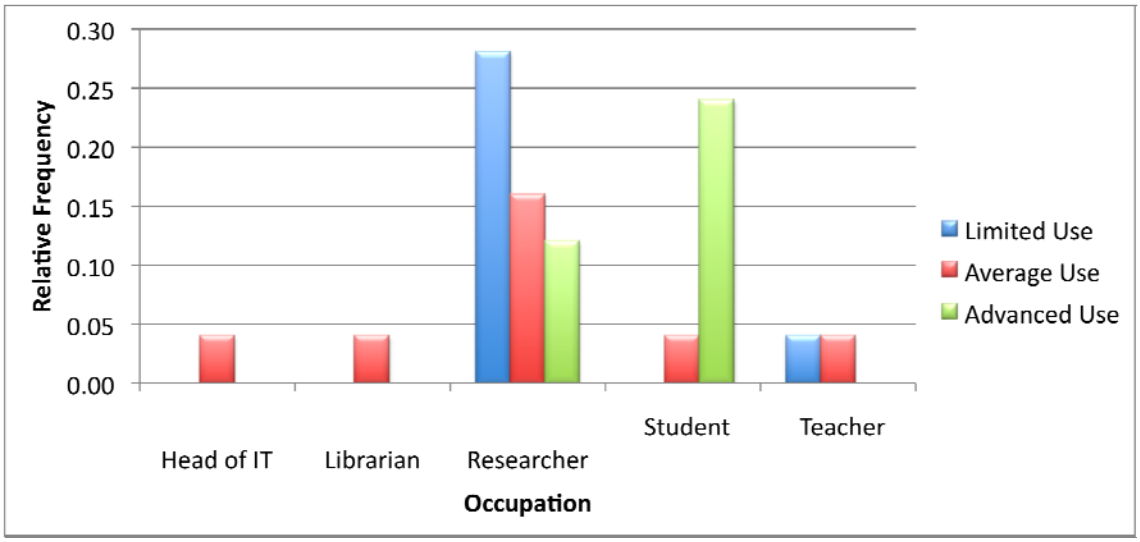

Figure 4.11: Occupation vs kind of usage of DL services.

Surprisingly, it seems that researchers have less familiarity with DL services. As anticipated in section 4.1.3, this suggests that many participants might have chosen researcher as occupation even if they were students and thus they have less familiarity with DL services.

However, this limited familiarity with DL services should be taken into proper consideration by Europeana – should it come from "real" researchers or students – since it calls for proper actions to make each category of users aware of what they can really do with a DL.

4.2 Multilingual Content Interaction

4.2.1 Experience with Multilingual Content

	Absolute Frequency	Relative Frequency
On the Web	25	0.25
Digital Libraries	11	0.11
Journals and newspapers	18	0.18
Books	11	0.11
Radio channels and music	15	0.15
Television and/or films	19	0.19
Other	1	0.01

The table above and the histogram of Figure **4.12** report the experiences with multilingual content that the participants in the questionnaire have.

It clearly emerges that the previous experience with multilingual content equally derives from Web browsing (all the 25 participants have this previous experience), multimedia documents, video, music, journals.

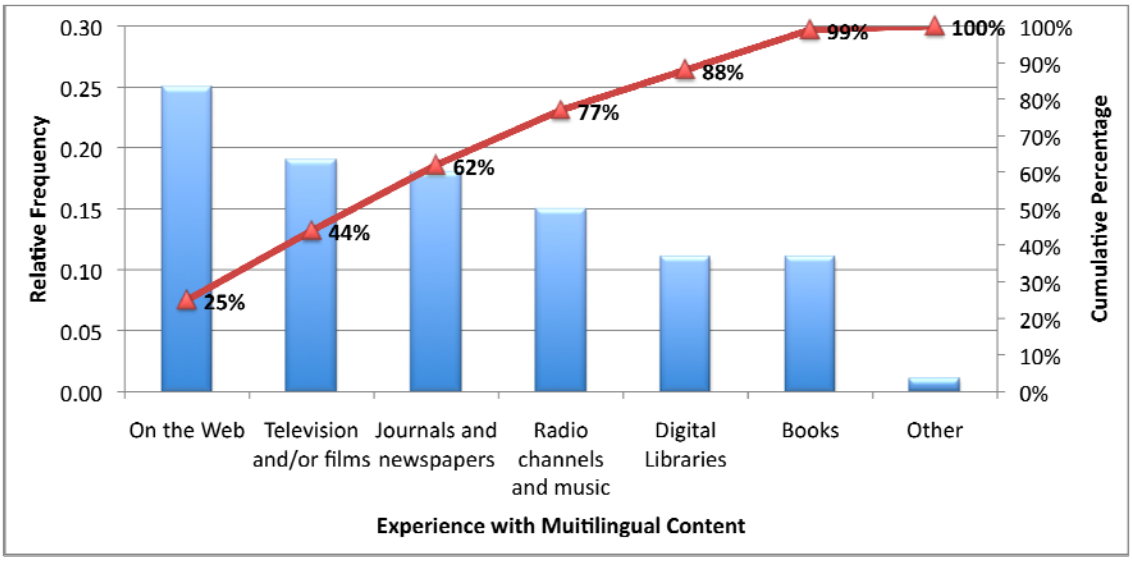

Figure 4.12: Distribution of the experience with multilingual content.

Figure **4.12** shows the cumulative percentages computed over each item of the table above; for example, if you consider the “On the Web” item you have 25 answers out of the 100 total answers, which means that 25% of the user has previous experience with multilingual content on the Web. Therefore, Figure **4.12** shows that 77% of the previous experiences with multilingual content comes from the infotainment area – Web, television, films, radio, journals; this poses a challenge to Europeana since it should offer access to multilingual content in multiple media (text, images, audio, video) in a easy, appealing, and possibly interactive way as the users may have already experience in this domain.

On the other hand, 22% of the previous experiences with multilingual content comes from the culture and cultural heritage area – books and digital libraries in equal measure; this is both encouraging for Europeana, since users are already experienced with multilingual content in the specific domain addressed by Europeana, and puts Europeana in the position of becoming a valuable dissemination and knowledge spreading channel for culturally-related multilingual contents being in the position of addressing about one fifth of the user needs with respect to multilinguality.

4.2.2 Multilingual Tasks

	Never		Seldom		Sometimes		Often		Always	
	Abs Freq	Rel Freq	Abs Freq	Rel Freq	Abs Freq	Rel Freq	Abs Freq	Rel Freq	Abs Freq	Rel Freq
Browsing	1	0.04	2	0.08	8	0.32	11	0.44	3	0.12
Searching	1	0.04	3	0.12	5	0.20	13	0.52	3	0.12
Bookmarking	4	0.16	9	0.36	7	0.28	3	0.12	2	0.08
Printing	4	0.16	9	0.36	8	0.32	3	0.12	1	0.08
Sharing	4	0.16	8	0.32	8	0.32	5	0.20	0	0.00

The table above and the histogram of Figure **4.13** report the rate of occurrence of typical tasks in the case of multilingual content.

The most common operations are Web *browsing* and *searching* multilingual content: 56% of the participants *often* or *always browse* multilingual content and it raises to 88% if we consider also participants who *sometimes browse* multilingual content; similarly, 64% of the participants *often* or *always search* for multilingual content and it raises to 84% if we consider also participants who *sometimes search* for multilingual content.

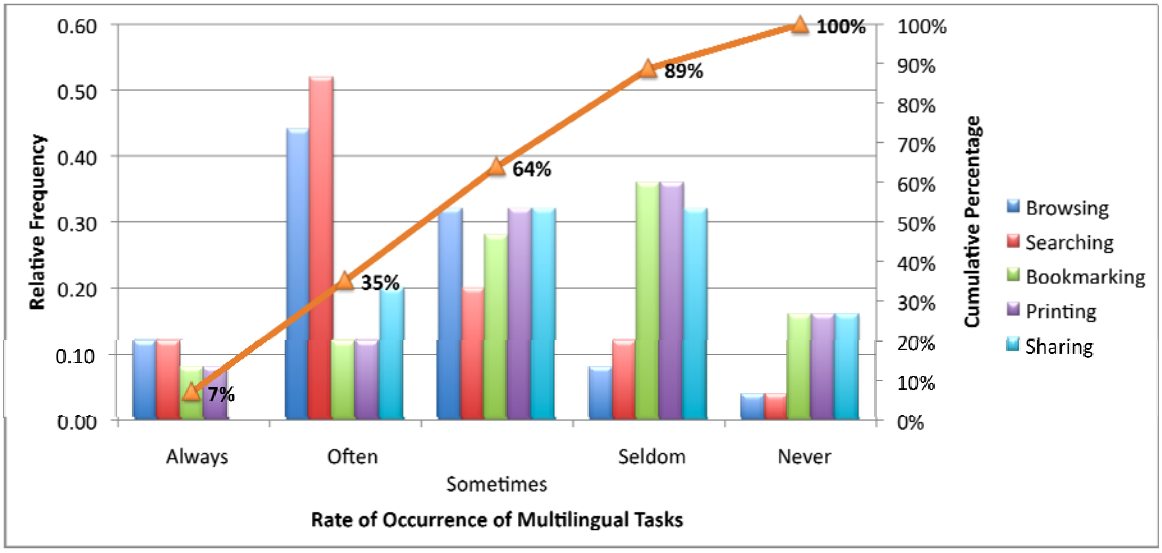

Figure 4.13: How often browsing, searching, bookmarking, printing, and sharing multilingual content are performed.

This represents a clear indication for Europeana about the need of providing effective means for accessing multilingual content. Figure **4.13** shows the cumulative percentages computed over each group of the table above; for example, if you consider the “Always” group you have 9 answers out of the 125 total answers, which means that 7% of the user “Always” performs those multilingual tasks. Therefore, as shown in Figure **4.13**, overall 64% of the participants perform all tasks in the range from *sometimes* to *always*; therefore, Europeana should carefully consider how to offer proper support for these tasks. For example, *bookmarking* requires a careful design of the URL/URI adopted to identify resources which should also take into consideration the fact that the same resource may be

represented in multiple languages; *printing* requires proper style sheets to render the content in a suitable way for the printed representation and this could also depend on cultural factors related to languages; finally, *sharing* calls for active collaboration means such as annotations.

4.3 Multilingual User Interface

4.3.1 Multilingual User Interface in the Native Language

	Absolute Frequency	Relative Frequency
Yes	18	0.72
No	7	0.28

72% of the participants in the questionnaire have expressed their preference for having the user interface in his/her own native language. This provides a clear indication about how much is worth putting effort and resources in internationalizing and localising the user interface of Europeana.

The comments provided by participant when answering “No” further stress the need for proper and accurate internationalization:

> “I always use English interface even if the interface in my native language is available. The reason for that is that anyway almost all development is done in English and thus there are no translation delays or errors.”

> “I'm fine with English. It feels awkward and funny to me have translated interfaces sometimes.”

4.3.2 How to Switch the User Interface to the Native Language

	Absolute Frequency	Relative Frequency
Automatically	7	0.28
Manually	18	0.72

72% of the participants in the questionnaire have expressed their preference for manually switching the Europeana user interface to his/her own native language. The main reason is the little control that users have with automatic switching, especially when they work from different locations.

The following comment explains well the feelings of the users with respect to automatic switching.

> “Automatic switching of the language interface based on geo-location is often embarrassing because it often leaves me no control! It should be accompanied by manual switching. And non-cookie-based one! Ideal way to do that is to use browser-supplied language preferences, because I specify there what are my preferred languages. For example, I'm a Russian native speaker, who uses English US language OS, with Italian or often UK locale, and I'm based in Italy. I set English UK, Russian, Italian in my browser preferences for language. Please, listen to this preferences :) To specify, I talk here about (in Firefox) Tools-Options-Content-Languages option.”

4.4 Information Access and Retrieval

4.4.1 Expected Search Functionalities

	Absolute Frequency	Relative Frequency
Search by author, year, publisher	25	0.30
Search by subject headings	20	0.24
Full Text Search	19	0.23
Additional search types, e.g. “more like this”	14	0.17
Other	6	0.06

The table above and the histogram of Figure **4.14** report the kind of search functionalities that the participants in the questionnaire expect. It emerges that there is no strong predominance of one search functionality with respect to the others. This might be due to the fact that the participants in the questionnaire declare to be “expert” in the field and so they might know equally well all these functionalities.

However, also in this case, it provides a clear indication to Europeana that equal effort and resources should be put in developing both exact match / metadata oriented search functionalities, like author or subject headings search, and best match / full text oriented search functionalities, like full text search or more like this.

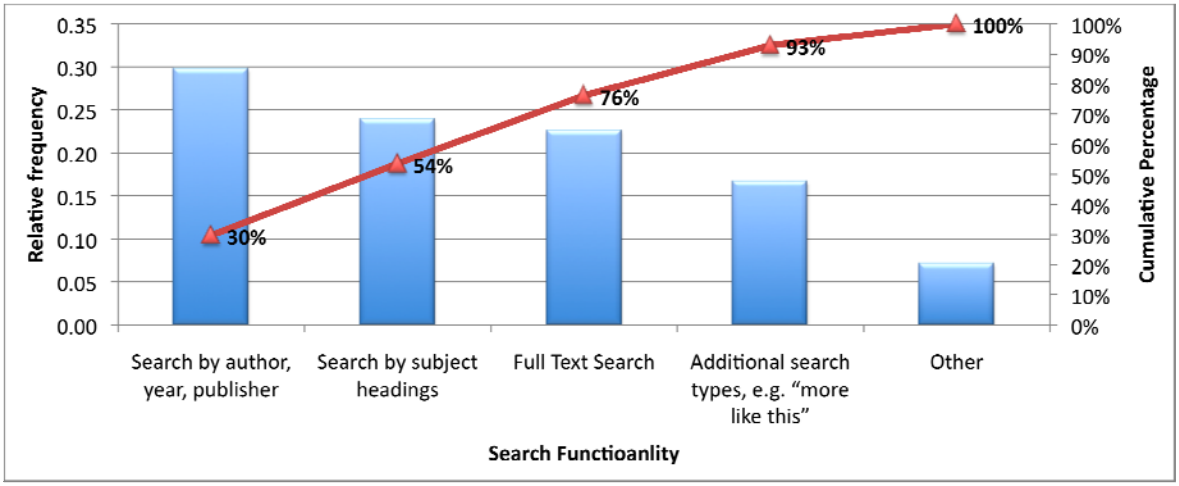

Figure 4.14: Distribution of the expected search functionalities.

As far as the “Other” answer is concerned, the participants provided the following suggestions – which basically fall into the broad categories (exact/best match) of search discussed above:

- Semantic search - related concepts / topics
- By tags
- By ISBN
- By ASIN
- By content type / media type
- Facets
- Relevant information about the topic or the subject but not full text

- By language
- Personalized search on the basis of my profile and preferences
- Links to Amazon and other online stores

Let us label the different search functionalities as follows: A = search by author, year, publisher; B = search by subject headings; C = full text search; D = additional search types, e.g. “more like this”; E = other. The following table reports combinations of obtained answers.

Combination of Expected Search Functionalities	Absolute Frequency	Relative Frequency
A	2	0.08
AB	2	0.08
ABC	5	0.20
ABCD	6	0.24
ABCDE	5	0.20
ABD	1	0.04
ABE	1	0.04
AC	1	0.04
ACD	2	0.08

It emerges that all the participants expect to be able to search by author, year, and publisher (answer A); 80% of the participants expect to be able to search by subject headings (answer C); 76% of the participants expects to be able to have full text search (answer C); finally, 60% of the participants expect also alternative search types, as for example “more like this”, “semantic search”, “faceted search” (answers D and E). This is also evident from Figure **4.14**.

This is quite a relevant information from an Europeana perspective since it indicates that there is much room for providing added-value to users via specific kind of searches beyond the Google-like search that users find on the Web.

4.4.2 Expected Ranking Types

	Absolute Frequency	Relative Frequency
By Fields	17	0.31
By Similarity	22	0.40
By Facets	14	0.25
Other	2	0.04

The table above and the histogram of Figure **4.15** report the kind of rankings that the participants in the questionnaire expect.

It emerges a small preference for similarity-based rankings. However, from the inspection of the questionnaires it comes out that participants have sometimes been inconsistent between this question and the previous one: e.g. they answered to prefer a search by author or year and to expect a ranking by similarity – which is not the case for this kind of searches. This might be due to the questions that were not explicative enough or to some misunderstanding in the participants about what the different kinds of ranking are.

The latter case, if correct, should be carefully addressed by Europeana by making aware users about what they should expect by each type of ranking, otherwise there is the risk to have user dissatisfied because they expect something different from what the system is supposed to do. Finally, it can be noted from Figure **4.15** that the ranking by metadata fields and by similarity cover 71% of the expectations which raise to 96% if we add also faceted ranking.

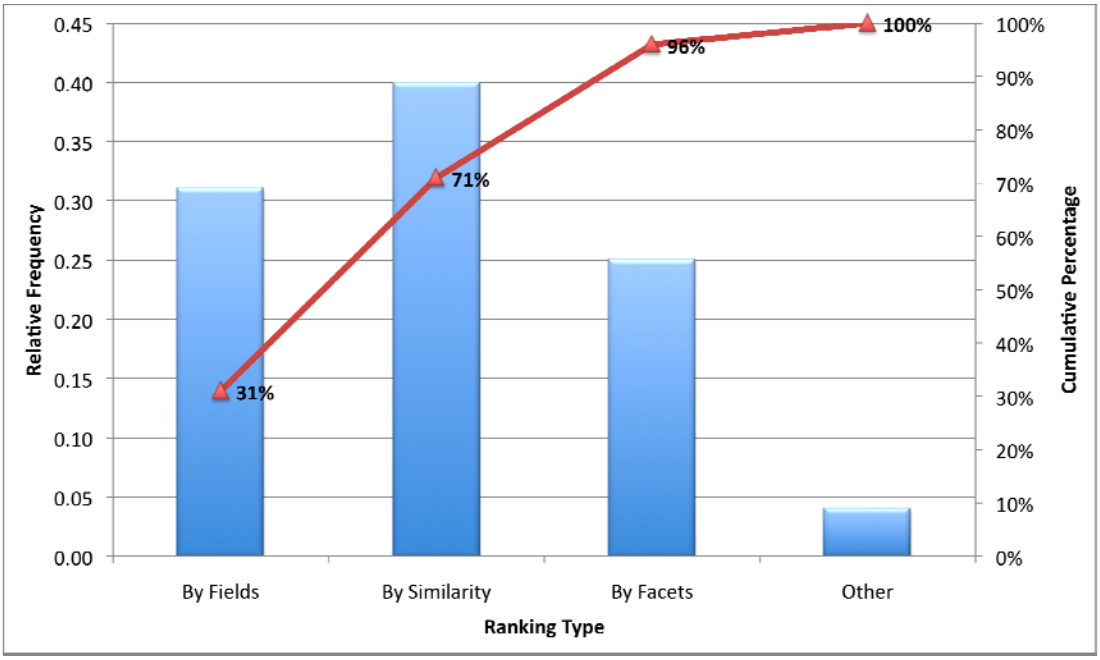

Figure 4.15: Distribution of the expected ranking types.

With respect to the "other" answer, participants specified:

- Ranking based on citation index or other popularity methods
- Results clustered by similarity: If I enter Victor Hugo in the query I would like his books to be identified in one group, biographies in another and critics of his work in another.

Let us label the different ranking types as follows: A = by fields; B = by similarity; C = by facets; D = other. The following table reports combinations of obtained answers.

Combination of Expected Rankings	Absolute Frequency	Relative Frequency
A	2	0.08
AB	2	0.08
ABC	10	0.40
AC	1	0.04
B	5	0.20
BC	3	0.12

It emerges that more than 70% of the participants expect to be offered with different and alternative ranking styles. This confirms the idea the participants are expert in the field and also poses a challenge to Europeana, since it should provide different ranking types and smooth transitions from one ranking type to another.

4.5 Multilingual Information Retrieval

4.5.1 Interest in Multilingual Results

	Absolute Frequency	Relative Frequency
A lot	11	0.44
A little	12	0.48
Not at all	2	0.08

The table above and the histogram of Figure **4.16** report the degree of interest of the participants in being offered with multilingual results in response to a query.

It emerges medium-high interest (92%) in having multilingual results and this provides a clear indication about how much rewarded will be the effort for adding multilingual information access functionalities to Europeana. It should also be noted that one of the "Not at all" answers is due to a participant who is native from United Kingdom who only speaks English.

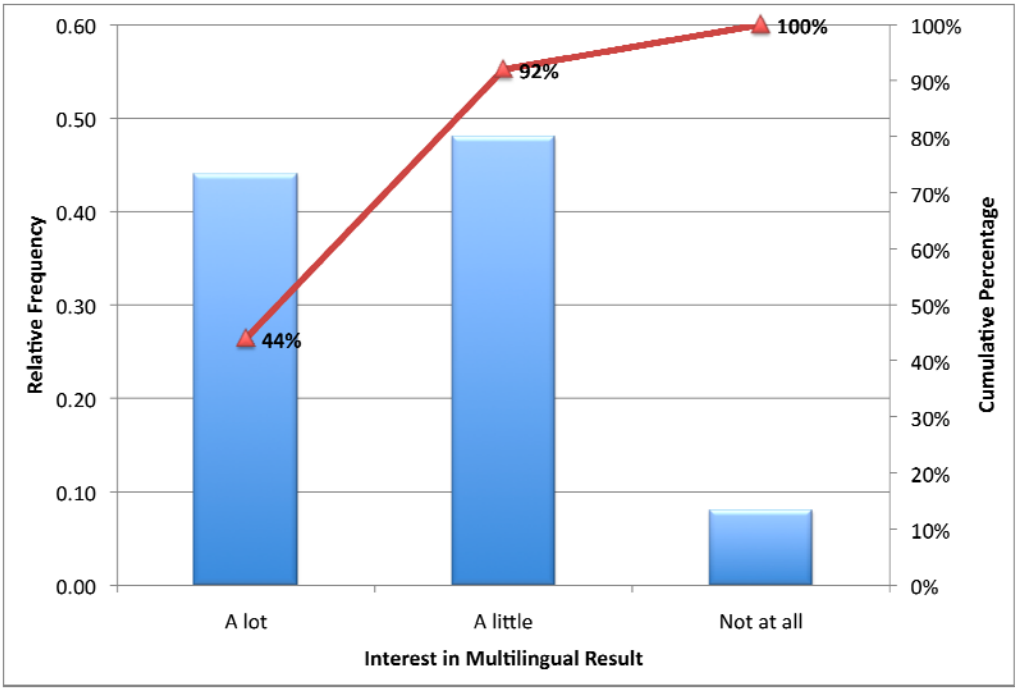

Figure 4.16: Distribution of the interest in retrieving multilingual results.

4.5.2 Multilingual Subject Headings Browsing

	Absolute Frequency	Relative Frequency
Yes but only in my language	3	0.12
Yes, and in multiple languages	20	0.80
No	2	0.08

The table above and the histogram of Figure **4.17** report the degree of interest of the participants in browsing subject headings.

It emerges medium-high interest (92%) in browsing subject headings and, especially in a multilingual way (80%). This provides a clear indication about how much rewarded will be the effort for adding multilingual subject headings browsing functionalities to Europeana. Moreover, it provides an additional indication of where Europeana can gain competitiveness by providing added-value functionalities with respect to the well-known Google-like search paradigm, as already outlined in

Section 5.13 discussing about the expected search functionalities. It should also be noted that one of the "No" answers is due to a participant who is native from United Kingdom who only speaks English.

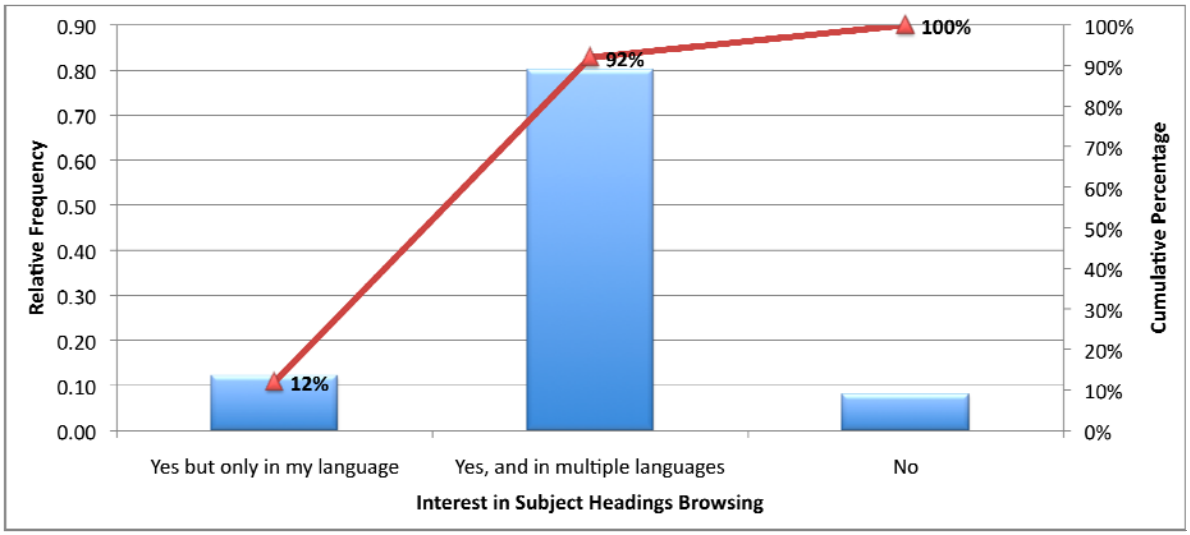

Figure 4.17: Distribution of the interest in subject headings browsing.

4.6 Multilingual Query Formulation and Expansion

4.6.1 Specification of the Desired Language of the Results

	Absolute Frequency	Relative Frequency
Never	7	0.28
Seldom	6	0.24
Sometimes	10	0.40
Often	2	0.08
Always	0	0.00

The table above and the histogram of Figure **4.18** report how often users specify the language desired for the results. The gathered results pose a big challenge for the development of multilingual information access functionalities since 48% of the participants *never* or *seldom* set the language; 40% *sometimes*; and, only 8% *often*.

This means that the developed multilingual information access functionalities should rely as few as possible on the user interaction for deciding the language of the results.

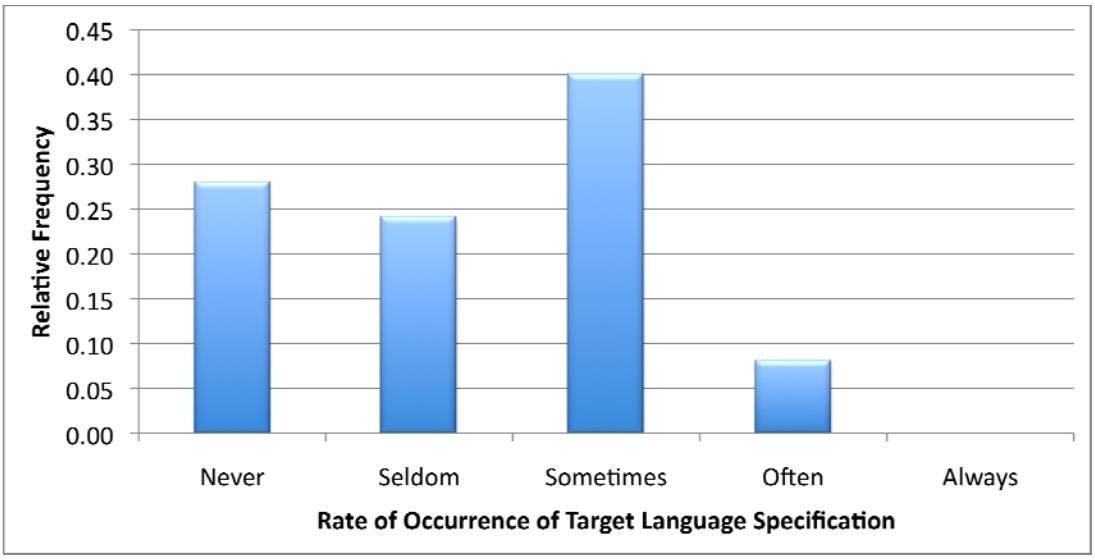

Figure 4.18: How often users specify the desired language for the results.

4.6.2 Number of Target Languages

	Absolute Frequency	Relative Frequency
Never	1	0.04
Seldom	6	0.24
Sometimes	11	0.44
Often	6	0.24
Always	1	0.04

The table above and the histogram of Figure **4.19** report how often users need to deal with multiple target languages for the results.

It emerges that 72% of the subject would need to have multiple target languages in the results. This should be carefully taken into consideration, especially in the light of the previous questions where participants have answered to mostly not specify the desired target languages. Indeed, the challenge becomes how to develop multilingual information access functionalities able to address complex user needs with limited user interaction.

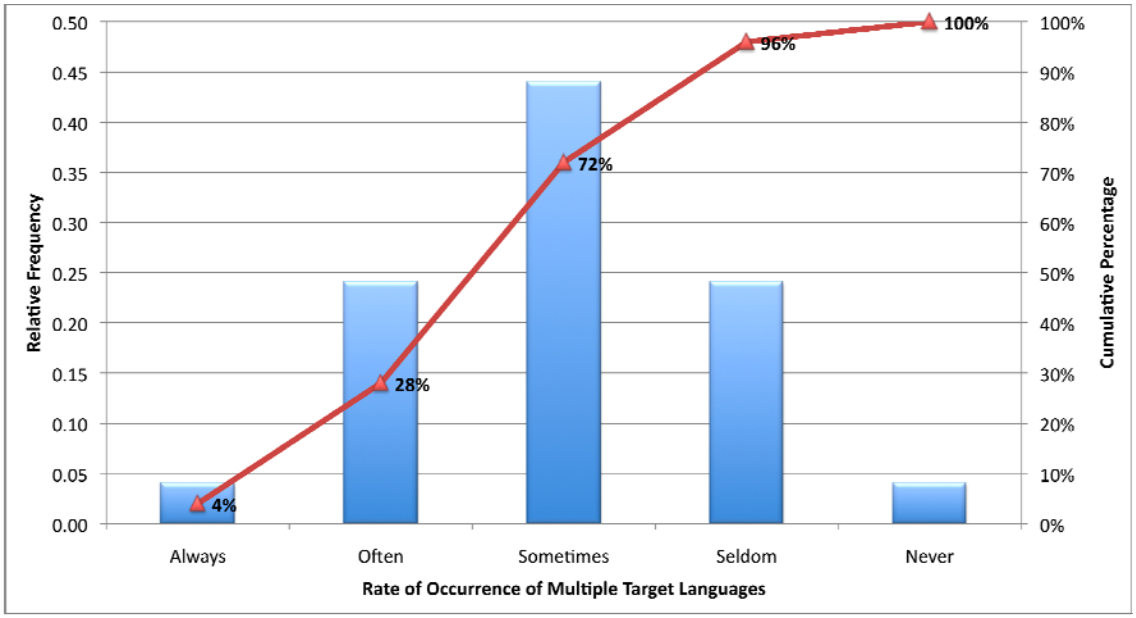

Figure 4.19: How often users need multiple target languages.

4.6.3 Authority Files in Multiple Languages

	Absolute Frequency	Relative Frequency
A lot	9	0.36
A little	10	0.40
Not at all	6	0.24

The table above and the histogram of Figure **4.20** report the degree of interest the participants have in browsing multilingual authority files for person and place names

It emerges medium-high interest (76%) in having multilingual results and this provides an indication about how much rewarded will be the effort for adding multilingual authority files to Europeana. However, with respect to the 92% in the case of multilingual subject headings, it seems that users are less aware of the importance of such kind of tools.

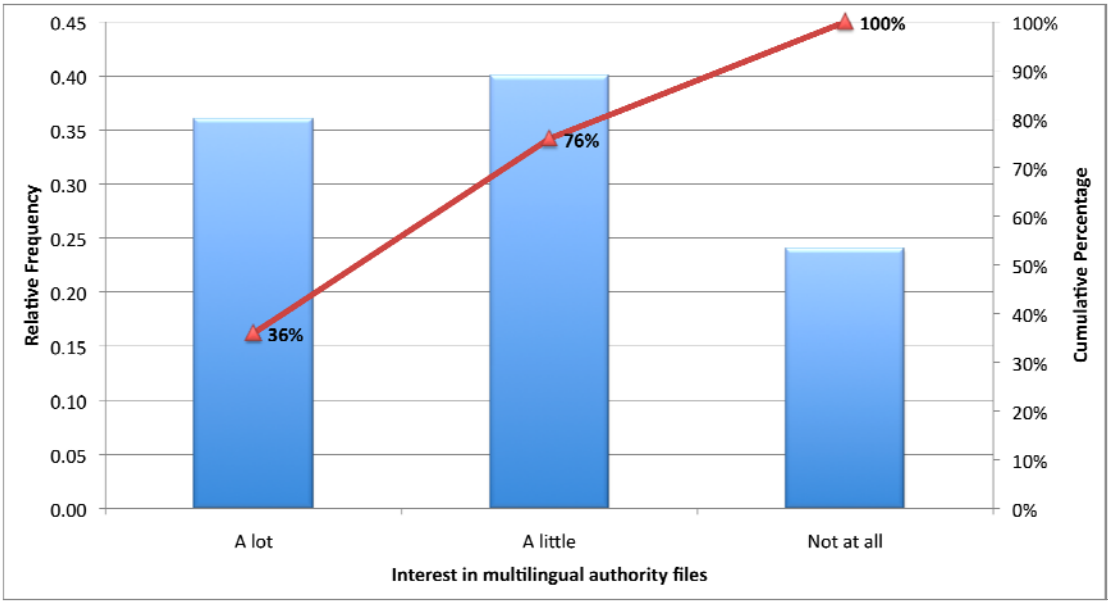

Figure 4.20: Distribution of the interest in multilingual authority files.

4.6.4 Interaction with the Multilingual Query Processing

	Absolute Frequency	Relative Frequency
Automatic	5	0.20
Interactive	20	0.80

From the table above it clearly emerges that 80% of the participants in the questionnaire would like to have the possibility to interact with the query translation process and to refine it, maybe also in an iterative way.

This is a relevant information for the development of multilingual information access functionalities in Europeana since it guarantees that the system can rely on some user interaction to focus the search, making the overall picture less problematic.

It can seem that this answer is partially in conflict with answers to questions 16 and 17 where the users say to not be willing to specify the target languages. This could be explained considering that selecting target languages is a kind of annoying task which the user would happily avoid while refining a query – being it multilingual or not – is generally perceived as in integral part of a search task which needs the expertise and the contribution of the user.

4.7 Multilingual Result Presentation

4.7.1 Multilingual Result Presentation

	Absolute Frequency	Relative Frequency
In relevance order, interleaving results in different languages	12	0.32
Grouping the results by language and, within each group, in relevance order	14	0.36
By highlighting results in different languages with different colours	12	0.32

The table above and the histogram of Figure **4.21** report preferences of the participants as far as the presentation of multilingual results is concerned. They both show a quite even distribution of the preferences, indicating that there is not a clear winner. Please note that, apart from the colour highlighting option, the other two are those actually implemented in the related projects discussed in this report, such as CACAO and MultiMatch.

From an Europeana point of view, this could mean that all these different strategies for presenting multilingual results should be supported and users should be offered with the means for smoothly passing from one to another in order to inspect the result lists from multiple viewpoints.

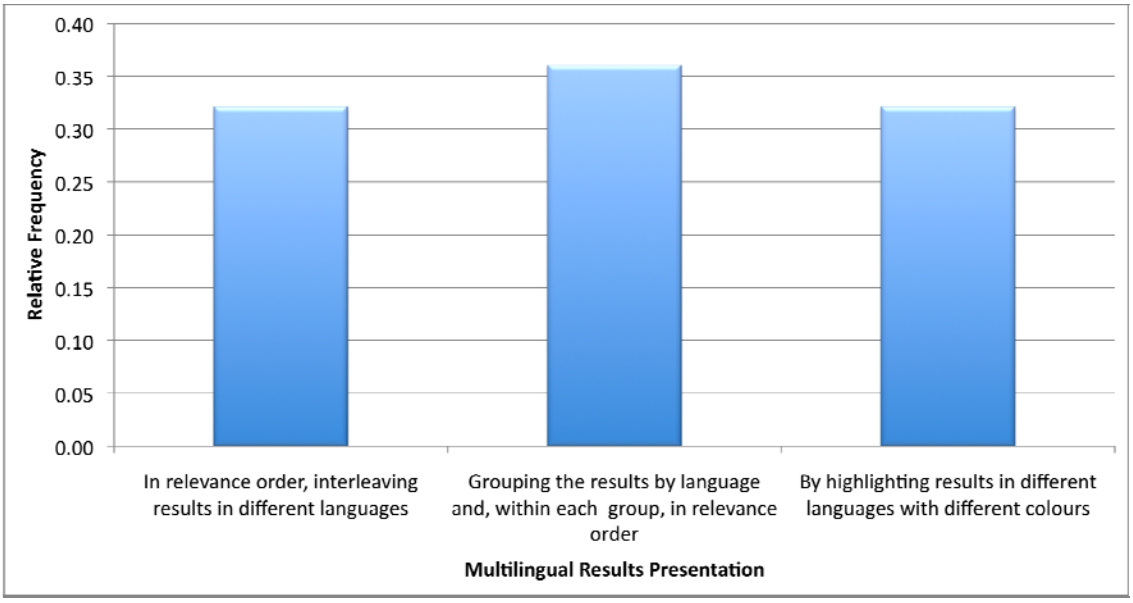

Figure 4.21: Distribution of the preferences for multilingual results presentation.

4.7.2 Multilingual Results Filtering

	Absolute Frequency	Relative Frequency
Yes	14	0.56
No	4	0.16
Not Sure	7	0.28

The table above and the histogram of Figure **4.22** report preferences of the participants as far as the filtering of multilingual results by language is concerned. They show a preference (56%) for filtering results by language, even if a good proportion (28%) of participants in the questionnaire is unsure about this possibility. This might seem somehow unexpected, since all the presentation options of question 20 can include some filtering possibility. This uncertainty of the participants might be due to a difficulty in figuring out this functionality should work, especially in relation with the alternative presentations of question 20; in other terms, this might be an indicator that a too sophisticated interface could be too complex or not motivated enough for some users.

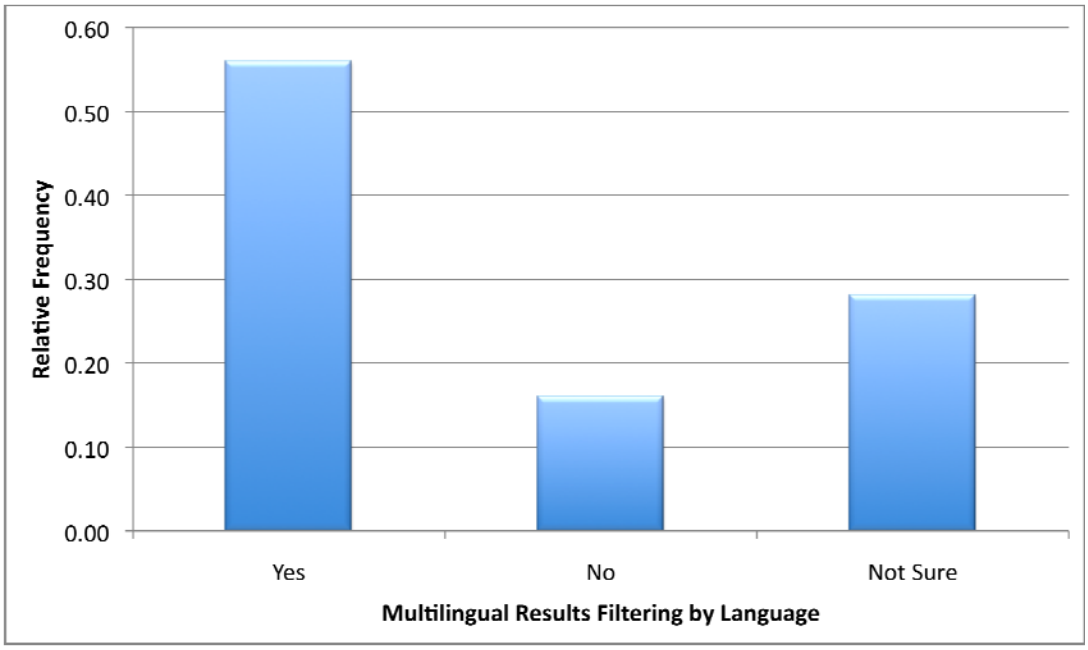

Figure 4.22: Preferences for multilingual results filtering by language.

4.7.3 Multilingual Results Translation

	Absolute Frequency	Relative Frequency
Yes	7	0.28
No	9	0.36
Not Sure	9	0.36

The table above and the histogram of Figure **4.23** report preferences of the participants as far as the translation of multilingual results is concerned. They show a moderate interest (28%) in this functionality, if not even the desire to not have this functionality at all (36%) or uncertainty about its real usefulness (36%). These answers could be somewhat surprising, since it is generally thought that results translation is a needed feature to make it possible to use the results, especially when the language skills are not very high. However, we should consider that the participants in the questionnaire have very good language skills, on average, and this could be a motivation for their lack of interest in result translation. Another cause might be a kind of scepticism for the actual effectiveness of machine translation.

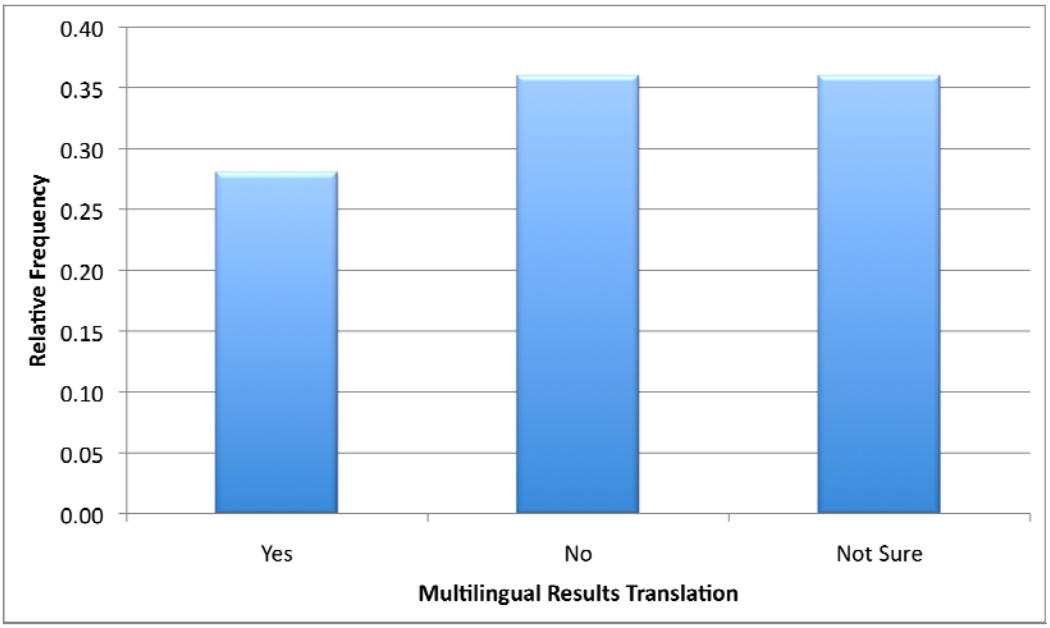

Figure 4.23: Preferences for multilingual results translation.

4.7.4 Expected Translation Quality

	Absolute Frequency	Relative Frequency
Linguistically and syntactically correct translation	6	0.24
Approximate translation	19	0.76

The table above shows how the majority of the participants (76%) does not expect too much from results translation, if provided. As discussed above, this might be a partial explanation of their lack of interest for results translation. However, it represents an important indication from an Europeana point of view since it means that Europeana can deliver a service of satisfactory quality even if the adopted translation engine produces only approximate translations.

5 Suggested Usage Scenarios for Multilingual Information Access in Europeana

Incorporating multilingual components seamlessly and effortlessly for the users poses some challenges, mainly for the user interface, scalability (number of languages) and performance.

This section provides suggestions for two components for multilingual access within Europeana: the query translation process and the representation of query translation options in the result set interface. The suggestions are based on the different aspects of multilinguality the user encounters in a digital library (Section 2), previous user studies which tested parts of these aspects (Section 3) and the user survey on MLIA we conducted (Section 4). The scenarios are designed to put forward ideas how multilingual access could be implemented into the Europeana retrieval process. The entirety of these work flows still needs to be tested on users.

The discussed user requirements of the multilingual interface are not taken into account in these scenarios as it is presumed that the interface will be in the native language of the user. Whether this is determined automatically or by user selection depends on technical conditions. It would be preferable to have an automatic detection of the user preferred language which than still can be edited manually by the user.

The following scenarios do not represent strict work flows but suggest scenarios which are based on the user requirements identified in section 4. It was deliberately avoided to generalize the results and translate them into recommendations for each component of multilingual information access. Each of the evaluated user studies is based on different user groups and profiles so that the result can not be generalized.

5.1 Query Translation

Query translation is one of the core features proposed for Europeana. Figure **5.1** displays a possible workflow for this scenario. It is proposed to perform this query translation process parallel to the first monolingual search (original query input) in order to minimize system response time. Section 5.2 suggests a way to incorporate the translation suggestions into the first result interface.

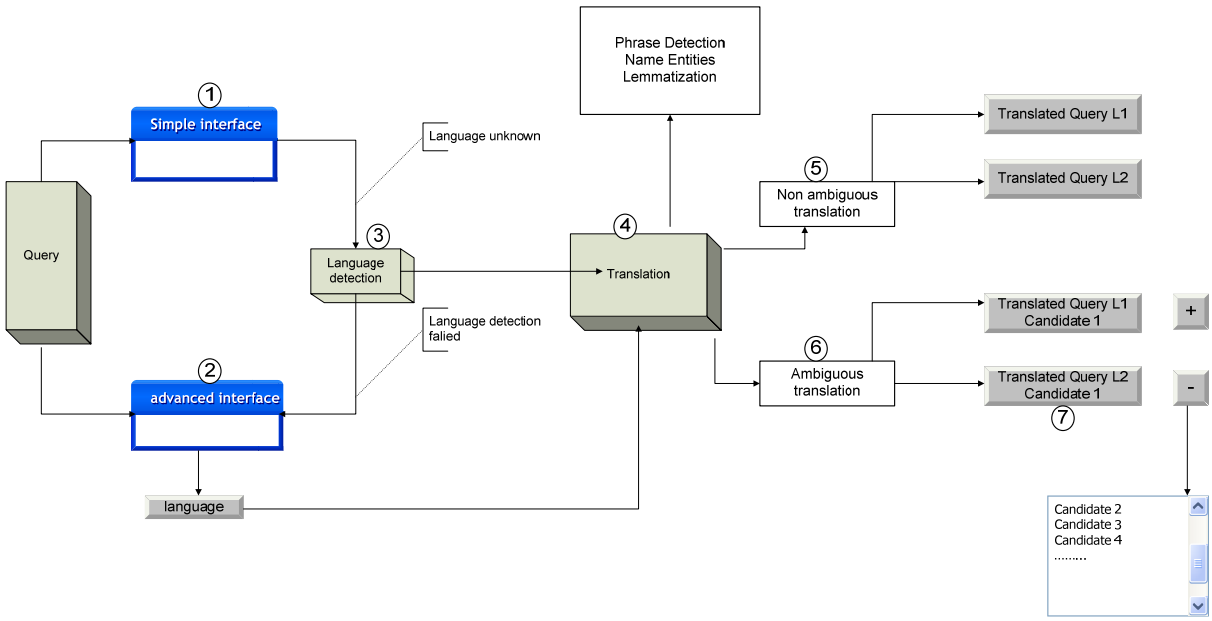

Figure 5.1.: Query Translation for Europeana

(1) Simple interface
- If the basic simple interface is used for querying, the system needs to detect the query language at first.

(2) Advanced interface
- An alternative access could be provided through the advanced interface where the user could indicate the query language, which makes the language detection process unnecessary. Clarity showed that the user would like to determine the query language.
- The advanced interface would also be accessed if the automatic query detection would have failed and the system would not have been able to identify the language.

(3) Language detection
- The evaluated studies showed that no conclusion on the query language can be drawn from the interface language. Language detection of the query term which is not derived from user agent settings or geographic location should be pursued.

(4) Query translator
- The query translation process which identifies queries or named entities in the query term and normalization of the query term should be concealed to the user. The query is translated and all translation candidates are put forward.

(5) Non ambiguous translation

- This should lead to translated queries in the different languages.

(6) Ambiguous translation

- If however, several translations are possible, it wouldn't be very effective to send all translation candidates to the search process, so another clarification step is needed. In this case, either an automatic component tries to determine the most likely desired translation or the user can be asked to provide more input in an interactive disambiguation process. As several user studies showed the user prefers supervision over the translation process to some extend. This could mean to give the user the possibility to select, deselect, add or transform translation candidates.

5.2 Result Set Presentation

If the query translation process is performed parallel to the first retrieval run of the original query as suggested, adding query translations would mean a second retrieval step after the first results are presented. This will add flexibility to the search process, as the user can determine the extent of query expansion through translation. The previous findings from the user studies concerning results representation are considered here.

Figure **5.2** shows a combination of already existing and desired functionalities for the result set representation and search refinement or expansion:

(1) First search:

As a first step, the user types the query into the search box. Europeana presents the list of results as we know it today, which contains several media types, like text documents, images or videos. This process performs a monolingual search (no query translation yet) and returns the results as determined by the Europeana search engine.

(2) Search refinement:

The user can refine his initial search by predefined facets like language, country, date, provider and type. The refinements the user chooses to narrow down the number of search results are accumulated. In any case, the refinement always performs operations on the existing result set and can only narrow the result set by the determined facets. The online survey conducted within Europeana found out that over 50% of all respondents or 69,6% of those reaching the search result page refined the search by language (IRN Research, 2009).

The functionalities described so far are already implemented into Europeana.

(3) Search expansion:

Search expansion means that the result set is expanded by documents that the user has not seen previously. This usually happens by changing or expanding the original query and resending the search to the search index for a new result set.

One possibility for expansion are added terms provided by the semantic enrichment based on the Europeana semantic layer. Possible examples for this idea are already prototypically implemented in the Europeana Thought Lab where they can be tested. Work package 1 within the EuropeanaConnect project is working on semantic representations of the Europeana content in order to provide semantic query enrichment.

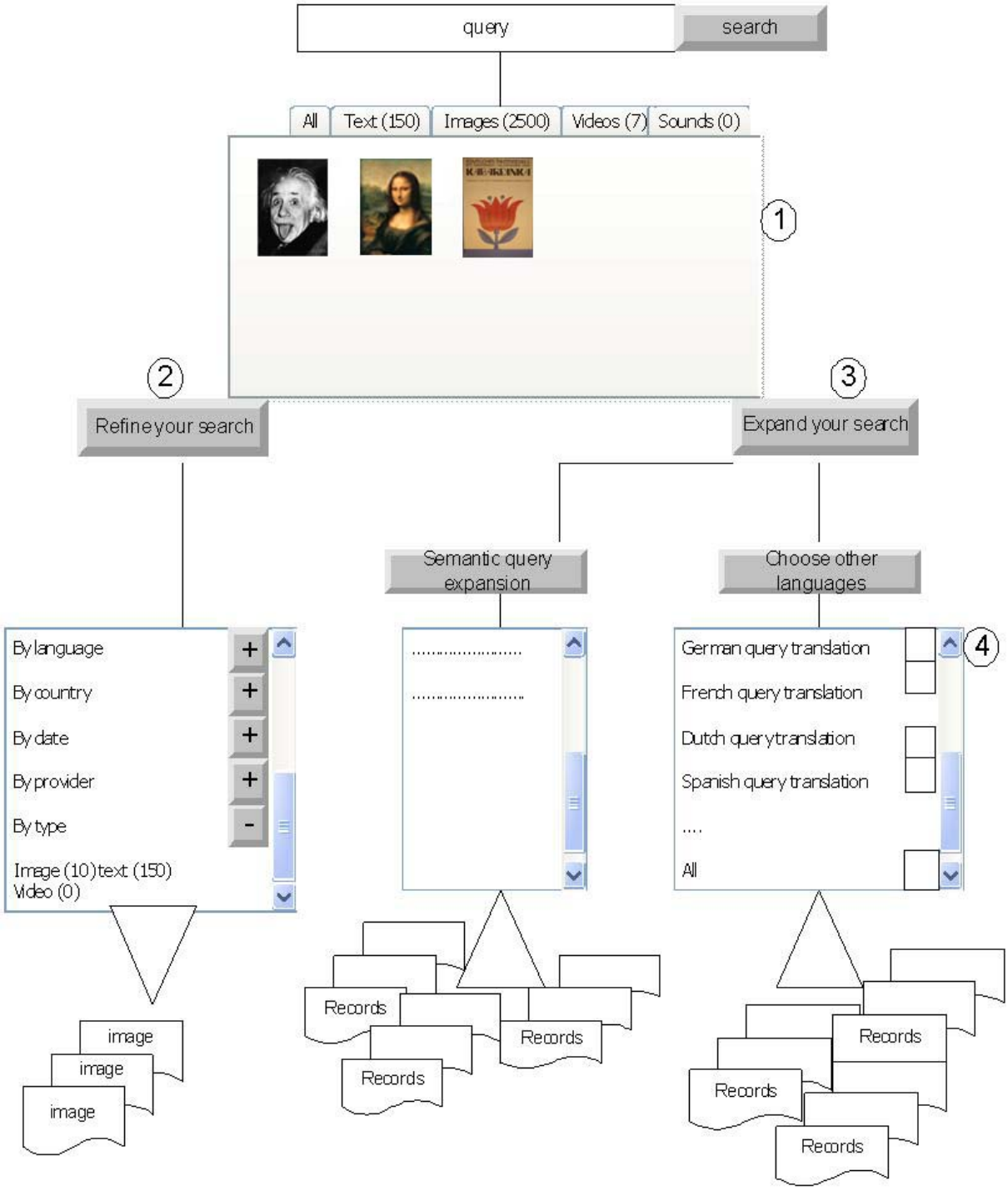

Figure 5.2.: Result Representation and Translation

(4) Adding query translations to the original search:

Another way of expanding the search is to choose a query translation for one or more of the languages Europeana supports (4). The user should be able to select to add translated queries in one language, more than one language or all languages. The search will be performed with those translated queries that were selected added to the original query. The list of search results can be presented in the same format as the initial list (1) with document languages according to the translations the user selected maintaining the interface he is used to.

If the translation candidates are presented for the user to choose from, the user can make an informed decision about the expansion of the query if he or she thinks the translation is correct or useful. This adds flexibility and user interaction to the search process, but also demands for careful inspection of visualization options.

The enrichment of documents with multilingual subject information with the help of knowledge organization system (KOS) mapping not only expands the search space by providing term translations already within documents, but can also aid the user in viewing decisions once the documents appear in a result list. As mentioned before most of the users only want to have translated subjects which helps them to identify whether a document is relevant or not. The translation of documents is not desired and not intended.

5.3 Open Questions and Challenges

The following section describes open questions and challenges deriving from the suggested scenarios. They are based on workshops and discussions in which multilingual issues in digital libraries were discussed. Some of these issues refer to the task Work Package 2 has to deal with in future. They also summarize challenges which might result from the implementation of the suggested scenarios and the listed user needs. In addition they also touch on the employment of an fully functional multilingual digital library.

Language Resources – Licensing, Updating, Maintenance

All multilingual access components depend on language resources that scale up exponentially with each new language that Europeana wants to support - not only adding the language processing tools like stopword lists and stemmers, but also bilingual dictionaries from the new language to each of the supported languages need to be considered.

Language resources are costly to develop and are therefore usually owned and licensed by organizations focused on distributing them professionally. Other tools are already made available open source but might not be fully operational or are not maintained. European initiatives like CLARIN and FlareNet are working on making aggregated language resources available to the scientific community, however due to the number of players and heterogeneity of resources, a standard solution for integrating language resources is not in sight.

In order to provide multilingual access, Europeana will need to find a solution, which language resources can be obtained, updated and maintained in a soluble way. Work package 2 is already working on creating a repository for available language resources but a wider community needs to be responsible for developing stable and agreeable licensing options for all stakeholders involved.

Knowledge Organization System Mapping

Knowledge organization systems (controlled vocabularies) are a special kind of language resource providing highly specialized, technical terminology targeted to describe the content of objects ingested in Europeana. Projects like Minerva have shown the heterogeneity and large number of terminologies being used in object descriptions in the cultural heritage sector and that many occur in several language versions.

However, several large KOS are still maintained in one language, but efforts are under way to connect them not only semantically but also multilingually (the MACS initiative mapping the subject headings lists of the national libraries of Britain, France and Germany to each other is a good example). One challenge is to decide which vocabularies can be mapped to each other and how they could be used

to enrich document descriptions in several language versions. Is multilingual mapping just another mapping layer, which can be added to the semantic mapping already under way in Europeana?

An even more difficult question is how a browsing interface can look like in the face of a multitude of different knowledge organization systems and how this can presented to the user of such an interface.

Document Enrichment – Scalability, Updating, Maintenance

Document enrichment through multilingual subject keywords or document translation can only work if the language of the document is known. This requires the metadata to contain language information not only on the document level but also more specifically on the "field" level. From experience we know that most object descriptions do not contain language information at this level of specificity.

Efforts could be undertaken to add language information during a processing step at document ingestion or indexing, however, questions of scalability and particularly updating remain. Whenever new information is added to the original metadata description delivered by the content provider, this information needs to be stored and handled separately so that updates from the content provider will not overwrite or expunge this information from the record.

Another question is how multilingual versions of the same document (be it only individual fields like subject headings or titles or the whole record) can be stored and handled both in the back-end system but also at the front-end interface. How can this be visualized in an user interface?

Document Translation – Scalability, Updating, Maintenance

As we discussed in Section 2, the main counterpart to query translation when approaching multilinguality is translating the whole collection of documents in multiple languages. While the less frequent choice compared to query translation in academic literature, document translation has been found to be very competitive (Braschler 2004) in some general settings. This is clearly an expensive task from both a computational point of view and a storage point of view and it has to be performed offline in a batch mode.

However, it should be noted that:

- document translation can improve the response time of the search, since all the translation are performed in advance at the collection level and does not happen at query time
- documents are usually longer than queries and this provides more context for disambiguating translations;
- document translation could happen together with other enrichment/augmentation/expansion steps that information resources managed by Europeana will undergo, as for example semantic enrichment; moreover, you should consider that many of these steps will already happen offline for performance reasons.

An approach of this kind has been demonstrated to be successful in the case of the bibliographic records managed by The European Library for a feasibility study that has been conducted in 2006/2007 (Braschler et al. 2006; Agosti et al. 2007; Braschler and Ferro 2007).

Therefore, document translation could be considered for a future development of Europeana, once the overall infrastructure will be more consolidated, more experience will be gained about its functioning and the expectations related to it, and more precise information about the actual volumes to be managed will be available.

Query Translation – Language Detection

Language detection of the query is an essential factor for the success of a query translation system. Section 5.1 proposes a combination of automatic and interactive methods for language identification. Automatic language detection will have to be supported by user interaction or query translation be suspended if the automatic mechanism fails to confidently identify the language of a query - a likely scenario as most language detection algorithms require a certain number of words to work effectively and most queries contain only very few words.

User interaction in the case of failed language detection could take several approaches: support in query reformulation, context disambiguation or a manual identification of the query language. As suggested in 5.1, this would require adapting the search interface to allow the user to input this information. Europeana also needs to develop an alternative option, if language detection and therefore query translation fails and the user is not willing or able to provide the needed information.

Query Translation – Named Entities

Named entities are yet another special case for translation. Multilingual lists of named entities are another language resource, which are very costly to develop, but are already available. Initiatives like VIAF or the Getty Union List of Artist Names for names or the Getty Thesaurus of Geographic Names or other multilingual gazetteers for place names are very good resources, which are already available to Europeana. Other resources will be needed to augment the service.

Incorporating named entity treatment also demands named entity recognition and disambiguation mechanisms during the search process. User interaction could be required if automatic methods fail. This is another challenge for user interface design during the search phase.

Query Translation – User Interaction & Visualization Questions

Section 5.2 proposes a mechanism how query translation options could be offered to the user during the search process. Other integration options are possible and need to be evaluated. In any case, different language versions are one query expansion option that needs to be incorporated into the user interface: a challenge for the interface design.

The disambiguation of translation candidates for the same language (ambiguous translations) poses another visualization and interaction challenge.

Query translation is not the only query expansion option that Europeana plans to offer. Query expansion based on semantic information is one of the core services that will be offered. How query translation and semantic expansion relate or can be integrated (e.g. what comes first) is an open research question.

Studies have found that users having at least a passive knowledge of a candidate language might require modifying or substituting a translation offered by the system. How user-driven modification of translation options can be integrated into the user interface, is an open design question.

Additionally, the user-provided translations should be treated as valuable information and archived to aid the translation process in later stages. This is a multilingual component not yet planned for the current development stage but should be addressed in the future.

Multilingual Browsing

Assuming that a browsing structure can be implemented for Europeana objects, adding multilingual information means that parallel versions need to be maintained in the system and updated whenever the knowledge organization systems are changed. This requires checking mechanisms and potential manual intervention on the system developer side.

Furthermore, the user interface for browsing needs to be adapted to represent multilingual options to the user. Default solutions for visualizing parts of the browsing structure that do not contain multilingual information need to be implemented.

6 References

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, M., Braschler, M., Ferro, N., Peters, C., and Siebinga, S. (2007). Roadmap for MultiLingual Information Access in The European Library. In Fuhr, N., Kovacs, L., and Meghini, C., editors, *Proc. 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2007)*, pages 136-147. Lecture Notes in Computer Science (LNCS) 4675, Springer, Heidelberg, Germany.

Aula, Anne; Kellar, Melanie (2009): Multilingual search strategies. In: CHI EA '09: Proceedings of the 27th international conference extended abstracts on Human factors in computing systems. New York, NY, USA: ACM, 3865-3870.

Ballesteros, L. A. (2000). Cross-Language Retrieval via Transitive Translation. In Croft, W. B., editor, *Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval*, pages 203– 234. Kluwer Academic Publishers, Norwell (MA), USA.

Braschler, M. (2004). Combination Approaches for Multilingual Text Retrieval. *Information Retrieval*, 7(1/2):183–204.

Braschler, M. and Ferro, N. (2007). Adding MultiLingual Information Access to The European Library TEL. In Thanos, C., Borri, F., and Candela, L., editors, *Digital Libraries: Research and Development. First International DELOS Conference. Revised Selected Papers*, pages 218-227. Lecture Notes in Computer Science (LNCS) 4877, Springer, Heidelberg, Germany.

Braschler, M., Ferro, N., and Verleyen, J. (2006). Implementing MLIA in an existing DL system. In Gey, F. C., Kando, N., Peters, C., and Lin, C.-Y., editors, Proc. International Workshop on New Directions in Multilingual Information Access (MLIA 2006), pages 73–76.

Braschler, M., Krause, J., Peters, C., and Schaüble, P. (1998). Cross- Language Information Retrieval (CLIR) Track Overview. In Voorhees, E. M. and Harman, D. K., editors, *The Seventh Text REtrieval Conference (TREC-7),* pages 25–32. National Institute of Standards and Technology (NIST), Special Publication 500-242, Washington, USA.

CACAO (2009): D4.2, "Advanced" interface (with report on usability and accessibility). http://www.cacaoproject.eu/fileadmin/media/Deliverables/CACAO_D4.2.pdf, Accessed 2009-10-

CACAO (2009): D7.4 User requirements for advanced features. http://www.cacaoproject.eu/fileadmin/media/Deliverables/CACAO_D7.4.pdf, Accessed 2009-10-30.

CACAO (2008): D.1.2, Definition of the structure and programmatic interfaces for components access. http://www.cacaoproject.eu/fileadmin/media/Deliverables/CACAO_D1.2.pdf, Accessed 2009-10-30.

CACAO (2008): D2.1, Assessment of Available Lexica. Accessed 2009-10-30.

CHI EA '09 (2009): Proceedings of the 27th international conference extended abstracts on Human factors in computing systems. New York, NY, USA: ACM.

Clough, Paul; Sanderson, Mark (2006): User Experiments with the Eurovision Cross-Language Image Retrieval System. In: *Journal of the American Society for Information Science and Technology*, 57(5), 697 – 708.

Di Bernardi, Raffaella Diego Calvanese Luca Dini Vittorio Tomaso Elisabeth Frasnelli and Ulrike Kugler et al (2006): Multilingual Search in Libraries. The case-study of the Free University of Bozen-Bolzano. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC 2006) http://www.inf.unibz.it/~calvanese/papers/bern-etal-LREC-2006.pdf, Accessed 2009-10-30.

EDLproject (2008): D1.2, Final Report on Usability Developments in The European Library. http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive/edlproject/downloads/EDLprojectD1.2Rep_usage_usability.pdf, Accessed 2009-10-30.

EDLproject (2007): M1.4, Interim Report on Usability Developments in The European Library.

http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive/edlproject/downloads/M1.4_Interim%20Report%20on%20Usage%20and%20Usability.pdf, Accessed 2009-10-30.

EuropeanaConnect (2009). Description of Work. ECP 528001.

Europeana (2009): D2.5 Europeana Outline Functional Specification. For development of an operational European digital library. http://version1.europeana.eu/c/document_library/get_file?uuid=a9e29cb4-a9b3-462a-a43d-0b480c677088&groupId=10602 Accessed 2009-10-30.

Ferro, N. and Peters, C. (2009). CLEF 2009 Ad Hoc Track Overview: TEL & Persian Tasks. In Borri, F., Nardi, A., and Peters, C., editors, Working Notes for the CLEF 2009 Workshop. Published Online at http://www.clef-campaign.org/2009/working_notes/CLEF2009-adhoc-final2.pdf.

FLaReNet (2009): D8.1, Action Plan. http://www.flarenet.eu/sites/default/files/D8.1_v.0.1.pdf, Accessed 2009-10-30.

He, Daqing; Oard, Douglas W.; Plettenberg, Lynne (2006): Studying the Use of Interactive Multilingual Information Retrieval. In *SIGIR workshop on new directions in multilingual information access.* http://terpconnect.umd.edu/~oard/pdf/sigir06ws.pdf, Accessed 2009-10-30.

IRN Research (2009): Europeana online visitor survey Research Report Version 3. http://version1.europeana.eu/c/document_library/get_file?uuid=e165f7f8-981a-436b-8179-d27ec952b8aa&groupId=10602, ACCESSED 2009-10-30.

Ishida, R. and Miller, S. K. (2006). Localization vs. Internationalization. http://www.w3.org/International/questions/qa-i18n.

Janssen, Olaf (2003): Gabriel 1997-2003 & Gabriel/TEL user survey (06.06.2003)

Koch, Walter; Scholz, Henning (2009): DISMARC and BHL-Europe: multilingual access to two aggregation platforms for Europeana. In *Proceedings of the Workshop on Advanced Technologies for Digital Libraries 2009*. 25-29. http://purl.org/bzup/publications/9788860460301, Accessed 2009-10-30.

Landry, Patrice (2009): Providing multilingual subject access through linking of subject heading languages: The MACS approach. In: *Proceedings of the Workshop on Advanced Technologies for Digital Libraries 2009*. 34-37, http://www.cacaoproject.eu/fileadmin/media/AT4DL/paper-09.pdf, Accessed 2009-10-30.

McNamee, P. and Mayfield, J. (2004). Character N-Gram Tokenization for European Language Text Retrieval. Information Retrieval, 7(1-2):73–97.

Marlow, Jennifer; Clough, Paul; Recuero, Juan Cigarran (2009): Exploring the effects of language skills on multilingual web search. In: *Lecture Notes in Computer Science*. Berlin: Springer, Bd. 4956, 126–137.

Minelli, S. H., Marlow, J., Clough, P., Cigarran Recuero, J.M., Gonzalo, J., Oomen, J. and Loschiavo, D. (2007): Gathering requirements for multilingual search of audiovisual material in cultural heritage. In: *Proceedings of Workshop on User Centricity – state of the art* (16th IST Mobile and Wireless Communications Summit, 2007) http://www.multimatch.org/docs/papers/minelli_gathering.pdf, Accessed 2009-10-3.

MLIA4DL Workshop (2009): Multilinguality in Information Access to Digital Libraries - User Needs and Evaluation of Multilingual Resources Use. Workshop at the International Conference on Digital Libraries and the Semantic Web 2009 (ICSD2009). http://www.europeanaconnect.eu/MLIA4DL09Workshop.php, Accessed 2009-10-30.

Minerva Plus (2008): Handbook on cultural web user interaction. http://www.minervaeurope.org/publications/Handbookwebuserinteraction.pdf, Accessed 2009-10-30.

Minerva Plus (2006): D6, Final Plan for using and disseminating knowledge and raise public participation and awareness Report on inventories and multilingualism issues: Multilingualism and Thesaurus. http://www.mek.oszk.hu/minerva/survey/delir20060130.pdf, Accessed 2009-10-30.

Multimatch (2008): D1.1.3 State of the Art Report. http://www.multimatch.org/docs/publicdels/1.1.3.pdf, Accessed 2009-10-30.

Multimatch (2006): D1.1 State of the Art Report. http://www.multimatch.org/docs/publicdels/sota-final-public.pdf, Accessed 2009-10-30.

MultiMatch (2006): D1.2, User Requirements Analysis. http://www.multimatch.org/docs/publicdels/D1.2Final.pdf, Accessed 2009-10-30.

Oakes, Michael; Xu, Yan (2009): Search Log Analysis at the University of Sunderland. Paper presented on the 10th Workshop of the Cross-Language Evaluation Forum.

Oard, D. W. (1997). Alternative Approaches for Cross-Language Text Retrieval. In Hull, D. A. and Oard, D. W., editors, *AAAI Symposium on Cross-Language Text and Speech Retrieval. Papers from the AAAI Spring Symposium*, pages 131–139. American Association for Artificial Intelligence.

Oard, D. W. (2006). Transcending the Tower of Babel: Supporting Access to Multilingual Information with Cross-Language Information Retrieval. In Popp, R. and Yen, J., editors, *Emergent Information Technologies and Enabling Policies for Counter-Terrorism*, pages 131–139. Prentice Hall, Upper Saddle River (NJ), USA.

Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G. J. F., Kurimo, M., Mandl, T., and Peñas, A., editors (2009). Evaluating Systems for Multilingual and Multimodal Information Access: Nineth Workshop of the Cross–Language Evaluation Forum (CLEF 2008). Revised Selected Papers. Lecture Notes in Computer Science (LNCS) 5706, Springer, Heidelberg, Germany.

Peters, C. and Sheridan, P. (2001). Multilingual Information Access. In Agosti, M., Crestani, F., and Pasi, G., editors, *Lectures on Information Retrieval – Third European Summer-School, ESSIR 2000*, pages 51–80. Lecture Notes in Computer Science (LNCS) 1980, Springer, Heidelberg, Germany.

Petrelli, Daniela; Beaulieu, Micheline; Sanderson, Mark (2002): User requirement elicitation for cross-language information retrieval. *The New Review of Information Behaviour Research*, 3, 17-35.

PrestoSpace (2007): D15.5, Cross-linguistic IE tools for metadata discovery. http://prestospace.org/project/deliverables/D15.5_public.pdf, Accessed 2009-10-30.

PrestoSpace (2006): D16.3, MPA3 Cross-language retrieval and access tools.

TEL-ME-MOR (2006): D3.4, Report on Cross-Language Subject Access Options.

TELplus (2009): D3.2, Improving full-text search in printed digital libraries' collections through semantic and multilingual functionalities - Technologies assessment & User requirements.

TELplus project (2008) D3.1, State of the art of semantic and multilingual engines or tools for digital libraries. Accessed 2009-10-30.

TELplus project (2008): D5.1, Report on User Requirements of the Target Library Services. Accessed 2009-10-30.

TrebleCLEF (2008a): D3.2, Workshop on Best Practices for the Development of Multilingual Information Access Systems: the User Perspective. http://www.trebleclef.eu/getfile.php?id=, Accessed 2009-10-30.

TrebleCLEF (2009): D3.3, Best Practices in System-oriented and User-oriented Multilingual Information Access. www.trebleclef.eu/getfile.php?id=249, Accessed 2009-10-30.

TrebleCLEF (2008b): D5.2, Best Practices in Language Resources for Multilingual Information Access

Srinivasarao, Vundavalli (2008): Mining the Behavior of Users in a Multilingual Information Access Task. Cross Language Information Forum. In: *Evaluation of Multilingual and Multi-modal Information Retrieval: 9th Workshop of the Cross-Language Evaluation Forum*. http://www.clef-campaign.org/2008/working_notes/vundavalli-paperCLEF2008.pdf, Accessed 2009-10-30.

Váradi, Tamás; Krauwer, Steven; Wittenburg, Peter; Wynne, Martin; Koskenniemi, Kimmo (2008): CLARIN: Common Language Resources and Technology Infrastructure. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08).

W3C (2009). W3C Internationalization (I18n) Activity. http://www.w3. org/International/.

Zhang, Jin; Lin, Suyu (2007): Multiple language supports in search engines. *Online Information Review* 31 (4), 516-532.

7 Appendices

7.1 Appendix A. Review of Initiatives Concerned with Multilingual Information Access

This section is a state-of-the-art report on existing projects in Europe and beyond which addressed multilinguality and multilingual information retrieval in digital libraries. The aim of this section is to give an overview of applications of multilingual features within digital libraries and on user surveys conducted.

7.1.1 International & European Projects

Minerva & Minerva EC

Minerva EC as one of the projects under the eContentplus program was very successful in creating a network of experts within Europe whose goal it is to improve accessibility to the European cultural heritage (http://www.minervaeurope.org/about/minervaec.htm). Its antecessor Minerva was focused on establishing best practices about digitization, metadata, long-term accessibility and preservation. Within this project a survey was conducted to discover the percentage of multilingual cultural websites within Europe and their use of information retrieval tools, especially multilingual thesauri (see Appendix C). Overall, the best practices derived from the results of the survey are more targeted on the implementation of multilingual user interfaces of a cultural website than on the multilingual features for information retrieval. The recommendations focus on the fact that multilingualism needs to be increased and therefore more multilingual thesauri should be employed. The question of how a multilingual user interface might serve the user needs and requirements or how to implement such a navigational interface in several languages is not touched. Nevertheless the survey delivers an overview of multilingual websites within the European cultural sector and a divers list of multilingual tools for information retrieval which are used.

Minerva EC also addresses multilingual issues in the “handbook on cultural web and user interaction” from September 2008. Here they underline multilinguality as an important aspect of the access to the European digital cultural heritage (Minerva: Handbook on cultural web user interaction, 2008). For information access without language barriers they envision a functionality which enables the user “to find information in foreign languages, read and interpret that information and merge it with information in other languages.” (Minerva: Handbook on cultural web user interaction, 2008). The handbook also offers a self-evaluation questionnaire for assessing user needs while developing web applications. One part of the questionnaire also addresses multilingualism.

Minerva is one of the first European projects which conducted an exhaustive survey on multilingualism. Their focus was on the tools used for information retrieval, especially controlled vocabulary and multilingual thesauri. The Minerva Plus project conducted a major survey to get an overview of the situation concerning language usage in cultural websites (Minerva Plus D6, 2006). In two periods they gathered in total 657 multilingual websites of which are 30% monolingual websites, 43% bilingual websites and 26% multilingual websites.

MICHAEL & MICHAELplus

The MICHAEL project was first a collaboration to establish a portal for digitizing cultural heritage between France, Italy and UK in 2004. MICHAEL plus is the follow-up project initiated in 2006 and additionally supported by Germany, Finland, Greece, Malta, Netherlands, Poland, Portugal, Sweden, Spanish, Czech Republic and Hungary. Aim is to consolidate the different national initiatives for

digitizing cultural heritage. The project realized a multilingual access portal to different collections of cultural heritage within Europe.

MICHAEL offers a multilingual user interface. It allows browsing options with reference to subject, geographic location, time, place and type of institution holding the collection. These are browsing features which are not only limited to different languages but also have a multicultural perspective. For example, the option to look for collections which have a geographic reference to the Middle East, does offer collections in several languages. A translation button leads to the Google translate service which translates the metadata of the collections in the user's preferred language. It is very interesting to see how multilingual access is implemented in this portal. The focus here is on the browsing interface which narrows the query before it is typed. Therefore the workflow of MICHEAL differs from other search interfaces as it only allows search via a browsing interface which requires the user to select a collection before typing a query.

MuSiL

MuSiL is a multilingual search mode which was developed to enhance OPAC searches (Bernardi et al., 2006) Its goal is to improve search in documents written in different languages within a library catalogue. MuSiL provides automatic translation of search queries into German, English and Italian.

An interesting feature is the possibility for the user to expand search results to similar queries which are found by analyzing the semantics of the query.

One possible workflow using the multilingual search interface (http://pro.unibz.it/opacdocdigger/index.asp? MLSearch=TRUE) would include the following steps:

1. User enters search term and specifies the language of the query (English, German or Italian),
2. the system automatically translates the query and offers results within the English language group independent of the query language. It also searches in the other two languages with the translated terms, the user can decide to expand this search field.

Multimatch

Multimatch is a project which aim it is to provide access to online content related to culture heritage. One ambitious objective is to eliminate language barriers for retrieval and give access to cultural content regardless of the media type and the language the content is presented in.

Multimatch delivers several state-of-the-art reports in work package 1 (user requirements and functional specifications) which constitute a basis for research in the area of online access to cultural content. The related report (Multimatch D1.1, 2006 and the later version Multimatch D1.1.3, 2008) offers a valuable summary of recent research and ongoing challenges in the area of multilingual and multimedia information retrieval. One part of the document also deals with user interaction and interface design and outlines recent research focusing on multilingual information access.

Furthermore, in this work package a survey was conducted which was based on the assumption that different user types expect features which are targeted on their needs (Multimatch D1.2, 2006). Based on this survey some user requirements were compiled but they only focused partly on requirements for multilinguality. Next to the survey, Multimatch also analyzed log files and competitors for an elicitation of user requirements (Minelli et al., 2007). Work package 5 in Multimatch is dealing with multilingual and multimedia information retrieval.

CACAO

CACAO offers an approach for multilingual access to textual content. The research within CACAO constitutes a fundamental basis for the development of a system which is able to retrieve textual documents in any language. The method CACAO chose to implement this is by query translation. In work package 1 features such a system should have to enable a correct translation of user queries were described. Modules the system needs to have were determined in D1.2 (CACAO D1.2, 2008) and were categorized as either corpus analyzing modules or query processing modules. The needed software components were presented.

In work package 2, the main goal was to establish language resources which enable multilingual information access across digital libraries. One deliverable which is also of interest for EuropeanaConnect is the list of all language resources available to the different partners of CACAO (CACAO D2.1, 2008). The report lists lexicon resources, multilingual dictionaries, thesauri and word indexes.

Work package 4 presents a user interface which enables the user to query library catalogues within CACAO. Key feature here is the combination of faceted navigation, tag clouds and personalization of the interface (CACAO D4.2, 2009). The paper also presents the different features related to multilinguality which enables to simulate a possible workflow for multilingual access.

TrebleCLEF

TrebleCLEF aims at consolidating knowledge and expertise in multilingual information access. The project evaluates multilingual features and retrieval results and continuously pushes research in this area. In annual campaigns three methods are used to evaluate multilingual information systems:

- test collections
- user evaluation
- log file analysis

One deliverable of TrebleCLEF is to provide best practices in system-oriented and user-oriented multilingual information access (TrebleCLEF D3.3, 2009).

The translation phase can be designed in 4 different ways:

1. query translation, i.e. the translation of the formulation of information need
2. document translation, i.e. the translation of the retrievable items
3. both query and document translation, usually by translating both into a common third language, a "pivot" language or "interlingua"
4. no explicit translation, but use of alternative techniques such as sub-word matching or reliance on cognates.

For the matching phase it is stated that “good monolingual matching in all languages to be covered is a prerequisite for effective multilingual retrieval” (TrebleCLEF D3.3, 2009)

For the user-oriented multilingual systems, it is essential to support interaction with the user across the search process. The paper summarizes results from iCLEF which evaluate the possibility of document selection and results exploration and query formulation, refinement and translation. TrebleCLEF gives best practices and required features for MLIA systems from a user perspective. Another result of TrebleCLEF was a deliverable collecting best practices in language resources (TrebleCLEF D5.2, 2009). Within the work package “Evaluation Packages and Language Resources” a survey was

conducted about existing MLIA language resources. Based on the results of the survey, requirements for MLIA resources were developed.

CLARIN

CLARIN's focus is to create, coordinate and make language resources available and usable (http://www.clarin.eu/). The project wants to build stable services by integrating the existing technology and resources (Koskenniemi et al., 2008).

In work package 5, CLARIN is dealing with tools, lexical resources and corpora. The aim here is to collect existing resources and their standards and integrate them into an infrastructure which offers interoperability between languages and domains. Therefore, CLARIN compiled repositories of language resources (http://www.clarin.eu/view_resources) and tools (http://www.clarin.eu/view_tools).

The compiled list is very essential and gives a good overview of existing LRs and tools. It is publicly available and can be expanded. Information about licensing agreements for these resources and tools were not collected.

FLaReNet

FLaReNet wants to develop a common strategy in the area of language resources and language technologies which consolidates the different efforts in Europe and worldwide. The project aims at building a community of language resource and technology experts, it wants to establish a roadmap and establish recommendations for the community (FLaReNet D8.1, 2009).

One milestone of the project, which has not been published, is the constitution of the state of the art about language resources and technologies.

7.1.2 Europeana Networked Projects

In the following section, projects from the Europeana network are examined for research related to multilinguality. Especially valuable are the projects which preceded Europeana and EuropeanaConnect as these results are the basis for future research within the Europeana network.

TEL

The European Library wanted to initiate a framework and set up a system for access to all European national libraries and their collections, digital or non-digital. Several other projects were initiated around TEL to explore and solve pressing issues which occur while converting the European Library into an European Digital Library (EDL)

(http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive_en.html#edl).

TEL-ME-MOR

The project integrated national library collections from ten New Member states in TEL. Goal of the project was also to develop multilingual interfaces in the languages of these member states. One work package (WP3) was dedicated to the task of making the collections of the national libraries available in the national languages through interfaces and search mechanisms. The first two tasks in WP3 were trying to identify possibilities in displaying data correctly across all languages, the remaining tasks dealt with CLIR (TEL-ME-MOR D3.4, 2006). Within this task an overview about projects and research on interoperability between subject tools was delivered. The projects listed are categorized in:

- interoperability between different subject headings (MACS),
- interoperability between different thesauri (Merimee, UMLS Metathesaurus, TermSiences)

- interopearbility between subject headings and classification, UDC (MSAC)
- interoperability between subject headings and classifications, DDC (WebDewey, CrissCross)
- interoperability between various subject indexing tools (HILT, Renardus).

EDLproject

The EDLproject is build on the project TEL and aims at providing access to electronic library services of the European National libraries. The project was able to enlist all EU countries in the European library service and develop multilingual features within the portal. This task was assigned to work package 2 which developed multilingual interfaces and search mechanisms.

Log files were analyzed to derive best practices for search interfaces. The subject of multilinguality was also touched within this analysis with the result that 84% of the users do not change the language of the interface no matter from which country they access the website (EDL D1.2, 2008).The log file analysis was also based on work done in work package 1 in an interim report on usage and usability developments in The European Library (EDL M1.4, 2007).

TELplus

TELplus' workplan has two work packages which are interesting for MLIA issues. WP3 is responsible for improving access by mapping vocabulary automatically and therefore improving multilingual subject access. In the deliverable D3.2 (TELplus D3.2, 2009), they investigate 5 functionalities with regards to cross-language capabilities:

1. Translation of request
2. Disambiguation of requests
3. Translation of result list
4. Translation of whole or parts of documents,
5. Translation of extracted entities, keywords and categories

For all these functionalities they list technical solutions and possible tools which offer these solutions.

Based on these multilingual features, a survey was conducted to explore the user's point of view regarding the implementation and development of these functionalities (TELplus D3.2, 2009). In addition, they studied semantic and multilingual technologies and features which are offered by companies or open source projects (TELplus D3.1, 2008). Work package 5 looked at services which enable user personalization. WP5 derived user requirements through analysis of log file data and user surveys. To reach this goal they also analysed user studies conducted by BNCF, KB and BNF. Furthermore, the University of Padua carried out surveys directly on the TEL portal (TELplus D5.1, 2008). Out of the data gathered only a few aspects deal with multilinguality and related features.

Athena

Athena's main goal is it to integrate digital content from museums into Europeana. It continues research from the previous projects MINERVA and MICHAEL. One issue here is also the variety of content which exists in different languages. Therefore there is a need to enrich the content so it is accessible via Europeana. This task is part of WP4 which tries to integrate existing language resources. A state-of-of-the-art report on existing multilingual, tools, thesauri and technologies is expected in the next months.

PrestoPrime

PrestoPrime was established to develop research and investigate solution for long-term digital media objects, programs and collections. For Europeana especially interesting is the task to provide access to audiovisual content ensuring cross-domain interoperability. Deliverable D16.3 (PrestoSpace D16.3, 2006) investigates the state of the art and tools for cross-lingual information retrieval. Based on this report, D15.5 (PrestoSpace D15.5, 2007) deals with tools for cross-linguistic information extraction and retrieval for metadata discovery.

DISMARC

DISMARC (DIScovering Music ARChives) aggregates metadata about audio fields in different languages (http://www.dismarc.org/). It is part of the Europeana network and is supposed to deliver metadata for Europeana. During the aggregation process web services are used for semantic enrichment. In DISMARC multilingual aspects are taken into account during different iterations of the aggregation and retrieval process, e.g. import of multilingual vocabulary, expansion of queries in selected languages user can chose interface language (Koch et al., 2009).

BHL

BHL wants to develop a multilingual portal providing access to biodiversity literature which is hold by European natural history museums and botanical gardens (http://www.biodiversitylibrary.org/). BHL will test mapping tools build on experience from DISMARC and will implemented them in the first prototype (Koch et al., 2009).

7.1.3 Commercial Search Engines

In this section, multilingual information retrieval in commercial search engines and library information system will be highlighted. In the area of search engines, Google has an interesting approach to MLIA – the retrieval is based on machine translation and user interaction. Also Yahoo! offers multilingual search.

Google

Looking at the implementation of multilingual web retrieval in commercial search, Google is one of the most important players. Google's approach to cross-lingual information retrieval has some valuable features related to language realized in its workflow for finding information in a language different from the query language.

In the study conducted by Zhang in 2007 (Zhang et al., 2007), Google is considered the best search engine in terms of machine translation, also in regards to supporting multiple languages in web retrieval. Google offers within its **"**Language Tools" the "Translated Search" which allows the user to search for documents in a language different from the query language. A workflow scenario for "Translated Search" could be:

- User enters a search query into the search field.
- With a drop-down menu, the user specifies the language of the query.
- User also determines the language of the websites.
- (User interaction possible if user wants to change or refine the query terms.)
- Google translates the query and performs a search with the translated term.

- After submission of the query two result sets are shown, the left side presents the translated results in the language of the query, the right side shows the original results in the language the user wanted to search documents in.

As mentioned in the list, the user is offered a query translation in the language he has chosen. User interaction allows the searcher to edit the translation of the suggested term and search with the translation entered by the user. This is an interesting feature which helps Google to improve its machine translation and the user to refine the search results. Another study was conducted (Marlow et al., 2008) to explore the effect of language skills of users on their web searching behavior. One result was that query refinement in a different language is only performed if the user has at least a passive knowledge of the language he is searching in. Unfortunately this translation service is rather hidden and probably not used very often by the majority of Google users.

Google's translation is still at a lexical level and it does not offer search across multiple languages at the same time. The user needs to know in which language the information he is seeking may be found.

Yahoo!

Yahoo!France and Yahoo!Germany offer a beta-version of a multilingual search functionality. It will automatically translate a search term in French, German or English (the search query needs to be in German or French) and translate the results back into the initial language the search term was in (http://de.docs.yahoo.com/translator/grund.html). The query translation is hidden and it is therefore not traceable with which translation terms the search was done. Additionally, the ranking and compounding of the results is not clear and the user has no possibility to determine in which language the search will be performed. (Multimatch D1.1, 2006)

7.2 Appendix B. Overview of relevant European Project Reports

ATHENA: Access to cultural heritage networks across Europe

- Leroi, Marie-Véronique; Holland, Johann (2009), Identification of existing terminology resources in museums. D4.1 (31.07.2009, 84 p.).
- McKenna, Gordon; de Loof, Chris (2009), Report on existing standards applied by European museums. D3.1 (30.04.2009, 126 p.). http://www.athenaeurope.org/index.php?en/149/athena-deliverables-and-documents

CACAO Project: Cross-Language Access to Catalogues and On-Line Libraries (eContentplus Programme of the European Commission)

- Rondeau, Gilbert; Roux Claude; Schiller, Anne (2009). "Advanced" interface (with report on usability and accessibility). D4.2 (29.05.2009, 13 p.). http://www.cacaoproject.eu/fileadmin/media/Deliverables/CACAO_D4.2.pdf
- Trojahn, Cassia; Siciliano, Luigi (2009). User requirements for advanced features. D7.4 (28.02.2009, 12 p.). http://www.cacaoproject.eu/fileadmin/media/Deliverables/CACAO_D7.4.pdf
- Bernardi, Raffaella; Bosca, Alessio; Chambers, Sally (2008), Final Report On Standards Compliance. D.3.4 (30.11.2008, 11 p.). http://www.cacaoproject.eu/fileadmin/media/Deliverables/CACAO_D3.4.pdf
- Bosca, Alessio (2008), Definition of the structure and programmatic interfaces for components access. D.1.2 (30.05.2008, 9 p.) http://www.cacaoproject.eu/fileadmin/media/Deliverables/CACAO_D1.2.pdf
- Roux, Claude (2008), Definition of programmatic interfaces (and their implementation) for multilingual access. D.2.2 (30.05.2008, 6 p.) http://www.cacaoproject.eu/fileadmin/media/Deliverables/CACAO_D2.2.pdf
- Roux, Claude; Schiller, Anne (2008), Assessment of Available Lexica. D.2.1 (18.04.2008,13 p.)
- Roux, Claude (2008), User Requirements. D.7.1 (18.04.2008, 15 p.) http://www.cacaoproject.eu/fileadmin/media/Deliverables/CACAO_D7.1.pdf

CLARIN: Common Language Resources and Technology Infrastructure (A European Research Infrastructure)

- Quochi, Valeria; Lemnitzer, Lothar; Kemp-Snijders, Marc (Ed.)(2009), Usage and Workflow Scenarios. D5R-2 (01.04.2009, 59 p

Europeana.net (EDLnet; Thematic Network)

- Dekkers, Makx; Gradmann, Stefan; Meghini; Carlo (2009), Europeana Outline Functional Specification. For development of an operational European digital library. D2.5 (01.03.2009) http://version1.europeana.eu/c/document_library/get_file?uuid=a9e29cb4-a9b3-462a-a43d-0b480c677088&groupId=10602

FLaReNet: Fostering Language Resources Network (eContentplus Programme of the European Commission)

- Calzolari, Nicoletta; Soria, Claudia; Baroni, Paola (2009), Action Plan. D8.1 (29.05.2009, 18 p.). http://www.flarenet.eu/sites/default/files/D8.1_v.0.1.pdf
- Calzolari, Nicoletta; Baroni, Paola (2009), Shaping the Future of the Multilingual Digital Europe. The European Language Resources and Technologies Forum (12.-13.02.2009, 105 p.). http://www.flarenet.eu/sites/default/files/Vienna09_Proceedings.pdf

MICHAEL: Multilingual Inventory of Cultural Heritage in Europe

- Christaki, Anna; Tzouvaras, Vassilis; Fresa, Antonella (2007), Achieving Interoperability in the MichaelPlus Project. (2007, 4 p.) http://www.delos.info/files/pdf/DELOS%20Multimatch%202007/Papers/8tzouvaras.pdf
- Fresa, Antonella (2005), MICHAEL: Multilingual Inventory of Cultural Heritage in Europe. (2005, 6 p.). http://www.michael-culture.eu/documents/fresaeva05.pdf

MINERVA Plus

- Handbook on cultural web user interaction. (September 2008, 170 p.). http://www.minervaeurope.org/publications/Handbookwebuserinteraction.pdf
- Final Plan for using and disseminating knowledge and raise public participation and awareness Report on inventories and multilingualism issues: Multilingualism and Thesaurus. D6 (07.02.2006, 121 p.). http://www.mek.oszk.hu/minerva/survey/delir20060130.pdf

MultiMatch: Multilingual/Multimedia Access to Cultural Heritage

- Peters, Carol; Oomen, Johan; Ibbotson, Carl (2008), State of the Art Report. D1.1.3 (December 2008, 157 p.). http://www.multimatch.org/docs/publicdels/1.1.3.pdf
- Jones, Gareth J. F; Zhang, Ying; Newman, Eamonn (2007), Multilingual/Multimedia Information Retrieval Engine for Prototype One. D5.1 Multilingual/Multimedia Information Retrieval Engine (September 2007, 22 p)
- Peters, Carol; Oomen, Johan; Ibbotson, Carl (2006), State of the Art Report. D1.1 (December 2006, 127 p.). http://www.multimatch.org/docs/publicdels/sota-final-public.pdf
- Minelli, Sam; Del Secco, Ilaria; Naldi Giovanna (2006) User Requirements Analysis. D1.2 (15.11.2006, 134 p.). http://www.multimatch.org/docs/publicdels/D1.2Final.pdf
- Oomen, Johan (2006), First Analysis of Metadata in the Cultural Heritage Domain. D2.1 (23.10.2006, 119 p.). http://www.multimatch.org/docs/publicdels/D2%201-FINAL-2006-23-10.pdf

TEL: The European Library

- Braschler, Martin; Ferro, Nicola (2007), Adding Multilingual Information Access to the European Library. (2007, 10 p.). (Lecture notes in computer science, 4877). http://www.springerlink.com/content/u6450w43wx5253vv/

TEL-ME-MORE: The European Library - Modular Extensions for Mediating Online Resources

- Clavel-Merrin, Genevieve; Žumer, Maja; Landry, Patrice (2006), Scenarios for multilingual access. D3.6 (17.07.2006, 26 p.).

- Landry, Patrice; Žumer, Maja; Clavel-Merrin, Genevieve (2006), Report on Cross-Language Subject Access Options. D3.4 (19.05.2006, 48 p.).
- Žumer, Maja; Clavel-Merrin, Genevieve; Landry, Patrice(2006), Report on subject access tools. D3.3 (30.01.2006, 75 p.)

TELplus project – The European Library Plus Project

- Mane, Laure (2009), Improving full-text search in printed digital libraries' collections through semantic and multilingual functionalities - Technologies assessment & User requirements. D3.2 (19.01.2009, 32 p.).
- Agosti, Maristella; Bergamin, Giovanni; Crivellari, Franco; Di Nunzio, Giorgio Maria; Ferro, Nicola; Ioannidis, Yannis; King, Ross; Lesquins, Noémie; Pomp, Rosemarie; Sadilek, Christian; Stamatogiannakis, Lefteris; Triantafillidi, Mei-Li; van Staveren, Elco; and Vayanou, Maria (2008), Report on User Requirements of the Target Library Services. D5.1 (31.12.2008, 72 p.).

The European Library (EDLproject; eContentplus Programme of the European Commission)

- Fuegi, David (2008), EDL: Final Report. D5.2b (20.03.2008, 9 p.). http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive/edlproject/outcomes.php,
- Angelaki, Georgia (2008), EDLproject. D1.2. Final Report on Usability Developments in The European Library (19.03.2008, 13 p.). http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive/edlproject/downloads/EDLprojectD1.2Rep_usage_usability.pdf
- Clavel-Merrin, Genevieve; Pisanski, Jan; Žumer, Maja (2008), Multilingual achievements in EDL and remaining challenges. D2.1 (28.02.2008, 22 p.). The European Library (eContentplus Programme of the European Commission). http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive/edlproject/downloads/D2%201final.pdf
- Angelaki, Georgia (2007), Interim Report on Usability Developments in The European Library. M1.4. (September 2007, 44 p.) http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive/edlproject/downloads/M1.4_Interim%20Report%20on%20Usage%20and%20Usability.pdf

TrebleCLEF

- Braschler, Martin; Gonzalo, Julio (2009), Best Practices in System-oriented and User-oriented Multilingual Information Access. D3.3 (15.09.2009, 41 p.). http://www.trebleclef.eu/getfile.php?id=249
- Moreau, Nicolas(2009), Best Practices in Language Resources for Multilingual Information Access. D5.2 (March 2009, 84 p.). http://www.trebleclef.eu/publications.php
- Gonzalo, Julio; Peñas, Anselmo; Verdejo, Felisa (2008), Workshop on Best Practices for the Development of Multilingual Information Access Systems: the User Perspective. D3.2 TrebleCLEF User Communities Workshop (31.12.2008, 16 p.). http://www.trebleclef.eu/publications.php
- Moreau, Nicolas (2008), Evaluation Resources for CLEF. D5.1.1 (September 2008, 55 p.). http://www.trebleclef.eu/publications.php

7.3 Appendix C. User Studies within a Multilingual Environment

Some of the projects mentioned above have already dealt with user needs in general. Few studies have addressed user or interface issues related to retrieval in a multilingual environment. Below we attempt to give an overview of work conducted concerning (multilingual) user needs. We introduce user studies and the different methods used to identify user needs and gather requirements with a view to find out what type of multilingual functionalities the users need or want. Individual users have different backgrounds and different skills and require multilingual systems adapt to individual requirements but also assist in helping the users help themselves (He et al., 2006). Based on the user needs the next challenge is to define requirements which feed into the design stage and especially multilingual features.

Multimatch

In Multimatch, the deliverable "User Requirements Analysis" (MultiMatch D1.2, 2006) identified users from three different sectors:

- Educational users
- Tourist users
- Cultural heritage users

Multimatch received its user data from questionnaires, interview focus groups and workshops. The results are presented separately for each user group, which allows insight in the different preferences of the groups and makes it also easy to compare.

Among other things the study found that "if multilingual search was available, (users) would like to have the results associated with descriptive snippets in their own language (preferably) or English (optionally)" (Multimatch D1.2, 2006).

Clarity

In connection with Multimatch, the Clarity prototype was used to perform another user study (Petrelli et al., 2002) to explore questions such as:

- Who wants to use a CLIR system?
- Why do they perform a multilingual search?
- Which features are needed for successful interaction with the system?

Clarity distinguishes between two user groups: those who want to obtain manually translated documents and bilingual users who wish to search for documents in all languages they know but all from a single query. Different user groups might have different requirements concerning the automatic translation of the query, or the document collection, or both. Key findings of the Clarity study were:

- Search multiple languages at the same time: query translation and with it access to different sources is more comfortable to users and will also save a lot of time.
- Choose your own query language: users should have the possibility to choose the language they want to search in depending on the individual skills and the task.
- Support user-created dictionaries: it seems to be very useful to share knowledge with the system and other users.

Eurovision

Eurovision combines existing translation resources to provide web access to an image archive provided by St. Andrews University Library. The Eurovision system was evaluated by multilingual users carrying out two search tasks with the system configured in English and five other languages (Clough & Sanderson 2006). Scenarios involving real users showed that it is possible to search multilinguality without having good knowledge of languages but English.

Eight participants were recruited who should be fluent in a language other than English. The pre-test questionnaire established that "63% of participants searched in a language other than their native language daily and regarded their command of English as fluent" (Clough & Sanderson, 2006)

The main findings with respect to multilingual information retrieval identified in the study are:

- The bilingual dictionary may not contain name entities.
- Dictionary terms and queries can be ambiguous and can add extraneous terms.
- "Bilingual searching is preferable where users can search in English and their native language and create mixed language searchers.
- Users would like to view the translated query in English and be able to modify it if not appropriate prior to searching." (Clough & Sanderson, 2006)

Tate Online

Another case study evaluated the services of Tate Online (Marlow et al., 2007). After an online survey and a log file analysis, a task-based user experiment involving 14 bilingual participants was conducted.

Based on 635 individual answers to the online questionnaire and the evaluation task, key findings were:

- 76.4% of those who did not prefer to view websites in English stated that they would be more likely to visit the collection site if it were translated into their preferred language.
- Priorities for translation were as follows: artists biographies (35,9%), general instructions (i.e., how to use the subject search) (22.4%), search (the ability to enter search terms in a language besides English) (22.1%)
- Most of the participants were willing to accept a text which they could understand but was not perfectly translated.

TrebleCLEF

TrebleClef organized a "Workshop on Best Practices for the Development of Multilingual Information Access Systems: the User Perspective to get more information about user requirements in a multilingual environment" (TrebleCLEF D3.2, 2008). The main goal was to identify the essential features that MLIA systems should offer from a user's perspective. As a result of this workshop, some general user requirements were formulated in another deliverable (TrebleCLEF D3.3, 2009).

The report presents best practice recommendations which were collected from experiments in iCLEF 2002-2003:

- Although users are in general not comfortable in foreign languages, the user should be able to improve the translations.
- The target term translations and its reverse translations should be displayed in a wider context, such as examples or definitions.

- In general, the translation of whole documents seems not be desired or sensible.
- The conclusion from the experiments is that the document translation and query translation / formulation / refinement must be designed together as a whole to produce an optimal translingual search assistant.

iCLEF 2008

iCLEF 2008 proposed a task, which consists of searching images in a naturally multilingual database, namely Flickr (Srinivasarao, 2008). They analyzed the behavior of users when facing strictly multilingual information access task in order to identify the differences in the search behavior according to the language skills. Two different tasks were part of the study, a search log analysis and interactive experiments. The behavior of the most successful users, the least successful users and the users with a success rate in between the two were studied.

Some of their findings were:

- Successful users look at the first two pages of search results and reformulate the query frequently instead of going through many result pages.
- All users reformulated the query very frequently while searching in their mother language as opposed to search in other languages.

CACAO

CACAO described features that are required by the different users of the system (CACAO D7.4, 2009). According to the level of knowledge of the users and their language skills, the interface should provide different services and presentation. Through user profiles information, about the user such as language preferences, query history and selected results is stored and processed.

A study of library catalog search logs found that:

- Within a university or research library, 40% of the queries were “duplicated” in at least two languages (usually in the local language and in English).
- In a library operating in a multicultural context, they observed that about 20% of the queries are written in three languages, namely Italian, German and English.

Google Translate

Google’s translation service (translate.google.com) was evaluated for non-native English speakers (Aula & Kellar, 2009). The main questions of the study were the search language decision, whether users switch languages within a search task and whether users change the lanuage setting on a search engine.

They found that:

- Most of the participants preferred to search in their mother tongue.
- Users formulate difficult questions first in their native language and - only if they are not succesful - then type queries in other languages.
- Users search in English when they want a broader result set.

Presuming that users with different language skills will have different needs another study with the Google web search engine and the Google Translate service focused on users´ language skills (Marlow et al., 2008):

- For monolingual users, query and document translation must be available.

- With bilingual users, document translation often is less desired.

Gabriel & EDLproject

The European Digital Library has also prepared user surveys and log file analysis to gather user requirements. They analyzed weblogs, the Gabriel guestbook and the Gabriel search engine queries (Janssen, 2003). In 2003, users of the Gabriel website were asked to participate in a TEL questionnaire which consisted of 5 parts, altogether 29 questions.

Two thirds of respondents (560 total) indicated they prefer using English on Gabriel. The remaining third wanted a German or French interface.

After the EDLproject did two user studies in the past to receive some feedback, they decided to get additional information from log files (EDL M1.4, 2007).

TEL & TELplus

The University of Padua analysed the IIS http traffic logs, divided in two: 1)the logs from the static part of The European Library portal - "ABOUT US, LIBRARIES, TREASURES" (http://libraries.theeuropeanlibrary.org) and 2) logs from the dynamic part, i.e. the search possibility under SEARCH and COLLECTIONS in the main navigation - http://www.theeuropeanlibrry.org.

One main characteristic about the sessions was that 77.44% involve only 1 query. In general sessions are short in requests and in duration.

One striking finding was that the majority of visitors to the portal do not perform any query. To find out whether users do not know how to search or whether they are not interested in searching a small scale user study was realized with a group of masters' students in information retrieval at the University of Padua.

The Max Planck Institute analyzed the verity server logs (action logs, user tracking) to research the user interaction behavior. In particular, they focused on the query and result-click history. Concerning the interface language selection, they found that the majority of users (84%) leaves the default interface language English. (M1.4, 2007). Another finding was that the most frequent keywords relate to European place names or subjects.

Work package 3 of the TELplus project conducted a user survey, where one of the results was that multilingualism is one of the biggest problems in accessing portals. However, users are suspicious in query translation and disambiguation of the requests. The full translation of documents is not required, only subject translation seems to be useful. (TELplus D3.2, 2009)

Europeana

Building on previous experiences, an online survey to identify specific user profiles for Europeana was designed (IRN Research, 2009). The study, commissioned by the independent research agency IRN Research, found that personal research is the dominant reason given for visiting Europeana: almost three-quarters visit for personal research activities and less than 20% visit for work-related research.

The online survey (3,204 participants) obtained user feedback on a range of issues including:

- Frequency of use of Europeana
- Use and ratings of features and functions
- Stages reached in a search
- Use of My Europeana
- Interest in additional services
- Likely future use of Europeana.

One results for multilingual needs was:

- Over 50% of all respondents or 69.6% of those reaching the search results page refined the search by language.

7.4 Appendix D. Questionnaire about Multilingual Information Access to Europeana

Thank you for taking time to answer questions about Europeana and multilinguality.

Your feedback will help us to develop Europeana and to better understand user requirements and expectations with respect to multilinguality.

Your answers will be treated anonymously and confidentially.

Fields marked with a star (*) are mandatory.

User Profile

1. Age (choose one)*

□ Under 15	□ 15 - 24	□ 25 - 34	□ 35 - 44
□ 45 - 54	□ 55 - 64	□ 65+	

2. Country*

3. Occupation (choose one)*

□ Student at school	□ Student at college / university	□ Researcher
□ Lecturer / professor	□ Teacher	□ Library / information specialist
□ Curator / archivist	□ Manager / administrator	□ Retired
□ Not employed	□ Other:	

4. Native Language*

5. What languages do you know and to what extent? (add as many as needed)

Language Name	Skill (Very good – Good – Basic)

6. How often do you perform the following activities in a language other than your native one? (choose one for each activity)*

	Never	Rarely	Sometimes	Often	Always
Reading	□	□	□	□	□
Writing	□	□	□	□	□
Speaking	□	□	□	□	□
Thinking	□	□	□	□	□

7. Which of the following digital library services are you familiar with? (choose one or more)*

□ Online library catalogs	□ Online journals	□ Literature databases
□ Digital repositories	□ Image archives	□ Audio archives
□ Video archives		
Other (please specify):		

Multilingual Content Interaction

8. Have you had previous experience with multilingual content? (choose one or more)*

□ On the web	□ Digital libraries	□ Journals and newspapers
□ Books	□ Radio channels and music	□ Television and/or films
Other (please specify):		

9. When you use multilingual digital content, how often do you perform the following tasks? (choose one for each task)*

	Never	Seldom	Sometimes	Often	Always
Browsing	□	□	□	□	□
Searching	□	□	□	□	□
Bookmarking	□	□	□	□	□
Printing	□	□	□	□	□
Sharing	□	□	□	□	□

Multilingual User Interface

10. Would you like to have the user interface of Europeana in your native language? (choose one)*

□ Yes	□ No

If not, why? (specify):

11. How should the user interface of Europeana switch to another language? (choose one)*

□ Automatically, e.g. based on the geographic location of the computer you are using	□ Manually

Information Access and Retrieval

12. What kind of search functionalities do you typically expect to find (on a digital library)? (choose one or more)*

□ Search by author, year, publisher	□ Search by subject headings	□ Full text search
□ Additional search types, e.g. "more like this", …		
Other (please specify):		

13. What kind of ranking of the results do you expect? (choose one or more)*

□ By fields: e.g., alphabetical order by author or publisher	□ By similarity among the query and the catalogue records	□ Grouped by different facets, e.g. media type of the results
Other (please specify):		

Multilingual Information Retrieval

14. Would you be interested in retrieving results in languages other than the one in which you formulated the query? (choose one)*

□ A lot	□ A little	□ Not at all

15. Would you be interested in browsing subject headings to look for more results in the same category? (choose one)*

□ Yes, but only in my own language	□ Yes, and in multiple languages	□ No

Multilingual Query Formulation and Expansion

16. How often do you specify the language of the documents you would like to retrieve? (choose one)*

□ Never	□ Seldom	□ Sometimes	□ Often	□ Always

17. Would you like to specify more than one language? (choose one)*

□ Never	□ Seldom	□ Sometimes	□ Often	□ Always

18. When querying for authors, places, ... would you be interested to have available their translation in different languages? (choose one)*

□ A lot	□ A little	□ Not at all

19. How much control would you like to have on the multilingual querying process? (choose one)*

□ It should be completely automatic and transparent to me	□ I would like to have the possibility to interact with it, e.g. by modifying the translation of the query

Multilingual Results Presentation

20. How should multilingual results be presented? (choose one or more)*

□ In relevance order, interleaving results in different languages	□ Grouping the results by language and, within each group, in relevance order	□ By highlighting results in different languages with different colours

21. Would you like to filter the results by language or language groups? (choose one)*

□ Yes	□ No	□ Not sure

22. When retrieving results in different languages, should they be translated to your native language? (choose one)*

□ Yes	□ No	□ Not sure

23. In the case the obtained results are translated, what quality do you expect from this translation? (choose one)

□ It must be a linguistically and syntactically correct translation	□ It could be a translation good enough to get an idea of the content of the results