

**ECP-2006-DILI-510003**

**TELplus**

## **Prototype integrating MACS initial data and new alignments into TEL framework**

<b>Deliverable number</b>	<i>D-3.4</i>
<b>Dissemination level</b>	<i>Public</i>
<b>Delivery date</b>	<i>4 January 2010</i>
<b>Status</b>	<i>V1.0</i>
<b>Author(s)</b>	<i>Antoine Isaac (VU), Sally Chambers (KB)</i>
<b>Contributors</b>	<i>Anna Gos, Willem Vermeer (KB), Genevieve Clavel, Patrice Landry (National Library of Switzerland)</i>



***eContentplus***

This project is funded under the *eContentplus* programme<sup>1</sup>,  
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

---

<sup>1</sup> OJ L 79, 24.3.2005, p. 1.

## TABLE OF CONTENTS

<b>1</b>	<b>INTRODUCTION .....</b>	<b>3</b>
<b>2</b>	<b>RATIONALE AND BACKGROUND .....</b>	<b>3</b>
<b>3</b>	<b>ACCESSING THE PROTOTYPE.....</b>	<b>4</b>
<b>4</b>	<b>SEMANTIC ALIGNMENTS EXPLOITED BY THE PROTOTYPE.....</b>	<b>4</b>
4.1	MACS MANUAL MAPPINGS .....	4
4.2	LEXICAL MAPPINGS, WITHOUT TRANSLATION (LNT).....	4
4.3	LEXICAL MAPPINGS, WITH TRANSLATION (LWT) .....	5
4.4	EXTENSIONAL MAPPINGS BASED ON DUALY INDEXED BOOKS (REALDUAL).....	5
4.5	EXTENSIONAL MAPPINGS BASED ON "INSTANCE MATCHING" (IBOMBIE) .....	5
<b>5</b>	<b>FUNCTIONALITIES OF THE PROTOTYPE – USER’S GUIDE .....</b>	<b>6</b>
5.1	MULTILINGUAL SUBJECT SELECTION .....	6
5.2	COLLECTION SEARCH .....	8
<b>6</b>	<b>COLLECTIONS IN WHICH THE PROTOTYPE ALLOWS TO SEARCH.....</b>	<b>9</b>

## 1 Introduction

This deliverable is the companion document to an implemented multilingual subject search prototype that is delivered as part of TELplus WP3.2 “Improving subject access”.

This deliverable presents the rationale for that prototype and gives the URL where it can be accessed. It also briefly describes the data that it exploits, how it exploits it, and features a small introduction for users.

## 2 Rationale and background

In the context of TELplus, WP3 “improve access” aims at proposing methods to enhance access to library documents, such as the ones made accessible in WP1. An essential constraint for WP3 is that collections are different languages, making it difficult to apply traditional search technology.

Task 2 in WP3, “improving subject access”, seeks to access these issues by focusing on the *semantic level of the subjects* of documents. This task has been conceived as a concrete case study: the methodological and technical investigations of that task are applied to the specific case of the collections from the English, French and German libraries.

In that case, which is also the one addressed in the MACS project<sup>1</sup>, the collections are described using subjects from the following languages: LCSH, RAMEAU and SWD. The key idea in that task is to use explicit connections (referred to as *mappings*) between concepts that are semantically connected even though they belong to description vocabularies in different natural languages.

Such a vision requires the following efforts, which are undertaken in various subtasks of WP3.2:

- enhance access to subjects in the first place, by means of standard formats and access mechanisms (T2.1);
- create mappings between the subjects (T2.2, which emphasizes on testing automatic mapping techniques);
- build a tool that can exploit the semantic mappings to allow users to search among collections in different languages (T2.3).

This deliverable is an outcome of T2.3. It presents a proof-of-concept prototype that use sets of mappings (alignments) that were manually created in MACS and automatically created in T2.2 to provide with a multilingual, subject-based access to selected collections in the TEL portal.

This prototype has been developed by the TEL Office with data input, advice and assessments from partners of the MACS project and the VU University Amsterdam.

---

<sup>1</sup> <http://macs.cenl.org>

It is important to mention that the prototype presented in this deliverable builds on a previous prototype, LVAT<sup>1</sup>, which also enabled a multilingual subject-based search. The main features of the new version are:

- ingestion and exploitation of automatic alignments next to the MACS alignments
- better integration with the TEL portal and the underlying software

### 3 Accessing the prototype

The prototype can be accessed at <http://search.tel.ulcc.ac.uk/portal/en/lvat.html>

At the time this report is being written, the prototype is password-protected. Please contact [aisaac@few.vu.nl](mailto:aisaac@few.vu.nl) or [Sally.Chambers@kb.nl](mailto:Sally.Chambers@kb.nl) to obtain the necessary username and password.

### 4 Semantic alignments exploited by the prototype

The prototype makes use of four sets of mappings (alignments) which are described in this section, and that relate concepts from the following vocabularies:

- Library of Congress Subject Headings (LCSH, in English)
- Répertoire d'autorité-matière encyclopédique et alphabétique unifié (RAMEAU, in French)
- Schlagwortnormdatei (SWD, in German)

For some automatic alignments, the quality as assessed from manual and automatic evaluations (cf. D3.5) was sometimes low. A filtering strategy was thus applied to select mappings with an expected level of precision of comprised between 50% and 60%.

Note that in the case of automatic mappings, a confidence measure is also attributed to all mappings. This allows ranking the mappings, putting in first positions those who are expected to have better quality.

#### 4.1 MACS manual mappings

This set of mappings manually created in the MACS project contains approximately 130,000 trilingual mappings, which involve approximately 100,000 RAMEAU subjects, 90,000 LCSH subjects and 15,000 SWD subjects.

#### 4.2 Lexical mappings, without translation (LNT)

This set of mappings is based on just comparing the labels of concepts to map. If they are equal, we consider we have a mapping. There are several comparison options, depending on

---

<sup>1</sup> <http://lvat.hoppie.nl>

the types of the labels compared and the pre-processing applied to it (lemmatization can be used). For more details on the process, the reader is referred to D3.5.

For this alignment no selection has been made. The alignment contains:

- 11,345 LCSH-RAMEAU mappings;
- 13,119 LCSH-SWD mappings;
- 11,215 RAMEAU-SWD mappings.

### **4.3 Lexical mappings, with translation (LWT)**

Those mappings were produced using lexical comparisons, similar to what was done for the previous alignment. But here, labels of subjects have been translated (using the Google Translate service) in the language of the vocabulary they should be mapped to, before the lexical comparison process. For more details on the process, the reader is referred to D3.5.

For this alignment no selection has been made. The alignment contains:

- 57,595 LCSH-RAMEAU mappings;
- 44,017 LCSH-SWD mappings;
- 42,869 RAMEAU-SWD mappings.

### **4.4 Extensional mappings based on dually indexed books (RealDual)**

This alignment is obtained as follows. The three collections have some books in common. So some books have been “dually” indexed, that is, with concepts from two vocabularies, and we can investigate, for two concepts in two different vocabularies, whether they have been used to index the same books. We can thus “measure the overlap” between two concepts. If the overlap is above zero, then the considered couple of concepts will be added to the list of RealDual mappings. For more details on the process, the reader is referred to D3.5.

For this alignment a selection has been made, resulting in the following amounts of mappings:

- 60,000 LCSH-RAMEAU mappings (corresponding approximately to an expected 60% precision level)
- 5,500 LCSH-SWD mappings (corresponding approximately to an expected 50% precision level);
- 7,250 RAMEAU-SWD mappings (corresponding approximately to an expected 50% precision level).

### **4.5 Extensional mappings based on "instance matching" (IBOMBIE)**

The “RealDual” technique has an important drawback: the number of shared books is small. IBOMBIE tries to improve it by creating a corpus of “virtually dually indexed books”. For each book 1 in collection 1, we compare its metadata to the metadata of books in collection 2. We take the most similar book, book2, which is of course indexed by concepts from vocabulary 2, and we add to the subjects of book 2 to the subject indexing of book 1. Book 1

is thus now dually indexed, and after we've done it for all books we can apply the overlap measure as done for ReadDual. For more details on the process, the reader is referred to D3.5.

For this alignment a selection has been made, resulting in the following amounts of mappings:

- 31,800 LCSH-RAMEAU mappings (corresponding approximately to an expected 60% precision level)
- 9,100 LCSH-SWD mappings (corresponding approximately to an expected 50% precision level);
- 5,500 RAMEAU-SWD mappings (corresponding approximately to an expected 50% precision level).

## 5 Functionalities of the prototype – user's guide

The prototype is organized into to logical steps for allowing a multilingual subject-based access to collections:

1. Multilingual subject selection: from an initial text search, the user can select subjects that have equivalents in other languages
2. Collection search: the selected (mapped) subjects are used in queries sent against the various collection the search engine is aware

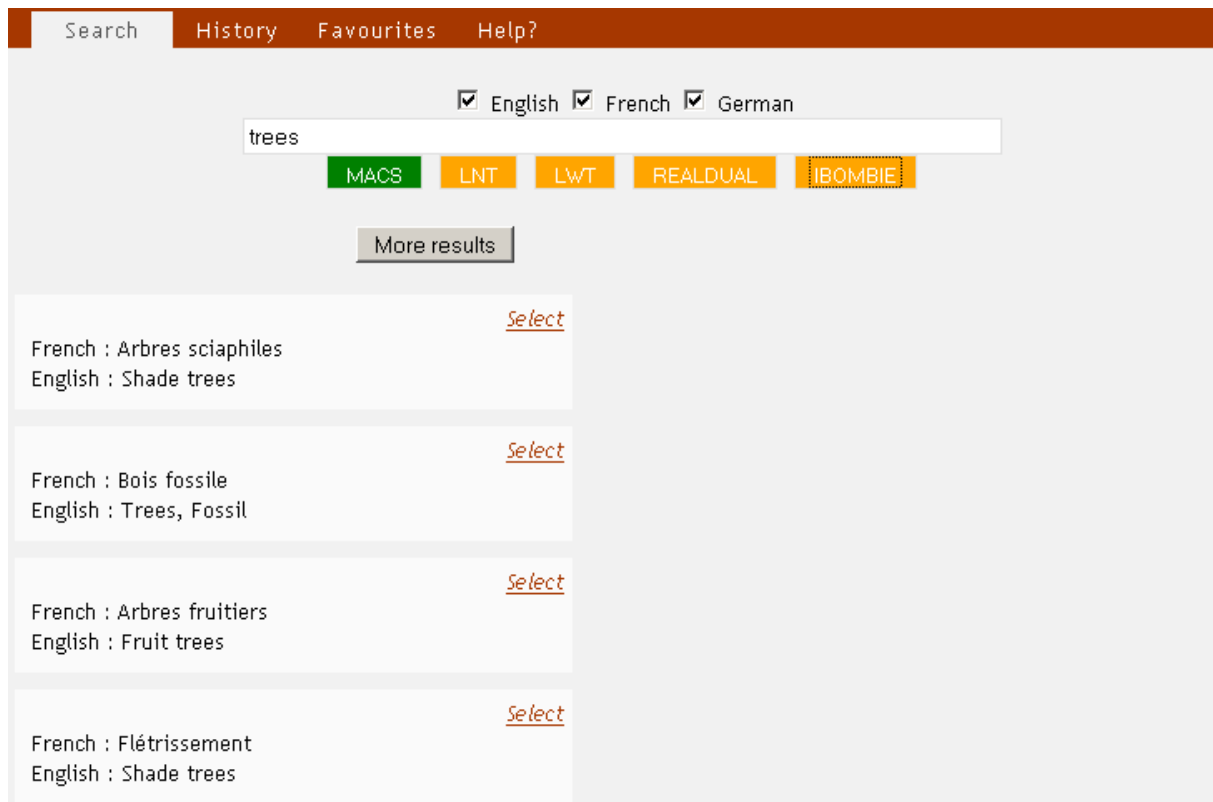
### 5.1 Multilingual subject selection

The process for this step is the following:

1. The user types a term she is interested in into the search box. For instance, she types "trees".



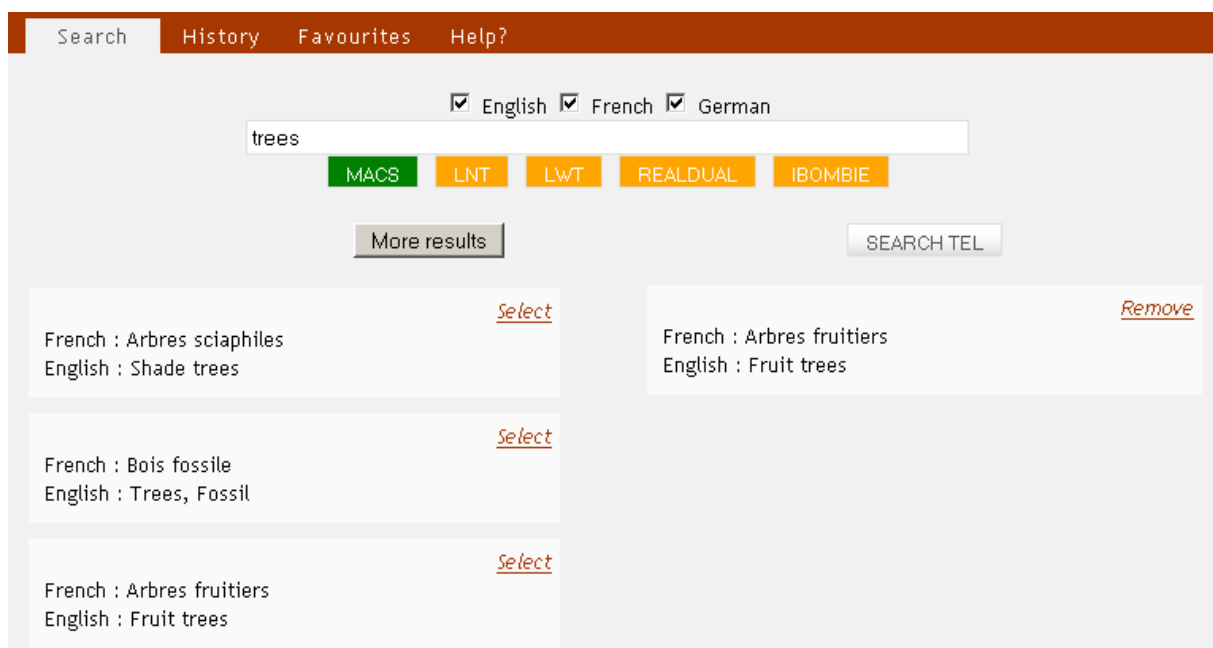
2. She clicks on the alignment she's interested to explore (MACS, LNT, LWT, REALDUAL or IBOMBIE). For instance, she clicks on "IBOMBIE".
3. The user gets a list of mappings on the left column (note: you can click on "more results" on top of the column) that matches her term of interest.



The screenshot shows the TEL framework search interface. At the top, there is a navigation bar with 'Search', 'History', 'Favourites', and 'Help?'. Below this, there are language selection checkboxes for 'English', 'French', and 'German', all of which are checked. A search input field contains the text 'trees'. Below the search field, there are five colored buttons: 'MACS' (green), 'LNT' (orange), 'LWT' (orange), 'REALDUAL' (orange), and 'IBOMBIE' (orange). A 'More results' button is located below these buttons. The search results are displayed in a list of four items, each with a 'Select' link:

- French : Arbres sciaphiles  
English : Shade trees
- French : Bois fossile  
English : Trees, Fossil
- French : Arbres fruitiers  
English : Fruit trees
- French : Flétrissement  
English : Shade trees

- The user can "select" the concept (and its mapping) that fits more her need. For instance, she selects "arbres fruitiers"/"fruit trees". User can select several of these concepts, resulting in a complex query. Note that users can even change the currently alignment to build queries using mappings from different alignments.



The screenshot shows the TEL framework search interface after a selection. The search input field still contains 'trees'. The 'SEARCH TEL' button is now visible. The search results are displayed in a list of four items, each with a 'Select' link. The first item, 'French : Arbres fruitiers / English : Fruit trees', is highlighted in a light blue box and has a 'Remove' link next to it. The other three items remain unchanged:

- French : Arbres sciaphiles  
English : Shade trees
- French : Bois fossile  
English : Trees, Fossil
- French : Arbres fruitiers  
English : Fruit trees

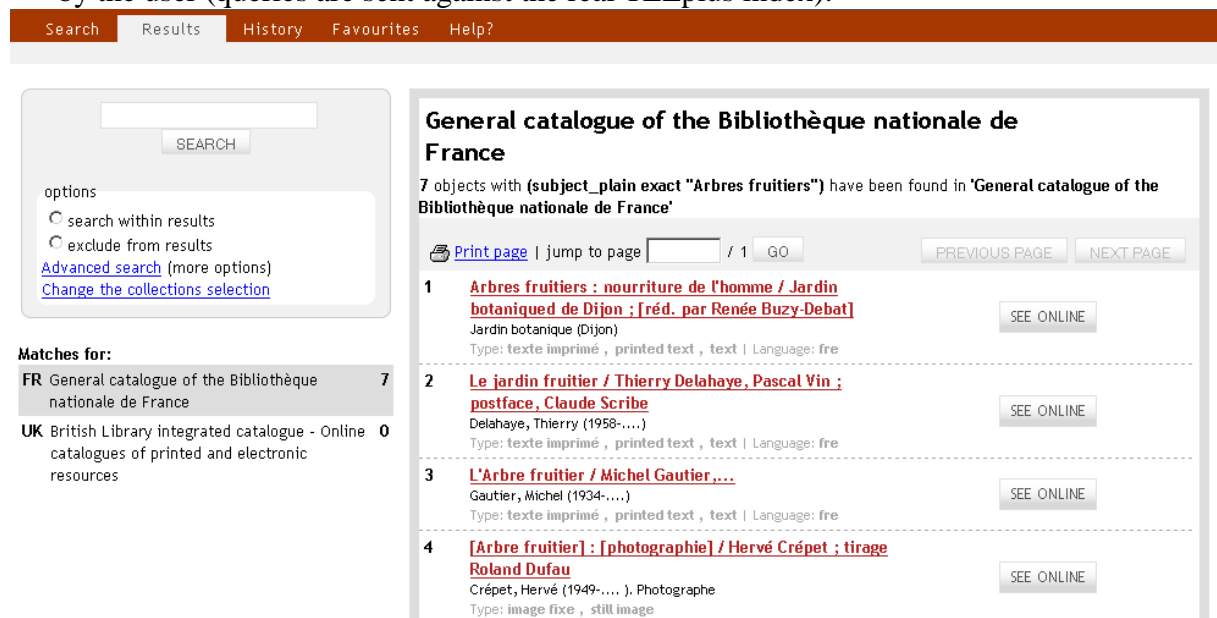
The user can repeat this process several times, for example to add subjects that correspond to other themes.

It is also possible to freely remove subject (mappings) from the query that the user is building, by clicking on the “remove” button on them.

## 5.2 Collection search

The process for this step is the following:

5. The user clicks on "search TEL"
6. This opens a new window that shows the result in the TEL portal for the query created by the user (queries are sent against the real TELplus index).



The screenshot shows the TELplus search results page. At the top, there is a navigation bar with tabs for Search, Results, History, Favourites, and Help?. Below this, there is a search box with a 'SEARCH' button and a list of options: 'search within results' (selected), 'exclude from results', 'Advanced search (more options)', and 'Change the collections selection'. The main content area is titled 'General catalogue of the Bibliothèque nationale de France' and displays 7 objects with the subject 'Arbres fruitiers'. The results are listed in a table with columns for item number, title, author, and a 'SEE ONLINE' button. The items are:

Item	Title	Author	Type	Language
1	<b>Arbres fruitiers : nourriture de l'homme / Jardin botanique de Dijon ; [réd. par Renée Buzy-Debat]</b>	Jardin botanique (Dijon)	texte imprimé, printed text, text	fre
2	<b>Le jardin fruitier / Thierry Delahaye, Pascal Vin ; postface, Claude Scribe</b>	Delahaye, Thierry (1958-....)	texte imprimé, printed text, text	fre
3	<b>L'Arbre fruitier / Michel Gautier,...</b>	Gautier, Michel (1934-....)	texte imprimé, printed text, text	fre
4	<b>[Arbre fruitier] : [photographie] / Hervé Crépet ; tirage Roland Dufau</b>	Crépet, Hervé (1949-....)	image fixe, still image	fre

7. Note that a complex query is treated as a disjunctive query ("OR-query", for instance “trees” or “physics”).

Search Results History Favourites Help?

SEARCH

options  
 search within results  
 exclude from results  
[Advanced search](#) (more options)  
[Change the collections selection](#)

**Matches for:**

AT	Online Catalogue of the Austrian National Library from 1992 onwards	36
CH	HELVETICAT : the catalogue of the Swiss National Library	Not responding
DE	Catalogue of the German National Library	
FR	General catalogue of the Bibliothèque nationale de France	49
UK	British Library integrated catalogue - Online catalogues of printed and electronic resources	0

### Online Catalogue of the Austrian National Library from 1992 onwards

Österreichische Nationalbibliothek

36 objects with (subject\_plain exact "Baum ") or (subject\_plain exact "Strassenbaum") have been found in 'Online Catalogue of the Austrian National Library from 1992 onwards'

[Print page](#) | jump to page  / 4 GO [PREVIOUS PAGE](#) [NEXT PAGE](#)

- [Zustandsanalyse von Jungbäumen im Stadtgebiet Wolkersdorf](#)  
Kraft, Mario  
Type: BOOK , BOOK , BOOK | Language: ger [SEE ONLINE](#)
- [Zustandsanalyse von Jungbäumen im Stadtgebiet Mödling](#)  
Schmid, Michaela  
Type: BOOK , BOOK , BOOK | Language: ger
- [Zustandsanalyse von Jungbäumen im Stadtgebiet Baden bei Wien](#)  
Pogats, Birgit  
Type: BOOK , BOOK , BOOK | Language: ger [SEE ONLINE](#)
- [Formalisierte Erfassung der Standortqualität von Straßenbäumen an Beispielen aus der Stadt Salzburg](#)  
Kutil, Bernhard C.  
Type: BOOK , BOOK , BOOK | Language: ger
- [Analysis of node isolation procedures and label-based parameters in tree structures](#)  
Kuba, Markus  
Type: BOOK , BOOK , BOOK | Language: eng [SEE ONLINE](#)

8. On the search result window users can browse in the collections from different national libraries, to see which results can be found in those libraries for your query as translated in the "language" of that library (left column). At the top of the right column, a line shows the current query that is being sent against the selected collection.

## 6 Collections in which the prototype allows to search

The following TEL-related collections will be made available through the prototype:

- German National Library (DNB)
- Austrian National Library (ONB)
- French National Library (BnF)
- British Library (BL)
- Swiss National Library (HELVETICAT)

At the time of writing this report, collections from BnF and ONB are fully available, and collections from BL and DNB are partly available.