



## **M 2.1 Report on Unicode**

*April 5, 2007*

**Report Authors:** Genevieve Clavel – Swiss National Library

**Report Contributors:** the national libraries of Greece, Iceland, Liechtenstein, Luxembourg, Norway, Spain, Sweden

## Table of Contents

Table of Contents .....	2
Summary .....	3
Summary .....	3
Context .....	3
Introduction .....	4
The UNICODE compliance questionnaire.....	5
Conclusions .....	8

## Summary

The national libraries of 6 of the 9 participant countries, whose collections are to be included in The European Library during the EDL project (Greece, Iceland, Liechtenstein, Luxembourg, Norway, Spain), responded to a questionnaire in order to identify issues brought up by portalling databases with different scripts, character set or individual character processing. The aim of this analysis is to help the European Library office and The European Library users make searches more accurate and avoid silence due to misunderstandings of how characters are used. The results of the survey were then compared to the results of a similar study conducted within TEL-ME-MOR project targeting the 10 then new member states. The compilers recommend that the questionnaire be filled by all The European Library partners with the goal to develop and maintain an overview of the Unicode compliance across The European Library.

## Context

Task 1 of Workpackage 2 of EDLProject examines the UNICODE capabilities of EDLProject partners. Task 2 is dedicated to establishing a clear picture of subject access mechanisms used by different partners. Task 3 will look at various cross language approaches to bibliographic subject data, based on Task 2 and findings and recommendations from TEL-ME-MOR, while Task 4 is to investigate possible solutions to multilingual access to full text. Task 5 will concern name authority control, whereas Task 6 should consider options of interoperability with other communities (museums) and test the possible use of new conceptual models of bibliographic universe.

## Introduction

Deliverable D3.1 from TEL-ME-MOR ‘Report on TEL UNICODE requirements’ gives a clear explanation of character sets and UNICODE which will not be repeated here: It summarized the main issues as

1. The presence of many different character sets.
2. The presence of special characters that are treated differently from country to country, language to language, or even database to database, changing the input accuracy needed, a fact that most users ignore, as well as user expectations in results presentation.
3. The impossibility or difficulty for users to input these characters.
4. Hurdles arising from the mixture of multiple scripts.

The results of the analysis in TEL-ME-MOR showed that Unicode was already widely used and the first issue above was not too serious. It confirmed the importance of the second and third issues and allowed to evaluate how strong the fourth is likely to be. The solutions envisaged to be as user-friendly as possible while dealing with these issues include

1. A move to Unicode, preferably in its UTF-8 encoding, everywhere it is possible but not yet done.
2. The resort to resources making a bridge between the different practices, such as shared authority files.
3. A web-based input help.
4. Considering an optional automatic transcription between scripts.

The report went on to make the following recommendations to The European Library Office:

1. To move the entire portal, or at least the navigation frame, to UTF-8, the best option for the web. On the search terms input and navigation frame, this will allow the input of all kinds of characters used by partners’ resources and correctly display labels and navigational texts. A software extended keyboard should be added to allow the input of characters significant for searching but absent from some hardware keyboard. On the results frame, moving to UTF-8 might improve performance, since UTF-8 is more frequently used in returned records and more compact than UTF-16. **This was done**

2. To address the questionnaire retroactively to current The European Library members, since these character set issues have been overlooked until now. For instance, a search on 'Böll', 'Boll' and 'Boell' on current default collections shows that the British Library integrated catalogue ignores the Umlaut ('Böll' and 'Boll' give the same result set, 'Boell' does not), whereas SNL's Helveticat takes it into account ('Böll' and 'Boell' give the same result set, 'Boll' does not). **This will be done in 2007; the current report covers the EDJ Project partners.**
3. To complement the collections descriptions by information about the use of special characters, where appropriate. **This is still pending**
4. To add to the search page 3 displayable soft keyboards, for Latin special characters, Greek and Cyrillic, with information about script prevalence, their characters' use and usual transliteration (prototype available). **This is in progress**
5. To study the feasibility of adding to the results page an automatic transcription option from Greek and Cyrillic to Latin, tailored to the interface language, respectively from Latin and Cyrillic to Greek on the interface in Greek. **After discussion it was considered that this was not feasible at this time.**

The European Library Office has now incorporated information about characters sets into its [Handbook](#) and links to the [TEL-ME-MOR UNICODE report](#)

## The UNICODE compliance questionnaire

A questionnaire of 9 questions, each completed by a short explanation about its aim, was sent to NMS partners to fill. This section provides the original questions and comments, the results and a posteriori comments where appropriate

1. For each resource you intend to make accessible through the TEL portal, please indicate in which character set(s) requests are accepted :

*The aim of this question is to know whether a conversion from UTF-8 is needed*

2. For each resource you intend to make accessible through the TEL portal, please indicate in which character set records are stored :

*The aim of this question is to know whether there would be a way to bypass a conversion loss, e.g. a database in Unicode accepting requests in a 8-bit character set because it is locally assumed that most PCs still do not yet have Unicode support*

3. For each resource you intend to make accessible through the TEL portal, please indicate in which character set(s) records are returned (if different from 2) :

*The aim of this question is to know whether a conversion to UTF-8 is needed*

Library resource	Accepted in requests	Used in storage	In returned records
Belgium	No reply		
Greece	ISO 8859-7/ cp 1253	ISO 8859-7/ cp 1253	ISO 8859-7/ cp 1253
Iceland	ISO-8859-1	ISO-8859-1	ISO-8859-1
Ireland	No erply		
Liechtenstein ALEPH web-OPAC ALEPH Z39.50	ISO 8859-1 (also UTF-8?) ISO 8859-1 (also UTF-8?)	UTF-8 UTF-8	UTF-8 ANSEL (but a server could be set up for UTF-8)
Luxembourg LUX01 LUX03 Luxemburgensia Online	Needs to be checked	UTF-8 UTF-8 ISO-Western	Needs to be checked
Norway	UTF-8	UTF-8	UTF-8
Spain	UNICODE and Latin I (ISO 8859-1)	Latin 1, moving to UNICODE	Latin 1, moving to UNICODE
Sweden	UTF-8	UTF-8	UTF-8

4. Indicate the majority language of your country and its minority languages when applicable

*The aim of this question is languages that are sopken in more than one country, in order a) to a posteriori confirm the sense of ‘portalling’ their libraries, b) possibly to help defining default profiles where resources of the same or similar languages would be offered together and c) to detect possible synergies, e.g. on indexing languages*

Library	Languages (and for minority languages: percentage of population speaking it)
Greece	Greek
Iceland	Icelandic
Liechtenstein	German (national langaue); Italian 3.5%, Serbian (incl. Bosnian and Croatian) 2.8%, Turkish 2.6%, Portuguese 1.7%, Spanish 1.3%
Luxembourg	Luxembourgish, French also German and English
Norway	Majority: Norwegian “bokmål”; Minority: Norwegian “Nynorsk”, Sami languages ( e.g. Inari Sami, Lule Sami, Northern Sami, Skolt Sami, Southern Sami )
Spain	Majority Spanish. Others: Catalan, Euskera and Galician
Sweden	Swedish

5. (for countries using Latin script) Do(es) the language(s) of your country and/or the resource(s) you intend to make accessible through the TEL portal contain special characters

that are not assimilated to a basic Latin letter (a-z), i.e. searched and sorted separately (e.g. ‘L’)? Please detail

Library resource	Special characters
Iceland	Diacritics for all the vowels like a, á, e, é etc. and we have ð, æ, ö, Þ All are used in sorting
Liechtenstein	Diacritical marks ignored in searching
Luxembourg	Luxembourg character set: (âââ ç ééêè ñ öö üùû ß ÄÅÂ Ç ÈÉÊÈ Ï Ò Ó ÙÚÛ) includes Luxembourgish, German and French
Norway	Norwegian bokmål and norwegian nynorsk: æ, ø, å Sami: á č đ ŋ š ț ž ı ö
Spain	Ñ
Sweden	Swedish characters: Å [00C5], å [00E5], Ä [00C4], ä [00E4], Ö [00D6], ö [00F6]

6. Does your country include a significant part of the population using a non-Latin script and personal computers with a non-Latin keyboard ? which script ?

7. Do(es) the resource(s) you intend to make accessible through the TEL portal contain a significant part written in that non-Latin script ?

Library	Non-Latin script	Easy Latin encoding	Frequent in resources
Greece	Greek		Yes
Iceland	No		
Liechtenstein	No		
Luxembourg	No, apart from a small number of East-European resources		
Norway	No		
Spain	No		
Sweden	No		

8. If you plan to make your collection(s) available via Z39.50, is your client UNICODE compatible? How does it treat characters with diacritics? If possible test using e.g. author Jančar, Dürrenmatt

Library	Z39.50 client
Greece	Our Z39.50 client didn't work well with the given test
Iceland	UNICODE compliant
Liechtenstein	I am unsure in this question. With our Aleph PC-GUI I can get an answer for 'Dürrenmatt' from Library a, but for Library b, I have to use 'Duerenmatt' for no silence answer. From our own Aleph-Test-Installation to our own Aleph-Production-Installation, I have to use 'Duerrenmatt'
Luxembourg	Needs to be checked

Norway	Not applicable
Spain	Yes it is UNICODE compatible
Sweden	Not applicable

## Conclusions

While the TEL-ME-MOR study focused on the 10 then new member states of the EU, who were in large part new to The European Library as well, the present study targeted the 9 partner libraries in EDL project. TEL-ME-MOR group consisted of predominantly Central and Eastern European countries, whereas EDLproject group includes Western European countries with rich histories of European integration, either in EU or EFTA. This however does not mean that UNICODE is more widely adopted or that searching will be easier: older systems in place mean that latin-1 is still in use though as Spain indicates moves to UNICODE are planned. System difficulties (Greece) mean that some questions remain unclear and will need to be clarified within the testing of access as countries join the European Library:

- Greece: test the Z39.50 client
- Iceland: searching with or without diacritics, sorting (are ‘common’ diacritics such as á, é really sorted separately)
- Liechtenstein, check searching and sorting with umlauts
- Luxembourg: check searching and sorting with umlauts
- Spain: check searching and sorting of Ñ

Norway and Sweden will provide data via OAI to the central index and thus are not concerned with the Z39.50 compatibility questions. However in each case they have special characters that are sorted differently and the European Library Office will need to test the sorting algorithms when incorporating data.

It is clear from the answers to the questionnaire that interoperability testing is required: all members are requested to carry out the [tests](#) as presented in the Handbook.

## Acknowledgments



The WP leaders wish to express their thanks to all the respondent libraries for taking the time to answer the questionnaire, providing further information as required and contributing to the preparation of this report.