L2/01-280

CEN TC304 N985
Subject/Title: Open Issues for EOR-2
Source: Marc Küster
Date: 16 July 2001
Note/Status: This document was presented 26 June 2001 at the TC304 plenary. A resolution was adopted on accepting the plan proposed (see document N986).

Source: Marc Wilhelm Küster, project editor
Action: For consideration in the TC304 plenary in Brussels, 2001-06-26

Open issues for EOR-2

The situation

A) European Ordering Rules, stage 1 (= ENV 13710:2000)

The European Ordering Rules, EOR-1 for short, define a set of rules how list should be ordered when they contain multilingual data from more than one country. Such lists exist all over the place: indices to pan-European legislation, phone books, job offerings, conferences, etc. The number of such lists will grow in the increasingly interconnected world of eEurope, and users expect consistent behaviour across such lists to speed up their access to relevant resources.
EOR-1 is based on the international standard 14651 and its equivalent version from the industry-driven Unicode Consortium. Both standards are currently being implemented on a large scale. Accompanying awareness campaigns for EOR-1 e. g. in the periodical Multi-Lingual have resulted in significant interest from IT companies to offer their customers this European ordering sequence. The European Commission can expect over the next few years broad support in mainstream software for EOR.

B) European Ordering Rules, stage 2 (ENV 13710:2002, EOR-2 for short)

EOR-1 is unfortunately not complete in two senses:

a) 14651 and its equivalent version from the Unicode Consortium are in the process of being updated to cover the full repertoire ISO/IEC 10646-1:2000 and its future amendments. The European Ordering Rules need to be aligned to parallel those updated versions. Only this will IT companies to use the specifications of ENV 13710 directly once those updates are available.
b) EOR-1 does not cover all European scripts, but only the most widely used characters from the Latin, Greek and Cyrillic scripts. EOR-2 will remedy this shortcoming by covering all letters from those scripts which are presently in Unicode and by adding the Armenian and Georgian scripts, thereby defining an ordering for all letters that are used in the European continent. Suitable contacts with Armenian standards and Georgian experts are in place, all the relevant research has been completed.

Steps towards EOR-2 – additional characters

Exact repertoire

EOR-2 is built on the *Multilingual European Subset No. 3* (cf. CWA 13873:2000) which is an open collection in ISO/IEC 10646. However, for the practical task of constructing the tables in the standard the repertoire must be fixed which can only be done through a “snapshot” of MES-3 against a well-defined reference edition.
There are three choices on which the snapshot for the exact repertoire can be reasonable based:

- The repertoire of IS 14651:2001,
- The repertoire of IS 10646-1:2000.
- MES-3B, the fixed collection specified in CWA 13873:2001.

Since EOR-2 is a profile of 14651 and references its tables, the first choice seems most natural, though already MES-2 contains letters not in the table of 14651[1] and a slight extension is inevitable.

Recommended action: Take the repertoire of IS 14651:2001 as the basis for the snapshot of MES-3.

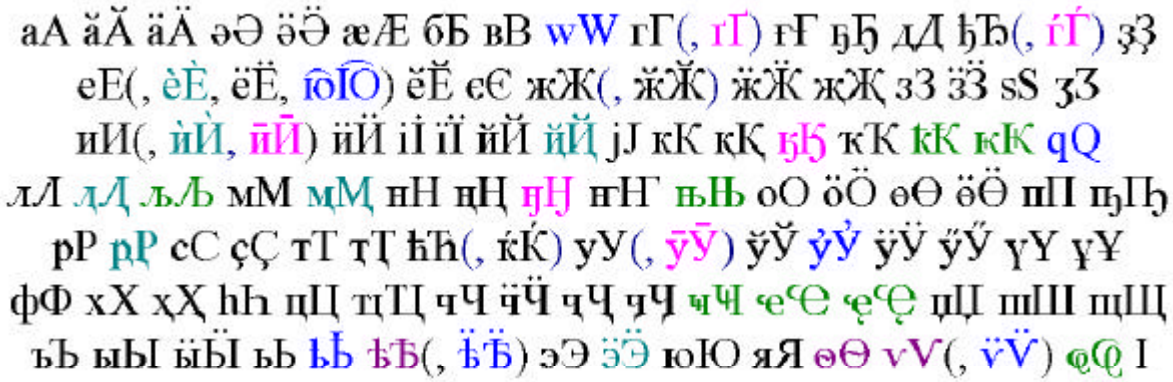
Armenian and Georgian

The addition of the Armenian and Georgian scripts are straightforward. They have a universally accepted ordering sequence that can be validated against experts from the Armenia and Georgia respectively. This sequence is virtually certain to be identical to the sequence specified in IS 14651.

Recommended action: Add the Armenian and Georgian scripts in the manner indicated

Cyrillic

The EOR-1 project team has established a close contact with the Russian standards body, GOST. GOST has specified a default order for the Cyrillic script that is:



(The letters in brackets are considered as second level variants of the letters before the bracket).
This order has been input as a European requirement into IS 14651 (cf. ISO/IEC JTC1/SC22/WG20/N681). It is recommended that EOR-2 also uses this table.

Recommended action: Continue to use the Cyrillic default order also for letters not yet covered in EOR-1.

Latin script

Currently, only the following letters are considered to be first level letters for the Latin script:
a A b B c C d D e E f F g G h H i I j J k K l L m M n N o O p P q Q r R s S t T u U v V w W x X y Y z Z þ Þ
All the other letters are considered to be either letters with diacritics or other second level variations. It is crucial for the homogeneity of the European Ordering Rules that this approach be kept. This is particularly virulent for all the letters of Latin Extended-B that got into MES-3 through the open collection approach. EOR-2 will align with the order of IS 12199

[1] In particular, these are U02EE (double apostrophe), the Greek letters small digamma, small stigma, small koppa and small sampi, the kai symbol, and the Cyrillic IE with Grave and I with grave.

"*Alphabetical ordering of multilingual terminological and lexicographical data represented in the Latin alphabet*".
Recommended action: Treat all Latin letters other than the current first level letters as second level variations and align with the order of IS 12199.

Other letters

The modifier letters modifier letter turned comma, modifier letter reversed comma, modifier letter apostrophe, and modifier letter double apostrophe are treated as special characters in ENV 13710 and have a weight only on level 4. It is suggested that the same strategy also be used for the other modifier letters in MES 3 but not in MES 2.
Recommended action: Treat all modifier letters as characters with weight on level 4 only.

Currency signs

Currency signs are given a weight on level 1 in IS 14651:2001, whereas they are treated as special characters in ENV 13710 and have a weight only on level 4. This is in line with dominant European practice. It is recommended that the same strategy also be used for the additional currency signs in EOR-2:

```
20A1 COLON SIGN
20A2 CRUZEIRO SIGN
20A5 MILL SIGN
20A6 NAIRA SIGN
20A8 RUPEE SIGN
20A9 WON SIGN
20AA NEW SHEQEL SIGN
20AB DONG SIGN
20AD KIP SIGN
20AE TUGRIK SIGN
```

Recommended action: Treat all currency as special characters with weight on level 4 only

Varia

The base repertoire of EOR-2, the *Multilingual Subset No. 3,* contains a variety of additional characters that need attention. In particular, these are characters that are in Unicode terms canonically equivalent to some letters. They should carry the same weight on levels 1 to 3 as their corresponding base letters (that is currently the norm also in 14651).
On the other hands there are a few symbols such as the *Greek pi symbol* and the *Greek ano teleia* that should both be treated as symbols with a level 4 weight only (in 14651 their treatment is inconsistent).
Deviations from the order of 14651 should be only made for good reasons and on a case-by-case basis to ensure maximum homogenity and ease of access. It should be the exception rather than the rule.
Recommended action: Give license to the project team to handle these various characters on a case-by-case basis.

Steps towards EOR-2 – technical issues

The task to align EOR-2 with IS 14651:2001 is a well-defined if complex task. During the ballot of ENV 13710 comments - notably those from Sweden - indicated that the revision should use the "minimal delta" version (currently Annex F) as the normative version. That was not possible in 1999 as the table in the then FDIS 14651, on which such a minimal delta would be based, was still in flux. With the acceptance in 2000 of 14651 as an international

standard that situation has changed and it is now feasible to use the minimal delta approach normatively. Technically, both approaches are equivalent and no change in the ordering sequence will result for letters in MES 2.

Recommended action: Make the successor to the current Annex F the normative table and drop the currently normatively default table (section 6).