

ISO
INTERNATIONAL ORGANIZATION FOR STANDARDIZATION
ORGANISATION INTERNATIONALE DE NORMALISATION
МЕЖДУНАРОДНАЯ ОРГАНИЗАЦИЯ ПО СТАНДАРДИЗАЦИИ

ISO/IEC JTC1/SC2/WG2
Multi-Octet Coded Character Set

ISO/IEC JTC1/SC2/WG2 N

1443

Date: 1996-03-04

Title: Allocating Ogham and Runic scripts to the BMP: a strategy for making the BMP maximally useful
Source: Michael Everson (IE) and Olle Järnefors (SE)
Status: Expert Contribution
Action: For consideration by WG2

The question as to whether the Ogham and Runic scripts should be included in the BMP of ISO 10646 must be resolved quickly, and the resolution should be YES. There are no substantive arguments as to why 108 characters (27 and 31), presented in mature script proposals (N1103R; N1330, N1417) should not be allocated to the BMP. Indeed, all evidence suggests that earlier research on the population of the BMP was correct, and that - while everything cannot fit there - many things can.

N947, based on N844 from the Unicode Consortium, proposed tentative allocations to a large number of contemporary and archaic scripts. N1370, Michael Everson's "Roadmap of the BMP" attempted to display the same basic material (a somewhat reduced set because of the doubtful nature of some archaic scripts). Except for CJK, whose ultimate needs cannot be met by the BMP even if there were nothing else there but Table 1, it is abundantly clear that there is plenty of space in the BMP for a great many scripts and characters.

The following facts may be observed:

After the Hangul expansion, the alphabetical zone has 14667 empty code positions. Of these at least 12544, being 48 empty rows and 2 empty half-rows, are not claimed for already coded scripts.

One new script has code allocations already accepted by WG2:

- 1) row 0F: Tibetan (174 characters)

Also accepted by WG2, without code allocations, are the following three scripts. (Here and in the rest of this overview, possible code allocations, mostly taken from N1370, have been indicated. "p" here means partial.)

- 2) row 15p: Cherokee (85 characters)
- 3) row 15p: Ogham (27 characters)
- 4) row 15p: Runic (31 characters)

WG2 has asked for more input on these proposals:

- 5) row 28?: Byzantine musical notation (220 characters)
- 6) row 34, 35, 36?: Mongolian (572 characters?)

Other scripts for which proposals are being prepared:

- 7) row 12, 13: Ethiopic (346 characters)
- 8) row 16, 17, 18p: "Canadian" syllabics (623 characters)
- 9) row 14p: Khmer (95 characters?)
- 10) row 14p: Burmese (85 characters?)
- 11) row 0Dp: Sinhalese (94 characters?)

Other scripts:

- 12) row 07p: Maldivian
- 13) row 07p: Maghreb
- 14) row 07p: Numidian
- 15) row 07p: Syriac
- 16) row 07p: Aramaic
- 17) row 07p: Samaritan
- 18) row 08p: Phoenician
- 19) row 08p: South Arabian
- 20) row 08p: Balti
- 21) row 08p: Meroitic
- 22) row 08p: Etruscan
- 23) row 08p: Ugaritic
- 24) row 08p: Pahlavi
- 25) row 10p: Cyrillic extensions
- 26) row 18p: Gothic
- 27) row 18p: Glagolitic
- 28) row 19p: Javanese
- 29) row 19p: Balinese
- 30) row 19p: Batak
- 31) row 19p: Buginese
- 32) row 1Ap: Kirat/Limbu
- 33) row 1Ap: Manipuri
- 34) row 1Ap: Rong/Lepcha
- 35) row 1Bp: Linear B
- 36) row 1Bp: Old Persian Cuneiform
- 37) row 1Bp: Tifinagh
- 38) row 1Cp: Tai Lu
- 39) row 1Cp: Tai Mau
- 40) row 1Cp: Tagalog
- 41) row 1Cp: Mangyan
- 42) row 2C, 2D, 2E, 2F: Yi (876 characters)

Even when solid proposals for all these scripts have been prepared and they have been included on the BMP, all of the following rows of the A zone will still be free to be used for symbol collections and further historic scripts:

1D, 28, 29, 2A, 2B, 37, 38
 39, 3A, 3B, 3C, 3D, 3E, 3F
 40, 41, 42, 43, 44, 45, 46
 47, 48, 49, 4A, 4B, 4C, 4D

These 28 rows have a capacity of 7168 characters.

WG2 really has no reason to unnecessarily delay the inclusion of any script into 10646 that happened to be excluded in the first edition, if the proposal is well designed, stable, and has been reviewed by competent representatives of the communities who use the script. On the contrary, WG2 should stimulate the development of such proposals by handling them swiftly.

The remaining coding space on the BMP should as soon as possible be allocated to the alphabetic extensions that are almost finalized, whether they are "living" scripts or "historic" scripts. For these scripts allocation on the BMP is

important, because it means that they can be easily and immediately implemented and used with all of the more or less experimental 10646 implementations that exist today, many of which can't handle UCS-4 or UTF-16.

The Ideographic Horizontal Extension could be allowed to use a part of the A-zone, rows 3C-4D, and a part of the O-zone, D8-DF, that is left empty after the Hangul reorganization. Further ideographic extensions should be referred to another plane. This means that the 25203 empty positions of the first edition of 10646 will be apportioned like this:

4516 (18 %) Hangul
6608 (26 %) ideographic
14079 (56 %) all other scripts + symbols.

This distribution cannot be said to be biased against the East-Asian countries.

Given this allocation of the BMP, there will be room for both contemporary and historic alphabetic scripts. To fill all the 14000 positions with characters according to well-prepared proposals will take such a long time that there are excellent chances that extension techniques to break out of the BMP (UCS-4 or UTF-16) are widely implemented at the time when the first previously unsupported script or symbol set will have to be allocated outside the BMP.

The immediate allocation of space on the BMP to proposals that are mature enough that WG2 has provisionally accepted the repertoire, order, and naming of the characters, will relieve WG2 of not very meaningful discussions about the relative importance of different scripts and create room for the more useful scrutiny of other proposals, that still are in earlier stages of development.

Once again, how precious is the BMP? In the long term, it is not precious. ISO 10646 will last for centuries. Why should it not? Subsets of ISO 10646 which are unable to make use of UCS-4 or UTF-16 will be abandoned to the fuller implementations which will be available. In the short term, the BMP is indeed precious - but not so precious that it should not be populated with alacrity when mature proposals are approved. The current trend to do nothing or to keep scripts or characters without a sufficiently large or important user community (no definitions have ever been given) in limbo serves no one. Neither have definitions of adequate size or importance of user communities ever been given - apart from the economic importance which won - rightly enough - more than five thousand characters for Hangul.

How big are the interested user communities? A quantitative investigation also proves interesting. The following hits were yielded after a search of the Alta Vista search engine for the World Wide Web:

Search: fonts
Hits: 315000
Search: cyrillic OR russian AND fonts
Hits: 10000
Search: chinese AND fonts
Hits: 4000
Search: Hebrew AND fonts
Hits: 2000
Search: arabic AND fonts
Hits: 1000
Search: runic OR runes AND fonts
Hits: 1000
Search: sinhalese OR sinhala AND fonts
Hits: 300
Search: tamil AND fonts

Hits: 600
Search: runic AND fonts
Hits: 500
Search: ogham OR ogam AND fonts
Hits: 400
Search: armenian AND fonts
Hits: 300
Search: khutsuri OR mkhedruli OR georgian AND fonts
Hits: 200
Search: cherokee AND fonts
Hits: 186
Search: thaana OR divehi OR maldivian AND fonts
Hits: 168
Search: cuneiform AND fonts
Hits: 133
Search: elvish AND fonts
Hits: 80
Search: gurmukhi AND fonts
Hits: 41
Search: etruscan AND fonts
Hits: 31
Search: deseret AND fonts
Hits: 19
Search: estrangelo AND fonts
Hits: 18

It can easily be recognized that this survey could be qualified by finding out exactly how many fonts are in fact available for each script - but the figures have a generally indicative value. In the case of Ogham for instance, there are a number of independent foundries producing Ogham fonts. Michael Everson, as convener of NSAI/AGITS/WG6, wrote to each of them and told them that Ireland was making a standard for encoding Ogham so that all implementations would be harmonized. Each and every one of the Ogham font providers responded that they will be delighted to agree to support the standard (good news for standards anyway). The character set used, whether in single or multiple octets, is identical to that proposed for inclusion in ISO 10646. The point is, that BMP support for scripts is important. If there were no interest in seeing Ogham or Runes in early implementations of ISO 10646, no one would have gone to the trouble to research and propose them. Most early implementations will undoubtedly be limited to the BMP.

If WG2 cannot, with - very conservatively counted - 15000 free character positions in the current standard, allocate 108 positions (0.7 %) in the BMP to Ogham and Runes, two mature proposals made by qualified experts of both script and coding and backed by member bodies who identify an immediate computing requirement for these scripts, there is something very, very wrong indeed.

The block U1580-159F is proposed for Ogham and the block U15A0-U15FF is proposed for the Runes. WG2 is requested to accept these allocations, or to give full, comprehensive, and detailed reasons as to why not.