ISO/IEC JTC 1/SC 2/WG 2
Multiple-Octet Codes
and

ISO/IEC JTC 1/SC 2/WG 3
7-bit and 8-bit codes and their extension

ISO/IEC JTC 1/SC2/WG2 *N 1706*

Date: **March 8, 1998**

Title: Re-visiting definitions of 'collection' in COR.2 of 10646

Source: USA (ANSI) and Canada (SCC)

Status: US and Canada - position

Action: For the consideration of WG 2 and For information to WG 3

References: Technical Corrigendum No. 2 to 10646 (ISO/IEC JTC 1/SC2/WG2 N 1664)

Distribution: ISO/IEC JTC 1/SC2/WG2 members

Introduction:

ISO

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION

ORGANISATION INTERNATIONALE DE NORMALISATION

ISO/IEC JTC1/SC2/WG2

Universal Multiple-Octet Coded Character Set (UCS)

ISO/IEC JTC1/SC2/WG2 *N 1807*

Date: 1998-07-06

Title	**Israeli Response to the Tetragrammaton Proposal, N 1740**
Source	**SII Committee 1109**
Date	**July 7th, 1998**
Compiled by	**Jonathan Rosenne**
Reference	**ISO/IEC JTC1/SC2/WG2 N1740 dated 1998-05-09**

The Israeli national body opposes the referenced proposal for the reasons explained below.

The proposed entity is not a character. Hebrew is an alphabetic rather than an ideographic script.

The basic premise of the proposal is incorrect. The proposal assumes that there are several textual representations or spelling variations of a single word. However, the several spellings of the name of god are not at all synonymous or equivalent. YY is not the same as YHVH and the two are not equivalent. The D' and H' are substitutes or euphemisms. Moreover, many people consider all the variations, including the various pointings, to be meaningfully different.

The various spellings are not even pronounced the same way.

Part of the justification for the proposal involves "plain text search". In Hebrew, plain text search is not a serious option anyhow, due to the lack of standard orthography, the extensive use of prefixes and suffixes, internal declinations, and partial pointing. Prefixes are used for functions that are considered separate words in most languages, such as and, to, as, from and the definite article, and even in the most superficial plain text search one would prefer to ignore these prefixes. Searching in Hebrew is an interesting and complex issue, but definitely not a character coding matter. The proposal does not provide meaningful relief to the search problem and even makes it worse because it mixes up two distinct words and their substitutes.

On a more practical level, were the proposal to be accepted, how would it be pointed and accented? How would one indicate to which of the four (or two) letters does each point and accent belong?

And who would use it? In biblical texts it is customary not to change the spelling (the example of Psalm 117 in part F is not the customary or Masoretic text used in Jewish Bibles). In prayer books too they would not change, because it is considered that the variations are meaningful. And in other texts the use is rare - most commonly the H' or D' substitutes are used.

To C.5b: This clause claims that the proposed entity is a unique sign. This claim is adequately contradicted by several examples in part E of the proposal.

To C.6a: If the proposal were to be accepted, the right place for it is in the FBxx block.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC 1/SC 2/WG 2

Universal Multiple-Octet Coded Character Set
(UCS)

ISO/IEC JTC 1/SC 2/WG 2 *N2002*
1999-03-08

Title:	**Principles and Procedures for Allocation of New Characters and Scripts and handling of Defect Reports on Character Names (Replaces N 1502)**
Source:	**Ad hoc group on Principles and Procedures (Edited by: V.S. Umamaheswaran)**
References:	**See refernces section in the document**
Action:	**To be considered by SC 2/WG 2 and all potential submitters of proposals for new characters the repertoire of ISO/IEC 10646, and for new collection identifiers**
Distribution:	**ISO/IEC JTC 1/SC 2/WG 2, ISO/IEC JTC 1/SC 2 and Liaison Organizations**

(Note: This document does not include replacement for the sections on Roadmaps – Annex D (document N1876). It will be completed based on discussion on the rodmap documents prepared by Mr. Michael Everson, at meeting M36. It incorporates all the other updates that have been approved by WG2 up to meeting M35 – Uma.)

1. Introduction

This document is a standing document of JTC 1/SC2 WG2. It consists of a set of Principles and Procedures on a number of items relevant to the preparation, submission and handling of proposals for additional characters or for identifying new collections in the standard (ISO/IEC 10646 and Unicode), for consideration by ISO/IEC JTC 1/SC 2/WG 2. Submitters should check the standard documents (including all its amendments and corrigenda) first before preparing new proposals. Submitters are also encouraged to contact the convener of SC 2/WG 2 to check and compare any similar proposals that may already have been considered by WG 2.

2. Allocation of New Characters and Scripts

Annex D of this document details a roadmap for allocation of characters in the Basic Multi-Lingual plane (BMP) and the supplementary planes (General purpose - GPSP, and Ideographic - ISP). The following sections describe the principles and procedures to be used for assessing whether a proposed script or character(s) could be a candidate for inclusion in the standard, and whether it should be encoded in the BMP or in the supplementary planes.

2.1 Goals for Encoding New Characters into the BMP

A. *The Basic Multilingual Plane should contain all contemporary characters in common use:*

Generally, the Basic Multilingual Plane (BMP) should be devoted to high-utility characters that are widely implemented in some form of communication system. These include, for example, characters from hard copy typographic systems that are awaiting computerization, and characters recognizable and useful to a large community of customers. The "utility" of a character in a computer or communications standard can be measured (at least in theory) by such factors as: number of publications (for example, newspapers or books) using the character, the size of the community who can recognize the character, etc. Characters of more limited use should be considered for encoding in supplementary planes, for example, obscure archaic characters.

ISO

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION

ORGANISATION INTERNATIONALE DE NORMALISATION

ISO/IEC JTC1/SC2/WG2

Universal Multiple-Octet Coded Character Set (UCS)

ISO/IEC JTC1/SC2/WG2 *N 1806*

Date: 1998-07-06

TITLE: DEFECT REPORT on ISO/IEC 10646-1 AMD.5 Hangul syllables, with Editor's response
SOURCE: Dr. Kyongsok Kim (Korea) & Bruce Paterson (project editor)
STATUS: Expert contribution
ACTION: For approval by JTC1/SC2/WG2
DISTRIBUTION: JTC1/SC2/WG2

Defect Report concerning:

ISO/IEC 10646-1 Information technology - Universal Multiple-Octet Coded Character Set (UCS) - Part 1: Architecture and Basic Multilingual Plane, AMENDMENT 5: Hangul syllables

Qualifier: Editorial error.

Reference in document: Page 7 to 181 (odd numbers), various entries in character name tables.

Nature of defects:

1. Defect type: AA -> A, (aa -> a).

On pages 107-117 (odd numbers), the character name entries for characters in the range hex C544 to C78F are incorrect in the following ways:
a.) Each character name in the range hex C544 to C78F contains an extra CAPITAL LETTER A at the beginning of the third component of the name. (This position in the name corresponds to the initial consonant of the syllable, and should be null for the characters in the stated range.)

b.) The transliterations in parentheses for this range of character names contain an extra SMALL LETTER A at the beginning of the annotation.

Examples:
C544 HANGUL SYLLABLE AA (aa) should read
C544 HANGUL SYLLABLE A (a).
C545 HANGUL SYLLABLE AAG (aak) should read
C545 HANGUL SYLLABLE AG (ak).
The number of character names affected is 21 * 28 = 588

NCITS-L2/98-042R
ISO/IEC JTC 1/SC 2/WG 2 N 1724
Date: 1998-03-05

ISO/IEC JTC 1/SC 2/WG 2
Universal Multiple-Octet Coded Character Set (UCS)

Title: **Formal Criteria on Disunification**
Source: **The Unicode Consortium, US National Body and Asmus Freytag**
Action: **For consideration and acceptance by SC2/WG2**
Distribution: **NCITS-L2, UTC, ISO/IEC JTC1/SC2/WG2**

Text version 003

(Retyped from Hard Copy and an older version of this document – Uma).

This document proposes a formal criteria for the evaluation of certain kinds of character encoding proposals. It is not intended for use with Han characters.

Background

There have been repeated proposals to disunify existing characters. These proposals cannot be fully evaluated without a more rigorous framework concerning the disunifcation / unification of characters. Without such formal criteria, all decisions are 'ad-hoc' and different proposals may get different level of review. Bot the Unicode Technical Committee and ISO/IEC JTC 1/SC 2/WG 2 need to spend some time in evaluating and possibly formalising the criteria that we use to decide these cases. This is similar to the formalization we have done for script prioritisation, but uses different criteria.

NOTE: The unification criteria used for the Han script are very thorough and quite sufficient. This document attempts to establish formal criteria for use in other scripts. There is no attempt to change the procedures used in Han unification.

What is disunification?

Disunification is the introduction of a new character which can also be encoded by an existing character. A strong case of disunification occurs where there is prevalent practice of using the existing character. A weak case of disunification occurs where there is little or no use of the existing character for the purpose for which the new character is intended.

> *Example*: Adding a period in a new script is a weak disunification if we assume that nobody has an existing implementation of that script using the regular period. Adding a clone of a Latin letter for use with Cyrillic script is a strong disunification as mixed Latin/Cyrillic character sets exist and have almost certainly been used for encoding the languages that the new characters are intended for.

Cost and Benefits

Proposals always claim that disunification brings benefits. Formal criteria attempt to critically evaluate those benefits, but also compare them to the costs. Any disunification, especially strong disunification, introduces several types of cost to *all* complete implementations of the Standard.

- First, any complete implementation will have to add and support both an additional entry in the properties as well as an additional glyph, or glyph mapping for the disunified character.
- Second, whenever the character in question has no appearance distinction, there is the cost of accidental confusion and mis-identification. All implementations will need sophisticated handling of

ISO

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION

ORGANISATION INTERNATIONALE DE NORMALISATION

ISO/IEC JTC1/SC2/WG2

Universal Multiple-Octet Coded Character Set (UCS)

ISO/IEC JTC 1/SC 2/WG 2 *N 1808*
1998-04-30

Title:	**Reply to "Proposal WG2 N1734" Raised at the Seattle Meeting Regarding Proposal WG2 N1711**
Source:	**China**
Action:	**Review and Feedback**
Distribution:	**ISO/IEC JTC1/SC2/WG2**

At the WG2 meeting in Seattle in March 1998, Mr. Ken Whistler made some comments and suggestions (N1734) concerning our Mongolian Encoding Proposal N1711. Later we had a special meeting in Hohhot, at which Mr. Whistler's proposal was discussed in detail. We would have submitted this reply of ours to WG2 after we reached common understanding with Standardization Department of Mongolia, if bad communications had not kept us from receiving any feedback from them though we had informed them of our views on April 14. Our reply is as follows:

1. MONGOLIAN SPACE.

Mr. Whistler suggested to use NO-BREAK SPACE instead of MONGOLIAN SPACE and requested us to further justify why both MONGOLIAN SPACE and NO-BREAK SPACE are adopted. In the Mongolian Encoding System, there is need for a unique space called MONGOLIAN SPACE which differs both in form and function from common SPACE (U+0020) and NO-BREAK SPACE. (U+00A0). Such a space has the following distinctive features:

(1) In form, it represents a gap. On the screen there should be a visual representation of a width different from that of SPACE. In print, there should be a regular gap of one third of a full character which differs from that of SPACE.

(2) This space also has the function of a VARIANT SELECTOR to determine the changed forms of the letters preceding and following it. That is, to determine that the word-final character of the given letter preceding it should be used. As for the form of the character that follows it, it involves a lot of special cases and has to be judged according to what suffix is concerned (For detail see Appendix III, 1, in N1711).

(3) It is used to separate a suffix from the letter's word stem, implying that the gap here is not the bound between character strings of the word.

(4) MONGOLIAN SPACE cannot be used to split a word or a line in two.

(5) MONGOLIAN SPACE appears at a very high frequency. Statistics shows that it appears 28117 times, or 28.12%, in a text of 100,000 words.

As for NO-BREAK SPACE, it remains to be used in the encoding of Mongolian word in its original function. Thus, NO-BREAK SPACE indicates how a word is formed, i.e., how several morphemes of a word are separated by it. For example, the Mongolian word ARADCILAL (Democracy) consists of four morphemes ARA-D-CILA-L, which is written as ARA(NBS)D(NBS)CILA(NBS)L in the word formation column in a computer's dictionary or in the language data. The form and function of NO-BREAK SPACE used in such cases differ from those of MONGOLIAN SPACE:

(1) In appearance, NO-BREAK SPACE does not indicate a gap, so it is NO-BREAK SPACE in the full sense of the term.

ISO

INTERNATIONAL ORGANIZATION FOR SANDARDIZATION

ORGANISATION INTERNATIONLE DE NORMALISATION

ISO/IEC JTC1/SC2/WG2

Universal Multiple - Octet Coded Character Set

(UCS)

ISO/IEC JTC1/SC2/WG2 N 1818?

Date: 1998-07-26

Title: ***Revised Proposal for Yi Characters and Yi Radicals***

Source: China

Action:

Distribution: WG2 members

Introduction:

At WG2 meeting 33, China's proposals ***N1608---Yi Characters*** and ***N1611---Yi Radicals*** were accepted by WG2 according to its RESOLUTION M33.18 (Yi Script and Yi radicals), " WG 2 accepts 1165 characters, their shapes and names in document N 1608 for the Yi script, and their assignment to code positions in the range A000 to A48C in the BMP, and 57 characters (Note: 18 more than in resolution M32.8), their shapes and names in document N 1611 for the Yi radicals, and their assignment to code positions in the range A490 to A4C8 in the BMP.", they were also accepted by SC2 according to its resolution M07.11 of the 7th SC2 meeting.

Since some foreign experts and Chinese Yi experts noticed that there are some errors in the proposals mentioned above, China votes "NO" on ***PDAM 14---Yi syllables and Yi radicals*** with this revised proposal. In this paper, Chinese Yi experts corrected some characters' names, removed duplicated character and changed their order in accordance with the habitual usage of Yi script.

The total number of Yi characters is 1165 and of Yi radicals is 57 as before. For details of coding and naming convention, please refer to the attachment.

The table below gives some of the correction.

Previous position and name	Current position and name	Change

ISO/IEC JTC 1/SC 2/WG 2/ *N 1876*
Date: 1998-09-20

Title:	**Proposed Replacement Text For Annex D in N1502R**
Source:	**Ad hoc on principles and procedures (V.S. Umamaheswaran)**
Status:	**For review**
Action:	**For consideration by WG 2 meeting 34**

This document contains only updated information for BMP. Plane 1 information is unchanged except for formatting. The principle of starting at half-row boundary has been included (per resolution M33.11). WG 2 position regarding allocation of '00' position in a block has been included (per resolution M33.12). Updated pictorial view of the BMP reflecting allocated code positions and guidelines on space needed and potential areas for candidate scripts that have not yet been processed by WG 2 has been included -- per action item AI-33.6.

Annex D
BMP and Supplementary Planes Allocation Roadmap

Overview

The intention of this annex D is to lay out a logical roadmap for further allocations of scripts in ISO/IEC 10646 (also in the Unicode Standard), within and beyond the BMP. This roadmap is a snapshot of known scripts and characters as of 1998-08-29. It is intended as a general guideline and does not attempt to make detailed allocations of characters. The roadmap consists of two parts.

- The first part addresses the BMP (Plane 0) in ISO/IEC 10646 (and the Unicode Standard). It locates all script and individual character additions accepted in amendments up to PDAM.27 (as of 1998-08-30) in WG 2 (and Unicode Technical Committee), plus all script additions currently foreseen to be reasonable candidates for future encoding on the BMP.
- The second part is for Plane 1 and Plane 2 (both accessible in ISO/IEC 10646 with Amendment No. 1, and in the Unicode Standard version 2.0 via UTF-16 and will be dedicated to all other future allocations, as follows:
    - Plane 1: General Scripts and Symbols Supplementary Plane (GSP)
    - Plane 2: Unified Ideographs Supplementary Plane (UISP)

For Plane 1, a proposed list of all additional known scripts is provided here, with rough estimates of the sizes of the scripts. In contrast to the roadmap for the BMP, no particular locations for scripts are proposed as yet. By current estimates (see details below), all remaining General scripts and symbol sets should fit within this one plane.

Plane 2 is envisioned as containing future Unified Ideographic character additions. The largest current Unified Ideographic character collections should fit within Plane 0 and Plane 2, as long as duplicate character encoding is avoided. No substructure for Plane 2 is proposed here.

The roadmap indicates that these three planes should suffice for all future encoding of characters having worldwide utility. However, note that 14 supplementary planes are available altogether for encoding (with an additional 2 planes reserved for private use). The planes described in this roadmap, as well as all other planes accessible by UTF-16, are explicitly enumerated in Table 1.

Note that WG 2 has under consideration a proposal for use of Plane 14 for encoding special characters -- such as alphabet used for language tagging in some Inernet (IETF) protocols.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC 1/SC 2/WG 2

Universal Multiple-Octet Coded Character Set
(UCS)

ISO/IEC JTC 1/SC 2/WG 2 *N 1877*
1998-09-21

Title:	**New Annex - Request for Collection Ids - in Principles and Procedures document**
Source:	**V.S. Umamaheswaran**
References:	**N726**
Action:	**To be considered by SC 2/WG 2 at M35**
Distribution:	**ISO/IEC JTC 1/SC 2/WG 2**

4. Collection Identification

Technical Corrigendum No. 2 to ISO/IEC 10646-1 defines collections (clause 4.11 collection, and clause 4.17 fixed collection). A *collection* is a set of coded characters which is numbered and named and which consists of those coded characters whose code positions lie within one or more identified ranges. If any of the identified ranges include code positions to which no character is allocated, the repertoire of the collection will change if an additional character is assigned in the standard to any of these positions in the future. However, it is intended that the collection number and name will remain unchanged (even if the repertoire increases). A *fixed collection* is a collection in which every code position within the identified range(s) has a character allocated to it, and which is intended to remain unchanged -- the repertoire remains fixed. Several collections -- some are fixed collections marked by an '*' in the positions column -- are defined in Annex A of ISO/IEC 10646-1.

WG2 has accepted the following recommendations from the ad hoc on collection identifiers at WG 2 meeting 34 (see N1726):

A. Annex A in Part 1 will be the home for all collection identifiers and their names for collections that are entirely within Part 1 (BMP) or span both Part 1 and Part 2 (BMP and supplementary planes) of ISO/IEC 10646.

B. Annex A in Part 1 will mark a block of numbers in it as reserved for identifying collections that are entirely within Part 2 (supplementary planes) of ISO/IEC 10646.

C. An Annex in Part 2 should be created, similar to Annex A in Part 1, containing the list of collection identifiers, collection names for collections that are entirely within Part 2. Also, some text should be added in this Annex to refer the readers to Annex A in Part 1 for the other collection identifiers in the standard.

A collection identifier and collection name are usually assigned whenever a new script gets added to the standard. A collection could be referenced in an application by its identifier or as a collection of collections by enumerating the collection identifiers or collection names. However, there may be situations where an application needs a single identifier for a specific collection that is not readily identified in the standard or a reference an enumeration of collections is not acceptable.

Source:	**WG2**
Title:	**ISO/IEC JTC 1/SC 2/WG 2 Comments on SC2 N 3144**, - Contribution from the Netherlands to JTC 1 on the Functioning of ISO/IEC JTC 1/SC 2, Coded Character Set (JTC 1 N 5449)

These comments concentrate on those aspects that deal with the functioning of WG 2.

ISO/IEC 10646 is called the "Universal Character Set" for a reason: it is intended to cover all the scripts of the world. This view is widely shared by the industry and the user communities, both of which participate both directly and indirectly in the work of WG 2.

Thus, the stated wish to create rational limitations to additions of new characters or scripts must not lead to arbitrary barriers for entry. The market relevance for inclusion cannot be defined in the terms proposed in the contribution by the Netherlands National Body, as long as the industry upholds the above stated stand on universality.

Admittedly, as the result of the above, it is not uncommon that any given National Body may find itself in a situation, where it has no expertise on a particular script in the process of being encoded. The value of the participation of such a National Body then lies in their expertise in the standardization process and in the general aspects of encoding characters for use in IT applications.

The process in SC 2/WG 2 is a transparent one and extensive efforts continue to be made to implement advanced planning for it. This process is being continuously improved and we welcome the Netherlands to participate in this process in co-operation with the other National Bodies, liaison organizations, and experts. It is unavoidable that many issues will be discussed and prepared outside the formal WG operation. In particular, WG2 has made successful use of ad-hoc meetings to resolve complex technical issues requiring specialized expertise. This would appear to be in line with the stated wishes of the Netherlands NB.

The suggestion by the Netherlands NB to turn over sections of the standard to relevant user communities is at best impractical for lack of mechanism and control and at its worst would result in a fractured standard, something that very much goes against the idea of a unified, universal character set which is the primary market requirement.

As to the particular technical issue raised in attachment B, WG 2 has gone out of its way to accommodate the Netherlands NB position in that matter (see WG 2 N 1789R2, of which an extract is attached to this document) and this issue would thus appear to be resolved, as of this meeting.

This has been a difficult process, not least because the relevant contribution of the Netherlands in Crete was withdrawn at the request of the Netherlands representative. Other than that, we believe that the representative of the Netherlands has been given

ISO/IEC JTC1/SC2/WG2 N1884R2
1998-09-25

Universal Multiple-Octet Coded Character Set

Doc Type: Working Group Document
Title: **Additional Characters for the UCS**
Source: Ad Hoc on Bucket 35
Status: Working Document

References: Cited separately in sections below.

Meeting Dates: 1998-09-22, 1998-09-23
Attendees: Joan Aliprand (USA)
Michael Everson (Ireland)
Klaas Ruppel (Finland)
Johan van Wingen (Netherlands)
Ken Whistler (USA)
Chris White (British Library)

Synopsis

This document constitutes the meeting report of the ad hoc committee on “Bucket 35”, the collection of characters from proposals for small additions of various characters.

From a procedural point of view, this collection is divided into two parts. The first part consists of a large number of characters from many TC46 standards proposed for encoding in 10646, along with 10 Cyrillic Sami characters discussed in the same set of documents. The second part consists of many small, unrelated proposals.

Characters Derived from TC46 Standards; Cyrillic Sami Characters

This collection of characters is dealt with in the following documents. The original TC46-related proposals are WG2 N 1741, N 1743, N 1744, N 1745, N 1746, N 1747, N 1748, N 1749. The Cyrillic Sami character proposal is N 1813 (there are earlier versions, but those documents are superseded by N 1813). WG2 N 1840 is the consolidated response of the U.S. national body to WG2 N 1741, 1743 - 1749. WG2 N 1885 is the consolidated response of the U.S. national body to WG2 N 1742, N 1813, and to those characters in several others of the TC46-related proposals which do not in fact derive from the TC46 standards. WG2 N 1847 is the consolidated response of the Irish national body to L2/98-292 (= WG2 N 1840) and to N 1885. The actual citations of particular characters are scattered all through this particular interrelated set of documents.

It is the opinion of the ad hoc that the best approach for WG2 to deal with these documents is to respond to the last in the sequence (N 1847). Any decision made on that basis of that document renders all the earlier documents moot. WG2 can then invite the Irish national body to develop superseding new proposals (if it so desires) regarding any remaining characters not satisfactorily dealt with by the response to N 1847.

Because a large number of the proposed characters proved to be controversial, they are divided here into three categories, as determined by the ad hoc: 1. Those that should be encoded, 2. Those that should not be encoded (for various reasons), and 3. Those that require further study. For those that the ad hoc determined

endum No. 2 to ISO/IEC 10646:1 - 1983 (E) was recently balloted on in JTC 1/SC 2, and most likely has been approved by SC 2. This contribution describes a deficiency in the standard for not being able to define sub-repertoires of 10646 which are *fixed over time,* and proposes that either the current definition of *'dense collection'* be modified or a new *'fixed collection'* be defined.

Reference:

(Note: the reference below is the version of DCOR.2 that was circulated to SC 2/WG 2 experts, which was further processed as COR.2 by an SC 2 ballot.)

Ref. ISO/IEC JTC 1/SC 2/WG 2 N 1664: Draft Technical Corrigendum No. 2 to ISO/IEC 10646-1:1 1983 (E)

> *4.11: collection: A set of coded characters which is numbered and named and which consists of those coded characters whose code positions lie within one or more identified ranges.*
>
> > *Note - If any of the identified ranges include code positions to which no character is allocated, the repertoire of the collection will change if an additional character is assigned to any of those positions at a future amendment of this International Standard.*
>
> *4.17: dense collection: A collection in which every code position within the identified range(s) has a character allocated to it.*
>
> > *Note - The repertoire of a dense collection can not be extended at a future amendment of this International Standard unless one or more of the identified ranges of code positions is extended.*
>
> Annex A of 10646 lists a number of sub-repertoires of 10646. The DCOR.2 has marked some of the listed collections (with an 'asterisk') as 'dense collections' based on the definitions cited above.

The Problem:

Among other things, the original request in SC 2/WG 2 N1512, for clarification of the term 'collection' in the standard was to remove the ambiguity as to whether the repertoire / sub-repertoire identified by the collection identifier was 'fixed' or 'variable' over time. Since some of the ranges of code positions used to identify some of the collections contain unassigned code positions, it was ambiguous as to whether a new collection identifier has to be issued when one or more of the unassigned code positions was assigned a character in the future.

SC 2/WG 2 N 1512 proposed the following definitions:

> *4.11 Collection: A set of coded characters which is numbered and named and which consists of those coded characters whose code positions lie within one or more identified ranges.*
>
> Note: If the identified range includes code positions that are unassigned, the repertoire of the collection will change if additional characters are assigned to one or more of those positions at a future amendment of this standard.

4.12 Specified collection: A collection which contains no code positions reserved for future coding.

The notion of a 'specified collection' was to state that the repertoire identified as fixed collections cannot be 'changed ' - expanded or contracted - over time, whether the expansion is by adding characters to unassigned positions within identified ranges or by extending the range of code positions. It identifies a fixed sub-repertoire of 10646.

The notion of a not-specified (or variable) collection was to give a convenient way of referencing a set of related characters with the possibility of being able to add to that repertoire without having to issue a new collection identifier and without any consequences to the users of the collection identifier. The proposal was to re-define 'collection' in the standard to permit variability over time.

The redefinition of term 'collection' in COR.2 - basically defines all collections to be potentially variable. The 'dense collection', as is defined, states that within the ranges of code positions identified by the collection identifier, if there are currently unassigned positions, they cannot be included in the identified collection in the future - and to that extent it is fixed over time. However, it leaves the possibility of 'adding by extending the range' and hence it is potentially a 'variable' collection. It does not fully meet the requirement for identifying sub-repertoires of 10646 that are fixed over time.

It was unfortunate that we did not catch the nuances of the definitions proposed in document SC 2/WG 2 N1556 at the WG 2 meeting in Crete, and in subsequent ballot on the Corrigendum No. 2.

Proposal:

Two options are proposed for a further corrigendum to the standard. Option 1 is preferred over option 2, since the current definition of 'dense collection' is believed to be really not needed.

Also, note that the words 'extended at a fture amendment' should be changed to 'extended by a future amendment' in both clauses 4.11 and 4.17 in COR.2 (quoted above).

Option 1:

Redefine the 'dense collection' in COR.2 as follows:

4.17: dense collection: A collection in which every code position within the identified range(s) has a character allocated to it.

Note - The repertoire of a dense collection can not be extended by a future amendment of this International Standard.

by deleting the following from the end of the 'Note' in 4.17 of COR.2:

"unless one or more of the identified ranges of code positions is extended."

Option 2:

Define another type of collection called *'fixed collection'* as follows:

4.1x: fixed collection: A collection in which every code position within the identified range(s) has a character allocated to it.

Note - The repertoire of a fixed collection can not be changed.

–