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

L2/05-030

Universal Multiple-Octet Coded Character Set (UCS)

ISO/IEC JTC 1/SC 2 *N*_____

ISO/IEC JTC 1/SC 2/WG 2 *N2902*

2005-01-14

Title:	**Principles and Procedures for Allocation of New Characters and Scripts and handling of Defect Reports on Character Names (Replaces N2652R, N2352R, N 2002 and N1876)**
Source:	**Ad hoc group on Principles and Procedures (Edited by: V.S. Umamaheswaran – umavs@ca.ibm.com)**
References:	**See References 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**

This document incorporates all updates that have been approved by WG 2 up to meeting M 45. and reflecting changes to clause numbers and annex numbers in ISO/IEC 10646: 2003.

Electronic versions of this document can be found at:

http://www.dkuug.dk/JTC1/SC2/WG2/docs/n2902.doc, or,
http://www.dkuug.dk/JTC1/SC2/WG2/docs/n2902.pdf.

Table of Contents

1. Introduction

This document is a standing document of ISO/IEC JTC 1/SC 2 WG 2. It consists of a set of Principles and Procedures on a number of items relevant to the preparation, submission and handling of proposals for additions of characters to the repertoire of the standard (ISO/IEC 10646 and the Unicode standard). The document also contains procedures and guidelines for adding new collection identifiers to the standard. Submitters should check the standard documents (including all the amendments and corrigenda) before preparing new proposals. Submitters are encouraged to visit the "where is my character" page on the Unicode web site for more information on checking whether a character or script is already encoded in the standard. Submitters are also encouraged to contact the convener of WG 2 (and the chair of the Unicode Technical Committee) to check if any other proposal on the intended character or script may have been considered earlier.

2. Allocation of new characters and scripts

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 information technology and communication systems. These include, for example, characters from hard copy publishing 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.

B. The characters encoded into the Basic Multilingual Plane will not cover all characters included in future standards:
It is not necessary, though it may often be desirable, that all characters encoded in *future* international, national, and industry information technology and communication standards are included *in the BMP*. The first edition used characters from pre-existing standards as a means of evaluating the established utility as well as ensuring compatibility with existing practice. Characters encoded in future standards may or may not have proven utility, and may or may not establish themselves in common use.

2.2 Character categories

WG 2 will use the following categories to aid in assessing the encoding of the proposed characters.

A Contemporary
There exists a contemporary community of native users who produce new printed matter with the proposed characters in newspapers, magazines, books, signs, etc. Examples include Myanmar (Burmese), Thaana (Maldivian), Syriac, Yi, Xishuang Banna Dai[1].

B.1 Specialized (small collections of characters)
The characters are part of a relatively small set. There exists a limited community of users (for example, ecclesiastical) who produce new printed material with these proposed characters. Generally, these characters have few native users, or are not in day-to-day use for ordinary communication. Examples include Javanese and Pahlavi.

B.2 Specialized (large collections of characters)

[1]Since the writing of this initial set of principles and procedures several scripts proposed following these guidelines have been reviewed and included in the standard.

The characters are part of a relatively large set. There exists a limited community of users (for example, ecclesiastical) who produce new printed material with these proposed characters. Generally, these characters have few native users, or are not in day-to-day use for ordinary communication. Examples include personal name ideographs, Chu Nom, and Archaic Han.

C Major extinct (small collections of characters)

The characters are part of a relatively small set. There exists a relatively large body of literature using these characters, and a relatively large scholarly community studying that literature. Examples include Old Italic and Linear B.

D Attested extinct (small collections of characters)

The characters are part of a relatively small set. There exists a relatively limited literature using these characters and a relatively small scholarly community studying that literature. Examples include Samaritan and Meroitic.

E Minor extinct

The characters are part of a relatively small set. The utility of publicly encoding these characters is open to question[2]. Examples are Khotanese and Lahnda.

F Archaic Hieroglyphic or Ideographic

These characters are part of a large set (for example, 160 or more characters) of hieroglyphic or ideographic characters. In general, for a large character set, it is difficult to obtain information or agreement on the precise membership of the set. Examples include Lolo, Moso, Akkadian, Egyptian Hieroglyphics, Hittite (Luvian), Kitan, Mayan Hieroglyphics, and Jurchin.

G Obscure or questionable usage symbols

The characters are part of a small or large collection that is not yet deciphered, or not completely understood, or not well attested by substantial literature or the scholarly community. Or they are symbols that are not normally used in in-line text, that are merely drawings, that are used only in two-dimensional diagrams, or that may be composed (such as, a slash through a symbol to indicate forbidden). Examples include Phaistos, Indus, Rongo-rongo, logos, pictures of cows, circuit components, and weather chart symbols.

As the standard evolved it was found necessary to provide guidelines on specific aspects of proposals for additional scripts and characters to the standard. See

Annex F: Formal criteria for disunification on page 25,
Annex G: Formal criteria for coding precomposed characters on page 28,
Annex H: Criteria for encoding symbols on page 30
Annex I: Guideline for handling of CJK unification and/or disunification error on page 34 and Annex J: Guideline for correction of mapping table error on page 35.

2.3 Procedure for encoding new characters and scripts

The following defines a procedure with criteria for deciding how to encode new characters in ISO/IEC 10646. This procedure shall be used for new scripts only after thorough research into the repertoire and ordering of the characters within the script.

See A.1 Submitter's responsibilities and the attached *Proposal Summary Form* in Annex A on page 10. Annex K: Levels of implementation in ISO/IEC 10646 on page 36 and Annex L: Character-naming guidelines on page 37 are extracts from the standard for convenience of users of the proposal summary form.

WG 2 evaluation procedure:

[2]The minor extinct category of characters may be secondary candidates for encoding elsewhere on the BMP or their limited scholarly communities may wish to encode them in the Private Use Area (PUA). Caution: Use of PUA is by agreement between sending and receiving devices and its content is NOT defined by the standard, and proposals for standardization should not include any of the PUA.

In assessing the suitability of a proposed character for encoding, WG 2 shall evaluate the credibility of the submitter and then use the following procedure:

1. ***Do not encode.***
   a) If the proposed character is a (shape or other) variation of a character already encoded in the standard and therefore may be unified, or
   b) If the proposed character is a precomposed character and does not pass the *formal criteria for coding precomposed characters* that is detailed in Annex G on page 28, or
   c) If the proposed character is a presentation form (glyph), variant, or ligature, or
   d) If the proposed character may be better represented as a sequence of standardized encoded characters, or
   e) If the proposed character is a non-Han character, and leads to disunification with an existing character in the standard, and does not pass the *formal criteria for disunification* that is detailed in Annex F on page 25.
2. ***Suggest use of the Private Use Area***

a) If the proposed character has an extremely small or closed community of customers, or
b) If the proposed characters are part of a script that is very complex to implement and the script has not yet been encoded in the standard (the Private Use Area - PUA, may be used for test and evaluation).

   (**Note:** Use of PUA is not standardized; its use is by agreement between sending and receiving devices, and its use should not be included in any proposal made to the standardization body for consideration.)

3. ***Encode on a supplementary plane***
   a) If the proposed character is used infrequently, or
   b) If it is part of a set of characters for which insufficient space is available in the Basic Multilingual Plane, or
   c) If the proposed character is part of a small number of characters to be added to a script already encoded in one of the supplementary planes (for example, the characters can be encoded at unallocated code positions within the block or blocks allocated for that script).
4. ***Encode on the Basic Multilingual Plane***
   a) If the proposed character does not fit into one of the previous criteria (1, 2, or 3 above), and
   b) If the proposed character is part of a well-defined character collection not already encoded in the standard, or
   c) If the proposed character is part of a small number of characters to be added to a script already encoded in the Basic Multilingual Plane (for example, the characters can be encoded at unallocated code positions within the block or blocks allocated for that script).

3. Handling defect reports on character names

In principle, the character names in the standard are not to be changed.

The main purpose of having this international standard is the interoperability of characters of all the world scripts represented by their assigned code points. Within each language version of the standard, the names of individual characters must be unique and fixed. The initially assigned names will be somewhat meaningful to the user community. However, it may be found to have some errors or found to be less satisfactory later on. Once standardized, these names must not be changed.

The short identifiers defined in the standard (in clause 6.3) can be used for identifying the standardized characters in a language-independent manner or between different language versions of the standard. The relevant text extracted from the standard is given below:

> *"Clause 6.3 Short identifiers for code positions (UIDs):*
> ISO/IEC 10646 defines short identifiers for each code position, including code positions that are reserved. A short identifier for any code position is distinct from a short identifier for any other code position. If a character is allocated at a code position, a short identifier for that code position can be used to refer to the character allocated at that code position."
>
> These short identifiers are independent of the language in which this standard is written, and are

thus retained in all translations of the text. The full syntax of the notation of a short identifier, in Backus-Naur form, is *{ U | u } [{+}(xxxx | xxxxx | xxxxxx) | {-}xxxxxxxx]*, where "x" represents one hexadecimal digit (0 to 9, A to F, or a to f).

Some examples -- U+DC00 identifies a code position that is permanently reserved for UTF-16, and U+FFFF identifies a code position that is permanently reserved. U+0025 identifies a code position to which a character is allocated; U+0025 also identifies that character (named PERCENT SIGN). The short identifier for LATIN SMALL LETTER LONG S may be notated in any of the following forms: 0000017F, -0000017F, U0000017F, U-0000017F, 017F, 017F, U017F or U+017F. Any of the capital letters may be replaced by the corresponding small letter.

One can view the names in each language version of the standard as unique long identifier of arbitrary character sequences *in that language*. Even in the English language version of the standard these names may not be very meaningful to casual readers of the standard. Such long identifiers are used to establish correspondences with names of characters in other character collections or standards in the same (and sometimes in a different) language.

The English language version, which is developed in WG 2, is also the reference document from which other language versions are created. This makes the invariance of names in the English version even more mandatory. Translated versions are generated by groups other than WG 2 - for example, the Canadian and French national bodies helped ITTF create the French language version of 10646.

If the names in the English language version of the standard are not suitable for clarity or accuracy for non-English users, these names can be translated in non-English versions of the standard, or in technical supplements in other languages. However, in all cases technical equivalence with the English version of the standard must be maintained from the viewpoint of all normative aspects of the standard including most importantly the interoperability of code points assigned to the characters.

There may be situations where annotations to names of characters in the English version of the standard may be warranted. Requests for such annotations to character names may be made by submitting a defect report. The principles of dealing with such defect reports by WG 2 are described in Annex B on page 18.

The following policy adopted by WG 2 at its meeting M41.11 in Singapore on 2001-10-31 captures the above paragraphs.

RESOLUTION M41.11 (Policy regarding acceptable changes to 10646):
WG 2 requests SC2 adopt the following policy regarding acceptable changes to ISO/IEC 10646 and convey the same to JTC1 for information and to SC2 membership to take note:

a. Once a character is assigned a code position in the standard it cannot be reassigned in the interest of ensuring interoperability of standardized characters.
b. The arrangement of the characters in the standard is fixed; sorting and collation of the characters is outside the scope of the standard.
c. The character names chosen by WG 2 for the English version of the standard are unique, fixed and may be arbitrary; once a character name is assigned, it *cannot* be changed even if additional information is provided later. These name strings are used, for example to establish correspondences with characters in other standards.
d. Any inconsistencies in names could be adjusted in other language versions either when the standard is translated or in supplementary external documentation.

4. Collection identification

ISO/IEC 10646 has the following definitions regarding collections:

> "*Clause 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. However it is intended that the collection number and name will remain unchanged in future editions of this International Standard."

The intent is to require a new collection identifier when that new collection either involves an expansion of identified range(s) or addition of new range(s) compared with an existing collection. Implementations may have associated a collection identifier using the outer bounds of defined ranges for an existing collection, and an expansion or addition of new ranges can negatively impact such an implementation.

> "*Clause 4.19 – Fixed collection:*
> 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 in future editions of this International Standard."

A number of collections -- some marked as *fixed collections* with an asterisk (*) in the positions column -- are defined in Annex on *Collections of graphic characters for subsets* in ISO/IEC 10646.

A collection identifier and a collection name are usually assigned whenever a new script is 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, and

- the required collection is not readily identified in the standard, or
- a reference to the required collection by an enumeration of standardized collections is not acceptable.

Annex E on page 23 provides a format and guidelines for requesting new collection identifiers in the standard.

When a new collection is a proper superset of an existing collection the name of the new collection should be chosen to be able to easily identify the superset subset relationship between these collections.

4.1 Enumeration of repertoires in other documents

There may be a need to enumerate a repertoire of characters in different documents such as national standards, resource definition documents or others. Such an enumeration can be in the form of:

- a listing of a sequence of one or more ranges of short identifiers (see section 3 on page 5), or
- a listing in the form of identifiers of one or more standardized collections, or
- a combination of the above - in the form of a list of one or more collection identifiers and a list of one or more ranges of short identifiers for the characters either removed from that collection or added to the listed collections.

4.2 Use of sequence identifiers

Where there is a need to identify a sequence of 'n' standardized characters that represents an element of a repertoire, the UCS Sequence Identifier (USI) (defined in clause 6.6 in the standard) should be used.

> "*Clause 6.6 UCS Sequence Identifiers*
> ISO/IEC 10646 defines an identifier for any sequence of code positions taken from the standard. Such an identifier is known as a UCS Sequence Identifier (USI). For a sequence of n code positions it has the form: <UID1, UID2, ..., UIDn>,
> where UID1, UID2, etc. represent the short identifiers of the corresponding code positions, in the same order as those code positions appear in the sequence. If each of the code positions in such a sequence has a character allocated to it, the USI can be used to identify the sequence of characters

allocated at those code positions. A COMMA character (optionally followed by a SPACE character) separates the UIDs. The UCS Sequence Identifier shall include at least two UIDs; it shall begin with a LESS-THAN SIGN and be terminated by a GREATER-THAN SIGN.

NOTE – UCS Sequences Identifiers cannot be used for specification of subset and collection content. They may be used outside this standard to identify: composite sequences for mapping purposes, font repertoire, etc."

Use of a combination of short identifiers, the collection identifiers, and UCS sequence identifiers in the manner described above provides a language-neutral way of enumerating a specific repertoire of characters. Whereas the USI definition permits code positions to which characters may not have been assigned, such a USI is not useful in defining character repertoires.

5. Work flow and stages of progression

To give the submitters of proposals for new scripts an understanding of how WG 2 deals with a proposal from its initiation to completion, Annex C on page 19 contains a description of the work flow and the various stages of progression of submissions to WG 2.

5.1 Checking the status of a proposal

The minutes and resolutions adopted by WG 2 at each of its meeting are made available at the WG 2's web site linked from the meetings.html page. The texts of any amendments in progress are also available from the WG 2's web site or through the national standard organizations that are the national member bodies of ISO. The Unicode consortium also maintains a document called pipeline.html listing all the characters that have been accepted for inclusion in the next version of the standard. These documents can be checked for the status of any proposal that has been submitted for consideration by the UTC and WG 2.

6. Roadmaps

A summary of the scripts and characters that have been included in the standard, and known scripts which are either work in progress in WG 2 (for which some initial discussion documents have been made available to WG 2), or scripts which are known for future possible inclusion in the standard but have not matured are addressed in Annex D on page 21.

7. Electronic submissions

Contributions for consideration by WG 2 (and to the Unicode Technical Committee) should be made in electronic form. The preferred formats are Word .DOC, or printable .PDF formats, with unprotected text portions and possibly copyrighted font portions. Whereas, files could be compressed to reduce the size, it should be noted that .EXE files may not be accepted in many organizations as part of their Security Policy and self-extracting .EXE files should be avoided.

8. Format of character additions in amendments to 10646

Per resolution M39.23, WG 2 has resolved that the format for amendments that involve character additions will be in the form of complete replacements of tables and character name lists where they exist, with an explanatory text listing the code positions to which new characters are assigned. If it is a new block it will be presented as a complete new table and names list.

9. On the relative ordering of characters

The repertoire encoded in the standard is intended for use by many languages. When characters of the same script are used in multiple languages they are unified. That makes it impossible to reflect the preferred alphabetical arrangement of characters for each language; a common arrangement of the characters is used instead. When scripts are encoded in the standard, the relative ordering of characters within that script is given due consideration. To ensure stability and interoperability, that arrangement remains fixed, even if additional characters of the same script are added at a later time.

Ensuring correct ordering of the characters encoded in ISO/IEC 10646 is outside the scope of the standard. ISO/IEC 14651 together with appropriate tailoring may be used to address the problem of ordering data encoded in ISO/IEC 10646 meeting the requirements of a given language or user community. The Unicode

Collation Algorithm (UCA) is synchronized with ISO/IEC 14651 and is freely available at http://www.unicode.org/unicode/reports/tr10.

10. Referencing ISO/IEC 10646

Referencing ISO/IEC 10646 should follow the format used for listing in the ISO directory. Note that the standard was published as two parts prior to end of year 2003.

Generic:
ISO/IEC 10646 Information Technology – Universal Multiple-Octet Coded Character Set (UCS).

For specific editions, for example:
ISO/IEC 10646: 2003 Information Technology – Universal Multiple-Octet Coded Character Set (UCS) -- Architecture and Basic Multilingual Plane, Supplementary Planes.

If you need to refer to versions of the standard that were published in two parts prior to the 2003 edition, for example
ISO/IEC 10646-1: 2000 Information Technology – Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane
ISO/IEC 10646-1: 2000 Information Technology – Universal Multiple-Octet Coded Character Set (UCS) – Part 2: Supplementary Planes
Please note that the 1993 edition of 10646-1 was amended significantly for the Hangul script by its Amendment 7 and must not be used.

11. WG2 web site

WG2 maintains its own web site at www.dkuug.dk/JTC1/SC2/WG2, hosted courtesy of Danish Unix User Group (dkuug). WG2 meeting notices, minutes, resolutions, document register, documents and standing documents are made available at this site. Since many of the documents at this site are referenced by many other documents WG2 has recognized the need for the stability of the URL of this site and has taken the following resolution:

> M45.31 (Stability of URLs for the web site):
> WG2 notes the critical importance of stable URLs for the existing collection of electronically available working group documents and authorizes its convener to take the necessary precautions and steps to ensure that all past, current and future URLs for working group documents remain stable.

Annex A: Information accompanying submissions

The process of deciding which characters should be included in the repertoire of the standard by WG 2 depends on the availability of accurate and comprehensive information about any proposed additions. WG 2, at its San Francisco meeting 26, designed a form (template) that will assist the submitters in gathering and providing the relevant information, and will assist WG 2 in making more informed decisions. This form has been revised over the years and the latest version is included in the following pages of this annex. The latest version of this form must be used in submissions. This form is also made available on line from the WG 2 web site – see http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html.

A duly completed proposal summary form must accompany each new submission. Such a form will assist WG 2 to better evaluate the proposal, and progress the proposal towards a speedier acceptance and inclusion in the standard. Submitters are also requested to ensure that a proposed character does not already exist in the standard.

Submitters are encouraged to visit the "Where is my Character" page on the Unicode web site for more information on checking if their proposed character or script is already encoded in the standard, or a similar proposal has already been made by someone else. There are also several electronic discussion lists maintained by the Unicode consortium that one could use to discuss with other experts internationally on various subjects related to the standard. Submitters are also encouraged to familiarize themselves with ISO/IEC TR15285 – Character Glyph Model (available on line from http://www.iso.org/iso/en/ittf/PubliclyAvailableStandards/c027163_ISO_IEC_TR_15285_1998(E).zip).

In addition to text extracted from the standard in Annex K: Levels of implementation in ISO/IEC 10646 and Annex L: Character-naming guidelines in the P&P document, the following definitions from the standard are also referenced in the proposal summary form:

> *Clause 4.12 Combining character*:
> A member of an identified subset of the coded character set of ISO/IEC 10646 intended for combination with the preceding non-combining graphic character, or with a sequence of combining characters preceded by a non-combining character (see also 4.14).
> NOTE – ISO/IEC 10646 specifies several subset collections which include combining characters.
>
> *Clause 4.14 Composite sequence*:
> A sequence of graphic characters consisting of a noncombining character followed by one or more combining characters (see also 4.12).
> NOTE 1 – A graphic symbol for a composite sequence generally consists of the combination of the graphic symbols of each character in the sequence.
> NOTE 2 – A composite sequence is not a character and therefore is not a member of the repertoire of ISO/IEC 10646.

A.1 Submitter's responsibilities

The national body or liaison organization (or any other organization or an individual) proposing new character(s) or a new script shall provide:

1. Proposed category for the script or character(s), character name(s), and description of usage.
2. Justification for the category and name(s).
3. A representative glyph(s) image on paper:
   If the proposed glyph image is similar to a glyph image of a previously encoded ISO/IEC 10646 character, then additional justification for encoding the new character shall be provided.
   Note: Any proposal that suggests that one or more of such variant forms is actually a distinct character requiring separate encoding, should provide detailed, printed evidence that there is actual, contrastive use of the variant form(s). It is insufficient for a proposal to claim a requirement to encode as characters in the Standard, glyphic forms which happen to occur in another character encoding that did not follow the Character-Glyph Model that guides the choice of appropriate characters for encoding in ISO/IEC 10646.
   Note*:* WG 2 has resolved in Resolution M38.12 not to add any more Arabic presentation forms to the standard and suggests users to employ appropriate input methods, rendering and font technologies to meet the user requirements.

4. Mappings to accepted sources, for example, other standards, dictionaries, accessible published materials.
5. Outline font resource:
   Prior to the preparation of the final text of the next amendment or version of the standard a suitable outline font resource is required. Outline fonts must be in TrueType format. Postscript fonts are acceptable if conversion to TrueType is possible using the tools available to the editors. Fonts submitted must have no license restrictions that prevent embedding into PDF documents. Because of synchronization between ISO/IEC10646 and The Unicode Standard, any grant of license must cover use in publishing both standards and related documents (see resolution M45.30). For technical reasons, the editors must be able under that license to freely modify or replace glyph outlines in their copies of the fonts. The submitter should be prepared to provide fonts of suitable quality and license conditions, unless equivalent fonts are already available to the editors.
6. List of all the parties consulted
   Submitters are encouraged to provide the email id-s of the submitters as well as other experts who have been consulted to facilitate any clarification queries.
7. Equivalent glyph images:
   If the submission intends using composite sequences of proposed or existing combining and non-combining characters, a list consisting of each composite sequence and its corresponding glyph image shall be provided to better understand the intended use.
8. Compatibility equivalents:
   If the submission includes compatibility ideographic characters, identify (per resolution M45.29):
   - the source which contains two distinct code positions that correspond to a single unified CJK Ideograph character of ISO/IEC 10646
   - the ISO 10646 unified CJK Ideograph
   - the code position in the source for the unified CJK ideograph
   - the code position in the source for the proposed compatibility ideograph
9. Properties that may affect the BiDi processing
   Any BiDirectional algorithm related properties associate with the characters should be spelled out (see UAX#9 - http://www.unicode.org/reports/tr9/).
10. Any additional information that will assist in correct understanding of the different characteristics and linguistic processing of the proposed character(s) or script.

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[3]
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for guidelines and details before filling this form.**
Please ensure you are using the latest Form from http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html**.**
See also http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** _____
2. Requester's name: _____
3. Requester type (Member body/Liaison/Individual contribution): _____
4. Submission date: _____
5. Requester's reference (if applicable): _____
6. Choose one of the following:
This is a complete proposal: _____
or, More information will be provided later: _____

B. Technical - General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters): _____
      Proposed name of script: _____
   b. The proposal is for addition of character(s) to an existing block: _____
      Name of the existing block: _____
2. Number of characters in proposal: _____
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary _____ B.1-Specialized (small collection) _____ B.2-Specialized (large collection) _____
   C-Major extinct _____ D-Attested extinct _____ E-Minor extinct _____
   F-Archaic Hieroglyphic or Ideographic _____ G-Obscure or questionable usage symbols _____
4. Proposed Level of Implementation (1, 2 or 3) (see Annex K in P&P document): _____
   Is a rationale provided for the choice? _____
      If Yes, reference: _____
5. Is a repertoire including character names provided? _____
   a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? _____
   b. Are the character shapes attached in a legible form suitable for review? _____
6. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? _____
   If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used: _____
7. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? _____
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? _____
8. Special encoding issues:
   Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)?
   _____
9. Additional Information:
Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see http://www.unicode.org/Public/UNIDATA/UCD.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

C. Technical - Justification

[3] Form number: N2902-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01)

1. Has this proposal for addition of character(s) been submitted before? _____

If YES explain _____

2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? _____

If YES, with whom? _____

If YES, available relevant documents: _____

3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? _____

Reference: _____

4. The context of use for the proposed characters (type of use; common or rare) _____

Reference: _____

5. Are the proposed characters in current use by the user community? _____

If YES, where? Reference: _____

6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? _____

If YES, is a rationale provided? _____

If YES, reference: _____

7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? _____

8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? _____

If YES, is a rationale for its inclusion provided? _____

If YES, reference: _____

9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? _____

If YES, is a rationale for its inclusion provided? _____

If YES, reference: _____

10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character? _____

If YES, is a rationale for its inclusion provided? _____

If YES, reference: _____

11. Does the proposal include use of combining characters and/or use of composite sequences? _____

If YES, is a rationale for such use provided? _____

If YES, reference: _____

Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? _____

If YES, reference: _____

12. Does the proposal contain characters with any special properties such as control function or similar semantics? _____

If YES, describe in detail (include attachment if necessary) _____

13. Does the proposal contain any Ideographic compatibility character(s)? _____

If YES, is the equivalent corresponding unified ideographic character(s) identified? _____

If YES, reference: _____

Example 1
ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[4]
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for guidelines and details before filling this form.**
Please ensure you are using the latest Form from http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html**.**
See also http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** ***Braille***
2. Requester's name: *Kohji Shibano, Japan*
3. Requester type (Member body/Liaison/Individual contribution): *Individual Contribution*
4. Submission date: *1994-10-10*[5]
5. Requester's reference (if applicable): *J2-94-xy*
6. Choose one of the following:
   This is a complete proposal:
   or, More information will be provided later: *Yes*

B. Technical - General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters): *Yes*
      Proposed name of script: *Braille*
   b. The proposal is for addition of character(s) to an existing block: *No*
      Name of the existing block:
2. Number of characters in proposal: *448*
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary *XX* B.1-Specialized (small collection) ____ B.2-Specialized (large collection) ____
   C-Major extinct ____ D-Attested extinct ____ E-Minor extinct ____
   F-Archaic Hieroglyphic or Ideographic ____ G-Obscure or questionable usage symbols ____
4. Proposed Level of Implementation (1, 2 or 3) (see Annex K in P&P document): *1*
   Is a rationale provided for the choice? *No*
   If Yes, reference:
5. Is a repertoire including character names provided? *Yes*
   a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? *No (will provide)*
   b. Are the character shapes attached in a legible form suitable for review? *Yes*
6. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? *Japan*
   If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used: *IBM Japan (ftp://ifi.jp/pub/font)*
7. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *ISO TC 173*
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of use of proposed characters attached? *No (will provide)*
8. Special encoding issues:
   Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)?
   No
9. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see http://www.unicode.org/Public/UNIDATA/UCD.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

C. Technical - Justification

[4] Form number: N2902-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-10, 2005-01)
[5] The date of this example is retained as originally created even though the form has been revised since that date.

1. Has this proposal for addition of character(s) been submitted before? *No*
   If YES explain
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? *No*
   If YES, with whom?
   If YES, available relevant documents?
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?
   Reference: *People with impaired vision (info will be provided)*
4. The context of use for the proposed characters (type of use; common or rare) *Common*
   Reference: *on-line database services for Braille-translated text (e.g. www: braille.dknet.dk)*
5. Are the proposed characters in current use by the user community? *Yes*
   If YES, where? Reference: *Worldwide*
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? *Yes*
   If YES, is a rationale provided?
   If YES, reference:
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? *Yes*
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? *No*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? *No*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character? *No*
    If YES, is a rationale for its inclusion provided?
    If YES, reference:
11. Does the proposal include use of combining characters and/or use of composite sequences? *No*
    If YES, is a rationale for such use provided?
    If YES, reference:
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?
    If YES, reference:
12. Does the proposal contain characters with any special properties such as control function or similar semantics? *No*
    If YES, describe in detail (include attachment if necessary)
13. Does the proposal contain any Ideographic compatibility character(s)? *No*
    If YES, is the equivalent corresponding unified ideographic character(s) identified?
    If YES, reference:

Example 2
ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[6]

Please read Principles and Procedures Document (P & P) from http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for guidelines and details before filling this form.**
Please ensure you are using the latest Form from http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html**.**
See also http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** ***Addition of two Latin characters***
2. Requester's name: *Danish Standards Association*
3. Requester type (Member body/Liaison/Individual contribution): *NB*
4. Submission date: *1995-03-10*[7]
5. Requester's reference (if applicable):
6. Choose one of the following:
   This is a complete proposal: *Yes*
   or, More information will be provided later:

B. Technical - General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters): *No*
      Proposed name of script:
   b. The proposal is for addition of character(s) to an existing block: *Yes*
      Name of the existing block: *Table 4 - Row 01: Latin Extended-B*
2. Number of characters in proposal: *2*
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary *XX* B.1-Specialized (small collection) ____ B.2-Specialized (large collection) ____
   C-Major extinct ____ D-Attested extinct ____ E-Minor extinct ____
   F-Archaic Hieroglyphic or Ideographic ____ G-Obscure or questionable usage symbols ____
4. Proposed Level of Implementation (1, 2 or 3) (see Annex K in P&P document): *1*
   Is a rationale provided for the choice?
      If Yes, reference:
5. Is a repertoire including character names provided? *Yes*
   a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? *Yes*
   b. Are the character shapes attached in a legible form suitable for review? *Yes*
6. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? *Michael Everson,*
   If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used: *Michael Everson,*
7. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *Yes*
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *Yes*
8. Special encoding issues:
   Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)?
   Specifications enclosed
9. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see http://www.unicode.org/Public/UNIDATA/UCD.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[6] Form number: N2902-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-10, 2005-01)
[7] The date of this example is retained as originally created even though the form has been revised since that date.

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? *No*
   If YES explain
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? *Yes*
   If YES, with whom? *Irish National Body, Oxford University*
   If YES, available relevant documents? *Enclosed*
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? *Yes*
   Reference:
4. The context of use for the proposed characters (type of use; common or rare) *Rare*
   Reference: *The Community of Gothic and Medieval English Literature*
5. Are the proposed characters in current use by the user community? *Yes*
   If YES, where? Reference: *Scholar Communities*
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? *Yes*
   If YES, is a rationale provided? *Yes*
   If YES, reference: *Enclosed*
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? *No*
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? *No*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? *No*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character? *No*
    If YES, is a rationale for its inclusion provided?
    If YES, reference:
11. Does the proposal include use of combining characters and/or use of composite sequences? *No*
    If YES, is a rationale for such use provided?
    If YES, reference:
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?
    If YES, reference:
12. Does the proposal contain characters with any special properties such as control function or similar semantics? *No*
    If YES, describe in detail (include attachment if necessary)
13. Does the proposal contain any Ideographic compatibility character(s)? *No*
    If YES, is the equivalent corresponding unified ideographic character(s) identified?
    If YES, reference:

Annex B: Handling of defect reports on character names

Since the first publication of ISO/IEC 10646 in May 1993, WG 2 has received several defect reports requesting changes to character names. In principle, the names in the standard are not to be changed. However, there may be situations where an annotation to the character name may be warranted.

B.1 Principles used by WG 2

The following paragraphs describe the principles of dealing with defect reports on character names:

A. Explanatory information in Annex P on *Additional information on characters* in the standard:
   If WG 2 decides that the request is justified, WG 2 will first consider accommodating the request by adding explanatory text to Annex P of the standard.
B. Non-normative parenthetic annotation of the name:
   If WG 2 considers that the request falls within the guidelines of Rule 12 in Annex L on *Character naming guidelines* in the standard, then an appropriate annotation will be added to the character name.
C. In instances where a name causes a potential problem for compliance by implementations of existing standard, and if the concern expressed in the defect report may be handled with a simple explanatory note, a note may be added.
D. Deprecation:
   If WG 2 considers that the character identified in the defect report should not have been in the standard, for reasons such as duplication, or incorrect inclusion in a block, then that coded character will be marked with the annotation *(deprecated character)* after its name. Note, however, that the character will never be removed from the standard.
E. Reject:
   In all other situations, where WG 2 considers that the request is not sufficiently justified or none of the above-mentioned measures is warranted, the defect report will be rejected with an explanation.

B.2 Some guidelines for submitters of defect reports

As a supplement to the above information on dealing with defect reports, the submitters can assist the working group by following the guidelines given below:

a) report all defects associated with characters from the same block or set of characters as a single defect report (for example, use a single one for all defects from within a character block such as Malayalam), instead of one for each character.
b) avoid including defective characters from different character blocks or sets in the same report.
c) please check if the defect has already been reported by some one else or considered earlier by WG 2. Copies of the dispositions of prior defect reports can be obtained from the SC 2 Secretariat.
d) if one or more new character(s) - with their own new name and glyph - is proposed to be added in conjunction with a defect report, please submit the addition requests separate from the defect report along with the Proposal Summary Form for the new characters.

Annex C: Work flow and stages of progression

This annex contains a description of the UCS workflow and stages in progression from initial proposal to final publication.

C.1 The UCS workflow

UCS workflow can be illustrated in a simplified form as follows:

Communication to WG 2 and communication inside WG 2 related to populating the standard	**Communication from WG 2 to the world outside**

Input		**Process**	**Output**	**Output**	
From whom	**What**	**Under meetings**	**After meetings**	**What**	**To whom**
• Convener • SC 2 • JTC 1 • ITTF	• Agenda; (see meetings.html). • Ballots	Resolutions; (see meetings.html) .	• Minutes (see meetings.html) . • Action Items	Result of request: • Acceptance • Rejection	Requester
• NBs • WG experts • IRG-group • Liaisons	Input documents: • Requests (e.g. N2555) • Defect reports (e.g. N1806) • Working documents • Liaison statements (see documents.html).			• Editorial corrigenda. • Technical corrigenda (e.g. N1393) • Amendments (e.g. http://std.dkuug.dk/jtc1/sc2/open/02n3760.zip) • Standards (e.g. ISO/IEC 10646: 2003)	• SC 2 • JTC 1 • ITTF
• Secretary • Editor	• Minutes (see meetings.html). • Action Items (included in latest Minutes) • Standing documents (see principles.html and roamdaps.html)				• IRG
	Types of Documents			**How**	
• Secretary • Editor	Standing documents: • WG 2 distribution list (e.g. N1351) • Document register (e.g. N1300) • Summary of WG 2 work (e.g. N1302) • • List of character names and code positions allocated (e.g. N1675) • Principles and procedures • Roadmaps to BMP and Supplementary Planes			Presentation forms: • Paper documents • Web site (the WG 2 web site at DKUUG and the IRG web site in HKSAR)	

C.2 Stages of work

Any new proposal for addition of new characters will pass a number of stages from initial proposal to finalized publication. The stages are:

- Initial proposal
- Provisional acceptance
- Final acceptance (Bucket)
- Hold for ballot

This terminology indicates the stage of maturity of the proposal and the WG's confidence in the proposal.

		In process within WG 2			Further progression			
Stage ⇒ ——— Item ⇓		**Initial proposal**	**Provisional acceptance**	**Final acceptance (allocation of bucket)**	**Hold for ballot**	**Progression/ Publication status**		
						SC 2 Ballot	**JTC 1 Ballot**	**ITTF Publication**
		1	**2**	**3**	**4**	**5****	**6****	**7****
1*	Character shapes	1.1	2.1					
2*	Character names	1.2	2.2					
3*	Code position allocation	1.3	2.3					
4*	Text to be included in the standard	1.4	2.4					
5*	Font**	1.5	2.5					
6	Other items from proposal summary form	1.6	2.6					

* Items 1 through 5 are mandatory for entering 'final acceptance' stage
** Outline fonts of publication quality are needed for charts prepared from SC2 ballot onwards. For information on the format of the font see the *Proposal summary form* in Annex A (item on *Outline font resource* on page 11).

- Stages 1 to 3 may contain provisionally allocated code positions. When a proposal enters stage 4 the code positions are final.
- The contents of the Buckets are reviewed at every meeting to decide whether the content shall progress for balloting (stage 4).
- The progress of each proposal is recorded in the WG 2 meeting minutes and resolutions.
- When a proposal reaches stage 4 its status is included in *List of character names and code positions allocated* (see also pipeline.html, which is in synch with 10646 repertoire additions).

C.3 Examples

List of character names and code positions allocated:

Code position	Status	Reference	Character name
...			
20AB	6	N1092	DONG SIGN
...			
012C			LATIN CAPITAL LETTER I WITH BREVE
...			
00E6	7	N1128	LATIN SMALL LETTER AE (ash)
...			
1E9B	6	N1132	LATIN SMALL LETTER LONG S WITH DOT ABOVE
...			
FFFC	2	N1365	OBJECT REPLACEMENT CHARACTER

Resolutions of each meeting captures the decisions on different proposals and minutes detail the discussions.

Annex D: BMP and Supplementary Planes allocation roadmaps

D.1 Overview

The intent of the *roadmaps* document is to show a visual layout of the coding space for further allocation of scripts in ISO/IEC 10646 (also in the Unicode Standard), in the BMP and in the Supplementary planes. The roadmap document is intended to be used as a general guideline – it *does not attempt to make detailed allocations of characters*.

The planes described in the roadmap document, as well as all other planes accessible by UTF-16 are explicitly enumerated in the following table.

Allocations for Planes in ISO 10646

Range of UCS-4 values (Hex)	Plane #	Name of Plane
00000000 ... 0000FFFF	**0**	Basic Multilingual Plane - BMP; envisioned for encoding all contemporary scripts and symbols including most frequently used ideographs.
00010000 … 0001FFFF	**1**	Supplementary Multilingual Plane for scripts and symbols – SMP; envisioned for encoding future non-Ideographic and non-Unified Ideographic scripts and symbols.
00020000 ... 0002FFFF	**2**	Supplementary Ideographic Plane (SIP); envisioned as containing future Unified Ideographic characters.
00030000 ... 0003FFFF to 000D0000 ... 000DFFFF	**3* to 13**	Reserved for Future Allocations.
000E0000 ... 000EFFFF	**14**	Supplementary Special-purpose Plane (SSP); envisioned for encoding special characters such as alphabet used for language tagging.
000F0000 ... 000FFFFF	**15**	Reserved for Private Use.
00100000 ... 0010FFFF	**16**	Reserved for Private Use.

* Should plane 2 prove to be insufficient for future Han character encoding, it is anticipated that further allocations may be provided on plane 3.

The roadmap layouts are maintained by an adhoc group on Roadmaps. This group's latest working document is located at http://www.unicode.org/roadmaps. A snapshot of these layouts is submitted for acceptance at each WG 2 meeting for the continued work on ISO/IEC 10646 and is closely coordinated with the work on the Unicode Standard in liaison with the Unicode Consortium. The latest snapshot of the roadmaps for the BMP and the Supplementary planes can be found at:

Roadmaps.html – http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html.

Please note that this roadmap consolidates into a single document information for each of the planes 0, 1, 2 and 14.

- The BMP or Plane 0 roadmap (a snapshot http://www.unicode.org/roadmaps/bmp/) locates all script and individual character additions published in ISO/IEC 10646: 2003 (and Unicode 4.0), plus all script additions currently foreseen to be reasonable candidates for future encoding in the BMP.
- The SMP or Plane 1 roadmap (a snapshot of http://www.unicode.org/roadmaps/smp/) locates all script and individual character additions included in ISO/IEC 10646: 2003 (included in Unicode 4.0), plus all script additions currently foreseen to be reasonable candidates for future encoding in the SMP. By current estimates all remaining general scripts and symbol sets not encoded or as possible candidates for the BMP should fit within the SMP.
- The SIP or Plane 2 roadmap (a snapshot of http://www.unicode.org/roadmaps/sip/) locates all script and individual character additions included in ISO/IEC 10646: 2003 (included in Unicode 4.0), plus all script additions currently foreseen to be reasonable candidates for future encoding in the SIP. This plane is envisioned as containing future Unified Ideographic character additions. The largest current Unified

Ideographic character collection should fit within the BMP and SIP, as long as duplicate character encoding is avoided.

The above layouts indicate that these three planes should suffice for all future encoding of characters having world-wide utility. In addition,

- The SSP or Plane 14 roadmap (a snapshot of http://www.unicode.org/roadmaps/ssp/) locates all script and individual character additions included in ISO/IEC 10646: 2003 (included in Unciode 4.0), plus all script additions currently foreseen to be reasonable candidates for future encoding in the SSP. This plane is used for encoding special characters such as alphabet used for language tagging, and variation selectors.

Note that additional 10 supplementary planes are available for encoding (with an additional 2 planes reserved for private use). Should plane 2 prove to be insufficient for future Han character encoding, it is anticipated that further allocations may be provided on plane 3.

The layouts show the different scripts in various stages of progression – published, accepted but not yet published, under evaluation in UTC and WG 2, exploratory having some preliminary documentation, or open with no proposal documents.

The status of script proposals and their progress at any given time can be found in the meeting resolutions, meeting minutes as well as from WG 2's document register (the document number for registers by convention is a multiple of 50 and will be the latest xx00 or xx50), available from WG 2's web site.

D.2 Guidelines for roadmap allotments

Some principles to be followed in assigning scripts in the roadmaps and for encoding in the standard are given below.

D.2.1 Block assignment starting on half-row boundary

When allocating code space to a block requiring fewer than 128 positions, these positions should not cross a 128-code position (half row) boundary. Wherever possible, if the number of positions is close to 128, it is preferable to start the collection at the half-row boundary. For blocks slightly larger than 128 positions the highest frequency characters should all be allocated within the first 128 positions. This highest frequency allocation principle may be overridden when there is justification to do otherwise. The purpose of this guideline is to insure greater compression ratios for run-length compression techniques. (See resolution M33.11). Further, for blocks requiring closer to 128 positions it is desirable to start at a half-row boundary.

D.2.2 1024 code position boundary for supplementary planes

Supplementary planes 1 to 16 are accessed using pairs of High and Low S-zone values employing UTF-16 transformation. Each High S-zone value corresponds to a block of 1024 code positions. When large blocks are considered for encoding in the supplementary planes it is desirable to start the block at the 1024-code position boundary. This facilitates range-checking operations for particular blocks in the supplementary planes by examining the High S-zone value alone.

D.2.3 Empty '00' position in a block

Proposals for code allocations should not leave position 00 unassigned in each block unless there are compelling documented reasons for doing so.

D.2.4 Gaps in ranges of assigned code positions

At the time of initial encoding of a script or a set of related characters, gaps may have been left in the range of assigned code positions. These gaps are reserved for future assignment of characters that are related in terms of its properties to the surrounding characters, for example a gap in a range of superscripted characters can be assigned a future superscripted character. In the supplementary planes, specifically in Plane 1, some gaps in the Math Alphanumerics and in the Western Musical symbols are left there for transient mappings, since some of the characters needed for these scripts were already encoded in the BMP before their encoding in Plane 1. Transient mappings permit more efficient processing of scripts that are split across the BMP and a supplementary plane.

Annex E: Request for new collection identifiers

(Source: ISO/IEC JTC 1/SC 2/WG 2 N1877 -1998-09-20 - modified based on discussion at M35; AI-M35-6b)

Request For Collection Identifier For a Sub-Repertoire Of ISO/IEC 10646	
	Date: _____
SOURCE:	
Email address of source:	
Phone number of source:	
Fax number of source:	
Address of source:	
WG 2 SPONSOR	
(Preferably a member body or liaison organization of ISO/IEC JTC 1 or its subcommittees and working groups)	
SUBMITTER'S REFERENCE:	

SUBMITTER AND THE SPONSOR SHOULD DO THE FOLLOWING:

A. Ensure that no existing collection identified with a Collection Identifier in ISO/IEC 10646 satisfies their needs. If a single collection does not exist, provide justification why an enumeration of two or more identified collections cannot satisfy the need.
B. Ensure that the proposed collection of characters is a true subset of the repertoire of characters of ISO/IEC 10646 (including all its amendments and corrigenda). The list of character names in Annex G of ISO/IEC 10646 can be used as an aid. If any character is NOT currently encoded in the standard, that character should be submitted for inclusion in the standard, following the guidelines documented in section 1 on page 3, and in Annex A on page 10 of this document.
C. Prepare a list of existing collections that are fully contained in the proposed collection. Ensure that you have considered all the approved amendments of the Standard while preparing this list of collections.
D. List any code positions that are included in the proposed collection, but are NOT included in the list of existing collections identified in step C above.
E. For each of the existing collection that is identified in step C above, list any code position that is to be excluded from the proposed collection.
F. If the proposed collection is to be marked as FIXED, provide a list of individual code positions that are NOT allocated in each of the collections identified in step C above and therefore to be excluded from the proposed collection.
G. Decide if the collection is to be marked as a FIXED collection (see section 4 on page 7 of this document).
H. Prepare a background document, including the rationale and intended use of the collection and forward it to the Convener of ISO/IEC JTC 1/SC 2/WG 2 for consideration, acceptance and assignment of a Collection Identifier by WG 2.
I. If the proposed collection is a superset or subset of an existing collection select a suitable collection name to easily identify the subset superset relationship. The following WG2 resolution is relevant in this context:
M45.32 (Stability of IICORE collection):
With reference to document N2780 on International Ideograph Core subset, WG2 accepts that any future amendments of the International Ideograph Core should be recorded as new collections with their own unique

collection names. The new collection must be a true superset of all existing IICORE and any of its amended collections in an upwardly compatible manner.

Format to be used for sub-repertoire submission

An example format of the proposal for collection definition is given below. The final form of documenting the sub-repertoire in the standard is at the discretion of the project editor(s).

Collection Name: **EXAMPLE COLLECTION**[8]

Collection to be marked as Fixed (Yes / No): **YES**

Plane *00*

Rows	Positions (Cells)
00	*20-7E, A0-FF*
01	*00-13 16-2B 2E-4D 50-7E*
02	*C7 D8-DB DD*
1E	*80-85 F2 F3*
20	*15 18 19 1C 1D AC*
21	*22 26 5B-5E 90-93*
26	*6A*

Collections containing the proposed sub-repertoire

The following UCS collections from Annex A of ISO/IEC 10646 contain characters of the above-proposed collection:

ID	*UCS-Collection Name / Code Positions*	*Positions to be included or excluded*
1	BASIC LATIN 0020-007E	All are included
2	LATIN-1 SUPPLEMENT 00A0-00FF	All are included
3	LATIN EXTENDED-A 0100-017F	Only 0114, 0115, 012C, 012D, 014E, 014F, and 017F are included.
6	SPACING MODIFIER LETTERS 02B0-02FF	Only 02C7, 02D8—02DB and 02DD are included.
32	GENERAL PUNCTUATION 2000-206F	Only 2015, 2018, 2019, 101C and 201D are included.
34	CURRENCY SYMBOLS 20A0-20CF	Only 20AC is included.
36	LETTERLIKE SYMBOLS 2100-214F	Only 2122 and 2126 are included.
37	NUMBER FORMS 2150-218F	Only 215B—215E are included.
38	ARROWS 2190-21FF	Only 2190—2193 are included.
47	MISCELLANEOUS SYMBOLS 2600-26FF	Only 266A is included.

Justification for a Single Collection Identifier Request

(For example) A single collection identifier is required to tag textual data in a particular protocol with a character set identifier.

[8]This example is based on an input document on Latin Characters based on ISO/IEC 6937:1994, from Mr. Johan van Wingen, Netherlands; the Euro Sign has been added; see WG 2 N2211 - Request for Collection Identifiers for European Repertoires.

Annex F: Formal criteria for disunification

(Source: ISO/IEC JTC 1/SC 2/WG 2 N1724 - 1998-03-05- adopted with revisions at M34 - action item M34-7d.)

There have been repeated proposals to disunify existing characters. These proposals cannot be fully evaluated without a more rigorous framework concerning the disunification / unification of characters. Without such formal criteria, all decisions are 'ad-hoc' and different proposals may get different levels of review. Both WG 2 and the Unicode Technical Committee need to spend some time in evaluating and possibly formalizing the criteria that we use to decide these cases. This is similar to the formalization we have done for script prioritization, 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.

F.1 What is disunification?

Disunification is the introduction of a new character that 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 been used for encoding the languages that the new characters are intended for.

F.2 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.

1. 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.
2. 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 equivalencies, especially, where disunification occurs on well-established characters (as opposed to among the characters of an entirely new script being fine-tuned in the proposal stage).
3. Keyboards that support the disunification need to be widely (and by default) available; this is especially troublesome for strong disunification of Latin characters as most keyboards have a Latin layer from which it is easy to type the existing and now-disunified character.

F.3 Criteria of analysis

I. Costs

The following questions are designed to evaluate the costs associated with the disunification.

1. Is there a glyphic distinction?
2. Is there a behaviour difference?
3. Is the use of the new character restricted to a new context (for example, use with a novel script)?
4. Is the use of the existing, ambiguous character instead of the proposed new character common, prevalent or established practice?
5. Does the character exist in ASCII (ISO 646 IRV)?

II. Benefits

1. Appearance: does disunification help to allow multilingual monofont text in an environment where this is commonly needed? In what way?
2. Layout: does disunification solve common layout differences (this would mostly be true for punctuation)?

3. Searching/sorting: Is there a *common* case where disunification allows better support for these?
4. Mapping to another standard: Is there a widely used standard that disunifies the characters in question? Are the characters in question the *only* ones that prevent cross mapping?

III. Alternatives
Finally, the analysis must explore whether other alternatives are possible.

1. Can the desired effect be achieved by changes to the display layer?
2. Can the desired effect be achieved by changes to protocols?
3. Can the desired effect be achieved by processing algorithms?

IV. Previously rejected proposals
WG 2 may have rejected previous proposals for a character on the basis of it being a glyphic variant of an already coded character. Any proposal, which later suggests that one or more of these variant forms is actually a *distinct* character requiring separate encoding, should provide detailed printed evidence that there is actual, contrastive use of the variant form(s). It is insufficient for a proposal to claim a requirement to encode *as characters* in 10646, glyphic forms which happen to occur in another character encoding that did not follow *TR 15285 - Character-Glyph Model* that guides the choice of appropriate characters for encoding in 10646.

(For example, the forms in the American Library Association / Latin Cyrillic Romanization tables were considered during the development of the original Cyrillic repertoire for 10646, and the variant glyph forms were explicitly unified, so that duplicate characters would not be encoded for Cyrillic. Later, a proposal was being prepared by TC46 on the basis that some of the variant forms were in an existing ISO standard, without due consideration for the Character Glyph Model - and hence Rejected.)

F.4 Some examples of precedents

Example 1:

Character: *Generic Decimal Separator Mark*

> In 1991 the proposal was made to add a new punctuation character in the General Punctuation block that would have the semantic property of decimal separator, but could be imaged as period, comma, space or apostrophe depending on the locale.

Asserted benefit: Solve the locale dependent display of numbers.

Costs: This new character would have disunified four widely used characters. Mapping from existing character sets would have become locale dependent. Users would have to turn on a special show-invisible-character mode to distinguish the new character from existing characters. Such modes exist, but are limited to word processing software, where numbers usually occur embedded in text, which in turn is 'frozen' into a given language. Database software, where locale dependent numeric displays are much more of an issue, does not normally need or support a show-invisible-character mode. Finally, in 1991 there were no keyboards supporting this new character, but it would be needed in *all* languages and applications, and *all* software would have to be specially adapted for it.

Alternatives: There already is an established technology to deal with locale differences, and in a way that is not limited to decimal numbers.

Result: **Rejected.** The costs far outweigh the benefits.

Example 2:

Character: *Angstrom Symbol*

Asserted benefit: Provide roundtrip mapping for East Asian character sets.

Costs: This character disunifies A WITH RING, which is in wide use in only a limited number of languages that all use Latin-1. In the Latin-1 context, it would be natural to use A WITH RING as the Angstrom Symbol. The Angstrom unit is not one of the preferred powers for the metric units of SI, but it is still commonly used in some disciplines, as it is convenient for atomic length scales. Disunifying the A WITH RING adds the important round trip mapping capabilities for East Asian character sets, but makes it harder to use the Standard as a pivot between these character sets and Latin-1. However, almost none of the other SI units that have explicit character codes in East Asian character sets can be mapped 1:1 with Latin-1, so the Angstrom Symbol adds little to that problem. Searching needs to support equivalencies; however, in the East Asian context the need for extended equivalencies (beyond simple case equivalence) is common.

Alternatives: None.

Result: **Accepted.** The benefits far outweigh the costs.

Annex G: Formal criteria for coding precomposed characters

**(Sources: ISO/IEC JTC 1/SC 2/WG 2 N1725 (1998-03-17) - adopted with revisions at M34 - action item M34-7e;
ISO/IEC JTC 1/SC 2/WG 2 N2176R (2000-03-07- adopted at M38 - action item M38-5d.)**

WG2 evaluation procedure 1b on page 5 addresses precomposed characters. This annex addresses in brief the criteria that support or rule against encoding of any specific proposed character as a precomposed character instead of as a combining character sequence. It also describes the impact of normalization of multiple representations of characters arising out of combining sequences in the standard on proposals for new precomposed characters.

G.1 Criteria

The positive criteria are of the form of necessary conditions, but not in themselves sufficient to make the decision. Proposals that meet the negative criteria should use composed character sequences instead. The cost criteria are provided as a help to gauge the impact of encoding new precomposed forms.

Positive:

- Existence in another character encoding standard (for the purpose of 1:1 character conversion)
- Existence of a precomposed letter in a well-established or official alphabet.

Negative:

- If it were to introduce multiple spellings (encodings) for a script where NO multiple spellings existed previously.
- If combining character sequences can be shown to meet the stated information processing needs (e.g. archival use)
- If solely intended to overcome short-term deficiency of rendering technology.
- If the intended use of the character is solely for transliteration purposes.

Cost criteria

- Incremental cost for each additional character
- Incremental cost for each new multiple spelling
- Declining benefit if immediate and widespread use is not anticipated.
- Effect on system / products that use pre-composed form as canonical (since addition of precomposed characters makes this set of canonicals unstable).

Note: some existing and widely available implementations of internal processes (collation) may use decomposed characters even where the editing interface does not support them. For these cases, additional multiple spellings provide explicit additional costs without *any* benefit.

- Short-term solution versus permanent cost

Note: the level of support for combining characters in Latin, Greek and Cyrillic documents is not as widespread as was anticipated when the first edition of the standard was published. It may be tempting to introduce precomposed forms as a short-term solution as long as the level of support for combining characters in Latin, Greek and Cyrillic documents is not yet widespread. Key font technologies with support for combining have been developed and at the same time, an increasing number of platforms routinely know how to handle combining marks for other scripts. Adding new precomposed characters could be a permanent unwarranted cost for such newer technologies versus the short-term benefit of being able to reuse not-so-new technologies. See also the discussion in the next section.

G.2 Implications of normalization on character encoding

As the standard has become more prevalent in implementations and in other standards, it has become necessary to produce very stable specifications for the comparison of text. In particular, a unique, normalized form of text is required for comparisons in domain names, XML element names, and other areas where a precise, stable, comparison of strings is required. Programs that require uniqueness also require forward compatibility: programs all over the web *must* be able to depend on the unique format not changing over time.

There are characters that are equivalently represented either as sequences of code points or as a single code point (called a *composite character*). For example, the *i* with 2 dots in *naïve* could be presented either as *i* + *diaeresis* (0069 0308) or as the composite character i-*diaeresis* (00EF). There are other cases where the order of two combining characters does not matter. For example, the pair of combining characters *acute* and *dot-below* can occur with either one first; both alternate orders are equivalent. In response to the need for a unique form, the Unicode Consortium has produced an exact algorithmic specification of normalized forms (see *UTR #15: Unicode Normalization Forms* - http://www.unicode.org/unicode/reports/tr15).

One of these forms, Normalization Form C, is designed to favour precomposed characters such as ã over combining character sequences such as a + ~. The W3C *Character Model for the World Wide Web (*http://www.w3.org/TR/charmod*)* requires the use of Normalization Form C for XML and related standards (this document is not yet final, but this requirement is not expected to change). See also the W3C *Requirements for String Identity Matching and String Indexing (*http://www.w3.org/TR/WD-charreq) for more background. We expect that the number of standards and implementations requiring normalization will continue to grow. Such implementations must produce precisely the same result for normalization *even if* they upgrade to a new version of Unicode / 10646. Thus it is necessary to specify a fixed version for the composition process, called the *composition version.* The composition version is defined to be Version 3.0.0 of the Unicode Character Database, which corresponds to ISO/IEC 10646-1:2000.

To see what difference the composition version makes, suppose that a future version of the standard -- Unicode 4.0 / 10646: 2003 adds the composite *Q-caron*. For an implementation that uses Unicode 4.0 / 10646: 2003, strings in Normalization Forms C or KC will continue to contain the sequence *Q* + *caron,* and **not** the new character *Q-caron*, since a canonical composition for *Q-caron* was not defined in the composition version. The implications for encoding new characters are that new precomposed characters are important to recognize. If Q WITH CARON were added to a future version of Unicode or 10646, then it would represent a duplicate encoding. This could be tolerated before Unicode 3.0 because canonical equivalence could be used to equate the two forms. But due to the need for stability in comparison by so much of the world's infrastructure, this situation cannot be tolerated in the future. For stability, characters that can be currently represented as sequences will always stay represented only as sequences. These include the following examples:

Character	**Code Point Sequence**	**Comments**
ch	<0063, 0068>	Slovak, traditional Spanish
tʰ	<0074, 02B0>	Native American languages
Ẋ	<0078, 0323>	
ƛ̓	<019B, 0313>	
ą́	<00E1, 0328>	LATIN SMALL LETTER A WITH OGONEK AND TILDE
i̇́	<0069, 0307, 0301>	LATIN SMALL LETTER I WITH DOT ABOVE AND ACUTE
ㇰ゚	<30C8, 309A>	Ainu in kana transcription

Moreover, the need for separate precomposed characters is diminishing quickly. The major GUI vendors are currently in the process of upgrading their systems to handle both surrogates and accurate positioning of combining marks, with such technologies as Open Type and AAT. By the time new precomposed characters could be added, there would be little need for them. It *is* possible to add future precomposed characters in the case where they cannot already be represented by combining character sequences. In such cases the situation is reversed; the component characters that would make up an equivalent combining character sequence cannot be added.

Annex H: Criteria for encoding symbols

(Source: ISO/IEC JTC 1/SC 2/WG 2 N1982 - 1998-02-26 - adopted at M36 - action item M36-6a.)

H.1 Symbols and plain text

The primary goal of ISO/IEC 10646 and Unicode is plain text encoding. Only a very limited class of symbols are strictly needed in plain text, if it is understood that an e-mail message is representative for plain text. A more expanded interpretation of plain text acknowledges plain text as the backbone for more elaborate and rich implementations. An example of such expanded use are the plain text buffer for a rich document, or searchable representation of text or notational system, such using character codes to access unit symbols in a CAD package, or to implement a complex notational system such as musical notation.

In the latter cases, the class of symbols for which encoding makes sense becomes much larger. It encompasses all symbols for which it is not enough to merely be able to provide an image, but whose identity and semantics must be able to be automatically interpreted and processed in ways that are similar to processes on text.

H.2 The 'symbol fallacy'

The 'symbol fallacy' is to confuse the fact that '*symbols have semantic content*', with '*in text, it is customary to use the symbol directly for communication*'. These are two different concepts. An example is traffic signs and the communication of traffic engineers about traffic signs. In their (hand-) written communication the engineers are much more likely to use the words *stop sign* when referring to a stop sign, than to draw the image. On the other hand, mathematicians are more likely to draw an integral sign and its limits and integrands than to write an equation in words.

H.3 Classification

Symbols can be classified in two broad categories, depending on whether a symbol is part of a symbolic notational system or not.

H.3.1 Symbols that are part of a notational system

Symbols that are part of a notational system have uses and usage patterns analogous to the notational systems used for writing. They feature a defined[9] repertoire and established rules of processing and layout. In computers they are treated similar to a complex script, i.e. with their own layout engines (or sub engines). Core user groups have shared legacy encodings, which allow at least their data to be migrated to the new encoding.

H.3.2 Symbols that are not part of a notational system

There are many distinct repertoires of non-notational symbols, some with very small frequency of occurrence. The design and use of many of these symbols tends to be subject to quick shifts in fashion; in many cases they straddle the realms of the informative and the decorative. Layout is usually quite simple and directly equivalent to an inline graphic. In computers they are treated as uncoded entities today: they are provided as graphics or via fonts with ad-hoc encodings, with no additional support for rendering. Because of the ad-hoc nature of the legacy encodings for these symbols, data migration is near impossible.

H.3.2.1 Legacy symbols

An important subclass of non-notational symbols is the class of technical symbols found in legacy implementations and character sets for which plain text usage is established. Prominent examples are compatibility symbols used in character mode text display, e.g. terminal emulation.

H.4 Kinds of symbols found in ISO/IEC 10646 and Unicode

1) Part of a notational system
   - Mathematical operators
   - Electrotechnical symbols
   - APL
   - Braille

[9] All large repertoires can have a sizeable 'gray zone', even if they can be called 'defined' here.

- Musical notations (accepted for Plane 1)

2) Compatibility for text mode display
   - Chess pieces
   - Forms and blocks
   - Control pictures
   - Integral pieces
3) Text ornaments
   - Dingbats
   - Enclosed/parenthesized
4) Traditional signs and icons
- Astrological symbols
- Religious symbols

5) Abbreviations or units used with text or numbers
   - Currency symbols
   - Units
   - Prescription etc.
6) Other
- Environment protection related symbols

H.5 Discussion

Any proposal to encode additional symbols must be evaluated in terms of what the benefit will be of cataloguing these entities and whether there is a realistic expectation that users will be able to access them by the codes that we define. This is especially an issue for non-notational, non-compatibility symbols.

The trend so far has not been encouraging there. The last few years have seen enormous progress in the end-user available support of ISO/IEC 10646 and Unicode as encoding for letters and punctuation. Instead of a collection of fonts with legacy encodings, system and font vendors now provide fonts with a common encoding, and, where scripts have similar typography, with combined repertoire. The most widely available fonts for symbols, however, have **not** followed that trend. Users of these symbols continue to use ad-hoc fonts in their documents.

Existing data encoded using legacy encodings for letters and punctuation can be converted to ISO/IEC 10646 and Unicode quite easily, and many systems and applications provide such translations in a transparent matter. A different story holds for symbols. Because almost all legacy data use ad-hoc encodings or even in-line images for non-notational symbols, one cannot easily convert existing data. Therefore there is more resistance to changing the status quo.

As a conclusion, any successful proposal would need to contain a set of non-notational symbols for which the benefits of a shared encoding are so compelling that its existence would encourage a transition.

H.6 Some criteria that strengthen the case for encoding

The symbol

- is typically used as part of computer applications (e.g. CAD symbols)
- has well defined user community / usage
- always occurs together with text or numbers (unit, currency, estimated)
- is required to be searchable or indexable
- is customarily used in tabular lists as shorthand for characteristics[10]
  (for example, check mark, maru etc.)
- is part of a notational system
- is used in 'text-like' labels (even if applied to maps and 2D diagrams)
- has well-defined semantics
- has semantics that lend themselves to computer processing
- completes a class of symbols already in the standard
- is letter-like
  (i.e. ordinarily varies with the surrounding font style)

[10] The typical camping, boating, or hiking symbols are often used in that way.

- itself has a name, (for example, *ampersand*, *hammer-and-sickle*, *caduceus*)
- is commonly used amidst text
- is widespread, i.e. actually found used in materials of diverse types/contexts by diverse publishers, including governmental

H.7 Some criteria weaken the case for encoding

There is evidence that

- the symbol is primarily used free-standing (traffic signs)
- the notational system is not widely used on computers (dance notation, traffic signs)
- the symbol is part of a set undergoing rapid changes (short-lived symbols)
- the symbol is trademarked (unless encoding is requested by the owner) (logos, Der grüne Punkt, CE symbol, UL symbol, etc)
- the symbol is purely decorative
- the symbol is an image of something, not a symbol for something
- the symbol is only used in 2-Dimensional diagrams, (e.g. circuit components)
- the symbol is composable (see diacritics for symbols)
- the identity of the symbol is usually ignored in processing
- font shifting[11] is the preferred access and the user community is happy with that (logos, etc.)

Or, conversely, there is not enough evidence for its usage or its user community.

H.8 Completion of a set

Mathematical operators are an example for an extensive set of symbols, which at the current time are incomplete. The existing repertoire is so incomplete that not only does it not meet the needs of the current user community, but even the use of the existing partial repertoire is precluded for many users. Therefore, completion of this repertoire has a high priority. Otherwise, for lack of usability, alternative encodings or mark-up will become the method of choice, stranding the large repertoire already encoded. In the particular example, this work is now being undertaken, and finishing it should be given a very high priority.

By extension, proposal that contain incomplete repertoires of a given category of symbol should be given a very low priority until they reach a level of completeness that makes a compelling case for a given user community.

H.9 Instability

The case has been made that either *rapid changes in the glyph representation*, or *changes in the meaning of the character* have nothing to do with encoding (defined as a purely positional assignment), as long as the general category of use of the symbol does not change.

The counter example to that is the recent decision to encode the Euro-Sign as a new character and not to reclaim the Euro-Currency sign based on a definite change in glyph. There are glyph changes that cannot be absorbed quietly since the new glyph bears so little relation to the old one that the change exceeds the implied range of glyphic variation.

It is normally allowable for a symbol (same glyph) to acquire some additional meaning(s) over time. However, for some symbols (part of a notational scheme) this could mean that the symbol would need to be processed differently (i.e. a change in operational semantics a.k.a. character properties). Such a change would necessarily affect coding.

In either case, rapid change means by definition that the situation is not settled, and reliable information on the range of acceptable glyphic variation or character properties is unavailable. Therefore it is a good reason to wait with coding.

H.10 Perceived usefulness

The fact that a symbol merely *seems to be useful or potentially useful* is precisely not a reason to code it. Demonstrated usage, or demonstrated demand, on the other hand, does constitute a good reason to encode the symbol. The Euro Sign is the classical example of the latter. It is a novel symbol for which there is demonstrated and strong demand.

[11] Shifting of fonts, however, is not a reliable method for the web.

It is important to distinguish the perception of 'usefulness' from the question of whether a symbol is in widespread use or not. ISO/IEC 10646 and Unicode cater to both general and specialized users, from modern world languages to historic and minority scripts. Widespread use will influence the prioritization, but should be somewhat independent from the decision of whether a symbol is an encodable entity in the first place. In order to be truly useful, an encoded symbol must be accessible to the user community in its encoded form. It requires implementers ready to supply implementations using the new encoding, and user community ready to migrate to those implementations.

Annex I: Guideline for handling of CJK unification and/or disunification error

(Source: ISO/IEC JTC 1/SC 2/WG 2 N2576R – 2003-10-21)

There are two kinds of errors that may be encountered related to coded CJK unified ideographs.

Case 1: *to be unified* error - Ideographs that should have been unified are assigned separate code points.

Case 2: *to be disunified* error - Ideographs that should not have been unified are unified and assigned a single code point. An example of this is the request from TCA in document N2271.

When such errors are found, the following guidelines will be used by WG 2 to deal with them.

I.1 Guideline for "to be unified" errors

A. The "*to be unified"* pair will be left disunified. Once a character is assigned a code position in the standard, it will not be removed from the standard.
B. If necessary, an additional note may be added to an appropriate section in the standard.

I.2 Guideline for "to be disunified" errors

A. The ideographs to be disunified should be disunified and should be given separate code positions as soon as possible (disunification in some sense, and character name change in some sense also). These ideographs will have two separate glyphs and two separate code positions. One of these ideographs will stay at its current encoded position. The other one will have a new glyph and a new code position.
B. For the ideographs that are encoded in the BMP, the code charts in ISO/IEC 10646 are presented in multiple columns, with possibly differing glyph shapes in each column. The question of which glyph shall be used for the currently encoded ideograph will be resolved as follows. In the interest of synchronization between ISO/IEC 10646 and the Unicode standard, the ideograph with the glyph shape that is similar to the glyph that is published in the "Unicode Charts" will continue to be associated with its current code position. For the ideographs outside the BMP, the glyph shape in ISO/IEC 10646 and the Unicode Charts are identical and will be used with its current code position.
C. The disunified ideograph will have a glyph that is different from the one that retains the current code position.
D. The net result will be an addition of new ideograph character and a correction and an additional entry to the source reference table.

I.3 Discouragement of new disunification request

There is a possibility of "pure true disunification" request. This is almost like the new source code separation request. This kind of request shall not be accepted disregarding the reasoning behind. Key difference between "TO BE DISUNIFIED" and "SHALL NOT BE DISUNIFIED is as follows.

a. If character pair is non-cognate (means different character), those pair are TO BE DISUNIFIED.
b. If character pair is cognate (means the same but different shape), those pair are SHALL NOT BE DISUNIFIED.

Disunification request with reason of mis-application (over-application usually) of unification rule should NOT be accepted due to the principle in resolution M41.11.

Annex J: Guideline for correction of mapping table error

(Source: ISO/IEC JTC 1/SC 2/WG 2 N2577 – 2003-09-02)

In principle, mapping table or reference to code point of existing national/regional standard (in the source reference tables) must not be changed. But once a fatal error is found it should be corrected as early as possible, under following guidelines:

J.1 Priority of error correction procedure

A. Consider adding new code position and source-reference mapping for the character in question rather than changing the mapping table.
B. If change of mapping table is unavoidable, correction should be done as soon as possible.

J.2 Announcement of addition or correction of mapping table

Once any addition or correction of mapping table is made, an announcement of the change should be made immediately. Usually this will be in the form of a resolution of a WG 2 meeting, followed by subsequent process resulting in an appropriate amendment to the standard.

J.3 Collection and maintenance of mapping tables that are not owned by WG 2

There are many mapping tables, which are included in national/regional standards or developed by third parties. These are out of WG 2's scope. Any organization (such as Unicode Consortium) that collects mapping information, maintains it consistently and makes this information widely available is invited and encouraged to do so.

Annex K: Levels of implementation in ISO/IEC 10646

The following is a summary of the three levels of implementation in 10646 – defined in clause 14 and other clauses in the standard. The levels have to do with how multiple spellings arising out of use of combining characters are to be dealt with. A Unicode implementation is a Level 3 implementation of 10646.

> ***"Implementation level 1***
> When implementation level 1 is used, combining characters and Hangul Jamo characters are not used."

> ***"Implementation level 3***
> When implementation level 3 is used, any character from the standard can be used. The implementation level 3 shall be used for the Hangul syllable composition method (from clause 26.1)."

Most of the proposals for new scripts or characters will use one of the above two levels.

> ***"Implementation level 2***
> When implementation level 2 is used, a set of combining characters (specified in clause B.2 of the standard) cannot be used. This set includes COMBINING DIACRITICAL MARKS (0300 to 036F), COMBINING DIACRITICAL MARKS FOR SYMBOLS (20D0 to 20FF), HANGUL JAMO (1100 to 11FF) and COMBINING HALF MARKS (FE20 to FE2F). An additional set of combining characters is also enumerated in clause B.2."

The standard also defines unique spelling rules applicable for levels 1 and 2 for certain scripts used in India and some other South Asian countries. Please reference the standard for more details. If you cannot get hold of a copy of the standard, the latest working draft in document N2578 on the WG 2 site (or the later SC2 N3699) may be referenced.

Annex L: Character-naming guidelines

(The following is the informative Annex L of ISO/IEC 10646 reproduced here for the convenience of users of this Principles and Procedures document.)

Guidelines for generating and presenting unique names of characters in ISO/IEC JTC1/SC2 standards are listed in this annex for information. These guidelines are used in information technology coded character set standards such as ISO/IEC 646, ISO/IEC 6937, ISO/IEC 8859, ISO/IEC 10367 as well as in ISO/IEC 10646. These Guidelines specify rules for generating and presenting unique names of characters in those versions of the standards that are in the English language.

NOTE – In a version of such a standard in another language:

a) these rules may be amended to permit names of characters to be generated using words and syntax that are considered appropriate within that language;

b) the names of the characters from this version of the standard may be replaced by equivalent unique names constructed according to the rules amended as in a) above.

Rules 1 to 4 are implemented without exceptions, unless mentioned in the rule itself (see Rule 4). However it must be accepted that in some cases (e.g. historical or traditional usage, unforeseen special cases, and difficulties inherent to the nature of the character considered), exceptions to some of the other rules will have to be tolerated. Nonetheless, these rules are applied wherever possible.

Rule 1

By convention, only Latin capital letters A to Z, space, and hyphen are used for writing the names of characters.

NOTE – Names of characters may also include digits 0 to 9 (provided that a digit is not the first character in a word) if inclusion of the name of the corresponding digit(s) would be inappropriate. As an example the name of the character at position 201A is SINGLE LOW-9 QUOTATION MARK; the symbol for the digit 9 is included in this name to illustrate the shape of the character, and has no numerical significance.

Rule 2

The names of control functions are coupled with an acronym consisting of Latin capital letters A to Z and, where required, digits. Once the name has been specified for the first time, the acronym may be used in the remainder of the text where required for simplification and clarity of the text. Exceptionally, acronyms may be used for graphic characters where usage already exists and clarity requires it, in particular in code tables. Examples:

Name: LOCKING-SHIFT TWO RIGHT

Acronym: LS2R

Name: SOFT HYPHEN

Acronym: SHY

NOTE – In ISO/IEC 6429, also the names of the modes have been presented in the same way as control functions.

Rule 3

In some cases, the name of a character can be followed by an additional explanatory statement not part of the name. These statements are in parentheses and not in capital Latin letters except the initials of the word where required. See examples in rule 12.

The name of a character may also be followed by a single * symbol not part of the name. This indicates that additional information on the character appears in Annex P. Any * symbols are omitted from the character names listed in Annex G.

Rule 4

Names are unique if SPACE and medial HYPHEN-MINUS characters are ignored, and if the strings “LETTER”, “CHARACTER”, and “DIGIT” are ignored in comparison of the names.

Examples of unacceptable unique names:

SARATI LETTER AA

SARATI CHARACTER AA

These two names would not be unique if the strings “LETTER” and “CHARACTER” were ignored.
The following two character names are exceptions to this rule, since there were created before this rule was specified.

116C HANGUL JUNGSEONG OE
1180 HANGUL JUNGSEONG O-E

Rule 5

The name of a character wherever possible denotes its customary meaning, for example PLUS SIGN.
Where this is not possible, names describe shapes, not usage; for example: UPWARDS ARROW.
The name of a character is not intended to identify its properties or attributes, or to provide information on its linguistic characteristics, except as defined in Rule 6 below.

Rule 6

Only one name is given to each character.

Rule 7

The names are constructed from an appropriate set of the applicable terms of the following grid and ordered in the sequence of this grid. Exceptions are specified in Rule 11. The words WITH and AND may be included for additional clarity when needed.

1	Script	5	Attribute	
2	Case	6	Designation	
3	Type	7	Mark(s)	
4	Language	8	Qualifier	

Examples of such terms:

Script	Latin, Cyrillic, Arabic
Case	capital, small
Type	letter, ligature, digit
Language	Ukrainian
Attribute	final, sharp, subscript, vulgar
Designation	customary name, name of letter
Mark	acute, ogonek, ring above, diaeresis
Qualifier	sign, symbol

Examples of names:

LATIN CAPITAL LETTER A WITH ACUTE
1 2 3 6 7

DIGIT FIVE
3 6

LEFT CURLY BRACKET
5 5 6

NOTE 1 – A ligature is a graphic symbol in which two or more other graphic symbols are imaged as a single graphic symbol.

NOTE 2 – Where a character comprises a base letter with multiple marks, the sequence of those in the name is the order in which the marks are positioned relative to the base letter, starting with the marks above the letters taken in upwards sequence, and followed by the marks below the letters taken in downwards sequence.

Rule 8

The letters of the Latin script are represented within their name by their basic graphic symbols (A, B, C, etc.).
The letters of all other scripts are represented by their transcription in the language of the first published International Standard.
Examples:

K LATIN CAPITAL LETTER K
Ю CYRILLIC CAPITAL LETTER YU

Rule 9

In principle when a character of a given script is used in more than one language, no language name is specified. Exceptions are tolerated where an ambiguity would otherwise result.
Examples:

И	CYRILLIC CAPITAL LETTER I	
І	CYRILLIC CAPITAL LETTER	BYELORUSSIAN-UKRAINIAN I

Rule 10

Letters that are elements of more than one script are considered different even if their shape is the same; they have different names.
Examples:

A	LATIN CAPITAL LETTER A
Α	GREEK CAPITAL LETTER ALPHA
А	CYRILLIC CAPITAL LETTER A

Rule 11

A character of one script used in isolation in another script, for example as a graphic symbol in relation with physical units of dimension, is considered as a character different from the character of its native script.
Example:

µ	MICRO SIGN

Rule 12

A number of characters have a traditional name consisting of one or two words. It is not intended to change this usage.
Examples:

'	APOSTROPHE
:	COLON
@	COMMERCIAL AT
_	LOW LINE
~	TILDE

Rule 13

In some cases, characters of a given script, often punctuation marks, are used in another script for a different usage. In these cases the customary name reflecting the most general use is given to the character. The customary name may be followed in the list of characters of a particular standard by the name in parentheses which this character has in the script specified by this particular standard.
Example:

□	UNDERTIE (Enotikon)

Rule 14

The above rules do not apply to ideographic characters. These characters are identified by alpha-numeric identifiers specified for each ideographic character (see clause 28.2).

History of changes

This document was originally prepared by Messrs. Mark Davis, Edwin Hart and Sten G. Lindberg, as document N946 (1994-10-11), based on N884 (1993-04-06) (authored by Messrs. Rick McGowan and Joe Becker). It has been enhanced by an ad hoc group on principles and procedures set up at the San Francisco WG 2 meeting no. 26. The result was presented as WG 2 document N1116 (1994-10-12). The following is a summary of changes made since that time:

1. At the Geneva WG 2 meeting no 27 (1995-04-07), where some enhancements were proposed. The result was presented as document N1202 (1995-06-26)).
2. At the Helsinki WG 2 meeting no 28 (1995-06-26), some enhancements were proposed and adopted. The result was presented as document N1252 (1995-06-27). The document was accepted, following Resolution M28.6 at that meeting.
3. At the meeting no 31 (1996-08-16) a new Annex C: Description of the UCS work flow and stages in progression from initial proposal to final publication was added. Furthermore a new question (C 10) regarding some properties of proposed characters has been included in the proposal summary form.
4. At the meeting no 32 (1997-01-24) a new Annex D: BMP and Supplementary Planes Allocation Roadmap was added. The Annex D is the inclusion of the US contribution N1499 (1996-12-27) only with minor editorial changes. Minor editorial changes have been made to align the different standing documents.
5. Principles regarding allocation of '00' position in a block (resolution M33.12) and regarding considerations for half-block boundary (per resolution M33.11) have been added from meeting M33 (1997-07-04).
6. The ad hoc report on collection identifiers for parts 1 and 2 (document N1726 - 1998-03-19) from meeting 34 (1998-03-20), and a form for submission of requests for collection identifiers (document N1735 - 1998-03-23, amended per AI-35-6-b) were consolidated into document N1877 - 1998-09-20; and has been incorporated in this document.
7. Formal Criteria for Disunification (per AI-34-7-d, based on document N1724 - 1998-03-05) was added.
8. Formal Criteria for Coding Pre-Composed Characters (per AI-34-7-e, based on document N1725 - 1998-03-17) was added.
9. The principle of '1K boundary for allocations in Plane 1 for ease of use with UTF-16' (per Action Item AI-35-6-a - 1998-09-25) has been added.
10. The unused 'WG 2 administration section D' has been removed from the proposal summary form (at meeting 36 - 1999-03-15).
11. A note has been added on the need for stronger justification for proposals to include 'Glyph Variants'.
12. A sample picture of the 'spread sheet' illustrating the skeleton format and column headings used in the parallel WG 2 standing document 'Status summary of WG 2 work items' has been removed, with the reference to that standing document.
13. The document has been reorganized slightly for better readability. This is presented as document N2002 at M36 (1999-03-15) (the revised Annex D is left as 'to do' pending acceptance of other roadmap contributions).
14. A new Annex on criteria for encoding symbols based on document N1982 (1998-02-26) has been added, per action item M36-6a (1999-03-15).
15. Annex on Pre-Composed characters has been enhanced with information on implications of Unicode normalization - based on document N2176R (2000-03-07), per action items M37-6a and M38-5d.
16. Information on use of UCS Sequence Identifier, based on document N2230 (2000-07-21) has been incorporated, per action item M39-5a.
17. Annex D has been updated to reference WG 2 standing documents containing the Roadmaps (documents N2316 - 2001-01-10, N2314 - 2001-01-10, N2215 - 2000-03-30, and N2216 - 2000-03-30) - details have been moved and updated from this document.
18. References to different clauses in 10646-1 in the document and in the Proposal Summary Form have been updated to the renumbered clauses and Annexes of 10646-1:2000.
19. References to relevant clauses and Annexes of 10646-2: 2001 have been added.
20. Refinements based on discussion at meeting M40 - 2001-04-02/05:

a) Section 3 on Character names was expanded.
b) Added a note about open collection identifiers when there is need to expand the ranges or add new ranges.
c) Section 9 on Relative Ordering of Characters was added with references to ISO/IEC 14651 and Unicode Collation Algorithm.
d) Under section B - General section of the proposal summary form, a new item 9 was added inviting more information regarding properties of the character(s) or script along with a condensed statement in section A.1.
e) Under technical justification section of the proposal summary form, a new question 9 was added along with a similar statement under A.1, renumbering questions 9.10, and 11 to 10, 11 and 12 respectively; new question 13 was added.
f) Added a new section in Annex D, explaining the use of reserved positions in the gaps in a range of assigned code positions.

g) Removed WG 2 administrative portion from Annex E on collection identifier submissions.
h) Numbers for sub items under item 1 of WG 2 Evaluation Procedure were corrected and reordered.
i) Footnote for bullet 3 under H.7 was replaced with a parenthetical phrase.
j) New footnote was added for last bullet on font shifting under H.7.
k) Deleted the note about allowing use of USIs in a collection submission
l) 96x96 bit-mapped format has been removed as one of the acceptable formats for printing the standard or its amendments - in section A.1, item 5 and in the submission form Section B, item 6.

21. The first HTML version of this document has been created in July 2001. The broken links have been repaired since then.
22. The following changes are made in this version dated October 2003:
    a) The HTML version of this document is discontinued. Only .doc and .pdf versions are generated.
    b) Changed all references to 10646-1 and 10646-2 to consolidated 10646 single part edition.
    c) Item 3c is added to section 2.4.
    d) Pointers to the roadmap annex from section 2 are removed.
    e) Resolution M41.11 – Policy regarding acceptable changes to 10646 - is reproduced in Section 3.
    f) Text referring to resolution M34.18 on documentation of collections spanning 10646-1 and 10646-2 has been deleted, in view of the consolidated edition of 10646: 2003.
    g) Section 5.1 on 'Checking the status of a proposal' is added.
    h) Section 10 on 'Referencing ISO/IEC 10646' is added.
    i) Annex I on 'Guideline for handling of CJK unification and/or disunification error' is added.
    j) Annex J on 'Guideline for correction of mapping table error is added.
    k) Annex K on 'Levels of implementation in ISO/IEC 10646', giving a brief summary of the levels 1, 2 and 3 is added.
    l) Annex L on 'Character-naming guidelines' (reproduced from the standard) is added.
    m) Pointer to "where is my character" on the Unicode web site is added in section 1 and in Annex A.
    n) Additional guideline paragraphs referencing TR15285 – Character Glyph Model, how to check the status of a proposal, and optional email ids of submitters and experts who were consulted, added in Annex A.
    o) Added extracted clauses 4.12 and 4.14 into section A.1 for reference from proposal summary form.
    p) Expanded item 3 in section B of summary form to a check list.
    q) Changed references from the standard to extracted annexes in P&P document for items 4 and 5 in section B. Updated reference to UCD.html in item 9.
    r) Updated links in UCS work flow in Annex C.
    s) Minor edits to section C items 6, 10 and 11 of proposal summary form.
    t) Updated references list, removing entries that are no longer relevant and fixing changed hyper links.
23. The following changes are made in this version dated January 2005:
    a) Added a sentence at the end of section 4.1 noting that the USI definition includes unassigned code positions.
    b) Reworded section 9 on Ordering of Characters to read better.
    c) Reworded section 10 on referencing of ISO/IEC 10646.
    d) Added new section on WG2 web site reflecting resolution on stable URL.
    e) Reworded item 5 in Annex A.1 to reflect updated font requirements, and reflecting resolution on font policy. Also updated the footnote in table in section C.2 reflecting the need for quality fonts to create amendment text.
    f) Expanded item 8 in Annex A.1 to reflect details regarding mapping information on CJK Compatibility Ideographs.
    g) Added new item 9 requiring BiDi property information, renumbering current item 9 to 10.
    h) Item c in B.1, changed … name change causes … ' to … 'name causes …'.
    i) Removed reference in section C.3, to spread sheet on 'Status of Summary of WG2 work items', which has not been kept up to date.
    j) Updated links in UCS work flow in Annex C.
    k) Added a footnote regarding plane 3 in table showing allocation of planes in section D.1.
    l) Added item i' regarding stability of collections such as IICORE in Annex E.
    m) Added a sentence to the first para in Annex G linking the annex to evaluation procedure 1b in the main body.
    n) Update Annex L, rule 4, to reflect changes from Amendment 1 to ISO/IEC 10646:2003.
    o) Made a number of minor editorial changes and updated links in different sections of the document.

The ad hoc group on principles and procedures had different members over time. The current members of the ad hoc group are:

Messrs. V.S. Umamaheswaran (Current editor of this document); Mike Ksar; and Ken Whistler.

References

Document numbers in the first column in the following table refer to WG 2 working documents (ISO/IEC JTC 1/SC 2/WG 2/ Nxxxx), except where noted otherwise. For those documents for which a link is not given, you may try http://www.dkuug.dk/JTC1/SC2/WG2/docs; some of the older documents are available only in paper form (contact the convener of JTC1/SC 2/WG 2 – Mr. Mike Ksar). Note that some of the documents may require a user id and password to access them.

Doc. No.	Title	Author(s)	Date
N884	Concerning Future Allocations	Joe Becker/Rick McGowan, Unicode Inc.	1993-04-6
N946	Proposed principles and procedures for allocation of new characters and scripts	Davis /Hart /Lindberg	1993-11-03
N947	A proposed initial list of character allocations	Davis /Hart /Lindberg	1993-11-03
N995	10646-1 Proposed Draft Amendment 3 (section 9-a-i.3)	Mark Davis WG 2 Project Editor	1994-03-03
N1002	Comments on N 947 - Proposed categorization and allocation of characters	Japan (TKS)	1994-03-28
N1061	IRG Comments to WG 2 N 946 (Proposed Principles and Procedures for Allocation of New Character and Scripts)	IRG	1994-09-14
N1137	Handling of Defect Reports on Character Names	Ad hoc group on Principles and Procedures - Messrs. V.S. Umamaheswaran, Sven Thygesen, Peter Edberg	1995-01-27
N1218	Comments on Character Addition Proposal Summary Form (N 1116)	Japan - TKS	1995-05-03
N1464	Guidance and Assistance in the Prioritization of the Allocation of Code Positions in ISO/IEC 10646	Sven Thygesen	1996-10-02
N1502	Update of N 1402 - Principles & Procedures of WG 2; N1502.xls and .doc	Sven Thygesen	1997-01-24
N1724	Formal criteria on disunification	US/Unicode - Asmus Freytag	1998-03-05
N1725	Formal criteria for coding precomposed characters	Expert contribution - Asmus Freytag, Ken Whistler	1998-03-17
N1726	Report of Ad Hoc on Collection Identifiers for Parts 1 and 2	Ad Hoc on Collection ID at M34	1998-03-18
N1735	Request for Collection Identifier in ISO/IEC 10646	Ksar / Uma	1998-03-21
N1791	Repertoire additions for 10646-1 - Cumulative List 7	Paterson	1998-06-08
N1876	Proposed replacement text for Annex D of N1502, Principles and Procedures document	Uma + ad hoc	1998-09-20
N1877	New Annex in Principles and Procedures document N1502 - Request for Collection Identifiers	Uma	1998-09-20
N1982	Towards criteria for encoding symbols http://www.dkuug.dk/JTC1/SC2/WG2/docs/n1982.doc	Unicode Consortium/US Member Body (Asmus Freytag)	1997-02-27
N2176R	Implications of Normalization on Character Encoding http://www.dkuug.dk/JTC1/SC2/WG2/docs/n2176.pdf	Unicode Technical Committee	2000-03-06
N2230	Proposal for Unique Sequence Identifiers (USI-s) and repertoire specifications including these USI-s http://www.dkuug.dk/jtc1/sc2/wg2/docs/n2230.rtf	US national body (Author: V.S. Umamaheswaran)	2000-07-21
N2271	Propose to amend two source code changes in BMP CJK Unified Ideographs block http://www.dkuug.dk/jtc1/sc2/wg2/docs/n2271.pdf	Tseng, Shih-shyeng, TCA	2000-09-15
N2576R	Annex I for N2352R(Guideline for Handling of CJK Unification and/or Disunification Error) http://www.dkuug.dk/jtc1/sc2/wg2/docs/n2576R.doc	T. L. Kobayashi , T.K. Sato, V.S. Umamaheswaran	2003-10-21
N2577	Annex J for N2352R (Guideline for correction of mapping table error) http://www.dkuug.dk/jtc1/sc2/wg2/docs/n2577.pdf	T. L. Kobayashi , T.K. Sato, V.S. Umamaheswaran	2003-09-02
WG 2 meeting minutes and resolutions	http://www.dkuug.dk/jtc1/sc2/wg2/docs/meetings.html Minutes and resolutions of past meetings of WG 2 are linked from the above page.	Mike Ksar, Convener	

Doc. No.	Title	Author(s)	Date
ISO/IEC 10646: 2003 (SC2 N3699)	Consolidated Text for ISO/IEC 10646: 2003 1st edition http://www.dkuug.dk/jtc1/sc2/def/02n3699c.htm (earlier WG 2 working draft is at http://www.dkuug.dk/jtc1/sc2/wg2/docs/n2578.pdf)	Michel Suignard, Project Editor	2003-11-06
TR152825	An Operational Model for Characters and Glyphs - http://www.iso.ch/iso/en/ittf/PubliclyAvailableStandards/c027163_ISO_IEC_TR_15285_1998(E).zip		1998
ISO/IEC TR 15285	Character Glyph Model http://www.iso.org/iso/en/ittf/PubliclyAvailableStandards/c027163_ISO_IEC_TR_15285_1998(E).zip	ISO Publicly Available Specifications	1998
ISO/IEC 14651	International string ordering and comparison – Method for comparing character strings and description of the common template tailorable ordering - http://wwwold.dkuug.dk/jtc1/sc22/wg20/docs/n731-fdis14651.pdf and http://wwwold.dkuug.dk/jtc1/sc22/wg20/docs/n991-14651-Amd-1.PDF	JTC 1/SC 22/WG 20	
UTR-9	Bidirectional Algorithm UAX10 - http://www.unicode.org/reports/tr9/		
UTR-10	Unicode Collation Algorithm - UTS#10 – http://www.unicode.org/unicode/reports/tr10		
UTR-15	Unicode Technical Report #15 – http://ww.unicode.org/unicode/reports/tr15		
Unicode Versions	Versions of the Unicode Standard: http://www.unicode.org/unicode/standard/versions/		
Unicode Database	Unicode Character Database http://www.unicode.org/Public/UNIDATA/UCD.html		
Unicode Pipeline	Proposed Unicode Characters http://www.unicode.org/unicode/alloc/Pipeline.html		
Roadmaps	http://www.unicode.org/roadmaps/		
w3c character model	Character Model for the World Wide Web: http://www.w3.org/TR/charmod	W3C i18N WG	
W3c-charreq	W3C Requirements for String Identity Matching and String Indexing - http://www.w3.org/TR/WD-charreq	W3C i18N WG	
Open Type	http://www.microsoft.com/typography/tt/tt.htm		
Apple Type Services - AAT	http://developer.apple.com/documentation/Carbon/Reference/ATSUI_Reference/		