-- DRAFT --

Document: 10646M/91-01
Date: 29 May, 1991

Subject: Summary of Results of Informal Meeting to Discuss Merging of DIS 10646 and Unicode into One Code

From: Edwin Hart, Moderator 10646M (Merger) *Ad Hoc* Group

Reply to: Edwin Hart
Johns Hopkins University
Applied Physics Laboratory
11100 Johns Hopkins Road
Laurel, MD 20723-6099
Electronic Mail: HART@APLVM.BITNET or
HART@APLVM.JHUAPL.EDU
Voice: +1 (301) 953-6926
Facsimile: +1 (301) 953-1093

This document represents the first draft of what we hope will become a proposal to merge DIS 10646 and Unicode into one code. The primary advantage of this proposal is that it is built on consensus of people supporting ISO 10646 *and* others supporting Unicode. We plan to submit a final consensus document to WG2 for consideration at the WG2 editing meeting planned for August, 1991 in Geneva, Switzerland. At that time, we plan to work within WG2 to refine the 10646 standard.

Summary

We affirm our strong support of the effort by ISO/IEC JTC1/SC2/WG2 to develop 10646. We believe that ISO with its open and responsive procedures will give careful consideration to our proposal to refine the DIS 10646. Moreover, we believe that the Unicode Consortium has provided valuable insight and technical solutions to newer requirements. We also believe that having a single international standard that incorporates the best features of DIS 10646 and Unicode as outlined in this proposal is far superior to having two incompatible standards with exactly the same goal.

Therefore, after the completion of the May, 1991 ISO-IEC JTC1/SC2/WG2 meeting in San Francisco, California in the USA, the delegates attended an informal meeting. At the meeting, we discussed requirements to merge the ISO-IEC DIS 10646 and Unicode codes. The people attending the informal meeting included some who favored the ISO 10646 code and others who favored Unicode. We believed that achieving consensus among these people would lead to a merger proposal more likely to be supported by ISO-IEC JTC1/SC2 *and* the Unicode Consortium.

In view of the diverse views represented at the meeting, the results are surprisingly positive. We succeeded in reaching a consensus of those present on major design issues that had previously separated the DIS 10646 and Unicode codes and made them incompatible. We believe that this proposal paves the way for a merger of the best features of DIS 10646 and Unicode into one multi-octet code standard. However, this is merely a first step; further work and consensus are required to produce a final proposal. In summary, although ISO and the Unicode Consortium have not yet endorsed this proposal, it is promising because it was the result of a consensus of a large number of people who represented both the ISO 10646 and Unicode Consortium efforts.

However, our work would have been almost impossible had it not been preceded by the excellent proposals submitted to WG2 by ECMA, Canada and China. To form our consensus, we used these proposals and new information on Chinese, Japanese and Korean Joint Research Group (CJK-JRG) announced at the WG2 meeting in San Francisco.

We believe this new proposal is very promising and those attending agreed to work to build support for it within their respective companies, and national and industry standard bodies, including ECMA and the Unicode Consortium.

General Objectives

We adopted the following objectives for the group:

1. Create a proposal to merge the best features of DIS 10646 and Unicode such that the proposal is acceptable to both ISO and the Unicode Consortium.

2. Increase cooperation between ISO-IEC JTC1/SC2 and the Unicode Consortium.

3. Define action items and the timing to complete them.

Participants

With the exception of Mr. Jenkins, the following people participated in the Wednesday afternoon discussions:

Jerry Andersen	IBM, USA
Lloyd Anderson	Ecological Linguistics, USA
Joseph Becker	Xerox, USA
F. Avery Bishop	Digital, USA
Willy Bohn	University of Hanover, Germany
Mark Davis	Apple, USA
Asmus Freytag	Microsoft, USA
Joachim Friemelt	Siemens, Germany
Edwin Hart	SHARE Inc./Johns Hopkins University, USA
Masami Hasegawa	Digital Japan
Huang, Weimin	CESI, China
Olle Järnefors	Royal Institute of Technology, Sweden
John Jenkins	Apple, USA
Bo Jensen	IBM Denmark

Mike Ksar	HP, USA
Takayuki Sato	HP Japan
Isai Scheinberg	IBM Canada
Karen Smith-Yoshimura	The Research Libraries Group, USA
Michel Suignard	Microsoft, France
J. G. Van Stee	IBM, USA
Kenneth Whistler	Metaphor, USA
Zhang, Zhoucai	CCID, China

On Thursday, Mr. Jenkins joined the group but Mr. Stee and Mr. Whistler were absent. In addition, Mr. Jenkins left prior to voting, and Mr. Hasegawa, Mr. Ksar, and Mr. Bohn were unable to stay for all the votes.

On Friday, with the exception of Mr. Friemelt (who had to leave before we concluded the meeting), the following participated in the voting: Mr. Anderson, Mr. Bishop, Mr. Bohn, Mr. Freytag, Mr. Friemelt, Mr. Hart, Mr. Hasegawa, Mr. Jenkins, Mr. Sato, Mr. Scheinberg, and Mr. Suignard.

Advantages of Having Only One Multi-Octet Code Standard

Here is a list of advantages to having one global multi-octet code standard:

1. Why should we be concerned about two standards?

! Inevitable requirement to support both
" 10646 because it is an international standard
" Unicode for compatibility with Unicode-based products

! Cost of supporting both
" The cost to do both is probably **very large**
" Must consider the costs to convert between the two

! Erosion of Asingle code standard@ mind-set
" If two, why not three? four? ten?

! Diminishes advantages of either alone without the other
" Single code standard solves many problems that would not be solved if we have two or more of them
" May introduce the requirement to switch between the two

2. The importance of *de-jure* standards

! Increasingly used as procurement requirements
" Gives customer more options for interconnection of products from different vendors

! Integral part of vast, interlocking family of standards, each assuming the others

! Better acceptance, because every country can participate
  " Not perceived as dominated by U.S.

3. Problems of code conversion

! Must identify both the source and the target code, but often no way to do this

! Conversion is application/subsystem dependent, and it often cannot be confined to one place (that is, it is much more expensive)

! Solving same problem in several places introduces probability of getting some solutions out of synchronization with others

! An uncontrollable, moving target (that is, you never own more than one of the two codes, you cannot control repertoires, etc.)

! Complicated by repertoire differences

! No Aright@ way to manage the differences
  ". Mismatch can range from minor irritation to catastrophe

! Further complicated by differences in character semantics
  " No tested solution is known
  " At best, makes translation even more difficult

4. The Costs of code conversion

! Monetary cost of developing, testing, maintaining, etc.
! Diversion of human and other resources by developers
! Performance and memory penalties (extra overhead)
! Errors and other problems are inevitable
! Customer dissatisfaction
! Customer conversion requirements will divert resources for creating local solutions
! Forces trade-offs between satisfying installed base and meeting new market requirements

5. Other advantages

! One reference source for the code

Areas of Consensus

1. Remove the AC0@and AC1@ restrictions.

We support the ECMA proposal, point 1, ATo remove the restriction on the so-called C1 space.@ This point is also included in the Canadian proposal, and other national body positions on DIS 10646 including the ones from China and the US.
Vote Thursday: 17 for/ 0 against/ 2 abstain (Davis, Freytag)

In addition, pending a careful review by computer communication, systems, and applications experts, from ISO, ECMA, CCITT, and within our enterprises, we believe it desirable to allow encoding graphic characters in the AC0@ space presently reserved in DIS 10646. This refines point 2 from the Canadian proposal. Annex ____ provides more details on this particular refinement (the ABohn@ refinement, named for Willy Bohn, who proposed it) of the ECMA proposal.
Vote Thursday: 16 for/ 0 against/ 3 abstain (Bishop, Hasegawa, Sato)

Removing the AC0@ restriction in addition to removing the AC1A restriction will provide for flexibility by allowing the encoding of more characters in the base multilingual plane that is the most important 2-octet plane for interchange and interworking. A consequence of removing the AC0@ restriction is that 10646 must change the way 1-octet control characters are encoded by placing the 1-octet control character into the least significant octet of the current compaction method and padding the most significant octets to the width of the current compaction method. In addition, the 1-octet compaction method must be adjusted to ensure that the control characters are correctly handled.

2. Create an International Repertoire of Unified Chinese, Japanese, and Korean Ideographs and Encode This Set of Ideographs into the Base Multilingual Plane.

   We propose a refinement to point 5 of the Canadian proposal. We believe that coding an international repertoire of unified Chinese, Japanese, and Korean ideographs in the base multilingual plane is mandatory for international interworking and processing efficiency. The encoding of the international C/J/K repertoire must be completed by the end of 1991. We propose to use the results of CJK-JRG if it is available in 1991; otherwise we propose to use the best information available at that time.
   Vote Thursday: 17 for/ 0 against/ 1 abstain (Ksar), 1 absent (Hasegawa)

   Recent statements by the Japanese delegates to WG2 indicate their strong support for the CJK-JRG. From this information, the group concluded that the unification of Chinese, Japanese, and Korean ideographs so highly desired by the international community is feasible. Providing that WG2 continues to recognize the stated Japanese requirement to encode its characters in its own 10646 plane, Japan recognized the need for an international repertoire of Chinese, Japanese, and Korean ideographs. A meeting of the CJK-JRG has been called (Tokyo, July, 1991) to start creating an international repertoire and ordering.

3. Allow the Option to Use Non-Spacing Marks.

   Pending careful review by ISO TC46 and CCITT, we propose to refine point iv) 2) of the ECMA proposal for floating diacritical marks as follows: The third Code Extension Level should specify:

a. In addition to diacritics, non-spacing marks should include stress marks, tone marks, and those used for text processing operations such as underlining or mathematical notation for the name of a vector.

b. Non-spacing marks should follow the base character for consistency.

c. Imaging and the order of multiple non-spacing diacritics should follow well-defined rules. (See Annex ____.)

d. To allow for compliance with future versions of 10646 which may encode additional pre-composed characters, allow both encoding a character as a pre-composed character or as a base character with one or more non-spacing marks. (That is, delete the ECMA statement Aif the accented letter is already coded as a single character, the alternative representation by means of floating diacrical marks is not allowed.@) This assumes that future revisions of 10646 will take certain characters that used floating marks in the current version of 10646 and encode them as pre-composed characters.

e. All sequences of codes should be allowed because of the difficulty of enforcing a legislation against certain sequences of code positions.

Vote Thursday: 16 for/ 0 against/ 1 abstain (Sato)/ absent (Bohn, Hasegawa, Ksar)

4. Define the merger (10646M) of DIS 10646 and Unicode as a 4-octet code.
Vote Thursday: 16 for/ 0 against/ 0 abstain/ absent (Hasegawa, Ksar, Bohn)

We support the 4-octet definition of the merger of DIS 10646 and Unicode. Using 4-octets provides the flexibility needed to expand the code repertoire to meet all foreseeable future requirements.

5. Location of Space for Presentation Forms

We would support a drastic reduction or elimination of the presentation forms in the base multilingual plane while retaining codes necessary to transcode existing standards in plain text. People were concerned that DIS 10646 reserved too much unused code space in the base multilingual plane. A final determination of the presentation codes will be made in consultation with Arabic and other experts.
Vote Thursday: 15 for/ 0 against/ 1 abstain (Becker)

6. Combine the Repertoires of DIS 10646 and Unicode into the Merged Code.

We propose that the repertoire of the base multilingual plane of the merged code, 10646M, be derived from a superset composed of the union of the repertoires of DIS 10646 and Unicode; for example, the superset should include pre-composed Latin, Greek, Hangul, Vietnamese, and additional symbols.
Vote Friday: 10 for/ 0 against/ 0 abstain

7. Simplify the Compaction Methods.

We are concerned about the complexity of the DIS 10646 compaction forms. For simplicity, we propose that there be several parts to the standard:

Part 1: General introduction, terminology, etc.

Part 2: The base multilingual plane (BMP). This part of the standard will specify the 2-octet implementation of the BMP. Other parts are not required for conforming implementations of the BMP. This part may be implemented without announcers.

Part 3: The full four-octet canonical form.

Part 4: Mechanisms for other compaction methods to be determined.

In the absence of other introducers for 10646 data, Part 2 should be assumed.

Vote Friday: 10 for/ 0 against/ 0 abstain

8. Make Annex H Part of the 10646 Conformance statement.

We recommend moving Annex H of DIS 10646 into the main body of the standard and making it a requirement for conformance.
Vote Friday: 9 for/ 0 against/ 0 abstain/ 1 absent (Bohn)

Due to time limitations we were unable to discuss and make recommendations to resolve the following differences between DIS 10646 and Unicode.

9. Coding of Semantics versus Shape.

For example, parenthesis, brackets and braces are coded as open/close in Unicode, and as left/right in DIS 10646.

10. Using Any Multi-Octet Coded-Character-Set Will Require Program Changes.

The following two examples show that neither DIS 10646 nor Unicode may be blindly used with the C programming language.

a. C Language Wide-Character (*wchar_t*) Model

Padding ISO 8859/1 characters with the decimal 032 value precludes the direct use (without conversion) of 10646 compaction forms 2-4 as the *wchar_t* data type in the C programming language. This is point 3 in the Canadian position statement.

b. NULL Characters in the C Language

Unicode may use 000 as the first or second octet of the 2-octet code. The C language uses the NULL (000) octet as a character string terminator for 1-octet character data. Therefore, C programs must be rewritten to use Unicode.

11. Other Issues

The above list of differences between Unicode and DIS 10646 is not exhaustive. Other issues seen by this group as being of lesser priority also need to be considered.

Action Items to Promote the Agreement

1. Participants will lobby for this proposal with their country and company constituencies. [All, immediately]

2. Ask the Unicode Consortium member companies to place a discussion of this document on the agenda of the next Unicode Consortium meeting on June 7. The Unicode Consortium should formally state that it agrees or disagrees with the general direction and state any of its concerns with specific points. [Whistler]

3. Form a joint editing committee to help draft the final 10646 merged standard. [Freytag provides updated code tables, Hasegawa provides undated structure and text, 15 Aug. list the areas of the DIS 10646 document that would require changes]

4. To achieve closer cooperation between ISO and the Unicode Consortium, we encourage the Unicode Consortium to pursue becoming a liaison member of JTC1/SC2, and for JTC1/SC2 to accept the Consortium as a liaison member. [Unicode Consortium, Sept., 1991]

5. Send this report to the national bodies and ask them to consider our consensus agreement in their votes on ISO-IEC DIS 10646. [Hart, 29 May]

6. Provide a list of the advantages of having one multi-octet code rather than two. [Andersen, done]

7. (Point 1) Coordinate an investigation of the impact of coding in C0. [Scheinberg, 15 Aug.]

8. (Point 2) Using formal minutes and other information, summarize the Tokyo CJK-JRG meeting. [Collins, 31 July]

9. (Point 3) Provide the Annex describing the rules to be used with multiple non-spacing marks. [Whistler, 9 June]

10. (Point 3) Coordinate review by ISO TC46 and CCITT of proposed use of non-spacing marks. [Smith-Yoshimura (TC46) and Friemelt (CCITT), Aug. 15]

11. (Point 5) Coordinate a review of the need to reserve so large an area for presentation forms for Arabic and other scripts on the base multilingual plane. [Ksar and Friemelt, 15 Aug.].

12. (Point 6) Investigate need for composed characters from Cyrillic and Polytonic Greek.

[Whistler, 15 Aug.]

13. (Point 7) Coordinate an investigation of which compaction methods to propose in APart 4@. [Järnefors, 15 Aug.]

14. Create 10646M electronic distribution list. Send electronic mail message to Hart to subscribe. [Hart, done]

[END OF DOCUMENT]