L2/98-281R

Unconfirmed Minutes – UTC #77 & NCITS Subgroup L2 # 174 JOINT MEETING
Redmond, WA -- July 29-31, 1998

Incorporating additions and corrections as described in Attachment 2

Chair Aliprand convened the joint meeting of the UTC and L2 (L2 Ad Hoc) at 9:45 a.m., Wednesday, July 29, 1998.

Administrative Items

Call for Proxies

The Chair announced that she had received proxy authorizations from Compaq Computer Corporation (formerly Digital Equipment Corporation) and Oracle Corporation appointing her to vote for them. No other proxy authorizations were presented.

UTC Membership Roll Call -- See Attachment 1 for list of Attendees

PRESENT: Apple Computer, Inc.; Hewlett-Packard Company; IBM Corporation; Justsystem Corporation; Microsoft Corporation; NCR Corporation; The Research Libraries Group, Inc.; Sun Microsystems, Inc.; Unisys Corporation
BY PROXY: Compaq Computer Corporation (formerly Digital Equipment Corporation); Oracle Corporation
(Total members represented: 11)
Quorum = 10

NOT PRESENT (at time of roll-call): Booz, Allen, Hamilton, Inc.; Mathema Software, GmbH; Novell, Inc.; Reuters, Ltd.; SAP AG; Silicon Graphics, Inc.; Sybase, Inc.; Xerox Corporation.
(Total not represented: 8)
Note: Representatives of Novell, Inc. and Sybase, Inc. arrived later.

Approval of the Minutes of the previous joint meeting and review of Action Items was deferred.

Invited Guest from the Orient Foundation

The Chair introduced Judith Lundberg, CEO and Founding Trustee of The Orient Foundation. Ms. Lundberg expressed thanks to Aliprand, Winkler, Sargent, and Ksar for arranging for her to attend the meeting. The Orient Foundation is an international non-government institution working primarily in Asia with libraries, museums, and archives. A major project is to provide Web based support for vendor neutral new media support. Work has been grass roots in nature.

Its chief sources of funding are the European Union and the Ford Foundation, and it has an array of institutional partners. The Bhutanese Government has appointed The Orient Foundation to represent it on standards issues, and the Foundation is working on Tibetan sets. It has set up a digitization center in Central India. It has set up focus groups throughout Asia and it has a strong working relationship with Harvard University.

A senior software engineer from the Orient Foundation will be present at the London meeting of the WG2, along with other Orient Foundation programmers. The Foundation is interested in a partnership with respect to the Tibetan encoding.

10:15 am Lloyd Honomichl (Novell, Inc.) arrived: 12 members represented.

UTC Administrative Procedures (I.G)

The Chair distributed the draft of revised UTC procedures to member representatives. Freytag suggested that UTC consider redefining the quorum as a specific number of members rather than 50%.

Action item 77-1 for Freytag: Provide suggestions re determination of quorum to UTC Chair.

Reports on Meetings (Agenda item II)

IRG Meeting #11 (II.A)
[Document L2/98-260]

Nelson Ng sent this report to "unicore". Comments from Michael Kung were incorporated.
IRG # 12 is to be held on December 7-11 at Redwood Shores, CA. The meeting is hosted by Oracle Corporation under auspices of the Unicode Consortium.

WG20 Meeting (II.B)

Winkler and Whistler attended the WG20 meeting in June. Winkler reported: WG20 threw out the current sorting standard, and is prepared to accept Unicode Collation. Their decision to use formal BNF notation is a good sign. Simondson will rewrite the cultural conventions standard using BNF notation.

Moore said Baldev said the two standards were different and would stay different even after this meeting, that is, they will not be 100% equal. Winkler said that further study might allow pre-processing, etc. The next WG20 meeting will be in October.

10:30 am Ken Whistler (Sybase, Inc.) arrived. 13 members represented.

IETF and W3C Issues (VI)

UTC statement of support for Language Tagging in Plain Text (VI.A)

Moved by McGowan, seconded by Freytag
[#77-M1] Motion: The UTC is still in full support of Language Tagging in Plain Text, and are glad to see that the IETF is taking it onto the standards track.
10 for; 0 against; 3 abstentions (Compaq, NCR, Oracle)
Motion approved.

Action item 77-2 for McGowan: Write letter to IETF reiterating UTC support for Language tagging in plain text (Motion #77-M1)

Request for Liaison between the UTC and the W3C I18N Working Group / Interest Group (VI.B)

The Chair pointed out that establishing a liaison relationship is the responsibility of the Officers, not of the UTC. This agenda item was included for information. Freytag said that he is in favor of a liaison relationship.

Action item 77-3 for Officers:
Pursue liaison relationship with W3C re common interests in I18N area

Arabic Script (New agenda item)

Davis submitted a proposal on Arabic script, co-written with Kamal Mansour.

Action item 77-4 for Aliprand/Winkler:
Put Arabic script proposal from Davis & Mansour [L2/98-274] on agenda of December joint meeting

UnicodeData Updates (III.A.1)

[Document L2/98-250]

Whistler summarized the changes. With respect to BiDi changes:

- The original collection of BiDi properties was based on the main range; compatible characters were subsequently re-examined.
- Combining marks inherit direct properties of base characters. (this is a clarification)
- The only really significant changes are under H and I (reflecting consensus of the BiDi Ad Hoc meeting) in March.

Quotation mark corrigendum 301F is covered in Version 2.1. This is split of categories to match already approved corrigenda.

Case Mapping (item III): Correction and oversight.

Re the Tibetan fixes (item IV in the document): Extensive discussion on TibEx list included expert opinion on these, with actual examples from Tibetan text.

Re V (compatibility with combining form rather than non-combining form): Davis commented that the change is more in accordance with the meaning of combining marks. McGowan said the characters are not designated as combining marks in the source standard. Davis said that functionally, combining marks are always relative to base character.

Item VII is an explicit addition. Davis noted that it does not include numerical values for Han characters. Whistler responded that this is not appropriate to the file, and those values should be treated separately.

Davis noted that there is a problem with case mapping changes. Whistler said that title mapping is addressed in a separate document, and was not included in this document.

Freytag raised a corrigendum issue re item II: Single and double lower quotation mark are unambiguously opening in current erratum. He proposed striking the line for single and double low quotation marks. Davis argued for no change to anticipate future discovery. Whistler was willing to accept the proposed change.

Moved by Whistler, seconded by Davis
[#77-M2] Motion: To approve document L2/98-250 with the changes identified during UTC discussion.
11 for, 0 against, 2 abstentions (Compaq, Oracle)
Motion approved.

Title Casing (III.A.2)

[Document L2/98-293]

Davis noted that there are some characters that have upper case mapping but not title case mapping. The cases in L2/98-293 fail the normal title casing implementation approach. He proposed introduction of a general rule: If there is an upper case mapping, then there must also be a title case mapping.

Whistler concurred that this is reasonable for consistency. Lunate sigma symbol should be added to the list. He pointed out a potential drawback: that unintended title casing results when one of these characters is used as a symbol.

Sargent pointed out that loss of information in casing movement can be solved by using a rich-text attribute. Davis has suggested this as a hint to implementers for Version 3.0. Freytag said that the description of casing in the book needs to be augmented to point out smarter handling of upper case and title casing, and that we need to note that this can't be solved in our database. Davis pointed out that, in most cases, you choose the words you will title-case.

Moved by Freytag, seconded by Sargent
[#77-M3] Motion: That the UTC formally decide that the UnicodeData file shall not allow an uppercase value to be present without a titlecase value, and vice versa. The UTC accepts the titlecase values for the six Greek symbols in document L2/98-263 and for the lunate sigma, with the titlecase values the same as the uppercase values for these characters. In future, there must always be a title case value when an uppercase value is present.
11 for, 0 against, 2 abstentions (Compaq, Oracle)
Motion approved.

Action Item 77-5 for Davis:
Supply detailed description of title casing including caveats to Editorial Committee.

Action Item 77-6 for Whistler:
Add titlecase values for the Greek symbols in L2/98-263 to UnicodeData.

Line-Breaking Properties(III.C.1.a)

[Document L2/98-267]

Freytag summarized the changes in this revision, including the division of the large class of "default" characters into separate categories. In addition to opening and closing marks, there are punctuation that may not start a line, non-punctuation script characters, and numeric punctuation. The Web has brought another class of character: small column texts to all "/" to allow break.

Whistler pointed out that the definition of line breaking property in Section 2 does not match that in Section 4. It was pointed out that the UTR should be based on Version 2.1 (with the Euro added to 4.15). A discussion of currency symbols should be added.

Whistler said that numeric expressions and currency expressions shouldn't divide across lines, unless you have to do so. There should be a collection of currency symbols. Davis said that equation breaking is interesting. Freytag replied that this is meant to be a default, to provide something simple.

Suignard said that the prefix vs. postfix is useful. Davis said that “gluing” across spaces can sometimes lead you into trouble. Sargent asked about hangul syllables. Freytag said this is covered in 4.6, with a discussion of default vs. customized.

Whistler and Davis pointed out that, since the rules 4.1 – 4.3 are intended to be applied in order, the ordering of the clauses is critical for deriving an algorithm. Whistler pointed out that a list is missing from 4.10, and is concerned that, in discussing the collection of categories, an unspecified model of breaking is being implemented. If the model depends on whether something is a prefix or postfix, we need a model on how this operates. Davis said there was a problem with the rule for NBS preceding rule for SPACE; that for char<space> <NBS><char>, the usual break position is between the space and NBS on "all word processors I know"

Davis said that the collection of data is useful, but it needs to be augmented by an algorithm specifying how it is to be used. Suignard said that only JIS CSS is available, and we need to have something. Much of his work was incorporated into this draft. Davis said that he could use an algorithm that is clearly described. Suignard said that so far we are referencing 405, and would like something else. McGowan would like to keep the algorithm simple, say; ten to twenty lines of code, and avoid the complexity of the

BiDi algorithm. Freytag said that the code to implement this stuff is pretty small, and the 2-D form needs to be worked out.

Action Item 77-7 for Freytag:
Update Line breaking properties: including:

- Update to Version 2.1 (add Euro to 4.15)
- Add discussion of currency symbols

Action Item 77-32 for Davis & McGowan: Validation of line breaking properties

11 members present and 2 by proxy

Math Symbols (IV.B)

[Document L2/98-093]

Sargent said that he had contacted Barbara Beeton and Patrick Ion at the American Mathematical Society, but had been unable to meet with them. The main problem for math symbols is the issue of a font change implying a semantic difference. The AMS experts think that all the symbols should be encoded separately, which seems to be an expensive way to do it. Sargent prefers a variant tag approach.

Because a meeting with AMS was not possible, it was not possible to prepare a formal proposal. Sargent said that math has been waiting for years, and is a complex script. He said that Beeton, expressing the AMS point of view, is interested in a good solution. The annotation proposal plus variant tagging may allow encoding of math in plain text.

Roberts pointed out that properties are different for math, and we need to think about this. Davis said there are lots of problems when something looks the same, but acts differently. He noted that Plane 14 has a set of characters -- presently only for language -- that could be used for other mark ups, but was not sure whether he liked this approach.

Ksar asked whether there are any documents for the WH2 meeting in London. Sargent replied that Patrick Ion is unavailable until the end of September, but it might be possible to submit a proposal. Aliprand advised against rushing development, and Ksar said any proposal should be reviewed before the WG2 meeting.

Whistler said that AMS should not be aiming to solve everything with their proposal, but should break up the problem, and address those that can be dealt with easily. Some are not easy, for example, the significance of font variants and scoping for plain text math. Others are easy, like finding missing subatomic symbols.

Freytag concurred, pointing out that if we don't address the issue of atomic symbols soon, we are presenting all but the most trivial issues. If at all possible, Sargent should create a set of atomics, in good time for the September WG2 meeting. Sargent said these were good ideas, and the symbols were available.

Davis asked how many math symbols are missing. Sargent did not know; McGowan thought not very many. Freytag said that the mathematical community has not given us a review after unification. Davis asked: What can reasonably be created with current Unicode characters? If something needs to be distinguished, why?

Whistler suggested abstracting information from the AMS web site, reviewing it, and then passing it back to them. Freytag proposed a target date of September 1 for a preliminary review draft. Roberts offered assistance.

Action Item 77-8 for Sargent:

Math Working Group is directed to work with Barbara Beeton and Patrick Ion of AMS to define a set of atomic math symbols, i.e., those symbols that cannot reasonably be generated by applying font information to existing Unicode characters. The first draft is to be available for internal UTC review by September 1.

Action Item 77-9 for Sargent:
Ask AMS to mirror proposal to Unicode site.

Proposals based on TC 46 Character Sets

Additional mathematical characters (IV.B.2)
[Document L2/98-212]

UTC opinion was that these characters are already in Unicode Standard and ISO/IEC 10646. Freytag said there was insufficient evidence to support disunification, and we haven't been presented with a clear requirement to create compatibility characters on the basis of transcoding.

Moved by Whistler, seconded by Roberts:
[#77-M4] Motion: The UTC opposes addition of these 6 characters because they are already present in the Unicode Standard and ISO/IEC 10646. The UTC does not have sufficient evidence to support disunification, and it has not received a clear requirement to create compatibility characters on the basis of transcoding.
9 for, 0 against, 4 abstentions (Compaq, Justsystem, Oracle, Sun)
Motion approved.

Additional Latin Characters (IV.C.1)
[Document L2/98-208]

Freytag said that his “resident expert” supported the Letter Z with hook, and he has documentation. Whistler concurred with this need for Middle High German.

Whistler said that traditionally Kra is caseless, and examples of its uppercase form need to be provided. Only the lowercase form is in the British Library’s font.

Whistler said that casing for Yr is not so problematical. Aliprand pointed out that U+0221 is upper case, and the lower case form should have been proposed.

The ou is also proposed as a Greek letter [in L2/98-210]. Aliprand pointed out that the proposal noted that ‘ou” was borrowed from Greek to Algonquin.

This initiated a discussion on the borrowing of letters from other scripts. Davis said that we are not going to disunify letters that are in some other script. Freytag asked about the (hypothetical example of) an Algonquin textbook on Greek, which would be an argument for disunification.

Whistler said that he could accept this for Cyrillic and Latin, but does not want to go down this road very far. McGowan pointed out that we split those [Cyrillic and Latin] because of source standards, but he still thinks that we should throw all characters into a bag. Davis said that we could have had three periods, which would make it easier for programming.

Ksar said that he thought the main reason for adding any characters from TC46 standards was for round trip integrity. Freytag pointed out that the ou is not from a TC standard. Umamaheswaran asked if there is there a requirement to disunify the ou. Freytag replied that, even more than Cyrillic, Latin and Greek tend to have different typographies. McGowan said that when there is a font issue, the question is where to draw the line. These two proposals show an endless stream of stuff to disunify.

Whistler summarized: No evidence for the LATIN CAPITAL LETTER KRA was presented.

LATIN CAPITAL LETTER YR is based mistaken mapping. The cased forms of Z WITH HOOK exist. The letter OU is used in Algonquin: the question is whether we need it separately encoded for Latin and Greek. As they stand, they are working documents. We should be making clear use of our technical expertise. By the time they get to PDAM we should comment them upon.

Ksar said that these proposals will be on the agenda of the WG2 meeting in London. Whistler recommended incorporating all technical feedback on the TC46 proposals into one document. Ksar said that he had asked for them to be treated a separate proposals. To avoid many separate votes, it was agreed to defer further discussion so that the proposals could be dealt with together.

Action Item 77-34 for Aliprand: Draft L2/Unicode position paper on applications for registrations.

Action Item 77-35 for Aliprand: Draft L2/Unicode position paper on proposed addition to 10646 of characters from TC 46 character sets.

Action Item 77-36 for Aliprand: Draft L2/Unicode position paper on proposed addition to 10646 of other characters (i.e., those not from TC 46 character sets).

Extended Tibetan (IV.D.1.a)

[Document L2/98-218]

Judith Lundberg asked permission to tape record this section of the meeting since Chris Fynn, the Orient Foundation’s expert, could not attend. The Chair asked whether anyone objected to this request. Since there were no objections, the Chair granted permission.

McGowan presented the proposal, which is based on the work of the Tibetan Extensions ad hoc group. The initial proposal to WG2 was from China and Ireland [WG2 N1660 = L2/98-024]. Extended Tibetan represents the parts that were removed from the original Tibetan script repertoire as “still controversial.” All proposed characters in this proposal are attested. Experts are able to agree on contents of this document. Unlike many proposals, this one has been thoroughly vetted and discussed with input from true experts.

Ksar asked about the positions of the proposed characters. Umamaheswaran said that we recommend that there shouldn't be any holes. Whistler pointed out that WG2 made a general exception for the organization of Brahmi-derived scripts, and Tibetan is one of these. McGowan said that nominal forms and subjoined forms are kept in parallel positions. It was also judged better to keep astrological signs together in a single block at the end, rather than breaking them up.

Ksar asked about the title "Basic Tibetan.” These are extensions, so the block should be “Tibetan.” Freytag pointed out that there is no need to continue to call it “Basic,” per 10646 decision on block names.

Ksar pointed out that this is a working document, not a PDAM. As a repertoire, there is no problem with the layout. McGowan said that this format was requested by the expert committee. Ksar asked whether the development of the proposal was all done on email. Whistler said there is 4.5 MB of email. Ksar asked whether this included GIFs. Whistler said it is mostly text.

Whistler said that the proposal has been thoroughly vetted and discussed with input from true experts who already have done implementations. He noted changes in some of the glyphs. Umamaheswaran recommended addition of explanatory comments for these, Ksar agreed. Whistler said that there is a whole series of things that need to be fixed, by the committee. Freytag said we don't need a corrigendum and a new amendment. Ksar agreed that a delta was needed, to assist Bruce Paterson, the editor of 10646.

Davis asked how this proposal related to the Orient Foundation’s proposed work. Lundberg said they are doing significant new work, and have talked with Microsoft about having a Beta site. Freytag recommended working with the Orient Foundation before the WG2 meeting to ensure that there are no problems.

Davis asked: Is there any more work needed on the character set that is going in? He stressed that what is being developed is the final version, and asked whether Orient Foundation work would be likely to change anything. McGowan said there is more out there, but we are extremely unlikely to retract anything. This is all rock solid. Freytag added that the WG2 meeting in London is the last opportunity before Version 3.0. We are very close to shut off. Lundberg said that the Orient Foundation could respond within the time frame if needed. McGowan said that Chris Fynn is on the TibEx list, so he should bring up any problems there. Whistler said that McGowan is the focal person for Tibetan, but he will be standing-in for him in London.

Ksar asked whether China had been involved. McGowan said only peripherally, but everything in this originated in the Chinese proposal on Tibetan. Whistler said that China has seen the proposal and their questions have been answered. It was suggested that the L2 comments note the review by China.

Whistler said that WG2 editorial issues should be kept separate from UTC issues. Umamaheswaran pointed out that properties for these characters had been dealt with as part of the UniData changes.

Lundberg asked whether changes could be added after a vote. Whistler replied that any time we accept characters into UTC, we have to do it provisionally based on what happens in WG2 balloting. McGowan added that operationally we can make some changes so that by time it goes to FPDAM everything is hammered down.

Moved by McGowan, seconded by Moore
[#77-M5] Motion: The UTC provisionally accepts the repertoire of new characters, code positions and names in document L2/98-218 as well as the glyph changes.
11 for, 0 against, 2 abstentions (Compaq, Oracle)
Motion approved.

Action Item 77-10 for Aliprand: Fax or give a copy of UTC/L2 98-218 to Judith Lundberg, The Orient Foundation
[Done at the meeting]

Action Item 77-11 for McGowan:
Work with Orient Foundation prior to September WG2 meeting to be sure there are no problems.

Action Item 77-12 for McGowan:
Write a cover memo describing

- the changes in the Tibetan extensions proposal, and
- distinguishing between editorial and technical changes from an ISO perspective.

Proposals based on TC 46 Character Sets (continued)

Greek characters (IV.C.2))
[Document L2/98-210]

Kai with varia can be encoded using composition. Since there are no case forms of Kai, it should be encoded as a letter-like symbol.

Hebrew cantillation characters (IV.C.3)
[Document L2/98-216]

Aliprand said that all characters in this proposal are from a TC46 standard. The Standards Institution of Israel was involved in development of this standard, but has recently been reported as wanting it to be withdrawn.

The proposal may be incomplete; for example, the *accent rafe* should not have been mapped to the *point rafe*. More work is needed to determine the complete repertoire of missing characters. Aliprand recommended that the whole issue of cantillation marks should be presented to experts in Israel and outside of Israel. The transcoding argument for table 2 in ISO 8957 is questionable because neither US nor Israeli library applications for Hebrew script include cantillation marks.

Since the proposals for addition of TC 46 characters are based upon mappings, it was the consensus of the UTC that access to those mappings is needed. Aliprand said that individual UTC members (herself, Whistler, Suignard, Carroll) have been involved. There are three projects to map existing library characters to Unicode/10646 and they need to be harmonized.

Cyrillic characters (IV.C.4)
[Document L2/98-211]

Whistler noted that addition of the right and left descender would cause problems for decomposition. Davis said that the question is whether there is a need for something that would be used for generative purposes. There are no examples other than characters that are already in the standard; our contention is that these are a closed set. Whistler said that all characters in ISO 10754 Annex A are in ISO 10646. It was noted that the character repertoire of ISO 10754 is not consistent with the character glyph model.

Action Item 77-13 for Aliprand:
Ask Randy Barry if there is a Russian name for the right and left descenders.

Characters from ISO 5426-2 (IV.C.5-6)
[Documents: L2/98-213, L2/98-214]

UTC consensus was that the superscript letters are not generative except for "e" and "o." The others are possibly Latin contractions ("r" definitely is). Freytag ruled out superscript e as a “ruby” annotation, but did not know about the others.

The question was whether such contractions belong in plain text. The UTC noted that SGML offers more options for text formatting, and is being used by the Text Encoding Initiative. Freytag suggested that these scholarly and manuscript forms belong in Plane 1 rather than Plane 0.

Umamaheswaran said that a representative of TC46 should be invited to explain why these characters need to be encoded. The UTC needs to understand what the user community really needs. What aspects of these standards are in actual use? Aliprand said that there had been no reports of use from her request for information on the IFLA-L list (run by the International Federation of Library Associations).

L2/215 Signature Marks (IV.C.7)
[Document L2/98-215]

McGowan said that signature marks aren't important, but Aliprand disagreed, as they are important for rare book identification. However, it is questionable whether plain text provides the level of information necessary for rare book scholarship.

It was noted that the reverse section sign is not a section sign, and is possibly spurious. Another question was whether dingbats from this proposal should be encoded in the “holes” in the Zapf Dingbats block.

Meeting closed 6:25 p.m.

THURSDAY JULY 30
Meeting called to order at 9:30 a.m.

PRESENT: Apple Computer, Inc.; Hewlett-Packard Company; IBM Corporation; Microsoft Corporation; Novell, Inc.; The Research Libraries Group, Inc.; Sybase, Inc.; Unisys Corporation
BY PROXY: Compaq Computer Corporation; Oracle Corporation
(Total members represented: 10)
Quorum = 10

NOT PRESENT (at time of roll-call): Booz, Allen, Hamilton, Inc.; Justsystem Corporation; Mathema Software, GmbH; NCR Corporation; Reuters, Ltd.; SAP AG; Silicon Graphics, Inc.; Sun Microsystems, Inc.; Xerox Corporation.
(Total not represented: 9)
Note: Justsystem Corporation; NCR Corporation; Sun Microsystems, Inc. arrived subsequently.

Lundberg sent apologies for being unable to attend.

Mongolian (IV.A)

[Document L2/98-268]

Whistler described the situation for Mongolian script (outlined in the introduction of L2/98-268).

Mongolian Space (gap within words)

Suignard said that using the NBSP would cause a problem in the HTML context, where NBSP is treated as true space.

Davis proposed as a motion: Acceptance of a new character, narrow non-break space, to be encoded in General Punctuation. There was no second, and Davis rephrased the motion to be a recommendation for the final motion.

Punctuation Character
Whistler said that the intent was to make this available for any other vertical rendering. He recommended that the UTC agree to accept the position proposed by both China and Mongolia.

Positional Format Controls
China's analysis shows that for all practical cases, there is no difference in the number of characters required. Freytag said that caution was needed, since this could depend on the structure of encoding in the text corpus. Whistler recommended adopting concurrence with the ZWJ/ZWNJ position (as China proposes), i.e.; no characters need to be added.

Free Variant Selectors
Davis: The only question is, are these free variant selectors? Whistler said that we should acquiesce to shared expert opinion. China/Mongolia have no preference for location.

Davis recommended the general punctuation block, but McGowan said they should be in the Mongolian block unless there is a need elsewhere. Davis argued that there is clearly a general-purpose use. Freytag pointed out why duplicate encoding is not bad: in Mongolian, these do not encode font variants. In the context of Mongolian, the font variant selectors come with default rules. Therefore they are different from general-purpose "font selection" variants. Freytag suggested acceptance of 3 variant selectors but to cover assignment of their location as part of variant tagging discussion.

Vowel Separator

Davis asked: If this happens with only one vowel, why not encode another form? Whistler said that we should not mess with repertoire. He recommended agreement with the Chinese position, and the name MONGOLIAN VOWEL SEPARATOR.

Todo Soft Hyphen

With respect to the proposed name, TODO HYPHEN, China is okay on this name, and Mongolia abstained. Freytag asked whether the character is a soft hyphen, and whether it is visible or not. Whistler did not know, but said that it applied only to the Todo language and he does not know of any implementations for Todo. He recommended acceptance as a Mongolian specific punctuation mark, and to request information from experts at WG2.

Davis asked: Does it ever occur anywhere but at the start? Umaheswaran asked whether any written examples were available. Whistler referred to the schematic in document L2/98-251. If the character appears only at the beginning of a line, then it may be soft. Freytag noted that acceptance of this character must be conditional, predicated on getting information on use from experts at WG2. Davis said we also need to know the context of its occurrence. He recommended making it a non-soft hyphen in absence of information.

Moved by Davis, seconded by Sargent:
[#77-M6] Motion: The UTC provisionally accepts the character repertoire, ordering and suggested names for Mongolian script in document L2/98-088, with the modifications outlined in document L2/98-268R. The preferred location for the free variant selector characters is to be determined in the discussion on variant tagging later in the meeting.
9 for, 0 against, 4 abstentions (Compaq, Justsystem, Oracle, Sun)
Motion approved

Davis noted that the "bang-interro" will have compatibility decomposition. Whistler said that it would be parallel to the double exclamation point.

Action Item 77-14 for Whistler:
Produce a revised version of L2/98-268R stating the UTC positions regarding Mongolian encoding issues for submission to WG2

Whistler said that Joe Becker is awaiting information from University of Indiana re implementation of Mongolian. We can always revisit the algorithm design if implementation experience shows ZWJ/WNJ is insufficient.

Variant Tagging (III.E.3)

[Document L2/98-277]

Hiura presented the proposal on variants. He said that the proposal is non-controversial because it covers the glyphic variations. The notion of user-defined characters has been added since the last version of this paper, to cover two cases: (a) corporate needs, both company-defined and for a customer, and (b) variants of existing characters created by the user. Kida said that the proposal provides way to systematically map into a conceptual base part.

Whistler said that there are likely to be more cases from historic scripts, which entail both changes over time and inherent variation during any particular period. Encoding becomes a problem because you have to determine the changes both historically and synchronically to define "character". Example: hieroglyphics.

Davis defined two fundamental distinctions in variants— (a) For normal purposes, I have a shape variant, and pay attention if I care; and (b) Character variant. A shape variant “indicator” following a character is easy to implement and fairly innocuous as long as it is well defined. As for the character variant proposal, he pointed out that we have the surrogate mechanism to deal with that. Freytag noted that there might be semantically important glyph variations as well as glyph variations that are not semantically important.

Davis noted that all of the characters are to represent glyph variants. One set will be standardized. In doing this, we are saying that we will reserve meaning for variants, whereas in user defined range, they are allowed to use the variant identifiers as they choose. Hiura said that with this methodology, you can at least guess the base character. The authors tried not to address any of the other types of variation identified.

Umamaheswaran asked whether the proposal applies to combining characters. Davis said that it should apply to combining character sequences. Whistler added that to be consistent with canonical equivalence has to be that way.

Umamaheswaran asked about the position of the new variant mark in a composed character sequence. Davis said it should be given the canonical equivalence of "0" to prevent disruption. Behaves like non-spacing mark of class 0 for processing.

Whistler asked whether we understand the intent/function of this proposal. He identified topics to consider:

- Is the number of tags correct?
- Is their function correct?
- Should they be encoded on Plane 14 or the BMP?
- Is there a possible unification issue for variant tags?
- What is their context of use?
- What are the implications for China, Taiwan, etc.?

He noted that the proposal had not been discussed by the IRG.

Davis suggested reserving a block in Plane 14 for format characters. Freytag said that it would need to be included in Version 3.0, and so needs agreement from WG2.

Whistler said that 256 code points seemed a little high in the context of CJK, and the IRG would need to be involved. Davis said that we would never encode vertical extension eligible characters in this space.

McGowan said that the results of research presented in the BUCS paper at IUC#8 in Hong Kong should help to answer some of these questions.

Action Item 77-15 for Oesterle:
Send copy of BUCS paper from IUC 8 (Hong Kong) to Winkler for distribution as L2 document.

Freytag noted a need for a semantic variation tag for math; must not overload with "boldness" attribute (for example). Treat the math notation in a similar way to musical notation, with generally useful characters encoded in the BMP, and really specialized forms in a higher plane. Could have duplicate alphabets in higher planes, with exact mapping for (say) TeX. Sargent confirmed that there is a need for combining marks in math.

Davis likes proposal to have 128 characters in BMP for variant tagging. These could be used to handle Mongolian, and the CJK proposal.

Moved by Davis, seconded by Sargent:
[#77-M7] Motion: The UTC accepts, in principle, reservation of the first 4K code positions of Plane 14 to be earmarked for format tag characters. This will be part of Version 3.0 of the Unicode Standard. This action of the UTC is to be communicated to WG2 as an initial request for the roadmap. The UTC will describe the expected behavior of this reserved block later.
10 for; 1 against, 2 abstentions (Compaq, Oracle)
Motion approved.

Freytag asked about the policy towards filtering variation marks, and pointed out that in Mongolian, it is utterly incorrect to filter them. He proposed a friendly amendment: filterable vs. non-filterable.

Davis said he considered the marks equivalent to compatibility mapping. Freytag pointed out that in Mongolian, they do not have compatibility mapping. Whistler said that this is debatable even in Mongolian, where the variants are of the same letter. Does removal of variant tagging make Mongolian illegible? Davis said that if the variant forms are fundamental parts of Mongolian, then the variant tags in the Mongolian context are <u>character</u> variant tags and not glyphic variants, and they don't belong in this group. Davis asked: When I transcribe Mongolian to Cyrillic to Mongolian, do I lose this character information? Whistler thought so, but admitted he was speculating.

Freytag said he was fairly convinced that there is a need in Mongolian for character tags. Removing the tags gives you incorrect spelling. Davis said he did not want character tags in general, and that this usage should be limited to Mongolian. Umamaheswaran said that if Mongolian requires character variants, these should be encoded on the BMP.

McGowan said that the UTC needs to study the BUCS paper, and get some examples. He pointed out an important distinction: that there are <u>character</u> variants as well as glyphic (shape) variants.

Davis withdrew his proposal on 128 glyph variant codes and all related amendments.

Moved by Davis, seconded by Sargent
[#77-M7] <u>Motion:</u> The UTC accepts, in principle, reservation of the first 4K code positions of Plane 14 to be earmarked for format tag characters. This will be part of Version 3.0 of the Unicode Standard. This action of the UTC is to be communicated to WG2 as an initial request for the roadmap. The UTC will describe the expected behavior of this reserved block later.
10 for; 1 against, 2 abstentions (Compaq, Oracle)
Motion approved.

Action Item 77-16 for Davis:
Write description of the use of characters from reserved 4K in Plane 14.

Action Item 77-17 for Unicode Liaison:
Inform WG2 of UTC action re Plane 14 as an initial request for the roadmap.

Davis said that acceptance could be done in two parts. User defined tags could be done right now. Format tags are reserved for Unicode. Could say that 256 are reserved as UDC variants. Umamaheswaran agreed that this should be added to the roadmap and forwarded to WG2. Winkler pointed out that this is a golden opportunity to bring the proposal forward as comments on Part 2 of ISO 10646. Umamaheswaran said that it should be an active proposal rather than comments.

Moved by Umamaheswaran, seconded by Davis:
[#77-M8] <u>Motion:</u> The UTC agrees, in principle, reserve 512 values in this 4K area. Specific ranges of 256 starting at U+E0400 are to be reserved for future allocation. Of these, the first 256 positions are to be for glyph variant tag characters, and the second 256 positions are to be for user-defined glyph variant tag characters.
8 for, 1 against, 4 abstentions (Compaq, HP, NCR, Oracle)
Motion approved.

Whistler pointed out that we are not accepting any characters for encoding yet. Here is the kind of situation where we are pre-assigning characters before the proposal has been approved. Davis responded that we are seeking a concrete proposal to encode tags for CJK variants, at which time tag values will be assigned. There is no practical import for these values now. Whistler suggested two ways to proceed: First is not asking WG2 to reserve; second, rewrite into roadmap. He objected to the next step, which is reserve particulars. Davis agreed in principle to add the reserved area to the roadmap.

BiDi Algorithm (III.B)

Davis reported on the BiDi Ad Hoc meeting held during lunch. He clarified the two reasons for embedding:

1. To change characters left to right and right to left.
2. When text is embedded, you can add a note that you can add LMR and RML around the boundaries of the block.

At the UTC meeting in April 1998, Paragraph T6 of the BiDi Algorithm was removed. Davis asked for this to be restored. Since this would be reversal of a UTC decision, the Chair noted that a 2/3 vote would be required to reverse the previous decision.

Moved by Davis, seconded by Moore:
[#77-M9] Motion: To reverse the striking of paragraph T6.
11 for, 0 against, 2 abstentions (Compaq, Oracle)
Motion approved by 2/3 vote (required to override previous UTC decision to strike paragraph T6)

Action Item 77-18 for Davis:
Roll the accumulated BiDi corrigenda (including Motion #77-M9) into a UTR for review by the BiDi Committee. If BiDi Committee approves, post on the public side of the Web for public review as a draft UTR using the next available UTR #

Action Item 77-19 for BiDi Committee:
Provide draft of UTR #9 BiDi reference implementation to Editorial Committee by Sept. 16

3:15 p.m. 10 members represented; Hewlett-Packard representative left.

EBCDIC-Friendly UTF-8 (III.F.4)

[Document L2/98-258]

Umamaheswaran described the EBCDIC-friendly UCS transformation format “EF-UTF,” which is designed to address basic interim transformation problems. He looked at particular problems which would prevent EBDIC from having a safe trip. IBM worked on the transformation format, and Oracle is implementing it. Umamaheswaran noted that surrogates must be transformed to UCS-4 form before this transformation is applied.

The transformation is intended to be used inside EBCDIC-based systems, and is not intended for use on the wire. It is recommended to convert to standard UTF-8 or UTF-16 for exchange

Davis pointed out the general name structure of UTF-<number>, and suggested “UTF-8-EBCDIC.”

Whistler asked: What does it mean that we accept the proposal? Davis replied that we describe the algorithm in the same fashion as we do UTF-8. Umamaheswaran said that it was his intent to treat it as a form of UTF-8 in ISO 10646. Whistler said that since Umamaheswaran is planning to go forward with a proposal to make this an annex in 10646, it has implications for Unicode Standard. Ksar said that since this is a specialized case, it may not need to be a normative annex in 10646.

Freytag said that UTF-8 is one of the number one forms of how to interchange data. He doesn't think that an informative annex in 10646 is an optimal publication strategy. This proposal should be published as a Unicode Technical Report and make available as widely as possible. The content of UTR should be a detailed and explanatory document.

Roberts asked how you could tell the difference between UTF-8 and UTF-8-EBCDIC. Umamaheswaran replied that you would look at the trailing bit patterns.

Moved by Davis, seconded by Moore:
[#77-M10] Motion: The UTC adopts proposal L2/98-257 with the name changed to "UTF-8-EBCDIC." It is to be published on the Web as a Unicode Technical Report (using the next available number) for public comment. The UTC has the expectation that UTF-8-EBCDIC will be incorporated into Version 3.0 of the Unicode Standard. Dr. Umamaheswaran is instructed to take document L2/98-257 forward to WG2 as a joint L2/Unicode contribution, for information at this time.
Unanimous (including 2 proxies per their instructions)

Action Item 77-20 for Umamaheswaran:
Submit text of UTR describing UTF-8-EBCDIC to Editorial Committee for review before posting it on the Unicode Web site.

Action Item 77-21 for Umamaheswaran:
After Editorial Committee review, post UTR describing UTF-8-EBCDIC on the Unicode Web site (using next available number), and solicit public comments

Action Item 77-22 for Umamaheswaran:
Submit document L2/98-257 to WG2 as a joint L2/Unicode contribution for information (at this time).

C1 Use (III.F.1)

The Editorial Committee was requested to give more guidance on this topic in version 3.0.

Meeting adjourned at 5:00 p.m.

FRIDAY JULY 31
Called to order at 9:25 a.m.

PRESENT: Apple Computer, Inc.; Hewlett-Packard Company; IBM Corporation; Justsystem Corporation; Microsoft Corporation; NCR Corporation; The Research Libraries Group, Inc.; Sybase, Inc.; Unisys Corporation
BY PROXY: Compaq Computer Corporation; Oracle Corporation
(Total members represented: 11)
Quorum = 10

NOT PRESENT (at time of roll-call): Booz, Allen, Hamilton, Inc.; Mathema Software, GmbH; Novell, Inc.; Reuters, Ltd.; SAP AG; Silicon Graphics, Inc.; Sun Microsystems, Inc.; Xerox Corporation.
(Total not represented: 8)
Note: Novell Corporation; Sun Microsystems, Inc. arrived subsequently.

Moved by Umamaheswaran, seconded by Whistler:
[#77-M11] Motion: To approve the Minutes of the UTC #76/L2 #173 joint meeting as amended.
6 for; 0 against, 5 abstentions (Apple, Compaq, HP, NCR, Oracle)
Motion approved.

Burmese (IV.E.1)

[Document L2/98-265]

Whistler summarized the situation. Burmese has been balloted, and resolution of comments will be done at the WG2 meeting in London. We should take Zaw Tut's additional information on Burmese as input to this, and have further e-mail discussion to formulate an expert contribution for London. The contribution from Zaw Tut is the first really significant input from a native Burmese speaker with contacts with the Myanmar Language Commission. One of his main concerns (Ken and Lee concur) is the potential for additions for other languages that use Burmese script (Shan, Mon, Karen) which would extend Burmese outside the block range, meaning that Burmese extensions would be separated from Burmese characters covering the Myanmar language. We should consider swapping Burmese and Khmer. Expert contributors could estimate the required space for Burmese with extensions.

Ksar asked whether we expect additions to Khmer. Ken replied that Maurice Bauhmann, who is an expert on Khmer, does not think so. Davis said that a script should not break across 128 block boundaries, because that causes inefficient compression. Whistler said he was aware of this and won't make a proposal that would affect compression.

Cyrillic (IV.C.4)

[Document L2/98-211]

Currently these are characters in the standard, so cannot be removed. Issue is canonical equivalencies in Unicode Data file (equivalent to saying they are basically the same characters).

Problems relating to Latin letters in Cyrillic alphabets were identified:

1. Perception, political problem.
2. Hard to make a principled case--Slavic-Cyrillic languages have similar cases.
3. Duplicate characters are problematic. Problem with arguing this is that we already have problem for normal Cyrillic characters.
4. Collation issue--unification across scripts makes it harder to define default sort order for certain languages.

Action Item 77-23 for Unicode Liaison:
Be aware of Unicode position on Kurdish Q and W at WG2 meeting

Towards a Model of Character Encoding

[Document L2/98-269]

Moved by Davis, seconded by McGowan:
[#77-M13] Motion: To take document L2/98-269 as a basis and ask Ken Whistler, with the assistance of the Editorial Committee, to produce a final version. To forward the final version to SC2 as a joint L2/Unicode contribution, and to also make it into a draft Unicode Technical Report to be posted for public comment.
11 for; 0 against; 2 abstentions (Compaq, Oracle)
Motion approved.

UTF-16 Registration/BOM Issue (Old business)

Davis said that "UTF-16BE" and "UTF-16LE" are designations for byte encodings serialized as sets of 16-bit entities. In this context, UTF-16 means UTF-16BE. It is much clearer to use UTF-16BE and UTF-16LE. In the absence of an explicit BE or LE, "UTF-16" is to be understood as UTF-16BE.

Whistler distinguished UTF-16 which is the encoding form, from UTF-16BE and UTF16LE which are encoding schemes.

Davis pointed out that we have serialized data in both of these, and need to need to make the situation clearer for everybody. McGowan question registering both UTF-16BE and UTF-16LE. Davis pointed out that they are closely related, and Whistler said that the point is to register both of them.

Davis identified a second issue: What is the UTC's recommendation for use of the BOM? There was consensus that the discussion to date hasn't changed anything about the interpretation or use of the BOM. If a BOM occurs, usage must agree with Unicode conventions. The RFC should include references to the appropriate parts of the Unicode Standard.

UTF-16 on the wire means big-endian. No use of BOM needed. Use "UTF-16LE" and "UTF-16BE" in context of "when serialized into bytes." "UTF-16" is not what you use to call a serialized form.

Moved by Davis, seconded by Freytag:
[#77-M14] Motion: The UTC recommends the use of the term "UTF-16BE" for all instances of serialized UTF-16 data in big endian form, and use of the term "UTF-16LE" for all instances of serialized UTF-16 data in little endian form. The term "UTF-16" is not to be used to refer to serialized byte forms. The UTC will actively promote use of these terms by other agencies.
10 for; 1 against; 2 abstentions (Compaq, Oracle)
Motion approved by 2/3 vote (required to override previous UTC decision to register "UTF-16").

Moore noted that XML wants to use IANA registration for UTF-16. There was consensus that the IANA registration should proceed with "UTF-16BE" and "UTF-16LE", and the RFC should specify the preference for the use of UTF-16BE in Internet protocols.

There was consensus that the RFC should state that either UTF-16BE or UTF-16LE might or may not have a BOM. Text using a UTF-16BE label may come from a source that uses the BOM as a signature. The RFC should also note that an inverted BOM is *illegal* in text labeled UTF-16BE or UTF-16LE.

There was consensus that the RFC should state that in the absence of a higher-level protocol, you may not assume that a BOM will either be present or not as the first character in the field, for text labeled as UTF-16BE or as UTF-16LE.

[#77-M15] Motion: The UTC explicitly does not take away the presently existing UTC definition of BOM.
Motion approved by consensus.

Sargent proposed a recommendation to Editorial Committee: Whenever text is serialized, BOM can be used to disambiguate. It was also recommended that the Editorial Committee include the form of BOM in UTF-8 encoding.

Serialization isn't the only time we need to talk about the data "UTF-16" is used in context. When we are not referring to byte-serialized data, use the term "UTF-16". The term UTF-16 can continue to be used when we are not talking about serialized data where we do not need to be specific about the byte order.

Moved by Whistler, seconded by Freytag:
[#77-M16] Motion: The UTC recommends that the term "UTF-16" should not be used when referring to a byte-serialized form with a specific byte order. The terms "UTF-16BE" or "UTF-16LE" should be used instead.
9 for; 1 against; 3 abstentions (Compaq, NCR, Oracle)
Motion approved

It was noted that “UTF-16” is valid terminology in other contexts.

Umamaheswaran asked about implications for UCS-4 and UCS-2. He suggested that additional action items were needed, to develop similar terminology for UCS-2 and UCS-4, as a contribution to WG2. Roberts suggested that it was more complicated than two forms. Ksar said there is a need to be on same track and be very clear. The current text always says, "when serialized as big endian." Umamaheswaran identified a potential problem for people who use IANA registrations in conjunction with ISO 10646.

Action Item 77-26 for Unicode Liaison:

Make it clear to WG2 that the UTC is now using the terms "UTF-16BE" and "UTF-16LE", and we encourage WG2 to use the terms as well.

Action Item 77-27 for Aliprand:
Instruct David Goldsmith that our previous motion to register "UTF-16" is superseded by registration of "UTF-16BE" (and also "UTF-16LE"). Include in the RFC that UTF-16BE is the preferred form for use "on the wire"

Action Item 77-28 for Goldsmith:
Work with the Editorial Committee to refine the wording of the RFC.

Action Item 77-29 for Goldsmith:
Circulate the RFC to the "unicore" and "ietf" lists for comment.

Meeting reconvened at 1:50 p.m.
9 members present + 2 proxies.

Unicode Normalization (III.F.2)

[Document L2/98-279]

Davis introduced the topic, which has been discussed this on email. He proposed progression of the proposed Draft to UTR. Freytag wants to have things done not glacially but systematically. The e-mail discussion was long and complex with subtle implications.

This UTR is urgently needed by W3C. Davis proposed a short UTC meeting at IUC #13 or a letter ballot to finalize this UTR.

Transmission of old data was discussed. McGowan said that you may have data that was once normalized and today is not normal. Whistler said that we know new decompositions and new characters will be added. Davis said that as long as Unicode only adds characters, then the normalization data is correct if based on Version 2.0. He added that we already refer to normalizations in standard, and supply the data but it is not well defined. What is missing is precise construction.

McGowan asked why are we defining Unicode normalization. Whistler said that it should be based upon our data files. The aim is to minimize chaotic confusion, with people making claims where they shouldn't be.

McGowan expressed concern about continual revision. Davis said that, even if it is the absolute worst case, it is better than nothing. McGowan said we must prevent ourselves from redefining pre-composed normalization.

Freytag said we need rules for composed; rules for decomposed; and rules for new characters. Implementers will have to carry two tables -- a new version and the frozen version. Whistler said that decomposition is on current version only, so the implementation actually needs to carry one table with the delta for characters that have decomposition.

Freytag asked: Does normalization force me into a particular stance re compatibility? Problem of recognizing new characters. Davis pointed out that the delta could conceivably override the mapping of some precomposed characters.

Roberts asked about interaction with variant tag proposal. Davis replied that we may conceivably add an empty tag. McGowan complained that it keeps getting more baroque. Davis replied that we are trying to narrow things and clarify the chaos. McGowan said: Words in the book are "weaseling" out of decomposition. Freytag agreed with this point, and cautioned about turning too baroque. However, complexity that is well defined is better than undefined complexity.

Whistler said that it is not possible to hold to the current form. This is a way to segment problem. Secondly, it gives us stronger ammunition for stopping the water torture. Someone commits to normalization -- stand firm against those who are interpreting it. Suignard said that there are those who want right away.

Umamaheswaran said that normalization needs to be anchored upon a specific version. Davis noted that we are looking at repertoire of Version 2.1. Ksar pointed out that WC3 is aiming at Version 3.0, and that is what he and Umamaheswaran recommend. Whistler said that if this is W3C's intent, then the need to progress this document to final Unicode Technical Report status is moot. Davis proposed progression to a draft UTR. Umamaheswaran said that if the draft UTR gets delayed, we need to let W3C know. Davis noted that the principal thing to change is the Unicode database.

[#77-M12] Motion: To progress document L2/98-279 to Draft Unicode Technical Report incorporating the editorial changes recommended by the UTC. The Draft UTR is to specify a database based upon Version 2.1. Work with WC3, to specify when they need this. Plan a UTC meeting in September and convene or cancel this based on W3C needs.
10 for, 1 against, 2 abstentions (RLG, Compaq)

Action Item 77-24 for Davis:
Revise document L2/98-279 and publish it as a draft Unicode Technical Report for public comment.

Action Item 77-25 for Davis:
Determine when W3C needs to receive UTC response on normalization. Convene or cancel special meeting during IUC 13 accordingly.

Action Item 77-30 for Moore:
Arrange meeting room and time, in order for the UTC to further discuss UTR #15 on Sept 8, 1998 in San Jose.

Action Item 77-31 for Davis:
Draft UTR #15 – Put in cross pointer between read me and document. Clarify policy and implications.

Line Breaking Properties (III.C.1.b)

[Document L2/98-267R]

A revised version of document L2/98-267 was distributed.

Freytag said that Version 2.0 describes an algorithm but with few details. Other implementations use different methods. Line breaking lists separately compiled by Suignard and by Freytag are almost identical. With the draft TR #14, Freytag wants to go with what is needed to make it acceptable. He noted that most of the line-breaking properties are informative with a few normative.

Davis said that there are some interesting implications. If I line break, but don't break at space after (as in FrameMaker), I would not be "non-conformant?" Whistler said that the UTC has never come to grips with what makes a character normative.

Davis suggested addition of a paragraph after conformance, to explain that these properties can be over-ridden by higher level protocols. Freytag noted other changes, such as Whistler's recommendation to substitute "ambiguously open or closing characters" for "paired characters", and McGowan's to include a table like the one in Suignard's paper/

Davis proposed that companies undertake validation of line-breaking properties. Freytag suggested that this be considered at the September meeting if it is held. McGowan said that there needs to be complete specification for all Unicode characters.

Action Item 77-32 for Davis & McGowan: Validation of line breaking properties

Freytag noted that there is no status change for this proposal, but he is approved to change the text of draft on the public side.

In-Line and Interlinear Annotation (III.C.1.c)

[Document L2/98-270]

Davis said that he agreed with Hiura’s comments in L2/98-270 (co-written with Kobayashi and Kida). He is worried about the case of not recognizing and filtering out the annotation characters.

Freytag said that the thrust of proposal is not to change interlinear annotation, but to recognize the codes and how they can be most useful.

Davis objected to their use in interchange, because of potential confusion. McGowan agreed, saying that the possibility of having serious misinterpretation is a real problem when data is turned into plain text without annotation. Sargent said that it happens with Ruby.

Davis said that if you cannot display an annotation as such, then you must support a distinct system for bracketing the annotation.

Freytag suggested holding an off line discussion of the issue, and asked: How shall we proceed? What should I do? McGowan said that it is worth going forward with the discussion. There is a sense of the community being in favor of a technique for annotation, but it hasn't been discussed enough.

Suignard said that the point of the proposal is that we want to use annotation in a plain text environment. You don't want these things in annotation or transmission. Whistler said that a minimal change to the proposal would be to state its intent.

Moore said she was hearing very mixed opinions. Freytag said that his preferred resolution would be to clearly spell out that this is a way for plain text transmission of annotations. There is a constituency for it, using 16 bits. ASCII doesn't allow for it. Enriched plain text is different. Annotations don't belong in plainest of plain text.

Without a clear set of directives, Freytag proposed that a willing co-author assume responsibility.

Action Item 77-33 for Suignard: Help Freytag with interlinear annotation document (proposed UTR #12) by preparing a new draft.

Umamaheswaran suggested that characters for annotation might be considered soft controls. If so, there should be a principle for soft controls and a defined fall back mechanism. McGowan said that the issues had been pretty well analyzed by Hiura, Kobayashi and Kida.

Freytag asked about writing user guidelines. Kida said that the problem lies in treating Ruby as part of the content of text, which it is not. Freytag responded that an annotation is parallel text, and annotations are not dividable.

McGowan said that Ruby text is interesting. Say you had Chinese with a gloss. Remove the gloss and you still have content. Remove only the Chinese, you still have content.

Whistler said that we have a consensus that there are a number of clarifications that we need to do, with Suignard’s help. McGowan proposed a special meeting involving the experts who know about this model, those who strongly support it and those who are strongly opposed.

TC 46 Proposals and Registrations (continued)

Umamaheswaran asked whether we should request an opportunity to review the mappings developed by TC46, and whether national bodies could be requested to review this work.

Freytag pointed out the difficulty of reviewing the proposals when the mapping which determined whether a TC 46 character was in ISO/IEC 10646 or not had not been made available. Aliprand said that she could provide a copy, but it was still a working document.

Action Item 77-34 for Aliprand:
Draft L2/Unicode position paper on applications for registrations.

Action Item 77-35 for Aliprand:
Draft L2/Unicode position paper on proposed addition to 10646 of characters from TC 46 character sets.

Action Item 77-36 for Aliprand:
Draft L2/Unicode position paper on proposed addition to 10646 of other characters (i.e., those not from TC 46 character sets).

Closing

Future Meeting Dates
Dec 2 - 4, UTC meeting hosted by Novell, Inc. in San Jose, CA, USA
Dec 7 - 12, IRG #12 meeting hosted by Oracle in Redwood Shores, CA, USA

Action item 77-37 for Aliprand:
Revise UTC Procedures incorporating additions from McGowan and herself. Arrange for redistribution to member representatives.

Action Item 77-37 for All UTC & L2 Members:
Send status reports on your action items by e-mail to Arnold.Winkler@unisys.com
UTC Meeting Adjourned.

ATTACHMENT 1

UTC #77 and L2 #174 Joint Meeting – Attendees

Wednesday, July 29, 1998

Joan Aliprand; RLG; Joan_Aliprand@notes.rlg.org
Julie Allen; Unicode, Inc.; julie@unicode.org
Asmus Freytag; Unicode, Inc.; asmusf@ix.netcom.com
Hideki Hiura; Sun Microsystems; hiura@eng.sun.com
Lloyd Honomichl; Novell; lloyd_honomichl@novell.com
Yasuo Kida; Apple Computer; Kida@apple.com
Tatsuo L. Kobayashi; Justsystem; Tatsuo_Kobayashi@justsystem.co.jp
Mike Ksar; Hewlett-Packard; mike_ksar@hp.com
Judith Lundberg; Orient Foundation; JLORIENT@compuserve.com
Rick McGowan; Apple Computer; rick@unicode.org
Lisa Moore; IBM; lisam@us.ibm.com
Julia Oesterle; Unicode, Inc.; julia@unicode.org
Gary Roberts; NCR; Gary.Roberts@SanDiego.CA.NCR.com
Murray Sargent III; Microsoft; murrays@microsoft.com
Michel Suignard; Microsoft; michelsu@microsoft.com
V. S. Umamaheswaran; IBM; umavs@ca.ibm.com
Ken Whistler; Sybase; kenw@sybase.com
Arnold Winkler; Unisys; Arnold.Winkler@unisys.com

Thursday, July 30, 1998

Mark Davis; IBM
Apologies: Judith Lundberg; Orient Foundation

Additional attendees for BiDi Ad Hoc Meeting

John McConnell; Microsoft
F. Avery Bishop; Microsoft
David C. Brown; Microsoft
Michael Jochimsen; Microsoft
Paul Nelson; Microsoft

Friday, July 31, 1998

Same as Thursday

ATTACHMENT 2

Corrections to Minutes

At the UTC #78 & L2 #175 joint meeting, it was pointed out that the Minutes of the preceding joint meeting (UTC #77 & L2 #174) in document L2/98-281 were incomplete.

The contents of document L2/98-281 have been revised using notes from Julia Oesterle and Arnold Winkler. The significant changes are:

- Motion M12 (on Unicode Normalization) corrected to capture full wording and accurate voting results.
- Discussion on Unicode Normalization added
- Discussion on Line Breaking Properties added
- Discussion on In-Line and Interlinear Annotation
- Concluding discussion on TC 46 Proposals and Registrations