Preliminary Minutes - UTC #74 & X3L2 #171 joint meeting
Mountain View, CA – December 3-5, 1997

1. Administrative Issues

1A. UTC Membership Roll Call

Quorum: 13

Present: Apple Computer, Inc.; Digital Equipment Corporation; Hewlett-Packard Company; IBM Corporation; Justsystem Corporation; Microsoft Corporation; NCR Corporation; Novell, Inc.; Oracle Corporation; The Research Libraries Group, Inc.; Sun Microsystems, Inc.; Sybase, Inc.; Unisys Corporation; Xerox Corporation.

Not Present: Booz, Allen & Hamilton, Inc.; Silicon Graphics, Inc.; Reuters, Ltd.

List of attendees: See appendix.

1B. Declaration of Joint Meeting

1C. Approval of Joint meeting Agenda [97-261R]

[#74-M1] Motion: To approve agenda as amended.
Moved by Jenkins, seconded by Carroll.
Unanimously Approved

1D. Approval of Minutes and Review of Action Items [97-255]

[#74-M2] Motion: To approve minutes of previous meeting as amended.
Moved by Jenkins, seconded by Sargent
Unanimously Approved

1E. Registration of new documents

1F. Meeting Calendar (see attachment)

10. Default Ordering

Guest: Alain LaBonté

SC22/WG20 Subcommittee for Programming Languages

WG20 deals with internationalization. International sort with default order. Based on method for Canada. There exist differences in perception of the Unicode consortium and WG20. There also exists questions about the mechanics of default ordering. Face-to-face meeting was arranged with Mr. LaBonté to resolve these differences, and seek common agreement.

Scope for international collation standard is at issue. What should be covered? See Document 97-280 for summary.

1. Want to emphasize “complete.” Part of criticism of 14651 was that some chunks had been left out.

2. Unicode requirement for resolution on canonical equivalencies and compatibility equivalencies.
3. No significant difference here. Exact details are different.
4. Rules relating string length and weight levels.
5. Tailoring: Scope and rules

Alain: Key construction is no longer normative – has been removed to informative annex.
Whistler: Critique of API
Alain: Part of standard. Could this be split later into separate standards?
Whistler: Speaking as a vendor, cannot use the API. Am implementing cross-platform.
Freytag: Shows parentage. The vendor community needs a standard that transcends a particular sub-context (though well established). Collation part of standard.
McGowan: Within the Unicode community, have typically stayed away from APIs because much work is cross-platform.
Ksar: Request for status?
Alain: FCD ballot ends in April 30, 1998. Has not yet been approved. Cannot modify draft at this stage. New version tries to answer all comments.
Ksar: Not a standard yet. Still under development. Is possible to utilize feedback on CD plus FCD – split into two different parts: collation and maybe a TR on needs of various APIs. Has been done in various cases.
Alain: Will be easier to have two different conformance standards.
Ksar: 1. Forces collation and 2. Not a standard. TR has APIs which are implementation specific. APIs are language dependent.
Freytag: There is a family of languages which are similar enough, that a binding can be written to cross but there are other languages that are different from their predecessors. Taking API and converting to Java is the wrong thing to do.
Uma: SC22 issue. We can say we suggest splitting into two documents. Separate API but don't go as far as TR. Ask SC22 to take a harder look at what is really needed by industry.
McGowan: Fundamental requirements: One single default collation for Unicode/10646. Fatal to have two standards. Corporate stance is that there can be only one standard. We have entered that age of diplomacy. We have to find a diplomatic solution.
Whistler: Different point of view. There is no way to conceivably get collating through ISO requirements.
Uma: What are requirements for default? Which are correctable – information needed vs. resolution needed.
(break)
Whistler: Three levels plus last level. Java uses four levels. No intent to go more levels.
Uma: Could be done with more levels.
Whistler: I'm not arguing for more levels. [Here's] the problem I'm having… p10-11 actual implementation specified ways of using table to convert for key cap construction for a Java implementation and Sybase.
Davis: Order is not "random." It is arbitrary after the fourth level.
Whistler: I want all Unicode characters weighted at all levels.
Alain: If we want to be compatible with SC22…
Freytag: The tail is wagging the dog. We don't want to think about language bindings, focus on ordering.
Not relevant to business practice. We are trying to put it on table to come up with an elegant solution.
Alain: Both are as elegant as the other.
Uma: There is a problem…
Whistler: Doesn't provide a complete solution.
Davis: Characters it doesn't know about it says it can ignore. This is wrong. Sets whole classes of characters as ignorable. Doesn't take canonical and compatibility mappings in Unicode.

Uma: Our disagreement seems to be standard and requirements needed.
Alain: The Scandinavians don't want. Only thing we can say is to avoid double coding.
Whistler: Sting ordering – can be ignored by those who don't care.
Alain: Separate that – doesn't belong to ordering
O'Donnell: Yes it does!
Freytag: For us to have a standard that does not describe sorting is unacceptable. It's absolutely fundamental.
Aliprand: European libraries use…Ignoring of their constituencies.
Davis: Need a standard to implement Unicode.
O'Donnell: Two standards?
Davis: No one I know is using 10646 without using Unicode.
Freytag: Let us try to agree on content of technical agreement, with how to ignore features. Can we in this room come up with an agreement? Instead of second guessing what others – Scandinavians – might want.
Alain: Coding issue does not belong in 14561.
Ksar: If we establish goal – "standard for Unicode/10646" – that is a noble goal. We need to come to agreement on contents of what that standard is. Specify all of the possible characters – including canonical and equivalence – in one standard. There is no reason to have two! Those who don't use… part of standard, not part of conformance.
O'Donnell: The fact of the matter is that it has these equivalencies and allows combining.
Alain: Does not describe equivalencies.
Aliprand: Should say something explicitly.
Alain: Coding issue – should be a different standard.
Davis: Is everyone up to speed? Levels are the same in the two standards once you deal with ignorables. Level 1 pass – alpha level; level 2 – accent; level 3 –case distinction…

Texin: Focuses on first level or all the detail… As an impelementer, I need more.
A class of ignorables makes my job simpler. UTC tables – meta table "here is precise description". Specification that has both.
Kernaghan: Compromise 10646. Value of table – very valuable. In Unicode 2.1 we could define.
Texin: Both are complete for scope.
Aliprand: ISO ??? is widely used in Europe. UniMark library exchange. As a compromise, must handle non-spacing marks.
Uma: Requirement for default?

- Script vs. language
- API split off
- Properties consistency
- Missing characters
- Incomplete weights for ignorables
- Incorrect characters (**Alain:** No wrong characters)
- Numeric issues sorted on first level.
- Canonical equivalences
- Weighting of 14651 of sorting to language

Davis: If we are talking about scope… specify how characters sort. Better to say we don't know how to handle, such as Denagari.
Alain: Tried to harmonize six languages. Not a default. They want "template" not "default."
Whistler: That is part of thing that bothers me. We get hung up upon terminology.
Methodology. Two core data files. Unidata.txt – is Unicode chars in primary sort.
Many ignorable characters need to be weighted with respect to each other.
Requirement for handling ignorables is not being met. Also disagreement re canonical equivalence.

Ksar: Concept of tailorability. Define collation to have in a neutral way for the implementers.
Freytag: Default of Unicode, you must use tailoring.
Alain: Now 14651, must do something in tailoring.
Whistler: Biggest issue are Cryrrilic and Latin. No particulars; not to any language on first level.
O'Donnell: Fundamental difference.
Whistler: Script neutral.
McGowan: Difference in default behavior.
Alain: 14651 has gone away from that.
McGowan: Need for default without tailoring
Freytag: Toggles needed.
Alain: You have to make a choice. The philosophy we took is better but we are moving away because are required choice.
McGowan: It is unacceptable to not have a defined standard.
Alain: If you don't make a choice you get an error.
McGowan: That is unacceptable.
Moore: Defined – specified neutral.
Joan: Script neutral.
Alain: From outside point of view – not neutral
Davis: I agree with Alain. Less important issue. Idea is complete, unambiguous specification.
Alian: Implementers are not users who are the key input.
Whistler: 14561 does not meet requirements of Unicode implementers. It may mean that we don't have an international standard.
Davis: I would rather have two standards with one of them right. Masssive tailoring of 14561. Unicode Standard can be massive tailoring.
Alain: Would be in favor of the later in a separate standard. Europeans don't want use of combining characters.
Whistler: Similar to Everson's recent comments on ?? comments on Hebrew accents. Can be ignored.
UTC opinion: Canonical equivalence is essential part of sorting standard.
Davis: Seems like only resource would be to publish a UTR for sorting for Unicode. Unless we can come up with something that correctly allows you to sort Unicode, the standard can't be used. Try to agree on contents of technical description. Can we in this room agree? Take up separate phase, persuasion of other countries to ignore features.
Ksar: Establish goal that we want to have one standard, not multiples. Encompass all of characters in Unicode, meets all needs. Should encompass all features.

Davis: Believe that once you resolve ignorables, levels are equivalent in both proposals.
Alain: That is so.
Texin: Needs complete and precise description. Needs to safisfy my markets. Good compromise: UTC tables as meta-level. Specifications should have meta-level, and then a practical default table.
Could we look at least in a little more detail. Which are correctable and which de we need to resolve. Which are which?
Uma: Split off API

L2 Write Contribution to SC22
[#74-M] Motion: That L2 write contribution to SC22 to make APIs as a separate work.
Moved by Umamaheswaran, seconded by O'Donnell
Approved by Consensus

Action L2-168 for (to be assigned): Write contribution to SC22 recommending that the APIs in FCD 14651 be made a separate work item.

Action 74-35 for UTC Members: When L2 does this, contract national body members to solicit support for the recommendation that the APIs in FCD 14651 be made a separate work item.

Davis: Let's produce a document that outlines the issues. Particular differences that we see. Suggest this be completely tabled – re scripts based vs. language based.
Whislter: Most important issue is canonical equivalence.

Alain: To make neutral must make choice and tailor…
Freytag: Clarify tailoring. Analogy: Pre-set toggle switches vs Feeding data tape with tables. Unicode case – tape you provide is never at zero length; plus you need to throw toggle switches. Table of information vs preset yes/no. (e.g. case/case insensitive) 14561 – throw toggle switches only.
Texin: What is ISO using for case rules?
Alain: These are based on names of characters.
Freytag: Would you fix any discrepancies between Unicode and 14561?
Alain: Absolutely.

Ksar: Wants to see feedback from WG20 re script-based principles.
Alain: Not an issue.

Unicode requirement -- by consensus. Contribution to WG20 – document important differences. Request that these must be rectified.

Freytag: Script based allows 1) agreement with Unicode properties, 2) decomposition, and be 3) automatable to allow maintenance with respect to weights in general.
Davis: Numeric version based upon earlier CD) needs fixes. Cases where 14561 disagrees with Unicode. To do so, we need a usable version of 14561. Have a program based upon an earlier version of 14561. Will Alain correct Marks version?
Alain: Will provide update with FCD changes.

Action 74-36 for Davis: Send numeric version of keys for 14651 based on earlier CD to LaBonté.

Action 74-37 for LaBonté: Provide update with FCD changes to Whistler.
Action 74-38 for Umamaheswaran: Check with Austin-IBM to see if they can generate set of numeric keys.

Regarding Properties:
Whistler: Sifter so there is a consistent treatment. Problem with arbitrary assignment of particular values can run up against properties. Example: Japanese NSB requirement. Not a problem for Kana. Also implies combining marks in Japanese punctuation. cf . earlier version of 10646 – cannot be amalgamation of national practices.
McGowan: Requirement to be consistent with Unicode properties.
Alain: Application dependent…
Whistler: Roman numerals are worst case situation.
Davis: Even if dumb, we don't care; but do if not consistent.

Digits: Anything in 14561 that appears numeric, should be sorted by 0 through 9.
Uma: Recommend to L2.

Recommendation to L2 for WG20:
Digits: 14561 sort in numerical order vs. sort 0-9 digits by default, but all other numeric values otherwise. Is partially related to equivalence issue.
Note that current treatment of digits is in violation of UTC requirement of consistency with Unicode properties. Therefore proposed amended set of weights for digits or numeric characters; and propose addition of informative annex to DC that gives existing weights that can be used for tailoring. Provide data for this.
Agreed by consensus.

Whistler: Problem of handing all this over to WG.
Alain: Someone has to do the job. If not you than me. Someone has to do the changes. If you do it, your errors, not mine.
Uma: All I hear is a request for help.

Missing Characters:
Input to WG20 re Indic scripts, Thai, Lao, Tibetan, etc. Chinese compatibility characters. Submission will include specified ordering for these. Requirement for syllabi forms and jamo canonical equivalence sequences. Separate submission Korean: needs to have a specified order.

Combining characters are all being treated as special in CD 14561.

Whistler: Indic is full of combining marks.
Alain: If it is a matter of moving from 4th level to first, then provide me the data and no one will object.
Whistler: The answer in Java is that we decompose so it's not a problem – after we decompose.

Pre-composed vs. Combined
If you don't recognize the combining characters, you will sort improperly.

McGowan: Standard has to specify correct algorithm.
Alain: ISO said they didn't want to deal with level 2 or 3.
Winkler: Decision in Copenhagen had nothing to do with data.
Whistler: You don't have to totally decompose.
Davis: We have to make a contribution that corrects the algorithm so that it works. What should be the algorithm's result? Specify same results no matter which algorithm is used.

Action 74-39 for Davis: Draw up, with Whistler and LaBonté:
1) new draft of algorithm that takes into account Canonical equivalence.
2)Look into conformance implications and draft additional text, if necessary.
By January 31, 1998 for review at Feb UTC. After done, L2 submission to WG20

Uma: Can ligatures be handled in data file – yes.
Alain: It is important that ignorables be predictable.
??? : Sharp and flat should be treated as special letters. True dingbats are ignorable. Needs more work. Tailoring can be done, but this group has skill and experience to do preliminary list.

Action 74-40 for Aliprand/Winkler: Put revision of international sorting algorithm on agenda for February UTC/L2 joint meeting.
Action 74-41 for Whistler: Look at miscellaneous symbols to do first cut re their sorting