Title: L2 Codes and Character Sets Liaison Report

Author: Dave Birdsall
Status: Informational
Date: February 2, 2001

1. Introduction

L2 is the NCITS committee that deals with codes and character set issues. More recently, L2 has expanded its focus into internationalization issues. L2 works in a symbiotic relationship (my term) with the Unicode® Technical Committee of the Unicode® Consortium. For all practical purposes, an H2 liaison to L2 is also one to UTC, and perhaps should be formally made so as well. An H2 motion to do this is suggested below.


Since the first version of this paper, I have exchanged more e-mail with L2 members, and attended one day of the most recent L2/UTC meeting. Changes in this version reflect what I’ve learned from these contacts.

My apologies for not including revision bars in this version. My word processor shows most of the paper as having them, so they seem of doubtful value. Therefore, I’ve chosen to omit them in the published version. I trust that this paper is short enough that lack of revision bars should not be an issue.

2. Related Documents

1. WG3: HEL-052 (also L2-01/057), Possible Problems with Characters, The Story Continues, J. M. Sykes.
2. Character Model for the World Wide Web 1.0, W3C Working Draft. This is a work in progress. The latest version can be found on the Web at http://www.w3.org/TR/charmod/.

3. Organization

L2 meets four times a year, co-located with UTC meetings. As the annual report puts it, this is “economical because most of the members of L2 are also members of the Unicode Consortium and the subject matter overlaps widely.” The normal mode of working is that technical papers are discussed and agreed upon under the UTC umbrella, then these results are ratified in an L2 plenary session on the last day as L2 positions. The L2 chair happens to be the UTC vice-chair, and vice versa.

1. Sometimes web search engines can be wonderful. A search on “L2 NCITS” reported this URL as its very first hit.
A listing of up-coming meetings can be found at the Unicode web site, http://www.unicode.org/unicode/timesens/calendar.html.

Within ISO, this work is spread over two committees, JTC 1/SC 2 for codes and character sets, and JTC 1/SC 22/WG 20 for internationalization. L2 holds US TAG responsibility for both. UTC is a Class C Liaison Body to ISO, and also makes recommendations to these ISO bodies.

L2’s web site currently lists 16 members. Among them are computer vendors (Apple, Compaq, Hewlett-Packard, IBM, Sun, Unisys), DBMS and other software vendors (Microsoft, Novell, Oracle, Progress Software, Sybase), and a diverse variety of other companies and industry groups (Bell Labs, which I assume is Lucent Technologies, Research Libraries Group, Reuters, Share, and Unicode).

Most of these (all except Novell and Lucent) are also Unicode Consortium members. Unicode Consortium members may all participate in UTC. There are several classes of membership. These classes vary in terms of voting power on UTC, access to materials and the nature of their participation. The list of Unicode Consortium members is quite long, and I will not repeat it here. It can be found at http://www.unicode.org/unicode/consortium/memblogo.html.

The L2 officers are:

Chair (and also UTC vice-chair) - Arnold Winkler, Unisys Corporation
Vice Chair (and also UTC chair) - Lisa Moore, IBM Corporation
International Representative - Michel Suignard, Microsoft Corporation

Papers are distributed electronically from their web site. Papers are zipped in sets of 20 for faster downloading.

4. Program of Work

4.1 Past Publications

L2 has a very long history. The annual report lists over 30 projects.

Besides the well-known character set standards (e.g. IS 646 also known as ASCII, various flavors of IS 8859, and IS 10646 also known as Unicode), L2 has standardized just about anything imaginable that involves characters. Examples include: Hollerith Punched Card Code (IS 6586:1980), Perforated Tape Code for Information Interchange (IS 1113:1979), Printing specifications for optical character recognition (IS 1831:1980), and Code Extension Techniques for Use with the 7-bit Coded Character Set for the ANS Code for Information Interchange (IS 2022:1994). The last example is one that H2 members may remember from SQL3 discussions on “form-of-use.”

4.2 Current Work

L2 is currently maintaining the ISO 8859 (8-bit character set) and 10646 (multi-octet character set) standards, and re-affirming other older standards. ISO 10646 is kept in synch with the Unicode Standard, as pub-
lished by the Unicode Consortium. L2 and UTC are also monitoring W3C’s work on Character Model [2], working to keep it and the Unicode Standard in alignment.

ISO JTC 1/SC 22/WG 20 has embarked on new standards in the internationalization area. These include string sorting, specification of cultural conventions (e.g. date and time formats), registration of cultural elements, and Internationalization APIs. UTC is also working on sort order specifications (i.e. UTS #10), keeping this in synch with the ISO effort. On the other hand, L2 would prefer that work on most of these other items stop.

Because of the large number of very stable standards that L2 maintains, L2 has driven an administrative innovation within NCITS. Standards can now be declared “stabilized,” meaning that they do not need to be continually re-affirmed.

### 4.3 Future Plans

A new 8-bit character set, 8859-16, has been proposed for Romanian, but L2 intends to oppose other new 8-bit standards. L2 prefers that new characters be added to ISO 10646 instead.

### 5. Relevance and Relationship to H2

Our SQL standards, ISO 9075, contain these normative references to the L2-developed standards:


In my contacts with L2/UTC members, I heard a variety of concerns.

One is that ISO 9075 refers to an out-of-date version of ISO 10646 (the current version is 2000). The SQL standard should be examined to update terminology and language where needed. Reference [1] was cited as a good source of discussion on terminology, though I’m told the terminology is based on Unicode 3.0, which has since been superceded.

Another was a concern that SQL vendors are not implementing Unicode support in a consistent way. SQL implementations differ in their approach to normalization, for example. I heard differing views on what kinds of normalization would be appropriate within a SQL engine, so this is a matter for further exploration.

Another concern is with SQL’s definition of identifiers. One member pointed me to XML as an example that aligns well with Unicode. One issue here is rules for case-folding. There is no adequate definition in an existing ISO standard. It may make sense to refer to Unicode data files instead.

This should not be considered a complete list. I was told of other L2/UTC members who could not be present that day that also had database issues.

I suggested that one way to proceed is to have an overview presentation within L2/UTC of current character set features in ISO 9075. A complete list of concerns could be gathered from this exercise. I volunteered to
put together such a presentation for the next L2 meeting, to be held May 21-24 in Pleasanton, CA. (Though because of conflicting travel plans, I will have to make this presentation by teleconference.) I would of course like H2’s input into the content of such a presentation. I plan to post a draft to H2 before the April H2 meeting.

Other L2-related standards and reports that may have future relevance to our work are:

TR 10176 — Guidelines for the preparation of programming language standards
TR 11017 — Framework for Internationalization
IS 14651:2000 — International string ordering

As a final matter, I ask H2 to consider the idea of making the L2 liaison also a formal liaison to UTC. Since I’m not an expert on NCITS by-laws, I can’t make a specific recommendation. But, I’ll formulate a possible motion below to start discussion:

Motion: That the H2 liaison to L2 also be considered an H2 liaison to the Unicode Technical Committee, given approval by the Unicode Technical Committee.

6. Contact Information

Arnold Winkler, the L2 chair and UTC vice-chair, can be contacted by e-mail at arnold.winkler@uni-sys.com. Lisa Moore, the L2 vice-chair and UTC chair, can be contacted by e-mail at lisam@us.ibm.com.

- End of paper -