Liaison report from SC2 to SC22
Character Set Standardization

August 2, 2001

In this report I will document the current status of character set standardization

1. Administrative information:

SC2 has currently 2 working groups, SC2/WG2 for the Universal Character Set (UCS), and SC2/WG3 for 8-bit character sets. SC2/WG2 has also the Ideographic Rapporteur Group (IRG) as advisor for related issues.

Chair of SC2:	Prof. Kohji Shibano, Japan
Convenor of SC2/WG2:	Mike Ksar, USA
Convenor of SC2/WG3:	Evangelos Melagrakis, Greece
Secretariat of SC 2:	Toshiko KIMURA

IPSJ/ITSCJ (Information Processing Society of Japan/Information Technology Standards Commission of Japan)*
Room 308-3, Kikai-Shinko-Kaikan Bldg., 3-5-8, Shiba-Koen, Minato-ku, Tokyo 105 JAPAN
Tel: +81 3 3431 2808; Fax: +81 3 3431 6493; E-mail: kimura@itscj.ipsj.or.jp; http://www.dkuug.dk/jtc1/sc2
SC2 documents are at http://lucia.itscj.ipsj.or.jp/servlets/ScmDoc10?Com_Id=02
*A Standard Organization accredited by JISC

2. Character set technology and standardization

2.1. ISO/IEC 10646-1:2000 – second edition

The second edition of ISO/IEC 10646-1:2000 has been published, it is available electronically on a CD. The repertoire of 10646-1 is equivalent to Unicode 3.0, the same code charts are used in both standards.

ISO 10646 is the only standard developed by SC2/WG2. It is intended as the universal character set, and is now seeing widespread implementation both as an interchange code and as a processing code on many platforms, in databases, and in many other applications.

ISO 10646-1 is used as the basis for many new standards activities, including internet and web standards by the W3C (World Wide Web consortium), the IETF (Internet Engineering Task Force), ECMA (European Computer Manufacturing Association), many JTC1 subcommittees, the Unicode Consortium, and other industry consortia.

Because of the universal nature of the character set in ISO 10646, the relationship between character encoding and character semantics is somewhat different for 10646 than for all other SC2 character encoding standards. SC2/WG2 specifies some character properties normatively as part of 10646, and the de facto implementations of 10646 based on the additional recommendations of the Unicode Standard go even further in connecting character properties firmly to the character definitions in the standard. SC22 committees need to take this change in how character standards are being viewed and developed into account when dealing with 10646.

Furthermore, because of the growing need for implementers to have good programming language support for 10646, the programming language standards need to find ways to embrace the universal character set in future revisions. Non-ISO specifications such as those for Java and XML are much further advanced than most SC22 programming languages in their adaptation to 10646.

2.2. ISO/IEC 10646-2

ISO/IEC 10646-2 codes characters in the Planes 1, 2, and 14 of 10646.
ISO/IEC FDIS 10646-2 has been approved, the standard will be published soon.

The new planes are:

- Supplementary Multilingual Plane (SMP) U+10000..U+1FFFF
- Supplementary Ideographic Plane (SIP) U+20000..U+2FFFF
- Supplementary Special-purpose Plane (SSP) U+E0000..U+EFFFF

The **Supplementary Multilingual Plane**, or Plane 1, contains several historic scripts, and several sets of symbols: Old Italic, Gothic, Deseret, Byzantine Musical Symbols, (Western) Musical Symbols, and Mathematical Alphanumeric Symbols. Together these comprise 1594 newly encoded characters.
The **Supplementary Ideographic Plane**, or Plane 2, contains a very large collection of additional unified Han ideographs known as Vertical Extension B, comprising 42,711 characters, as well as 542 additional CJK Compatibility ideographs.
The **Supplementary Special-purpose Plane**, or Plane 14, contains a set of tag characters, 97 in all.

The repertoire of ISO/IEC 10646-2 has been added to the Unicode Standard to define **Unicode 3.1**

2.3. ISO/IEC 10646-1:2000, Amendment #1

The first amendment to the second edition of ISO 10646-1:2000 (BMP) is in the final process of approval. This amendment adds characters to the BMP, mainly:

- 500 mathematical symbols, as recommended by the Mathematical Society and the Mathematical working group of the W3C
- 14 additional ZAPF Dingbats characters
- 4 additional Recycling Symbols, and
- many additional symbols needed for inter-working with the new Japanese standard JIS X 0213

The repertoire of Amendment #1 will be added to the Unicode Standard to define **Unicode 3.2**

2.4. ISO/IEC 8859 family of 8-bit character set standard

All ISO/IEC 8859-x standards have been revised to synchronize the character names with the ones in ISO/IEC 10646.

Currently existing members of the 8859 family:

	Name	**Used in:**
8859-1	Latin alphabet no. 1	English countries, Western Europe, South America
8859-2	Latin alphabet no. 2	Eastern Europe, former Yugoslavia
8859-3	Latin alphabet no. 3	Esperanto, Malta, South Africa, Catalan, Turkey
8859-4	Latin alphabet no. 4	Scandinavia, Estonia, Greenland, Latvia, Lithuania
8859-5	Latin/Cyrillic alphabet	Bulgaria, former USSR, Macedonia, Serbo-Croatia
8859-6	Latin/Arabic alphabet	Arabic countries
8859-7	Latin/Greek alphabet	Greece
8859-8	Latin/Hebrew alphabet	Israel, Hebrew script
8859-9	Latin alphabet no. 5	Western Europe, Turkey, Faroese
8859-10	Latin alphabet no. 6	Scandinavia, including Sámi (Lappish)
8859-11	Latin/Thai	Thailand
8859-12	unassigned	
8859-13	Latin alphabet no. 7	Baltic Rim countries
8859-14	Latin alphabet no. 8	Celtic
8859-15	Latin alphabet no. 9	Modified part 1 for the EURO and additional characters for Finnish and French
8859-16	Latin alphabet no. 10	Romania

2.5. ISO/IEC 6937 – Coded graphic character set for text communication, Latin alphabet

This 8-bit character set, that allows specified combining characters, is currently being revised, mainly to add the EURO and also to synchronize the character names with ISO/IEC 10646. The standard is in FDIS ballot and will most likely be approved.

3. Additional information

3.1. Unicode Technical Reports

I am including this information about the Unicode Technical Reports, because many implementations quote the Unicode standard for compliance. Unicode Technical Reports contain valuable supplementary information that complements WG2's work of defining characters and their coding.

All Unicode Technical Reports can e found at http://unicode.org/unicode/reports/index.html

The UTC (Unicode Technical Committee) has decided to classify the UTRs into 3 groups:

> **A Unicode Standard Annex (UAX)** forms an integral part of the Unicode Standard, carrying the same version number, but is published as a separate document. Note that conformance to a version of the Unicode Standard includes conformance to its Unicode Standard Annexes.
>
> **A Unicode Technical Standard (UTS)** is an independent specification. Conformance to the Unicode Standard does not imply conformance to any UTS. Each UTS specifies a base version of the Unicode Standard. Conformance to the UTS requires conformance to that version or higher.
>
> **A Unicode Technical Report (UTR)** may contain either informative material or normative specifications, or both. Each UTR may specify a base version of the Unicode Standard. In that case, conformance to the UTR requires conformance to that version or higher.

AFW