

Source: Mike Ksar
Title: Liaison report from SC2 to SC22
Action: For information to SC 22 plenary
Distribution: ISO/IEC JTC 1/SC 22
Date: 2003-08-19

This report documents 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 a sub-group of SC2/WG2.

- Chair of SC2: Prof. Kohji Shibano, Japan
- Convener of SC2/WG2: Mike Ksar, USA
- Project Editor: Michel Suignard, USA
- IRG Rapporteur: Zhang Zouchai, China
- Convener of SC2/WG3: Evangelos Melagrakis, Greece
- Secretariat of SC 2: Toshiko Kimura, IPSJ/ITSCJ, Japan

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;

SC2 web site is at <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. SC2/WG2 - Amendments to ISO/IEC 10646-1:2000 and ISO/IEC 10646-2:2001

Since the publication of ISO/IEC 10646-1:2000 and ISO/IEC 10646-2:2001 one more additional amendment was published for 10646-1 on 2001-11-01 and two are in FDAM stage, Amendment 2 for 10646-1 and Amendment 1 for 10646-2. It is expected that the latest two FDAM ballots will close in mid-October 2003. ISO Central Secretariat has not moved on the processing of these two amendments for several months because ISO Central Secretariat claimed they had printing problems even though the SC2 Secretariat and other national bodies were able to print them without any problem.

Here are the two last FDAM ballots:

Ref. No.	Title	Due Date
----------	-------	----------

ISO/IEC 10646-1: 2000/FDAM 2 Doc No: SC 2 N 3671	ISO/IEC 10646-1: 2000/FPDAM 2, Information technology -- Universal Multiple-Octet Coded Character Set (UCS) -- Part 1: Architecture and Basic Multilingual Plane AMENDMENT 2: Limbu, Tai Le, Yijing and other characters	2003-10-15?
ISO/IEC 10646-2/FDAM 1 Doc No: SC 2 N 3673	ISO/IEC 10646-2: 2001/FPDAM 1, Information technology -- Universal Multiple-Octet Coded Character Set (UCS) -- Part 2: Supplementary Planes AMENDMENT 1: Aegean, Ugaritic, and other characters	2003-10-15?

Both of these amendments add additional scripts to the repertoire of ISO 10646 to the Basic Multilingual Plane and to the Supplemental Planes.

SC2/WG2 is also working on publishing ISO 10646 as a single standard combining parts 1 and 2 into a single document before the end of 2003. This publication is to be known as ISO/IEC 10646: 2003. All changes to create the merged document are editorial in nature. The repertoire of the merged ISO 10646 parts 1 and 2 is fully synchronized with the latest publication of Unicode 4.0.

The URL for the program of work of SC2 is at: <http://www.dkuug.dk/JTC1/SC2/open/pow.htm>

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 programming languages, in databases, and in many other applications.

ISO 10646 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, 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. SC2 has been normatively defining some of these properties and is cooperating with the much more extensive efforts by the Unicode Consortium to continue to do so. Now that SC2 is much more concerned about character properties for ISO 10646, it is now emphasizing the need to define all character properties in contributions requesting additional repertoires. There is a document on the SC2/WG2 web site which contains the [Principles and Procedures](#) for Allocation of New Characters and Scripts and handling of Defect Reports on Character Names. A [form for submitting proposals for addition to the repertoire of ISO/IEC 10646](#) is also available. In addition, experts who submit script proposals should also consult the [roadmap documents](#) that show the placement of currently encoded scripts and candidate scripts for possible encoding in the standard.

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. It is noteworthy to see that new versions of COBOL and Fortran provide support for ISO 10646. SC2 is also aware of extensive efforts in C and C++ to do the same. ISO specifications, such as C#, and non-ISO specifications such as Java, HTML and XML are much further advanced than most SC22 programming languages in their adaptation to 10646.


The total repertoire of ISO/IEC 10646 now contains 96,447 characters of which 96,248 are graphic characters, 134 are format and 65 are control. Again the character repertoire of ISO 10646 corresponds exactly to Unicode 4.0.

2.2. Past and future meetings of SC2/WG2:

- Meeting 41 – Singapore; October 2001
- Meeting 42 – Dublin, Ireland; May 2002 – Resolutions at URL: <http://www.dkuug.dk/JTC1/SC2/WG2/docs/n2454r.pdf>
- Meeting 43 – Tokyo, Japan; 9-12 December 2002 – Resolutions at URL: <http://www.dkuug.dk/JTC1/SC2/WG2/docs/n2554.htm>
- Meeting 44 – San Francisco Bay Area, U.S.; 20–23 October 2003
- Meeting 45 – Europe (tentative – call for host sent) – US (Backup); June/July 2004

2.3. SC2 ballot – Transfer of ISO/IEC 14651 from SC22 to SC2/WG2

An SC2 ballot to accept the transfer of ISO/IEC 14651 from SC22/WG20 to SC2/WG2 started in 2003-07-01 based on the recommendations of the JTC1 ad hoc meeting, attended by SC2, SC22, SC35, and SC35 and several national body representatives. (SC2 ballot results will be available in early September after the SC2 ballot closes on 2003-09-01). Here is the ballot posting as it appears in SC2's web site:

Doc No: SC 2 N 3687 	Transfer of Project 14651 (JTC 1.22.30.02.02) from JTC 1/SC 22 to SC 2	2003-09-01
---	--	------------

2.4. SC2/WG3 - ISO/IEC 8859 family of 8-bit character set standard

ISO/IEC 8859-7 (Latin/Greek) was successfully balloted on 2003-07-28 and a new edition will be published by ISO Central Secretariat (ITTF) shortly.

ISO/IEC 2375 – Registry of 8-bit character sets has been approved and was published in February 2003 by ISO Central Secretariat (ITTF).

The program of work of SC2/WG3 projects is available at the following URL: <http://www.dkuug.dk/JTC1/SC2/open/pow.htm>

2.5. *Past and future meetings of SC2/WG3:*

The last time SC2/WG3 met was in October 2000. Currently there are no plans for future meetings.

3. **Additional information**

3.1. *Unicode 4.0*

The Unicode Consortium recently announced the availability of Unicode 4.0. It is a major release and is available in hard copy as well as on-line. Many SC2/WG2 members reviewed and contributed to the development and publication of Unicode 4.0. SC2/WG2 and the Unicode Consortium continue to cooperate effectively in ensuring the complete synchronization of the repertoire of both publications. Here is the URL that gives a summary of Unicode 4.0 and its various components:

<http://www.unicode.org/versions/Unicode4.0.0/>.

Here is the final draft of press release announcement as of August 19, 2003:

Mountain View, CA, August dd, 2003 -- The Unicode® Consortium announces publication of The Unicode Standard, Version 4.0, the fundamental IT specification for the representation of text. Widely used in modern software, Unicode enables internationalized domain names in all geographies, and allows increasingly more locations and cultures to have on-line global accessibility, putting them on the right side of the digital divide.

Version 4.0 encodes twice as many characters, strengthening Unicode support for worldwide communication, software availability, and specialized publishing. The text has been extensively rewritten, and incorporates specifications that were previously only available as separate documents. The clarified specification of conformance requirements incorporates the most highly developed character encoding model in existence.

Record-breaking character content

Version 4.0 encodes over 96,000 characters, twice as many as Version 3.0, and includes two record-breaking collections of encoded characters. The largest encoded character collection for Chinese characters in the history of computing has doubled in size yet again to encompass over 2000 years of Chinese, Japanese, Korean, and Vietnamese scholarly and literary usage, including all the main classical dictionaries of these languages. Version 4.0 also encodes the largest set of characters for mathematical and technical publishing in existence.

Eliminating the "digital divide"

The Unicode Consortium, the major consortium of computer companies that create software for world-wide use, has the goal of extending the Unicode Standard to meet the needs of linguistic minority communities and academic specialists, to prevent their threatened permanent exile on the wrong side of the digital divide. The ideal is for all major software, right out of the box, to provide the necessary specialized characters and writing systems, so that the expensive custom software or ugly hacked-up ASCII text workarounds currently used can be replaced with mainstream technology.

Minority indigenous scripts added in Version 4.0 include Limbu from Nepal, Tai Le from Southeast Asia, Osmanya from Somalia, and several from the Philippines. A large set of symbols used for Western musical notation is complemented by a set of

symbols for Byzantine music. Newly-added archaic scripts include Old Italic, Gothic, Linear B, Cypriot, and Ugaritic.

Technical enhancements

Technical enhancements in Version 4.0 include new specifications for text boundaries and casing, the encoding of supplementary characters, and a major expansion of the properties defined by the Unicode Character Database. Formalized policies for the stability of the Unicode Standard are defined. The semantics of special characters, including the byte order mark, have been clarified. Version 4.0 is fully synchronized with the third version of International Standard ISO/IEC 10646.

Version 4.0 is published by Addison-Wesley (ISBN 0-321-18578-1), and is available from the Unicode Consortium or through the book trade. The text and code charts of Version 4.0 are also available on the Consortium's Web site www.unicode.org.

About the Unicode Consortium

The Unicode Consortium is a non-profit organization founded to develop, extend and promote use of the Unicode Standard, which specifies the representation of text in modern software products and standards.

Members of the Consortium are a broad spectrum of corporations and organizations in the computer and information processing industry. Full members are: Adobe Systems, Apple Computer, Basis Technology, Government of India (Ministry of Information Technology), Government of Pakistan (National Language Authority), HP, IBM, Justsystem, Microsoft, Oracle, PeopleSoft, RLG, SAP, Sun Microsystems, and Sybase.

Membership in the Unicode Consortium is open to organizations and individuals anywhere in the world who support the Unicode Standard and wish to assist in its extension and implementation.

For additional information on Unicode, please contact the Unicode Consortium.