ISO/IEC JTC 1/SC 2/WG 3
7-bit and 8-bit codes and their extension
SECRETARIAT : ELOT

DOC TYPE :	Proposed NP
TITLE :	NP for TR on Character set model
SOURCE :	DS
PROJECT:	NP
STATUS :	NB contribution
ACTION ID :	For discussion in WG3 and SC2 letter ballot
DUE DATE :	-----
DISTRIBUTION :	P, O and L Members of ISO/IEC JTC 1/SC 2 WG Conveners, Secretariats WG 3 Members ISO/IEC JTC 1 Secretariat ISO/IEC ITTF
MEDIUM :	P, Open*
NO OF PAGES :	12

Contact 1: Secretariat ISO/IEC JTC 1/SC 2/WG 3 ELOT Mrs K.Velli (acting)
Acharnon 313, 111 45 Kato Patissia, ATHENS – GREECE
Tel: +30 1 22 80 001 Fax : +30 1 22 86 219 E-mail : kkb@elot.gr

Contact 2 : Convenor ISO/IEC JTC 1/SC 2/WG 3 Mr E.Melagrakis
Acharnon 313, 111 45 Kato Patissia, ATHENS – GREECE
Tel: +30 1 22 80 001 Fax : +30 1 22 86 219 E-mail: eem@elot.gr

NP FOR A TR ON CHARACTER SET MODEL

1. TITLE

Character set model

2. SCOPE

To specify one or more models on how character related terms relate to each other based on SC2 standards.

3. PURPOSE AND JUSTIFICATION

A Danish comment on the NP ballot on TR 15285 recommended a clause on a model of character sets related terms.

SC 2/WG 2 recommended in its disposition of comments that a separate TR be produced on this issue.

SC 22/WG 20 has been working on a character set mo9del in order to provide support in APIs for character set related information, so that programming languages can utilize character set information in a consistent way.

DEN/TC 304 has been working on a model for character set transformation.

The C liaison IETF has been working on character set models for communication on the Internet.

So that SC 2 standards can be used in a coherent manner it is much needed that a model be present to quide the use of SC2 standards in other standardization for a.

4. PROGRAMME OF WORK

A technical report is expected to be developed for this project.

5. RELEVANT DOCUMENTS TO BE CONSIDERED

ISO/IEC 8859
ISO/IEC 10646
ISO/IEC 10367
ISO/IEC 2022
ISO/IEC 6429
ISO/IEC TR 11017 Framework for Internationalization
IETF RFC 2130
CEN/TC 304 project 9.1

6. COOPERATION AND LIAISON

ISO/IEC JTC1/SC2/WG2
All ISO/IEC JTC1/SC22 WGs
Other ISO/IEC JTC1 SCs where appropriate
CEN/TC304/WG3
DEN/TC304/WG4
IETF

7. PREPARATORY WORK OFFERED WITH TARGET DATE (s)

A PDTR document will be ready for registration and ballot 12 months after approval of the project by JTC1.

A DTR document will be ready 20 months after approval of the project.

The TR will be ready for publishing 26 months after approval of the project.

Keld Simonsen (Denmark) is the proposed editor for this standard.

There will not be a need for a maintenance agency nor a registration authority for this project.

There are no known requirements for coding for this proposed technical report.

The proposed technical report does not concern known patented items.

Reference number of working document : **ISO/IEC JTC1/SC2/WG 3 N**

Date : 1998-09-14

Reference number of document : **ISO/IEC WDO NP CSM**

Committee identification : ISO/IEC JTC1/SC2

Secretariat : JITSC

Information technology – Character set models

Technologies de I'information – Modelles pour de jeux des characteres

Document type : International standard
Document subtype : if applicable
Document stage : (40) Enquiry
Document language : E

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

The main task of a technical committee is to prepare International Standards but in exceptional circumstances, the publication of a Technical Report of one of the following types may be proposed :

- type 1, when the required support cannot be obtained for the publication of an International Standard, despite repeated efforts ;

- type 2, when the subject is still under technical development or where for any other reason there is the future but not immediate possibility of an agreement on an International Standard;

- type 3, when a technical committee has collected data of a different kind from that which is normally published as an International Standard (“state of the art”, for example).

Technical Reports of types 1 and 2 are subject to review within three years of publication, to decide whether they can be transformed into International Standards. Technical Report of type 3 do not necessarily have to be reviewed until the date they provide are considered to be no longer valid or useful.

Technical Report ISO/IEC CSM was prepared by Joint Technical Committee ISO/IEC JTC 1, “Information Technology”, subcommittee 2, “Character sets and their encoding”.

Suggestions and comments for improvement of this document are welcome. They should be sent to :

Keld Simonsen
Sankt Jørgens Alle 8
DK-1615 Copenhagen V
Email : keld@dkuug.dk

CONTENTS

Page

TECHNICAL REPORT

Information technology – Character set models

1 Scope

This Technical report specifies one or more models on how character related terms relate to each other based on ISO/IEC standards.

2 Conformance and testing

As this specification is a Technical Report, there cannot be any conformance claimed to this Technical Report.

3 References

The following normative documents contain provisions which, through reference in this text, constitute provisions of this Technical Report. For dated references, subsequent amendments to, or revisions of, any of these publications do not apply. However, parties to agreements based on this Technical Report are encouraged to investigate the possibility of applying the most recent editions of the normative documents indicated below. Or undated references, the latest edition of the normative document referred to applies. Members of ISO and IEC maintain registers of currently valid International Standards.

ISO/IEC 2022, “Information Technology – ISO 7- bit and 8-bit coded character sets – Code extension techniques”.

ISO 2375, “Data processing – Procedure for registration of escape sequences”.

ISO/IEC 6429, “ Information Technology – Control functions for coded character sets”.

ISO/IEC 6937, “ Information Technology – Coded graphic character set for text communication – Latin alphabet”.

ISO/IEC 8859, “ Information Technology - 8-bit single-byte coded graphic character sets (all parts)”.

ISO/IEC 9945, “Information Technology – Portable Operating System Interface (POSIX (all parts))”.

ISO/IEC 10646, “Information technology – Specification Method for Cultural Conventions”.

ISO/IEC 15897: 1998, “ Information Technology – Procedures for Registration of Cultural Elements”.

4 Definitions and abbreviations

For the purpose of this technical report the following definitions and abbreviations apply.

4.1 character : a member of a set of elements used for the organization, control, or representation of data, (10646)

NOTE : Characters are conveyers of some sound or other symbolic representation of some meaning, such as special characters, or control characters. Characters do not have an encoding to them, some also call them abstract characters.

4.2 control character : A control function the coded representation of which consists of a single bit combination (6937)

4.3 control function : An element of a character set that controls the recording, processing, transmission or interpretation of data, and that has a coded representation as one or more bit combinations (6937)

4.4. graphic character : A character, other than a control function, that has a visual representation normally handwritten, printed or displayed (10646)

4.5 character set : A specified set of characters.

4.6 coded character set : A set of unambiguous rules that establishes a character set and the relationship between the characters of the set and their coded representation. (10646)

NOTE - 1. In ISO 2022 this known as a G0/G1/G2/G3 or C0/C1 set for graphic and control characters respectively.T he control CCS are 32 cells each, while the G0 is 94 and G1 is 96 cells respectively for 7/8 bit CCS. There are also a size for 14-bit CCS, and also specific sizes for special CCS of 256 cells. This could be called a simple coded character set.

NOTE - 2. In the ISO 8859 series a CCS can consist of, for example, two registered CCS of 2375. For example ISO 8859-1 is defined as ISO 2375 register number 6 and number 100- it does not have control characters. ISO 646 IRV (US-ASCII) is defined as the graphical CCS registration number 6 and the control CCS registration number 1. ISO 6937 is defined as ISO 2375 registration number 2 and then registration 101, with some specific rules for combinations of two octets into one character. This could be called a composite coded character set, as it may consist of several simple coded character sets.

NOTE - 3 In ISO 9945 (POSIX) standards, a CCS is taken to be a description of the whole datastream, that is including both the control characters and the graphic characters. This is close to the MIME “charset” definition, but POSIX does not (yet) have specification techniques to define state-dependent encodings like 2022 based encodings. POSIX can handle ISO 6937 and Shift-JIS and UTF-8 with its charmap specification technique. This could be called a fully coded character set, as all bit combination are defined.

4.7 encoding : the relation from the binary representation via coded character sets to (abstract) characters. The encoding defines the meaning of a binary data stream. It can consist of more than one coded character set, and an encoding scheme can be applied to regulate how these coded character sets are encoded. Also symbolic characters can be encoded in the encoding.

4.8 encoding scheme : A set of unambigeous rules that establishes a character set and the relationship between the characters of the set and their coded representation, by combining more simple coded character sets or encoding a simple coded character set in another way. Examples are 2022 or UTF-8, the 2022 coding system works on one or more simple CCS as registered with ISO 2375. The procedures of forming characters in ISO 6937 is not a coding system, this is a (composite) coded character set.

4.9 repertoire : A specified set of character that are represented in a coded character set. (10646)

NOTE - 4 A repertoire contains a finite set of elements in the form of chracters.

4.10 symbolic character : a character represented by a name or nu7mber. Examples are html/sgml character entities like õ and É RFC 1345 mnemonics like &a: TeX and groff character names.

4.11 transfer-encoding (MIME) : a general transformation applicable to a binary stream, to obtain some specific properties of the byte stream. Examples are base64, uuencode, quoted-printable. UTF-8 is not a transfer-encoding.

5 Relation between a character repertoire and its encoding

The encoding of a character repertoire can consist of at least three parts :

1. The (simple, composite or full) coded character sets, for example ISO/IEC 8859-1 combined with the control character set of ISO/IEC 6429.
2. The rules for combing or coding one or more simple coded character sets, for examp0le ISO/IEC 2022 or UTF-8.
3. A symbolic character notation, like SGML entities of the type $aacute;

On top of the encoding, general transformation schemes that are applicable to any binary representation may be applied. These generally applicable schemes include:

- compression schemes, (examples : zip, gzip)
- encryption schemes, (examples PGP)
- safer passage schemes, such as avoiding the 8^{th} bit or byte values 0-31, (Examples : base64, uuencode)

Bibliography

- RFC 1345 : Character Mnemonics and Character Sets. K. Simonsen, July 1992
- RFC 2045 : Multipurpose Internet Mail Extensions (MIME) Part one : Format of Internet message Bodies. N. Freed, N. Borenstein, November 1996
- RFC 2070 : Report from the IAB meeting on treference for character sets.
- IBM CDRA
- C3