INTERNATIONAL ORGANIZATION FOR STANDARDIZATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC 1/SC 2/WG 2

Universal Multiple-Octet Coded Character Set
(UCS)

ISO/IEC JTC 1/SC 2/WG 2 *N 2696*
2004-01-22

Title:	Presentation Foils from National Workshop on Unicode, New Delhi, Sept 24-26, 2003
Source:	V.S. Umamaheswaran – umavs@ca.ibm.com
References:	
Action:	For information to WG2
Distribution:	ISO/IEC JTC 1/SC 2/WG 2

At the request of our convener Mr. Mike Ksar, I have packaged the set of foils (modified slightly) that I had presented at the National Workshop on Unicode, New Delhi, Sept 24-26, 2003, organized by the Ministry of Information and Communication Technology, India. Some of you involved with JTC1/SC2/WG2 and the Unicode Technical Committee may find it of some use.

In particular, slide number 4 of the second presentation – on page 14 – titled 'Framework for Discussion' was also used in WG2 meeting M44 during our ad hoc on Tibetan. It is a gist of the principles to follow while proposing additions or changes to the standard.

Unicode and ISO/IEC 10646

V.S. Umamaheswaran
umavs@ca.ibm.com
IBM Toronto Lab, Canada

Topics

- **Unicode and ISO/IEC 10646**
- **UCA and 14651**
- **Processes**
- **Guidelines for Proposals**
- **Organize the Expertise**

Unicode and ISO/IEC 10646

	Unicode	10646
Code Space	0 to x10FFFF	0 to x10FFFF*
Repertoire	Same	Same
Supp. Planes	Same	Same
BMP non CJKV	Same	Same
BMP CJKV	*Single Col*	*CJKV Cols*
Chart Creation	*Common DB*	*Common DB*

Unicode and ISO/IEC 10646

Row/Cell Hex code	C G- Hanzi	-T	J Kanji	K Hanja	V ChuNom
078/000	——	——	——	——	——
4E00	0-523B 0-5027	1-4421 1-3601	0-306C 0-1676	0-6C69 0-7673	1-2121 1-0101

Unicode and ISO/IEC 10646

	Unicode	10646
Publication	Web; Book Dot Release	Edition + Amds (1 volume end of 2003)
	Book Style	ISO Style
Conformance	=Level 3	Levels 1, 2, 3 (*use 3 for Indic*)
BiDi	Defined	*Refers to Unicode*
Normalization	Defined	*Refers to Unicode*

Unicode and ISO/IEC 10646

	Unicode	10646
Combining	Property + TRs+ Text	List + Minimal Info
Format Chars	Property	Some Listed
Script Info	Lot of Detail	Minimal
Annotations	Many more	Some in Annex
Naming Rules	*uses 10646*	Defined

Unicode and ISO/IEC 10646

	Unicode	10646
Properties + Processing Rules	Defined	*Out of scope*
UTF-8,-16, -32/UCS4	Same	Same
Compressions	Defined	*Not included*
….	…..	…..

Conforming to Unicode will automatically conform to 10646 Level 3 plus lots more

Unicode Collation Algorithm and ISO/IEC 14651

- Synchronized with Each Other
- Share same Concepts for Weights Categories and Tailoring
- Tailoring Required in Both
- Default Weights and Repertoire Identical in Both – generated from the same data base
- 14651 Editions + Amds versus UCA Versions

Conforming to UCA will also conform to 14651 plus more functions

Processes

	In process within UTC				Further progression
Stage ⇒	Initial proposal	Provisional acceptance	Final acceptance	Pipeline and Unicode Database	Beta Publication Web plus Book Database

WG2 Stage ⇒	Initial proposal	Provisional acceptance	Final Acceptance	WG2 Bucket	SC 2 Ballot	JTC 1 Ballot	ITTF Publication
	In process within WG2					Further progression	

Processes

	In process within UTC				Further progression
Stage ⇒	Initial proposal	Provisional acceptance	Final acceptance	Pipeline and Unicode Database	Beta Publication Web plus Book Database

2 Ballots Draft, Final 12-18 months

WG2 Stage ⇒	Initial proposal	Provisional acceptance	Final Acceptance	WG2 Bucket	SC 2 Ballot	JTC 1 Ballot	ITTF Publication
	In process within WG2					Further progression	

Processes

- UTC has additional procedures for preparing and processing Technical Reports
- See FAQ page at Unicode site

Processes

- Membership in SC2
  - National Bodies
    - Ex: INCITS in USA, SCC in Canada, BIS in India
    - Roster on SC2 site www.dkuug.dk/JTC1/SC2
- Membership in UTC
  - Review by all members and experts
  - Voting by Corporate Members
    - Government of India is a Corporate Member
    - Roster on Unicode site.

Proposal Guidelines

Do your homework

? Check if Already encoded ?

(see http://www.unicode.org/standard/where/)

- Check Charts in Unicode V4
- Also charts in TRs –
  - TR15 Normalization charts
  - TR10 Collation charts
  - TR21 Case map charts
  - TR24 Script charts
- or for legacy sets ICU Charmaps or equivalents

Proposal Guidelines

- May be in a block with recognized name ..
- Search Nameslist file in Unicode Database

  Name could be in Annotations
- Shape in standard can be a variant

  (see handout page 2)
- Is it a Glyph (from a Font for example?)

  http://www.unicode.org/reports/tr17/#Characters vs. Glyphs

  and TR 15285 – Character Glyph Model

  http://isotc.iso.ch/livelink/livelink/fetch/2000/2489/Ittf_Home/PubliclyAvailableStandards.htm??Redirect=1

Proposal Guidelines

- Character may be under consideration
  - Look in Unicode Pipeline
    http://www.unicode.org/alloc/Pipeline.html
- Check if previously considered and rejected -
    http://www.unicode.org/alloc/rejected.html
- Also for any accepted pending scripts:
    http://www.unicode.org/pending/pending.html

Proposal Guidelines
Do your homework

For entire script - check out the *ROADMAPS*:

http://www.unicode.org/roadmaps
http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html

Already encoded- **Bold text** in Roadmap
proposal accepted
- **(Bold text between parentheses)**
under consideration (Text between parentheses)
exploratory ¿Text between question marks?
possible future – no suggestions ???
hot links for latest proposal included

http://www.unicode.org/roadmaps/bmp/

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
00		Basic Latin								Latin 1 Supplement						
01	Latin Extended-A								Latin Extended-B							
02	Latin Extended-B				IPA Extensions				Spacing Modifiers							
03	Combining Diacritics								Greek							
04	Cyrillic															
05	Cyrillic Sup.				Armenian						Hebrew					
06	Arabic															
07	Syriac				(Arabic Ext.)				Thaana				¿N'Ko?			
08	(Avestan and Pahlavi)					¿Mandaic?			(Aramaic)		(Tifinagh)				¿Samaritan?	
09	Devanagari								Bengali							
0A	Gurmukhi								Gujarati							
0B	Oriya								Tamil							
0C	Telugu								Kannada							
0D	Malayalam								Sinhala							
0E	Thai								Lao							
0F	Tibetan															
10	Myanmar										Georgian					
11	Hangul Jamo															
12	Ethiopic															
13	Ethiopic								(Eth.Ext.)		Cherokee					
14	Unified Canadian Aboriginal Syllabics															

Proposal Guidelines

Do Your Homework

? Can the character be represented as sequences ?
Remember no Duplicate Representation

- Indic conjuncts fall into this category
- Check out Chapter 9 of Unicode 4.0 (Examples in handout last 3 pages)
- http://www.unicode.org/standard/where/ , and
- http://www.unicode.org/faq/char_combmark.html

Proposal Guidelines

Other proposals may exist elsewhere in draft form especially with archaic / minority scripts

Ex: Kharoshthi, Brahmi, Surashtrian .. proposals

Ask / network on the public discussion lists
http://www.unicode.org/consortium/distlist.html

indic@unicode.org is set up for Indic

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC 1/SC 2/WG 2

Universal Multiple-Octet Coded Character Set
(UCS)

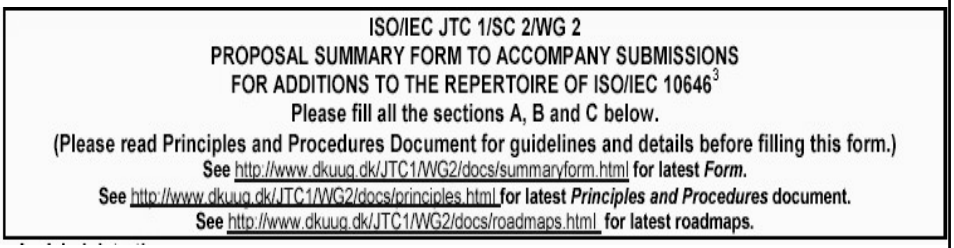
ISO/IEC JTC 1/SC 2/WG 2 **Use Latest**

Title:	Principles and Procedures for Allocation of New Characters and Scripts and handling of Defect Reports on Character Names (Replaces N2352, N 2002 and N1876)
Source:	Ad hoc group on Principles and Procedures (Edited by: V.S. Umamaheswaran)
References:	See references section in the document
Action:	To be considered by SC 2/WG 2 and all potential submitters of proposals for new characters the repertoire of ISO/IEC 10646, and for new collection identifiers
Distribution:	ISO/IEC JTC 1/SC 2/WG 2, ISO/IEC JTC 1/SC 2 and Liaison Organisations

www.dkuug.dk/JTC1/SC2/WG2/principles.html

Annex A: Information Accompanying Submissions
Annex F: Formal criteria for disunification
Annex G: Formal criteria for coding precomposed characters
Annex H: Criteria for encoding symbols

WHEN YOU ARE CERTAIN A NEW PROPOSAL IS WARRANTED



ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[3]
Please fill all the sections A, B and C below.
(Please read Principles and Procedures Document for guidelines and details before filling this form.)
See http://www.dkuug.dk/JTC1/WG2/docs/summaryform.html for latest *Form*.
See http://www.dkuug.dk/JTC1/WG2/docs/principles.html for latest *Principles and Procedures* document.
See http://www.dkuug.dk/JTC1/WG2/docs/roadmaps.html for latest roadmaps.

Prepare the Proposal Summary Form

www.dkuug.dk/JTC1/SC2/WG2/summaryform.htm

Proposal Guidelines

Proposal Summary Form

- Contains several questions to be answered
- See *Submitter's Responsibilities* in Form
- Most related to the previous checking steps
- Additional Information to assist in evaluation by UTC and WG2
  - Unicode Properties, Evidence of use, References
  - Information about submitters & others consulted
  - Preferred location, Glyphs/Font for publications

Facilitates evaluation by UTC, WG2 and other experts worldwide

Organize the Experts

Some Observations / Suggestions

- Workshops are Educational
- Formal review and *Consensus Process* helps in consolidated national positions
- Participation by Regulators (Governments), User Communities and Industry – is important
- Possibly re-activate BIS working group
- Be present at UTC and ISO committees with some Continuity of Participation
- Maximize use of e-discussion lists – free dialog
- Continue to Prepare and disseminate Resources and Education material

Unicode Issues
Dravidian Group
Kannada, Malayalam, Tamil & Telugu

V.S. Umamaheswaran
(umavs@ca.ibm.com)
IBM Toronto Lab, Canada

Characters added in V4.0

(in response to latest request from India)

0CBC KANNADA SIGN NUKTA
0CBD KANNADA SIGN AVAGRAHA

(from TNG Keyboard Layout)

0BF3 TAMIL DAY SIGN (Naal)
0BF4 TAMIL MONTH SIGN (Maatham)
0BF5 TAMIL YEAR SIGN (Varudam)
0BF6 TAMIL DEBIT SIGN (Patru)
0BF7 TAMIL CREDIT SIGN (Varavu)
0BF8 TAMIL AS ABOVE SIGN (Merpadi)
0BF9 TAMIL RUPEE SIGN (Rupai)
0BFA TAMIL NUMBER SIGN (Enn)

Additions in V4.0

Additions to text of Chapter 9 to address several of the requests in latest input from Gov of India and from other inputs.

Some examples:

Added text - where users are to look for the DANDA and DOUBLE DANDA characters (in the Devanagari block).

0CCD KANNADA SIGN VIRAMA
*** preferred name is halant**

See handout charts and names list for Annotations added.

Framework for discussion

- **Respect Stability Policy**
  - **No removal of existing character**
  - **No relocation / reordering of existing code positions**
  - **No name changes**
  - **No changes to existing canonical equivalences / normalization**
  - **No new multiple spellings**
  - **No new encoding model**
  - **If sequences satisfy the requirement no new character needed (Ch 9)**
- **Suggestions that can be entertained**
  - **Text for FAQ, Tech Note, Standard - for better understanding**
  - **Possible new sequences**
  - **Annotations where appropriate**
  - **New characters only with evidence**
  - **Deprecation only with strong justification**

Packaging Results of Discussion

For each Dravidian Script Categorize issues as:

- Proposal for FAQ material
- Proposal for Unicode Technical Note
- Proposal for Explanatory text
- Proposal for Annotation
- Proposal for Deprecation
- Proposal for New Character

Assign an Owner for Each