Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation internationale de normalisation
Международная организация по стандартизации
國際標準化組織

Doc Type: Working Group Document
Title: Proposal to add Ideographic Description Characters (IDC) to the UCS
Source: Richard S. Cook
Status: Expert Contribution
Date: 2002-05-18

A. Administrative

1. Title
Proposal to add Ideographic Description Characters (IDC) to the UCS
2. Requester's name
Richard S. Cook
3. Requester type
Expert contribution.
4. Submission date
2002-05-18.
5. Requester's reference
6a. Completion
This is a complete proposal.
6b. More information to be provided?
No.

B. Technical – General

1a. New script? Name?
No.
1b. Addition of characters to existing block? Name?
Ideographic Description Character
2. Number of characters
10.
3. Proposed category
Category B1.
4. Proposed level of implementation and rationale
Base characters.
5a. Character names included in proposal?
Yes.
5b. Character names in accordance with guidelines?
Yes.
5c. Character shapes reviewable?
Yes. See below.
6a. Who will provide computerized font?
Cook.
6b. Font currently available?
Yes.

6c. Font format?
TrueType, and PostScript Type 1 formats are available.
7a. Are references (to other character sets, dictionaries, descriptive texts, etc.) provided?
Yes.
7b. Are published examples (such as samples from newspapers, magazines, or other sources) of use of proposed characters attached?
Yes.
7c. Proposed Unicode and ISO/IEC 10646 bibliographical citations:
See bibliography below.
8. Does the proposal address other aspects of character data processing?
No.

C. Technical – Justification

1. Contact with the user community?
Yes.
2. Information on the user community?
Chinese document processing.
3a. The context of use for the proposed characters?
Description of unencoded hanzi in a legacy (Big-5-based) system.
3b. Reference
See bibliography below.
4a. Proposed characters in current use?
Yes.
4b. Where?
Symbols originated in and have strong usage in Taiwan, Japan, US.
5a. Characters should be encoded entirely in BMP?
Yes.
5b. Rationale
Keep together with existing IDC.
6. Should characters be kept in a continuous range?
Yes, if possible.
7a. Can the characters be considered a presentation form of an existing character or character sequence?
No.
7b. Where? / 7c. Reference
N.A.
8a. Can any of the characters be considered to be similar (in appearance or function) to an existing character?
No.
8b. Where? / 8c. Reference
N.A.
9a. Combining characters or use of composite sequences included?
No.
9b. List of composite sequences and their corresponding glyph images provided?
N.A.
10. Characters with any special properties such as control function, etc. included?
No.

D.0. Background

The Chinese Document Processing (CDP) Laboratory in the Institute of Information Technology at Academia Sinica in Taiwan, ROC, developed a Big5-based system employing ideographic description characters ("operators") to represent Chinese characters not encoded in Big5. This extended Big-5 system and the syntax of its operators is described in the Chuang (1989) paper appended to this proposal, and also available here:

<http://linguistics.berkeley.edu/~rscook/images/CDM-jpg/index.html>
<http://linguistics.berkeley.edu/~rscook/pdf/CDM-HanziMissingChar.pdf>

D.1. Attested Usage

In the years since its initial development, a considerable amount of data has been amassed employing various versions of this system. The following two examples of usage amount to more than 81,000 database records (multiple fields per record):

• 羅鳳珠 Prof. Luo Fengzhu <gefjulo@saturn.yzu.edu.tw> and her students (at 台灣桃園縣元智大學中語系) produced a 25,000 record index of the head-entries and phonological *fanqie* notation in the Middle Chinese rhyming dictionary *«Song Ben Guangyun»* (《新校正切宋本廣韻》，台灣黎明文化事業公司出版，林尹校訂 1976 年出版).

• Indices and component descriptions for the complete <Hanyu Da Zidian> (56,097 records) were produced in the CDP lab. Portions of this data were used in the production and proofing of the "kHanyu" field of the current Unihan database (see the description in the "kHanyu" section of the Unihan header).

D.2. CDP Operators

The full set of 13 visibly displayed CDP Operators is as follows:

D.3. Proposed IDC

The first 3 of these,  , are infixal operators, functionally equivalent to 3 of the 12 Unicode/10646 "Ideographic Description Characters" (IDC): [U+2ff0], [U+2ff1], [U+2ff4]. Despite this functional equivalence, the differing syntax suggests that these 3 signs be assigned unique codepoints.

The remaining 10 CDP Operators do not have functional equivalents. The first 8 of these,  , are prefixal multipliers, which is to say that when they are prefixed to a single component, they specify the manner of multiplication of that single character component. The final two,  and  , are employed for bracketing purposes, to delimit component description sequences of >2 components in the stream of Big 5 text. We would like to propose that for the purpose of future Unicode conversion of existing CDP data, all 13 of these characters be encoded in Unicode/10646.

D.4. Proposed Codepoints and Names

	U+2FE0	IDEOGRAPHIC DESCRIPTION CHARACTER HORIZONTAL CONJOINER
	U+2FE1	IDEOGRAPHIC DESCRIPTION CHARACTER VERTICAL CONJOINER
	U+2FE2	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUNDING CONJOINER
	U+2FE3	IDEOGRAPHIC DESCRIPTION CHARACTER HORIZONTAL DOUBLE MULTIPLIER
	U+2FE4	IDEOGRAPHIC DESCRIPTION CHARACTER VERTICAL DOUBLE MULTIPLIER
	U+2FE5	IDEOGRAPHIC DESCRIPTION CHARACTER TRIANGLE MULTIPLIER
	U+2FE6	IDEOGRAPHIC DESCRIPTION CHARACTER HORIZONTAL TRIPLING MULTIPLIER
	U+2FE7	IDEOGRAPHIC DESCRIPTION CHARACTER VERTICAL TRIPLING MULTIPLIER
	U+2FE8	IDEOGRAPHIC DESCRIPTION CHARACTER SQUARE MULTIPLIER
	U+2FE9	IDEOGRAPHIC DESCRIPTION CHARACTER HORIZONTAL QUADRUPLE MULTIPLIER
	U+2FEA	IDEOGRAPHIC DESCRIPTION CHARACTER VERTICAL QUADRUPLE MULTIPLIER
	U+2FEB	IDEOGRAPHIC DESCRIPTION CHARACTER LEFT SPAN DELIMITER
	U+2FEC	IDEOGRAPHIC DESCRIPTION CHARACTER RIGHT SPAN DELIMITER

G.0. Acknowledgements

This proposal was prepared by Richard S. COOK <rscook@socrates.berkeley.edu> of the STEDT Project, with contributions from 莊德明 Derming CHUANG <derming@gate.sinica.edu.tw> and Christian WITTERN <wittern@kanji.zinbun.kyoto-u.ac.jp>.

STEDT Project research, in the Department Linguistics at the University of California at Berkeley, is supported in part by grants from:

- The National Science Foundation (NSF), Division of Behavioral & Cognitive Sciences, Linguistics, Grant Nos. BNS-86-17726, BNS-90-11918, DBS-92 09481, FD-95-11034, SBR-9808952 and BCS-9904950;
- The National Endowment for the Humanities (NEH), Preservation and Access, Grant Nos. RT-20789-87, RT-21203-90, RT-21420-92, PA-22843 96 and PA-23353-99.

For more information, please visit STEDT on the web at <http://stedt.berkeley.edu/> or send email to <stedt@socrates.berkeley.edu>.