ISO/IEC 10646:2017-12 (E)

Information technology - Universal Coded Character Set (UCS)

Contents

Page

Foreword vii

Introduction viii

1 Scope 1

2 Normative references 1

3 Terms and definitions 2

4 Conformance 8
4.1 General 8
4.2 Conformance of information interchange 8
4.3 Conformance of devices 8

5 General structure of the UCS 9

6 Basic structure and nomenclature 10
6.1 Structure 10
6.2 Coding of characters 11
6.3 Types of code points 11
6.4 Naming of characters 12
6.5 Short identifiers for code points (UIDs) 12
6.6 UCS Sequence Identifiers 13
6.7 Octet sequence identifiers 13

7 Revision and updating of the UCS 14

8 Subsets 14
8.1 General 14
8.2 Limited subset 14
8.3 Selected subset 14

9 UCS encoding forms 14
9.1 General 14
9.2 UTF-8 14
9.3 UTF-16 15
9.4 UTF-32 (UCS-4) 16

10 UCS Encoding schemes 16
10.1 General 16
10.2 UTF-8 16
10.3 UTF-16BE 16
10.4 UTF-16LE 16
10.5 UTF-16 16
10.6 UTF-32BE 17
10.7 UTF-32LE 17
10.8 UTF-32 17

11 Use of control functions with the UCS 17

12 Declaration of identification of features 18

12.1 Purpose and context of identification 18
12.2 Identification of a UCS encoding scheme 19
12.3 Identification of subsets of graphic characters 19
12.4 Identification of control function set 19

13 Structure of the code charts and lists 20

14 Block and collection names 21
14.1 Block names 21
14.2 Collection names 21

15 Mirrored characters in bidirectional context 21
15.1 Mirrored characters 21
15.2 Directionality of bidirectional text 21

16 Special characters 22
16.1 General 22
16.2 Space characters 22
16.3 Currency symbols 22
16.4 Format characters 22
16.5 Ideographic description characters 23
16.6 Variation selectors and variation sequences 23

17 Presentation forms of characters 24

18 Compatibility characters 25

19 Order of characters 25

20 Combining characters 25
20.1 Order of combining characters 25
20.2 Combining class and canonical ordering 26
20.3 Appearance in code charts 26
20.4 Alternate coded representations 26
20.5 Multiple combining characters 26
20.6 Collections containing combining characters 27
20.7 Combining Grapheme Joiner 27

21 Normalization forms 27

22 Special features of individual scripts and symbol repertoires 28
22.1 Hangul syllable composition method 28
22.2 Features of scripts used in India and some other South Asian countries 28
22.3 Byzantine musical symbols 28
22.4 Source references for pictographic symbols 29

23 Source references for CJK ideographs 29
23.1 List of source references 29
23.2 Source references file for CJK ideographs 32
23.3 Source reference presentation for CJK Unified ideographs 34
23.4 Source references presentation for CJK Compatibility ideographs 37

24 Source references for Tangut ideographs 37
24.1 List of source references 37
24.2 Source reference file for Tangut ideographs 38
24.3 Source reference presentation for Tanguts ideographs 39

25 Source references for Nüshu characters 39
25.1 List of source references 39
25.2 Source reference file for Nüshu characters 39

26 Character names and annotations 40

26.1 Entity names 40
26.2 Name formation 40
26.3 Single name 41
26.4 Name immutability 41
26.5 Name uniqueness 41
26.6 Character names for CJK ideographs 42
26.7 Character names for Tangut ideographs 42
26.8 Character names for Nüshu characters 42
26.9 Character names for Hangul syllables 43

27 Named UCS Sequence Identifiers 44

28 Structure of the Basic Multilingual Plane 46

29 Structure of the Supplementary Multilingual Plane for scripts and symbols (SMP) 48

30 Structure of the Supplementary Ideographic Plane (SIP) 51

31 Structure of the Tertiary Ideographic Plane (TIP) 51

32 Structure of the Supplementary Special-purpose Plane (SSP) 51

33 Code charts and lists of character names 52
33.1 General 52
33.2 Code chart 52
33.3 Character names list 52
33.4 Summary of standardized variation sequences 53
33.5 Code charts and lists of character names 54

Annex A (normative) Collections of graphic characters for subsets 2611

A.1 Collections of coded graphic characters 2611
A.2 Blocks lists 2617
A.3 Fixed collections of the whole UCS (except Unicode collections) 2620
A.4 CJK collections 2623
A.5 Other collections 2624
A.6 Unicode collections 2628

Annex B (normative) List of combining characters 2629

Annex C (normative) Transformation format for planes 01 to 10 of the UCS (UTF-16) 2630

Annex D (normative) UCS Transformation Format 8 (UTF-8) 2631

Annex E (normative) Mirrored characters in bidirectional context 2632

Annex F (informative) Format characters 2633

F.1 General format characters 2633
F.2 Script-specific format characters 2635
F.3 Interlinear annotation characters 2636
F.4 Subtending format characters 2636
F.5 Shorthand format characters 2637
F.6 Invisible mathematical operators 2637
F.7 Western musical symbols 2637
F.8 Language tagging using Tag characters 2638

Annex G (informative) Alphabetically sorted list of character names 2640

Annex H (informative) The use of "signatures" to identify UCS 2641

Annex I (informative) Ideographic description characters 2642

I.1 General 2642
I.2 Syntax of an ideographic description sequence 2642
I.3 Individual definitions of the ideographic description characters 2643

Annex J (informative) Recommendation for combined receiving/originating devices with internal storage 2645

Annex K (informative) Notations of octet value representations 2646

Annex L (informative) Character naming guidelines 2647

Annex M (informative) Sources of characters 2650

Annex N (informative) External references to character repertoires 2674

N.1 Methods of reference to character repertoires and their coding 2674
N.2 Identification of ASN.1 character abstract syntaxes 2674
N.3 Identification of ASN.1 character transfer syntaxes 2675

Annex P (informative) Additional information on CJK Unified ideographs 2676

Annex Q (informative) Code mapping table for Hangul syllables 2679

Annex R (informative) Names of Hangul syllables 2680

Annex S (informative) Procedure for the unification and arrangement of CJK ideographs 2681

S.1 Unification procedure 2681
S.2 Arrangement procedure 2685
S.3 Source separation examples 2685
S.4 Non-unification examples 2690

Annex T (informative) Language tagging using Tag Characters 2692

Annex U (informative) Characters in identifiers 2693