L2/04-396

Title: Status Report on TC 46 Coded Character Set Standards
Source: Joan M. Aliprand (Senior Analyst, RLG)
Status: Expert Contribution
Action: For consideration by ISO/TC46
Date: 2004-09-28

1. Background

In 1999, ISO/TC 46/SC 4 and ISO/IEC JTC 1/SC 2 reached agreement on the transfer to SC 2 of standards that had been developed by TC 46. SC 2 agreed that it would make no changes to the transferred standards nor withdraw them without TC 46/SC 4 approval. That is, at each revision of the standards, TC 46/SC 4 would be asked to provide a position statement (documented in: ISO/IEC JTC 1/SC 2 N 3370). SC 2 resolved to delegate responsibility for any transferred standards to its WG 3.

SC 4/WG 1 stated in its document N 235:

> There does not appear to be any procedural problem with transfer of responsibility for the maintenance of a standard from one ISO subcommittee to another. WG1 hopes that within the next few years characters from the other sets it helped develop are accommodated by acceptable mappings to ISO/IEC 10646. WG1 is willing to accept as a new policy the transfer to JTC1/SC2 of responsibility for the maintenance of an existing TC46 character set once acceptable mappings in ISO/IEC 10646 have been found for all characters in the TC46 set.

TC 46 authorized the transfer of six character set standards to ISO/IEC JTC 1/SC 2 in Resolution 4 of the 28th Plenary meeting of ISO/TC 46 (Paris, May 1999). Altogether, eight character sets have been transferred to SC 2 – Appendix A lists the standards.

2. TC 46 Coded Character Set Standards Not Yet Transferred to SC 2

Graphic Characters

ISO 5426-2:1996 *Information and documentation -- Extension of the Latin alphabet coded character set for bibliographic information interchange -- Part 2: Latin characters used in minor European languages and obsolete typography*

ISO 6861:1996 *Information and documentation -- Glagolitic alphabet coded character set for bibliographic information interchange*

ISO 8957: 1996 *Information and documentation -- Hebrew alphabet coded character sets for bibliographic information interchange*

ISO 10754:1996 *Information and documentation -- Extension of the Cyrillic alphabet coded character set for non-Slavic languages for bibliographic information interchange*

Many of the characters in these standards have ISO/IEC 10646 equivalents; ISO 6861 is now completely mapped. The unmapped characters in these standards are discussed in detail below.

Control Characters

ISO 6630:1986 *Documentation -- Bibliographic control characters*

3. Subsequent Developments Related to Standards for Bibliographic Use

3.1. Additional Mappings for Characters in TC 46 Standards

For several of the remaining standards, new characters added to ISO/IEC 10646 and the Unicode Standard have increased the number of character mappings beyond those approved by ISO/TC 46/SC 4/WG 1 at its meeting in May 2000.

TC 46 Standard	Characters	WG 1 Approved	Additional	No equivalent
ISO 5426-2:	70	34	5	31
ISO 6630	15	not evaluated	not evaluated	ineligible?
ISO 6861 Table 1	53	not evaluated	all	0
ISO 6861 Table 2	37	not evaluated	all	0
ISO 8957 Table 1	90	all	n/a	0
ISO 8957 Table 2	51	34	0	17
ISO 10754	93	49	16	28 (26 eligible)

All additional mappings except one were due to new characters proposed and documented by experts on scripts and standards. See Appendix B for details of the new characters, including references to the proposals.

Only 26 characters of ISO 10754 are eligible for consideration for addition to ISO/IEC 10646. In 1998, WG2 decided that the combining right descender and combining left descender (characters 24 and 34) are not needed because Cyrillic characters with descenders are encoded solely as precomposed characters in ISO/IEC 10646.

The control characters in ISO 6630 may be ineligible for encoding in ISO/IEC 10646 (see below for discussion).

3.2. ISO/IEC JTC 1/SC 2/WG 3 disbanded

At its 13th Plenary Meeting, ISO/IEC JTC 1/SC 2 decided to disband its Working Group 3 (responsible for 7 and 8 bit codes and their extensions):

> RESOLUTION M 13-06: Disbandment of WG 3
> Noting that WG 3 currently has no active projects and that the term of office of Mr. Evangelos Melagrakis as WG 3 Convener is ending at this plenary, SC 2 resolves to disband WG 3. The maintenance of WG 3 standards will be transferred to SC 2.
> SC 2 expresses its appreciation to Mr. Melagrakis and the Greek National Body for administrating WG 3 for the past seven years.
> *Unanimous*

As a result of this action, SC 2 now is directly responsible for all coded character set standards that were formerly the responsibility of WG 3.

3.3. "Stabilized" Status an Option for SC 2 Standards

In 2001, ISO/IEC JTC 1 introduced the concept of a *stabilized standard*, defined as:

> A stabilised standard is one that has ongoing validity and effectiveness but is mature and insofar as can be determined will not require further maintenance of any sort. While a standard is in stabilised status it will no longer be subject to periodic maintenance but will be retained to provide for the continued viability of existing products or servicing of equipment that is expected to have a long working life.

Source: *ISO/IEC Directives: Procedures for the technical work of ISO/IEC JTC 1 on Information Technology*, 15.6.

The stabilized category reduces the likelihood that a standard will be revised since it no longer circulated for review. Clause 15.6.4 describes the circumstances under which a stabilized standard would be withdrawn.

> Where a Sub Committee, National Body or other standards owning body within JTC 1 becomes aware that a stabilised standard is no longer in use or its use has been superseded or it now unsafe to continue to use the standard, the Sub Committee ,National Body or other standards owning body within JTC 1 may request JTC 1 to issue an immediate 60 day letter ballot to reclassify the standard as withdrawn.

Since SC 2 is a subcommittee of JTC 1, standards for which SC 2 is responsible can be proposed for the stabilized category at the time of a standard's regular review. (This includes the standards listed in Appendix A.)

4. ISO 6861:1996 Now Eligible for Transfer

The complete mapping of the characters of ISO 6861:1996 to ISO/IEC 10646 equivalents (currently under ballot) has been verified by Joan Aliprand and Michael Everson. (The mapping has been submitted for information as a separate document.)

Recommendation: That TC 46 transfer ISO 6861:1996 *Information and documentation -- Glagolitic alphabet coded character set for bibliographic information interchange* 6861 to ISO/IEC JTC 1/SC 2 without delay, with a recommendation that ISO 6861 be stabilized as soon as possible.

5. Remaining Graphic Standards

The remaining TC 46 standards comprise three partially-mapped graphic character sets, and ISO 6630, which defines control characters and is discussed separately.

ISO 5426-2 (Extended Latin: Part 2)

The 31 characters that still lack mappings are all Latin contractions found in medieval manuscripts and early printed works that emulate the orthographical traditions of manuscripts.

These 31 contractions are but a tiny subset of documented Latin contractions that number in the hundreds (perhaps thousands). Coding only a few Latin contractions will cause frustration among those who want to use them.

From the library perspective, it is questionable whether any of these characters are needed. Modern cataloging practice is to spell out Latin contractions as full text, and indicate what letters the cataloger has supplied. Cataloging experts at the two leading databases of antiquarian books – the English Short Title Catalogue, and the Hand Press Books database – did not see a need for the contractions.

The only known bibliographic use of some of the contractions found in ISO 5426-2 is by the British Library. However, the British Library does not use ISO 5426-2; it uses a different set of contractions encoded in its own "Specials" character set. In the CHASE Project (managed by the British Library), the Latin contractions in the "Specials" character set were mapped to code points in the Private Use Area.

If contractions were used instead of full text, a searcher would have to "or" contractions with spelled out text to achieve full retrieval. Compounding this problem is the fact that some of the contractions common to ISO 5426-2 and the British Library's "Specials" character set have a different definition in each source. If retrieved records contained contractions, the searcher would have to be able to interpret them (but most users are not expert in scribal handwriting).

IT advances have eliminated the need for the contractions encoded in ISO 5426-2. With the ability to link a cataloging record to an image of the source of information (and the support for this in MARC 21 and UNIMARC), there is less need to transcribe exact typographical details in the cataloging record.

ISO 8957 (Hebrew)

ISO 8957 has two tables of coded characters.

- Table 1 contains the letters of the Hebrew alphabet, plus commonly-used vowel points and other marks of pronunciation, and punctuation marks. Except for one difference due to misinterpretation of a character in the preparation of this standard, Table 1 matches the Hebrew character set developed by RLG and specified for use in MARC 21 records by the Library of Congress. Table 1 is completely mapped to characters of ISO/IEC 10646.

- Table 2 contains cantillation marks, etc.(“pointing”) used for specialized purposes. Since all pointing (including the Tiberian vocalization and other marks of pronunciation encoded in Table 1) is omitted from Hebrew script cataloging, both in Israel and in other parts of the world, the characters of Table 2 are not needed for library applications.

ISO 10754 (Extended Cyrillic for Non-Slavic Languages)

The characters in this set are from the Cyrillic script alphabets of minority languages spoken in Russia and other parts of the former Soviet Union. They are of potential use not only for original script cataloging, but to speakers of these languages and language scholars.

ISO 10754 was developed before publication of the Character-Glyph model (ISO/IEC TR 15285:1998) that underpins the character content of ISO/IEC 10646 and the Unicode Standard. What ISO 10754 encodes as characters may not be regarded as such under the Character-Glyph model. Four of the “letters” in ISO 10754 may be typographical ligatures, and six may be variants of already encoded letters. Further investigation is needed.

Eight of the unmapped letters may be considered clones of Latin letters. These characters are already encoded if they are mapped to Latin letters, but there is a sharp difference of opinion as to whether “borrowings” of Latin letters into Cyrillic contexts should be encoded as Latin letters or as Cyrillic letters.

6. ISO 6630 (Bibliographic Control Characters)

Four of these characters are sanctioned for use in UNIMARC: PLU, PLD, NSB, NSE. MARC 21 has also approved use of NSB, NSE, but only for certain specific purposes initially.

The other 11 are not sanctioned as characters for bibliographic exchange in MARC 21 or UNIMARC. Whether any are sanctioned for use in national MARCs is not known.

Control codes represent a mixing of plain text with protocol functions. While this was acceptable at the time that ISO 6630 was developed, IT has moved on. The encoding of characters has generally been divorced from the higher-level protocols that manipulate and format the characters. Under this view, use of control characters is generally regarded as a higher level protocol divorced from the encoding of characters.

It should be noted that the use of ISO 6630 in existing implementations using individual 8-bit and multi-byte character sets is not being questioned. These implementations are not imperiled even if ISO 6630 were to be transferred immediately to SC 2 (with a recommendation for stabilization).

7. Future Prospects

7.1. Graphic character sets (ISO 5426-2, ISO 8957, ISO 10754)

All of the unmapped characters in TC 46 graphic character set standards – Latin contractions, Hebrew pointing, and Cyrillic letters from alphabets used for minority languages of the former Soviet Union – are in highly specialized areas of knowledge. When these characters were proposed for addition to ISO/IEC 10646 in 1998, SC 2/WG 2's opinion was that further study was needed if they were to be added to ISO/IEC 10646. Do members of TC 46/SC 4 have the necessary expertise and time to carry out this further study?

In addition, all of the unmapped graphic characters have to be assessed with respect to the Character-Glyph model (ISO/IEC TR 15285:1998) that underpins the character content of ISO/IEC 10646 and the Unicode Standard.

Progress on providing more mappings of TC 46 characters (documented in Appendix B) is entirely due to proposals or information from language and script experts. Except for Michael Everson, JTC 1/SC 2's liaison to TC 46/SC 4, none of the other experts was directly involved with TC 46.

No evidence has been presented that that the library community uses any of the unmapped characters in the TC 46 graphic character sets. Current cataloging practice rules out the use the Latin contractions and of Babylonian, Palestinian and Samaritan pointing. If there is a need for the unmapped characters in a plain text environment (for example, in e-mail text), the scholars that use them will make the case for their addition to ISO/IEC 10646 and the Unicode Standard.

For example, proposals to encode characters of the Babylonian, Palestinian and Samaritan pointing systems have been submitted to SC 2/WG 2 by Elaine Keown, a scholar of Hebrew and Aramaic scripts. These proposals hold out the prospect that the unmapped characters of ISO 8957 may be included in ISO/IEC 10646 and the Unicode Standard some day.

The Medieval Unicode Font Initiative (MUFI) http://helmer.aksis.uib.no/mufi/ was established by scholars to coordinate work on encoding special characters found in medieval Latin manuscripts and to develop fonts to display the characters. Approximately 1/3 of the unmapped Latin contractions appear in the *MUFI Initial Character Recommendation.*

For the unmapped characters of ISO 10754, Cyrillic script experts should determine which of the unmapped characters are properly characters according to the Character-Glyph Model, document their use, and submit a proposal to SC 2/WG 2 and the Unicode Standard. The proposal for Komi Cyrillic by Everson, Ruppel, and Trosterud is an excellent model. The experts would also propose mappings for any of the unmapped TC 46 characters that they consider to be already encoded (i.e., ligatures or variants).

7.2. ISO 6630 (Bibliographic Control Characters)

The meta-question with regard to ISO 6630 control characters is whether the functions of control characters have been superseded by subsequent developments in IT and the

structure of bibliographic data (e.g., the increasing use of XML). For example, how does the ISO 6630 character CUS (Close-up for sorting) relate to what can be done with the Unicode Collation Algorithm? ISO 6630 itself states (clause 6, paragraph 1):

> Alternative techniques [to convey control information] are possible and are not ruled out by this International Standard.

If TC 46 is going to propose that these control characters be added to ISO/IEC 10646 and the Unicode Standard, there is a fundamental question that must be answered in the proposal to justify the addition (because SC 2/WG 2 and the Unicode Technical Committee will surely ask it): What functionality exists that can only be met by the use of ISO 6630 control codes in plain text and not by any other technical solution?

8. Conclusion

The goal of having every character in all of the TC 46 standards represented in ISO/IEC 10646 and the Unicode Standard cannot be met. As noted above, WG 2 decided not to add the combining descenders of ISO 10754 to ISO/IEC 10646.

Latin contractions and the ancient Hebrew pointing systems are undergoing study by experts with the goal of seeing them encoded in ISO/IEC 10646 and the Unicode Standard. There will be no progress on the addition of the unmapped Extended Cyrillic characters unless a language expert undertakes the further study requested by SC 2/WG 2 and prepares a proposal.

There are several advantages to transferring ISO 5426-2, ISO 6630, ISO 8957, and ISO 10754 to SC 2 (with stabilization as soon as possible after transfer):

- These standards are due for review in 2006. Transfer to SC 2 before the next review would save the TC 46 Secretariat from the effort and expense of a pointless exercise (since TC 46 has already determined that these standards should never be revised and should not be withdrawn for now).
- The “stabilized” category that SC 2 can assign to the standards ensures their unchanged existence as ISO standards for the foreseeable future.

In particular:

- ISO 6630 (Bibliographic Control Characters) should be transferred to SC2 without delay (with a request for stabilization).
- How to implement the functionality of the ISO 6630 control characters in a Unicode environment should be investigated by library system designers in conjunction with the technical experts on the Unicode Technical Committee. Many library system vendors (including RLG and OCLC) are members of the Unicode Consortium.

APPENDIX A

Standards Originally Developed by TC 46 and Formally Transferred to SC 2

ISO 5426:1983 *Extension of the Latin alphabet coded character set for bibliographic information interchange*

ISO 5427:1984 *Extension of the Cyrillic alphabet coded character set for bibliographic information interchange*

ISO 5428:1984 *Greek alphabet coded character set for bibliographic information interchange*

ISO 6438:1983 *Documentation -- African coded character set for bibliographic information interchange*

ISO 6862:1996 *Information and documentation -- Mathematical coded character set for bibliographic information interchange*

ISO 10585:1996 *Information and documentation -- Armenian alphabet coded character set for bibliographic information interchange*

ISO 10586:1996 *Information and documentation -- Georgian alphabet coded character set for bibliographic information interchange*

ISO 11822:1996 *Information and documentation -- Extension of the Arabic alphabet coded character set for bibliographic information interchange*

APPENDIX B

Additional Mappings of Characters from TC 46 Standards to ISO/IEC 10646 (UCS)

ISO 5426-2:1996

Code	ISO 5426-2 Name	UCS Mapping	UCS Name
41	COMBINING LATIN SMALL R ABOVE	036C	COMBINING LATIN SMALL LETTER R
45	COMBINING LATIN SMALL A ABOVE	0363	COMBINING LATIN SMALL LETTER A
46	COMBINING LATIN SMALL E ABOVE	0364	COMBINING LATIN SMALL LETTER E
47	COMBINING LATIN SMALL O ABOVE	0366	COMBINING LATIN SMALL LETTER O
63	LATIN CAPITAL LETTER KRA	004B 02BC	LATIN CAPITAL LETTER K followed by MODIFIER LETTER APOSTROPHE

- The first four mappings come from the 13 medievalist diacritic marks added to ISO/IEC 10646 and the Unicode Standard as a result of the proposal ISO/IEC JTC 1/SC 2/WG 2 N 2266 *Diacritics for medieval* studies, by Marc Küster and Isabel Wojtovicz, dated 2000-09-14.
- The mapping for the LATIN CAPITAL LETTER KRA is an additional mapping identified by INCITS/L2 after discussion with an expert on Greenlandic.

ISO 6861:1996

The 94 Glagolitic characters proposed in ISO/IEC JTC 1/SC 2/WG 2 N2610R *Final proposal for encoding the Glagolitic script in the UCS*, by Michael Everson and Ralph Cleminson, dated 2003-08-24 were accepted by WG 2 at Meeting 44 (Mountain View, CA, USA, October 2003) and by the Unicode Technical Committee in August 2003. (The addition to ISO/IEC 10646 us under ballot.)

During the WG2 meeting, Michael Everson (Ireland) and Joan Aliprand (USA) verified that all the characters in ISO 6861 could be mapped to characters in the proposal. The mapping table has been submitted to TC 46 as a separate document.

ISO 10754:1996

Code	ISO 5426-2 Name	UCS Code	UCS Name
2B	CYRILLIC SMALL LETTER KOMI DE	0501	CYRILLIC SMALL LETTER KOMI DE
2C	CYRILLIC SMALL LETTER KOMI DJE	0503	CYRILLIC SMALL LETTER KOMI DJE
2E	CYRILLIC SMALL LETTER KOMI DZE	0507	CYRILLIC SMALL LETTER KOMI DZJE

Code	ISO 5426-2 Name	UCS Code	UCS Name
2F	CYRILLIC SMALL LETTER KOMI ZJE	0505	CYRILLIC SMALL LETTER KOMI ZJE
3B	CYRILLIC CAPITAL LETTER KOMI DE	0500	CYRILLIC CAPITAL LETTER KOMI DE
3C	CYRILLIC CAPITAL LETTER KOMI DJE	0502	CYRILLIC CAPITAL LETTER KOMI DJE
3E	CYRILLIC CAPITAL LETTER KOMI DZE	0506	CYRILLIC CAPITAL LETTER KOMI DZJE
3F	CYRILLIC CAPITAL LETTER KOMI ZJE	0504	CYRILLIC CAPITAL LETTER KOMI ZJE
48	CYRILLIC SMALL LETTER KOMI ELJ	0509	CYRILLIC SMALL LETTER KOMI LJE
4D	CYRILLIC SMALL LETTER KOMI NG	050B	CYRILLIC SMALL LETTER KOMI NJE
58	CYRILLIC CAPITAL LETTER KOMI ELJ	0508	CYRILLIC CAPITAL LETTER KOMI LJE
5D	CYRILLIC CAPITAL LETTER KOMI NG	050A	CYRILLIC CAPITAL LETTER KOMI NJE
64	CYRILLIC SMALL LETTER KOMI ESJ	050D	CYRILLIC SMALL LETTER KOMI SJE
65	CYRILLIC SMALL LETTER KOMI TJE	050F	CYRILLIC SMALL LETTER KOMI TJE
74	CYRILLIC CAPITAL LETTER KOMI ESJ	050C	CYRILLIC CAPITAL LETTER KOMI SJE
75	CYRILLIC CAPITAL LETTER KOMI TJE	050E	CYRILLIC CAPITAL LETTER KOMI TJE

The 16 characters were added to ISO/IEC 10646 and the Unicode Standard as a result of the proposal ISO/IEC JTC 1/SC 2/WG 2 N2224 *Archaic Komi Cyrillic characters for the BMP of the UCS*, by Michael Everson, Klaas Ruppel, and Trond Trosterud, dated 2000-06-09.