A Tool Supporting the Internationalisation of the Generic Network Services

Borka Jerman-Blažič <jerman-blazic@e5.ijs.si>

Abstract

This paper describes an attempt for provision of internationalised network services by introduction of a tool for transformation of character set codes in different network applications. The tool helps the end user with the conversion of the received or sent data encoded in character set codes that are not supported by the user system. The tool considers the user cultural environment and provides conversion customised to the user needs and cultural requirements. The tool was developed within the Project C3 of the RARE WG-CHAR Task Force. Currently, the tool caters for the scripts used in Europe, but the model developed and the API implemented can be used for any other cultural environment.

I. General Introduction and Overview

Internationalisation as a broad IT platform can be understood as a response of the IT standardisation community to the business requirements for providing and developing products in terms of hardware, software, documentation, packaging and distribution customised to a different cultural environments. The purpose of the internationalisation framework is to enable an easy and internationally agreed approach for products development on various platforms which will support the communication in natural languages and the cultural conventions used by end users, systems administrators, operational personnel and application developers world-wide.

The products developed in line with the internationalisation framework and later localised for particular cultural environment are expected to be integrated with, utilised by and/or supported by services wherever there is a need for support of cultural conventions in the world-wide markets.

The primary benefits of internationalised products and services are expected to be:

- ***the support of cultural conventions in different environments,***
- ***interchange of multilingual data between processing units in a correct, consistent and efficient manner and in conformance with the requirements of the local user.***

There are many cultural items which are relevant to the internationalisation and are considered as an integral part of internationalised products and services. Among them we can find: textual representation of numbers, date/time and calendar data representation, currency, number formatting, measurement systems, postal address formatting, icons etc., but no one of these items can be compared with the importance of the basic subject of the internationalisation-communication in the user's chosen language. This includes all relevant items needed for handling letters, literals and all other data required for representing and processing of the written language. The link between the elementary characters of the written language and the bit patterns representing them in the computer system is provided with the character set code supported by the system. Handling of character set codes is the main issue in the process of internationalisation or its counter part i.e. localisation of a particular IT product or service.

II. Character Sets Codes and the Generic Network Services

Today more then 3 000 languages are spoken throughout the world. A few more than a hundred of these languages are actually written. Half of the world's population uses some version of the Latin alphabet. The other half uses 23 different significant alphabets (i.e. one million or more users). The International Organisation for Standardisation has developed coded character sets for Latin, Greek, Cyrillic, Arabic and Hebrew Alphabets. These coded character sets use either 7-bit combinations for representing letters (the most known is the 7-bit ASCII and the national versions of its international reference version) or 8-bit combinations (most known are coded character sets from the ISO 8859 series). Besides these standard character set codes there are many others with similar repertoire (i.e. the set of contained characters) developed and used in the equipment of well known manufacturers of computer technology. If only scripts used in Europe are considered and the character set codes that cater for those scripts implemented in different vendor's equipment then the number of different character set codes comes close to 40. When data are represented with different coding schemes and are interchanged between non-compatible vendor's equipment then usually the data are either misinterpreted or misunderstood. That is why the 7-bit ASCII or its international version known as

ISO 646 was used so long as the only common denominator in the generic services of the international networks.

However, a character set code which caters for almost all known languages and employs universal coding system based on 16 to 32 bits combinations has appeared recently as the international standard ISO 10 646 (UCS) or its manufacturer's version i.e. Unicode. The implementation of UCS in the generic network applications can provide the required integrity of the characters over the network as well as the required support for multilingual services. However, the implementation of UCS is not fast enough as desired due to different reasons. Some of them are:

- ***the existing applications still running can support only character sets used in their systems,***
- ***some of the widely used protocols over the international networks are hostile to 8-bit coded data,***
- ***there is still lack of mechanism for easy and default recognition of a data stream coded by the UCS scheme,***
- ***products implementing UCS are still very rare on the market.***

Despite that, some properties of the existing network applications and used protocols are encouraging i.e. some of the current and most important network service applications have the capability for carrying a large number of different character sets. The X.400 application supports the use of 8-bit coding systems defined in two standard code documents: ISO 6937 and ISO 8859. The Directory service according to X.500 Recommendation applies the T.61 coding which cater for almost all characters used in languages written by the Latin Alphabet using special extension technique defined in ISO 2022. The Japanese networks make use of the same slightly changed technique for interchange of Internet mail written by Japanese ideograms. Since January 1992, the Internet also has the means of exchanging messages containing multiple character sets by using the methods defined in the RFC for MIME which is a good mailing tool allowing exchange of data coded in different coded character sets in addition to the audio-visual data.

We may say that the current networking world is in transition between the 8-bit coded characters sets scheme which provides support of the required national letters by use of the switching technique and the forthcoming 16-bit coding provided within the UCS scheme.

The major problem with different services providing transport mechanisms for carrying large number of different character sets is in the inconsistency of the applied data interchange standards and in the input/output (rendering) problems of the character sets in use due to the diversity of the user's equipment. We may say that the network services dealing with different character sets are mature enough and that additional effort is required at the user side. In this paper we discuss an attempt for provision of a solution on the user side in the form of a tool for character sets transformation and conversion. The tool helps the end user by providing conversion of the received or sent data which are coded in a character set codes that are not supported by the user system. The tool considers the user cultural environment and provides conversion customised to the user needs and cultural requirements. The tool was developed within the Project C3 of the RARE WG-CHAR Task Force and in co-operation with KTH, Stockholm, Sweden, DKUUG, Denmark. Currently, the tool caters for the scripts used in Europe but the model developed and implemented (in the tool) can be used for any other cultural environment.

III. The C3 System for Coded Character Set Conversion

The C3 systems are designed to support conversion between any pair of a wide selection of coded character sets used in Europe in most efficient way.

**Coded character set conversion* is the kind of data conversion in which a *source text*, encoded in one coded character set, the *source character set*, is transformed to a *target text*, encoded in another coded character set, the *target character set*. Ideally, the text should be the same, only the method of encoding being changed.*

*By a *coded character set* is meant a function (in the mathematical sense) which for any sequence of *encoding units* tells if it is legal and in that case gives the sequence of characters it represents. Encoding units are any ordered group of bits of a certain length. They are normally *bytes*, groups of 8 bits. The set of characters that can be represented are called the *character repertoire* of the coded character set.*

The tool is intended to be used in the following fields of applications:

- ***Conversion of different coded character sets in text for integration in MIME and X.400 MUA/MTA systems, netnews, X.500 directory services and network information retrieval services (NIR),***
- ***Conversion in terminal communication, when the terminal program runs on a computer with another coded character set than the host,***
- ***Conversion in transfer of text between incompatible word-processing or text-processing systems by means of plain text files.***
- ***Conversion of character coded data within one computer system in connection with a change of that system's principal coded character set, e.g. from a 7-bit character set to an 8-bit character set.***

III.A. Basic Features of the System

When converting text from one coded character set to another, the first necessary task is to transform to the target character set the encoding units of those characters of the source text that are also present in the target character set. This only involves having accurate information for all possible encoding units of both character sets about which character it stands for in that character set. The conversion in this case is *exact*: The source characters are re-encoded in the target character set without distortion.

This information about the definition of a character set is included in the *elementary table* for this character set in the C3 system. By combining the elementary tables for the source character set and the target character set by a simple algorithm, a *working table* is created that is used for the direct transformation of the sequence of encoding units in the source text to the sequence of encoding units in the target text.

This design makes possible conversion of any coded character set of those supported by the C3 system to any other such character set. How this is to be done is fully specified by N elementary tables, compared to the N(N-1) conversion tables usually used in brute force approach.

In the character sets world it is impossible to provide full coverage of a 8-bit target character set of the full repertoire of any source character set. For these cases a *non-exact* representations for the source character sets designed in the C3 system takes place. The non-exact conversion is implemented with a choice of three different types of conversion:

Conversion type 1 -- *one-to-one conversion*: Preserves the length of lines and data fields by always transforming one source character to one target character. Example:

LATIN CAPITAL LIGATURE AE	**>**	**A**
COPYRIGHT SIGN	**>**	**C**
MICRO SIGN	**>**	**u**
AMPERSAND	**>**	**&**

(included in all character sets)

Conversion type 2 -- *legible conversion*: Gives the user as much information about the original character as possible by means of a legible representation, sometimes using more than one character. Example:

LATIN CAPITAL LIGATURE AE	**>**	**AE**
COPYRIGHT SIGN	**>**	**(c)**
MICRO SIGN	**>**	**u**
AMPERSAND	**>**	**&**

Conversion type 3 -- *reversible conversion*: Uses a one-to-many-representation of characters not available in the target character set which is designed to make re conversion to the original character set possible without any information loss or distortion. Example:

LATIN CAPITAL LIGATURE AE	**>**	**&AE**
COPYRIGHT SIGN	**>**	**&(C)**
MICRO SIGN	**>**	**&mu**
AMPERSAND	**>**	**&&**

The C3 system does not change the interpretation of a character or how it is converted because of its context in the source text. For the most important case in which source context sensitive conversion is needed (the conversion between different ways of representing line breaks) a special mechanism is provided. In addition, there is a mode of using the C3 system

in which special byte sequences in the source text, inserted by a pre-processor, will change on the fly the conversion options used.

When the target character set does not contain the letters of the script used in the source text, e.g. if the source text is written by Cyrillic letters but the target character set is purely Latin-script, the non-exact conversion applied by the C3 system is a transliteration or transcription of the source script in the script available in the target character set. The C3 system thus includes transliteration from the Cyrillic script to the Latin script.

The best choice of non-exact representations depends on several factors, among which the most important ones are the language of the source text (e.g. transliteration rules are usually not identical for Russian and Serbian) and the cultural environment for which the target text is intended (e.g. different transliteration systems are used in English-speaking countries, in Germany, and in the Scandinavian countries). To cater for the different conventions preferred by different user groups, a set of parameters is provided, for which the user can set the values in order to get the most adequate conversion for a particular cultural environment. This constitutes a run-time localisation feature of the C3 system.

In the elementary table of a character set is included full information about the non-exact conversions of all characters outside the character set in the three different conversion types. The modifications of the inon-exact conversions needed for different values of the conversion factors are specified in *overlay tables*, which are applied to the elementary tables before the construction of the working table.

The design of elementary tables for the coded character sets supported by the C3 system is facilitated by the *prototype* *table*. This table gives the preferred representation of every character included in any of the supported character sets by means of a string of characters in the common sub repertoire of all character sets.

III.B. Control Characters and Extended Coding

A special problem in character set conversion is the treatment of control characters. The basic view taken in the design of the C3 system is that the abstract control characters (with encoding units in the intervals 0-31 and 127-160 of ISO 8-bit character sets) may be used to represent different control functions in different contexts. Since control functions are very often represented by sequences of control characters and graphic characters instead of simple control characters, making source context sensitive character set conversion necessary, the C3 system does not attempt conversion of control function representations. Instead, only the encoding units of individual abstract control characters are converted.

In most coded character sets the encoding units are simply bytes, but for more complicated character sets like T.61 (or ISO 6937 known by its non constant length of coding per character or application of ISO 2022 extension technique) the characters can be best treated by being represented as combination of one state (depending on the previous encoding units in the text) and one byte. The C3 system fully supports the conversion of such *stateful* coded character sets.

IV. Parameters for User Control of the Type of Conversion

This section describes the most important parameters by which conversion operations in the view of the user cultural requirements are specified in the C3 system.

Basic parameters:

A1. Conversion system: Indicates which of the available complete conversion systems is to be used. A conversion system is well-defined in all details, so e.g. a new version of the C3 system will constitute a new conversion system, as well as a locally modified variant of the C3 system.

A2. Source character set

A3. Target character set

A4. Conversion type

A5. Line-break string(s) in source text: Indicates all character sequences that are to be treated as line breaks in the source text.

A6. Line-break string in target text: Indicates the character sequence to which line breaks in the source text shall be converted in the target text.

Additional parameter for conversion type 3:

B1. Conversion direction: Determines if an ordinary type 3 conversion is to be done or the reverse conversion.

Parameters for special conversion facilities:

C1. Conversion selector: This parameter determines if the conversion shall affect the whole source text, or if special control character markers in the source text shall control which parts of the text is subject to which kind of conversion.

C2. Indication of non-exact conversions: This parameter determines whether the non-exact conversions will be indicated in the target text or not.

C3. Data error handling: This parameter controls how source data errors are treated.

Parameters affecting established (default) conversion rules:

D1. Target cultural environment: This parameter indicates the general cultural and linguistic conventions according to which the non-exact conversions have been chosen. Values:

0: *Don't bother*

1: *Pan-European*

2: *Swedish*

3: *Slovenian*

4: *Danish*

5: *Others ...*

D2. Cyrillic source language: This parameter indicates the dominant language using Cyrillic letters in the source text. This affects the transliteration of these letters to Latin letters. Values:

0: *Don't bother*

1: *Russian*

2: *Serbian*

3: *Macedonian*

D3. IBM PC control character handling (source text)

D4. IBM PC control character handling (target text)

These parameters indicates to which extent encoding units in IBM PC character sets are to be treated as control characters.

D5. This parameter indicates which variant ISO 646 characters, if any are to be regarded as the corresponding ASCII character and thereby protected against conversion in many cases. Affects the source character set tables.

V. The Internal Operation of the C3 System

The C3 system supports a character repertoire formed by including all characters of the 39 coded character sets which are used in Europe. (This includes the 65 control characters possible in an ISO 8-bit character set in the C0, DEL, and C1 areas.) A limited number of additional characters from the European subset of ISO/IEC 10646, currently being designed by CEN/TC304. This repertoire is called the United Repertoire.

The Common Sub-repertoire of the coded character sets consists of the 83 invariant graphic characters (including SP) of ISO 646:1992.

The elementary table in the C3 system for a coded character set indicates for each character of the United Repertoire if it is included in the character set or not. If it is, the table specifies its encoding unit. If it is not, the table specifies what non-exact representation of the character is to be used, when converting text to this character set. This is done for each of the three conversion types. The characters of the United Repertoire are specified by means of their encoding unit in UCS-2 (ISO 10646-1:1993).

The prototype table is an aid in the design of elementary tables. It specifies for each character in the United Repertoire that is outside the Common Sub-repertoire a prioritised list of one or more non-exact representations of this character. The alternative with the lowest priority shall use only characters of the Common Sub-repertoire. One such list is given for each conversion type. The characters are specified by means of their encoding unit in UCS-2.

To automatically construct a full elementary table for a coded character set, the prototype table can be used together with the partial elementary table, where the encoding units of all characters in the character set are included. For all remaining characters one non-exact representation of it in the prototype table is chosen, namely the highest prioritised one all characters of which exist in the character set. When this representation is included in the elementary table, the encoding units in the coded character set are used for all characters.

Prototype tables and elementary tables of the C3 system are specified using a special textual table format. It is designed to satisfy these requirements, among others:

- ***The table files should be directly human-readable text files and also serve as control files for general conversion programs.***
- ***Facilities should be provided in the format to make it easy for a user to express what he means, without making the parsing of the table by a program too complicated.***
- ***The table files should be transferable between systems with different coded character sets. They should work without conversion in all computer systems as control files for a conversion program.***
- ***One shall not have to write the same information twice in a table file.***

The working table directly applied by the C3 system in the actual conversion of the source text has an entry for each possible encoding unit in the source character set. It specifies the sequence of encoding units in the target character set that the source encoding unit shall be replaced by.

VI. Concluding Remarks

There are many other conversion systems in the character set codes world but they differ from C3 system in their coverage of the scripts and the functionality. In general they provide conversion between a set of selected 8-bit character sets with the same or similar character repertoire, the missing characters in the converted data are represented with some "error" characters or some "similar" replacements. The C3 system is different and unique because it provides conversion between any pair of the supported character sets, enables three conversion types: 1-to-1, legible, reversible and re conversion (if required) to the source code. The main property of the C3 system which makes it close to the requirements for an internationalised product is the transliteration facility and the selectivity in representations of the characters defined or chosen by the end user.

Additional important features of the C3 system reflect the requirements for portability of such system and its integration in applications such as SMTP, PP mail gateway system and terminal emulation. This is provided by the system API which enables simultaneously a conversion several source texts. The API is designed to support also context-sensitive texts. We can also say that the major advantage of the C3 system is the handling of the defined subset of UCS (implementation level -1), UTF-2 format of data, Quoted-Printable, Base64 and other byte-oriented data formats. Some features in the API are also designed for the support and handling of representation-ambiguous character sets and flexible treatment of data errors.

The API for the C3 system is written in C for UNIX. The concept used follows the reference descriptions used in "Xlib Reference Manual for version 11" of the X Window System. The implementations for MSDOS and MAC as well as the integration in SMTP, PP mail system gateway and terminal emulation are underway.

The future extensions of C3 systems are planned towards handling bi-directional text, conversion between logical and visual encoding of Thai/Lao text and implementation of level 3 of UCS.

VII. References

[1.] ISO 646, 3 ed, 1992: "ISO 7-Bit Coded Character Set".

[2.] ISO 2022, 3ed, 1986: "Code Extension Technique".

[3.] ISO 4873, 2ed, 1986: "ISO 8-Bit Code Structure and Rules".

[4.] ISO 6937, 1991: "Coded Character Sets for Text Communication".

[5.] ISO 8859-x, 1ed, 1987: "8-Bit Single-Byte Coded Graphic Character".

[6.] ISO 10646: "Multiple Octet Coded Character Set".

[7.] Unicode : "16-bit multi-lingual character set code".

[8.] ISO 6937, 1991: "Coded Character Sets for Text Communication".

[9.] ENV 41 503: "European graphic character repertoires and their coding".

[10.] ENV 41 508: "East European graphic character repertoires".

[11.] Working documents of the RARE WG-CHAR and C3 TF, 1993

[12.] Olle Jarnefors: Status Report: "Overall design of C3", 1993.

[13.] Nathaniel Borenstein, Ned Freed, RFC, "MIME -Multipurpose Internet Mail Extensions", 1991.

[14.] Keld Simonsen, RFC on Mnemonics, 1992

[15.] B.Jerman-Blažic, "Character handling and computer communications, in User needs in information technology standards", eds C.D.Evans, B.L.Meek, R.S.Walker, Butterworth Heinemann 1992.

Acknowledgement:

This paper is the result of the collective efforts of the C3 TF RARE WG-CHAR members: Olle Jarnefors, KTH, Peter Svanberg, KTH, Keld Simonsen, DKUUG and Borka Jerman-Blažic, IJS.

Author Information

Borka Jernan-Blažič is a chair of the Laboratory for open systems and networks at Jožef Stefan Institut, Ljubljana, Slovenia. She is teaching postgraduate course on Telecommunication Services at the Faculty for Economics, University of Ljubljana. She is a member of the ARNES (Slovenian academic network) Steering Committee and member of RARE Technical Committee. She is convenor of RARE WG on Character Sets and national representative to ISO JTC1 SC2, JTC1 SC22WG20 and CEN TC 304.

Her research currently focusses on networks applications and issues related to internationalisation of software applications. She is also interested in development of security services and CA infrastructure. She has been co-ordinator of the ex-Yugoslav part of the project COSINE and General Secretary of YUNAC. She is author of more then 150 published scientific papers.