Recommendation 6: Processing multiple languages, character sets and encoding

Within the global trading environment, the supplier chain will inevitably cross national boundaries and lead to the use of local character sets and cultural conventions. Existing international standards do not provide full support for such character sets and conventions, which by their nature should be independent of individual applications. In addition, the inclusion of multiple character sets in any one set of product information becomes an essential requirement.

Character sets

The multi-octet character sets in the ISO/IEC 10646 standard do not support the full range of character sets used internally in nations, and the extension mechanisms to be used by ISO should be consistent with that used by UNICODE.

Recommendation 6.1: It is recommended that ISO/IEC establish a consistent policy for encoding multi-octet character sets. In addition, ISO/IEC 10646 should be reviewed to incorporate more national character sets, and to separate out merged national standard character sets. This must be undertaken in cooperation with the UNICODE organisation to ensure compatibility for the code extensions, and to establish a migration path for existing information.

Application independence

The use of the content of character strings to control the processing of the information can limit the use of multi-octet characters sets.

Recommendation 6.2: It is recommended that all ISO digital data standards that may include character strings should be modified if necessary to be independent of the content of the string.

Cultural issues

The basic philosophy of supporting local cultural variations is described in ISO/IEC TR 11017, using the methods of internationalisation and localisation.

Recommendation 6.3: It is recommended that an additional study be undertaken of the feasibility and implications of applying the principles of ISO/IEC TR 11017 to support cultural variations in ISO standards.

ISO HLSG ON CALS

Processing multiple languages, character sets and encoding

1. Introduction

Many countries become understand the importance of the CALS systems in order to increasing agility, productivity and reliability of the products and development processes of the industries in the countries. So many companies in many countries are becoming want to use CALS systems. In other word CALS systems should be usable by every companies all over the world.

The business activities become more and more internationally. Also the industries activities which are developing products become more and more internationally. Moreover the relations among the contractors, the venders of assembly and the suppliers of the parts become more and more complex and such relations will become like the chains and so it is called 'Suppliers' Chain'. It is very important to establish the suppliers' chain internationally, to maintain it, and to change the structure or members of that chain easily and timely. The international suppliers' chain will be usually containing many scale of companies, very small ones to very large ones, and also it will be established from many companies in many countries, and in those countries they will be using alphabetic languages or non-alphabetic ones. And is it easy or not to establish, maintain and control the international suppliers' chain will become fatal issues of the success of the businesses. In other word, the success of the international business and the multiple national projects will depend how to easily and timely establish, maintain and control the international suppliers' chain.

There may be sometimes difficulty to establish such international suppliers’ chains because of the difference of cultures among the members countries. Especially the handling mechanisms of 'languages' and 'cultural items’ related issues are very important. Up until now CALS related standards has been developed mainly in United State and European countries. So there are, sometimes, lack of the handling mechanisms of multiple-octet characters, and usually no consideration about how to handle the cultural issues, this function is known as the internationalization and localization. Many countries have their own languages which are not expressed by Latin alphabet characters and many languages are able to expressed only by the multiple-octet characters. Because of these reason, there may be some difficulties how to use these international standards in many countries.

So in this section we want to clarify the problems and what kinds of cares are necessary when developing the international standards which can be used by many countries, especially the countries which have their own languages and which is not expressed by Latin alphabet characters .

2. Scope

In this section, 'processing of multiple languages' means only the documentation related processing and exchange of documents via communication mechanisms such as mailing systems, file transfer systems, etc..

Such as the automatic translation between different languages, recognition of voice, hand written characters, and other natural languages related issues will be not discussed in this section. Because techniques of the automatic translation of the natural languages are pre-mature to internationally standardize but if the translation is limited to only between item names or restricted to the specific industry and commercial sectors then there may be some techniques such as using the BSR or the Mediator techniques, or other specific techniques. Of course the automatic translation techniques of the natural languages are very important so research activities of the techniques are recommended. And also the techniques of the recognition of voice, hand written characters are pre-mature to internationally standardize for the natural languages which have many pronunciations and many characters. Moreover it is possible to process these matters in the front end and the back end processors, and the main processor only handle coded characters stream as the data.

There are many issues expressed by 'processing of multiple languages', but if those issues will be able to handled by the front end processors and the back end processors and they may be able to transfer only encoded characters to/form the main processors which will be processing CALS standards related functions. And if encoded characters data stream from multiple character sets will be able to processed by the main processor then there should some way to input and output, these are containing mailing, file transfer, data store and so on, these characters. For examples:

- processing of documents which contains characters of multiple character sets;

  Main difficulties of this case is how to input and output the characters and so if the front end processors and back end processors have suitable input and output mechanisms for every character sets and those processors can transfer encoded characters to/from the main processor it will be enough to consider the functions which will be done in the maim processors and these functions are mainly how to handle the documents which are written in multiple languages.

  In Asian area there are research activities how to handle the multiple languages but it is just started and pre-mature to internationally standardize.

- store of characters of multiple character sets in same data store, database and file;

If the main processor will be able to process these data stream it will be not so difficult to make these issues possible because there are some mechanisms to identify which characters are member of which character sets

- many other issues will be treated same way mentioned above

So we can focus our discussions on how to handle the data stream which is constructed from characters belonging multiple-octet character sets.

And in this section 'encoding' means only encoding scheme of characters. So encoding of images, graphics, and voice, etc. are excluded. There are also other types of encoding such as encryption, data compression and so called document transfer encoding, etc., and they are also excluded in this section. Because they are not so deeply depending different cultures of each country, especially character sets, or some are pre-mature to internationally standardize.

3. Analysis of problems

As mentioned in Introduction, a clear problem is lack of the handling mechanisms of multiple-octet characters in the international standards. ISO/IEC JTC 1/SC2-WG2 had developed an international standard ISO/IEC 10646(UCS: Universal Multiple-octet Coded Character Sets), especially in the BMP(Basic Multi-lingual Plane) there are many countries characters are encoded. Many countries hoped newly developed standards will have the mechanisms to handle the UCS, especially supports of the BMP, and also existing standards will be modified to add such mechanisms. But to support current BMP of UCS is not enough to processing all languages because it is merely a collection of the various, but not exhaustively, character sets used throughout the world and takes up the national standards character sets, as is, in reconstructed form. Usually many countries have other character set standards which are not merged into the BMP, for an example in the Japanese kanji case only so called level-1 standard character set is merged into the BMP but there are so called level-2 and level-3 standards character sets in Japan. There are same problems for the Chinese, the Korean and other many countries character sets. Moreover there are many countries, their character sets are not merged into the current BMP. So at this moment there are no internationally standardized solution for processing of multiple languages because there are no real universal character coded sets which contains all characters in the world. And it is needed to review the encoding scheme of character sets if it is preferable to process multiple language by one character code set, currently it will be ISO/IEC 10646.

It is not so difficult to standardize functions to handle the UCS if in the standards the definition of processable character strings are made by ASN.1 notation. In this case, it may be enough to refer the latest version of ASN.1 standard or only add the UniversalString to definition of character strings. It is out of scope how to implement the actual systems based on these new international standards. But by a report of working group for Ideographics Character Interoperability of AOW(Asia and Oceania Workshop), it was not so difficult to modify existing implementation of OSI FTAM to support the UCS. If the standards are not using ASN.1 definition, it will be needed some considerations but which may be not so difficult.

So it seems enough for handling multiple-octet character sets to support the UCS. But there are two big issues for this solution. One is migration from currently used character sets to the UCS because they are not compatible. It is very difficult, more exaggerately saying it will be impossible to do such migrations because there are already exist much volume of data in data store. Because data are mixture of single byte coded characters and two bytes encoded characters and usually systems can not identify which is a one byte code and also which is a two bytes code. Usually users' application can identify them. Users do not want migration of their data to the UCS or to the BMP. So some countries, who already developed the national standards for multiple byte character sets according to the ISO 2022 character sets enhancement mechanism and which are widely used in that countries, are heavily discussing how to realize the implementations of the UCS in the countries. Another is no more enough space to add new characters into the BMP. This issue will be solved by using other planes when to add many characters to the UCS. It is basic policy of the UCS to use multiple plane to supports many characters. But there is a serious problem that current BMP is compatible to the UNICODE, the de-facto standard of the multiple-octets and multi-lingual character set, and the UNICODE is also now under the consideration of expansion to add new characters, but the expansion mechanism is not same to above mentioned UCS expansion. So it should be continued to close contacts and discussions between ISO/IEC JTC 1/SC2 WG2 and the UNICODE. Inc..

There is one more critical problem to processing of multiple languages. It is how to treat the countries specific cultural items. For an example, when to use date in documents according to the international standard and in which only yy/mm/dd type of the date presentation is allowed, many countries may claim and against that standards. There are many culture dependent items. The example of cultural items are listed in annex 1. The treatment of cultural issues is known as the internationalization and the localization and this basic philosophy is described in ISO/IEC TR 11017. POSIX group already discussed these issues. So in this paper we only point out the importance of the mechanisms of the internationalization and the localization.

4. Current status

It is very difficult for many countries to use only English based(it means ASCII code based) information processing environments because it is very difficult to enforce the usage of only English based system for every users if they are in non alphabetic countries. In many countries if the international standards do not have handling mechanisms of their own languages and the cultural items, each country will develop the national standards to add the handling mechanisms of the own national language and cultural items. According such national standards almost every countries have information processing environments to process their own national languages and cultural items. But unfortunately sometime there are miss understanding of original international standards and then the national standards become incompatible to the international standards. It is needed very heavy efforts to modify the international standards to the national standards and also to develop the national standards based information processing environments. Even if in such cases almost every national standards keep the original functions in the international standards, it means almost every national standards are compatible to the international standards when processing ASCII code(ISO/IEC 646). So if there are needs of data interchange between different countries it will be able to do so by using ASCII code based data. It is enough for the persons who are very good at English but it is usually difficult to request English ability for every persons. In such case many companies in non alphabetic countries usually translate English documents into their own languages and after translation they will distribute the documents to appropriate sections and workers. It will be sometimes become very big time loss and very expensive for companies, especially for the small and medium scale enterprises. It will also become the reason of difficulties to establish international suppliers' chain.

It is sometimes happen that the national standards are compatible but actual implementations based on those standards are not compatible with implementations which are developed based on the international standards. There are some reasons why such cases occur. One is the nationals standards only seem compatible with international standards but actually they are not compatible. Another is the implementation techniques problems. Because the national standards which are developed based on the international standards are including expanded functions, in this case the handling mechanisms of own countries language processing, so at when implementing the system implementers sometimes add special flags or other variations to the systems even if when users are only using ASCII code. There are actual examples of such cases in some countries. In such cases many companies are using international versions for exchanging documents with foreign countries and to use domestic versions for in their countries. It usually becomes very big over head and also users will confusing which systems they should use.

There is another important problem which is the handling of each vender's specific expanded characters and each user's developed characters. Usually standardized character coded sets have some spaces in where no characters are encoded and each vender and user now freely using this spaces in order to add new characters which they want. This problem will be only solved to prohibit to use these characters if user want to exchange their documents with persons who are using the different systems. These characters should be only used between same venders and users system.

There are appearing the systems which are supporting the UNICODE. But in many cases the UNICODE is only used as an internal processing code set and when do the functions of the input and output of the characters they are usually converted to/from the national standards character codes. Because the existing input and output devices can be only processing characters encoded number of the national standards character sets. Almost every character encoded numbers are different between the national standards and the UNICODE. It arises even if in the case of ASCII code if the pure UNICODE is used in the systems(in this case lower 7 bits are completely same to ASCII encoded number so only ignore upper 8 bits). It is possible to develop new input and output devices to process the UNICODE but many users already have such devices and it is very difficult to enforce them to buy new devices because of the change of using character sets of the systems.

There are also appearing the systems which have some degrees of localization functions. But the supported functions are very limited so issues of handling of the localization still open issues for the computer venders.

5. Recommendations

It is very important for non alphabetic languages countries to have capabilities of processing of multiple languages in every international standards. So ISO/IEC should clearly express their opinions how to treat this issue. In other words, ISO/IEC should establish consistent policy of how to define and develop multiple octet character sets and how to treat the occurrences of different encoded characters in the same data store and how to exchange such data stream via network systems. From the technical view points, following recommendations will be important and helpful to establish the policy.

(1) The international standards, in which character strings will be processed, should be developed as character coded sets independent where the users will make their own data. But such as key words and some other fixed terms, expressions and items are excepted, it means such items may be defined in ASCII code.

To realize such standards and systems it is needed some mechanisms how to designate the using character set, for example, in the case of file transfer it will be needed a new file type definition, in the case of MHS it will be needed to specifying a new contents type, and sometimes it will be needed to add negotiation process to identify the using characters set before sending actual data.

(2) If it is difficult to realize above recommendation (1), the international standards should be developed to usable ISO/IEC 10646 as character coded sets. Moreover escape sequences in the data streams, which will change the character sets names following these sequences, should be appropriately handled in every implementations.
(3) ISO/IEC 10646 should be reviewed to include more countries character coded sets and to expand merged national standards character coded sets. It is necessary much contributions from every countries who want to have their own language processing mechanisms in internationally standardized information processing environments. It is also needed to continue close discussions and cooperation with the UNICODE Inc. because currently the UNICODE seems de-facto standard in the fields of two-byte coded set.
(4) It is preferable for the international standards to have the internationalization and the localization capability. So the more exhaustive studies and more concrete investigations of this technical area and the philosophy of the internationalization and the localization described in the document ISO/IEC TR 11017 will be needed.