ISO/IEC JTC 1/SC 35 N 0749

DATE: 2005-01-31

ISO/IEC JTC 1/SC 35

User Interfaces

Secretariat: Association Française de Normalisation (AFNOR)

TITLE: Name-writing in Swedish Authorities' data processing - Background information

SOURCE: Karl Ivar Larsson, Swedish Expert

STATUS: FYI

DATE: 2005-01-31

DISTRIBUTION: P and O members of JTC1/SC35

MEDIUM: E

NO. OF PAGES: 16

Secretariat ISO/IEC JTC 1/SC 35 – Nathalie Cappel-Souquet – 11 Avenue Francis de Pressensé
93571 St Denis La Plaine Cedex France
Telephone: + 33 1 41 62 82 55; Facsimile: + 33 1 49 17 91 29 e-mail: nathalie.cappel-souquet@afnor.org

Note:

This document was handed out in the SC35 Stockholm meeting 2004-11-24.

Contents

Name-writing in Swedish Authorities' data processing – Background information

(Note: Since the contents of this document may be of interest in some Authorities' international contacts, it has been written in English.)

1 General background

Electronic interchange of information between Public Authorities in Sweden is increasing in importance extremely rapidly, as is interchange between Authorities and citizens as well as companies. The Swedish Government has decided to promote the "24-hour Public Administration" concept, to provide continuous access to public information.

Information interchange is however often made difficult by a lack of compatibility between different Authorities' data processing and/or communication systems. A multitude of such systems exist in the marketplace, some conforming to formal standards, and others based on "proprietary" – i.e. company-specific – principles. The Government has therefore found it essential:

"To decide on a minimum of binding rules and standards necessary for a well functioning electronic communication within the public administration and with its customers and to provide a supporting set of basic functions as a common infrastructure for the communication and co-operation between the different public agencies." *(Extract from speech by Gunnar Lund, Swedish Minister for International Economic Affairs and Financial Markets).*

The Ministry of Finance has appointed "Nämnden för elektronisk förvaltning" (The Swedish Board for electronic administration, e-nämnden) to initiate necessary work on guidelines, later to be turned into formal regulations. The Board started its work in January 2004, and has commissioned a number of specific tasks.

One such task is to develop a set of guidelines to satisfy the needs in data processing for correctly representing names of individuals, companies and geographical places. With today's international situation – including the Swedish EU membership – this requires the handling of a set of letters vastly larger than was earlier needed.

The Board has commissioned Statskontoret (The Swedish "Agency for Administrative Development" – to investigate the matter and propose guidelines to be referred for comments to Authorities and others concerned. Statskontoret, in turn, has commissioned part of this investigation to the company LWP Consulting. This document describes various factors of relevance in the matter, including:

- An overview of languages and their writing systems
- Factors relating to representation of letters and other characters in data processing
- Proposed guidelines for the representation of names in Authorities' data processing

2 Languages and writing systems

2.1 General aspects

A very large number of languages exist in the world, although the number is rapidly decreasing as languages spoken by small minorities become extinct. Considerably more than 3.000 languages could however probably still be identified, although some of them can be considered dialectal varieties of a common language. Many of these languages/dialects however lack an established writing system, and are therefore not candidates for commonplace data processing. The number of languages in the world

used by a larger number of speakers and also possessing a writing system could be put at around 200.

The number of characters needed to correctly write even this smaller number of languages is extremely large. This is mainly because the writing systems of the three important languages Chinese, Japanese and Korean use an "ideographic" system of writing, with tens of thousands of characters. (The common term "ideographic" is used here loosely and not strictly correctly; also it should be noted that the indigenous Korean system Hangul is really syllabic, although used in a way necessitating data processing similar to the truly "ideographic" systems.)

In comparison, the number of characters actually used in international exchanges of computerised information is small. Most such exchanges use the Latin, the Cyrillic or the Arabic script. Information originating from languages with other writing systems is then "transliterated" into one of these scripts (the term "transcription" is also used in the connection, but here only "transliteration" is used, without a strict definition).

2.2 Minority language issues

2.2.1 General factors

One factor to be considered for the data processing of names is majority- vs. minority-language use. In most countries, public administrations permit writing of names only in the writing system of the "official" language, although this is often not clearly stated in national legislation. In Sweden, for instance, the law governing the taking of a new personal name (SFS 1982:1134) simply states that a surname cannot be accepted if its "composition, pronunciation or spelling" is "not suitable". This text is, naturally, not a very precise basis for interpretations in a specific case.

The matter becomes even more complicated by the fact that some countries do not explicitly declare a "state" language. A survey of the matter in the case of Europe (see References) however concludes that: "Most European constitutions declare one or more languages as the state, national or official language. Generally these three terms are used with one and the same meaning, i.e. that the languages declared are those used by the authorities, particularly in legal connections."

Out of the 47 states recognised by the Council of Europe (CoE) as European – i.e. being either CoE members or eligible for membership – 28 explicitly declare one or more state/national/official language(s) in their constitutions, and another two do it indirectly. However the remaining countries, amongst them all of the Scandinavian states, do not declare a state language.

The matter of language minorities, and therefore also of their requirements on name-writing, is now becoming internationally recognised because of the CoE "European Charter for Regional or Minority languages" (CETS No. 148). This charter has at present (September 2004) been ratified by 17 CoE member states, including Sweden, and further signed – although not yet ratified – by another 13.

2.2.2 The Sami situation in Sweden

Part of the Swedish ETS 148 ratification is an undertaking

"to allow and/or encourage **. . .** the use or adoption, if necessary in conjunction with the name in the official language(s), of traditional and correct forms of place-names in regional or minority languages."

and further

"to allow the use or adoption of family names in the regional or minority languages, at the request of those concerned."

These parts of the Swedish ratification apply to Sami, Finnish and Meänkieli (Tornedal Finnish), and are supported in legislations SFS 1999:1175 and 1176. While Finnish and Meänkieli are well-defined languages, "Sami" is somewhat unspecific, since several varieties exist, generally considered separate languages, with partly differing alphabets. In the absence of a clear specification, the Swedish legislation should be considered to apply to all Sami varieties.

The data processing requirements of Sami have been specified in documents issued by the joint Finnish-Norwegian-Swedish Sami Parliamentary Council; see References.

2.3 EU working languages

After the latest EU enlargement the number of official working languages in the union's institutions is 20. Irish Gaelic and the Luxembourg language Letzeburgesh – which is really a German dialect – were not requested as working languages; and Belgium, Austria and the EU-recognised part of Cyprus use the majority languages of other EU countries.

At present two scripts are in use within EU, the Latin and the Greek. With the possible future entry of candidate country Bulgaria the Cyrillic script will also come into use.

It should be noted however that the Swedish situation for personal names already makes handling of Cyrillic-script names necessary, on a very much larger scale than Greek-script ones. Also Turkish-origin names are important in Sweden.

The EU working language situation therefore not really influences needed decisions on name-handling in Swedish Authorities.

2.4 Writing systems

2.4.1 Concept of characters and "glyphs"

A problem emerging with the introduction of data processing was the need to differentiate between characters as "conceptual entities" and their corresponding rendering, i.e. their written, printed or screen-displayed representation. For such representation the term "glyph" is used. (Strictly speaking the rendering is a "glyph image", while a "glyph" is a conceptual shape. In this document the term "glyph" – in Swedish "glyf" – is however used with the meaning of "glyph image", for simplicity.)

The reason for this problem is that writing systems have "borrowed" glyphs from each other. For instance, the first letter of the names Aachen, Архангельск and Αθήναι have the same shape, but are different characters**:** Latin A, Cyrillic A and Greek Alpha, respectively. The difference is of no consequence to a human reader, but it is to computers.

On the other hand, differences in fonts are generally unimportant in data processing, except in output. Whether a name is rendered in a specific case as e.g. Aachen or *Aachen* is irrelevant to any foregoing processing of the name.

2.4.2 Letters in Swedish data processing

It would of course be desirable that all official data processings of names that are in any way in use in Sweden could employ their correct original representations, whatever their writing system used. This is however unrealistic – except possibly in some very specialised bibliographical data applications – both because of limitations in computer systems and because officials could not in general be expected to be familiar with more than the Latin-alphabet writing system.

It therefore appears necessary that data-processed name information from Authorities is always made available in the Latin script, i.e. transliterated. This does not exclude

Authority-internal processing in other scripts; and/or the storing of names in a form complementary to their Latin form, e.g. in the original writing system, if possible and suitable.

2.4.3 Latin-script considerations

The Latin script, being the most widely used writing system in the world today, had originally 24 letters, the letters J and W being later additions. The present 26-letter alphabet is in principle sufficient only for English and Dutch, and for all other languages additional letters are needed.

Such additional letters are either formed by adding a diacritical sign to a base letter, e.g. the acute accent in é, or by making a small variation on an existing letter, like the Maltese ħ. In some cases a letter variation traditionally used for phonetical purposes has been added to a script, e.g. the letter ŋ used in Sami.

Diacritical signs generally indicate some change of the phonetical value of a letter, e.g. the French e, é and è corresponding to three different pronunciations. In some languages however, an acute accent may be used to indicate stress. (A Swedish EU-parliament candidate spells his name *Sacrédeus*, indicating desired stress on the second vowel**;** the common pronunciation in Sweden of names ending in -eus is with the stress on the next-to-last vowel.)

Conversely for instance another diacritical sign, the diaeresis, is used in e.g. French not for phonetical purpose, but to indicate specific pronunciation, namely that two adjacent vowels shall be articulated separately (like in *Citroën*). In some other languages the diaeresis is phonetical, indicating an "umlaut" (like in *Köln*). To further complicate matters, in Swedish and Finnish the letter ö – as well as the ä – are considered complete and separate letters, not phonetical variations on o (and a); and therefore also ordered separately, at the end of the alphabet.

The original way of employing diacritics in data processing was taken over from typewriting, by "building" a letter through two successive operations. This is still the method for entering unusual letters from keyboards – e.g. é and ü from Swedish keyboards – but the internal representation of such letters in practically all computer systems is nowadays "pre-composed", i.e. all occurring variations on a base letter have their own unique representations.

Therefore in data processing the Latin script has a need for many times the basic 26 letters. With both the special-shape letters and the letters with diacritics there are more than a hundred Latin letters in European languages. Also, these letters must be available in both lower-case and upper-case form – even if some of them may not occur as the first letter of a name – thereby doubling the number of letter-type characters that must be handled by computer systems.

3 Character data processing

3.1 Fundamental requirements

Computers are intrinsically number-processing devices. For their use in processing character-based information three fundamental requirements must be satisfied:

1. It must be possible to unambiguously represent the needed characters internally in the computers, as well as on external storage media and in data communication.

2. It must be possible to input the needed characters.

3. It must be possible to output the needed characters.

These requirements may appear self-evident, and therefore trivial. What is often not sufficiently realised, however, is that solutions satisfying these three requirements are largely independent of each other, and must consequently sometimes be tackled separately.

In particular, character input – for which the keyboard is still the most important component of computing systems – must be given special attention, since factors like ergonomics are so important. Character output, requiring fonts, is in comparison a smaller problem, and is not treated in this document.

3.2 Character representation

The early computers, emerging in the forties, were purely computational, i.e. working with arithmetic. Input and output was only of numbers and mathematical operators.

With the second-generation computers in the fifties, some limited processing of characters was introduced. Since the computers were basically intended to perform arithmetic, their "architecture" was based on comparatively large units of information, to achieve sufficient precision in computations**:** "words" consisting of 24 bits was a common solution. Representations of characters, when needed, were then packed into the words according to some computer-specific scheme.

With the third-generation computers in the sixties, processing of characters was recognised as a highly important field, along with the traditional arithmetic. This led to a new architecture for computers, with machine instructions adapted to handling character units represented by seven or eight bits. The machines' arithmetic instructions then worked on connected multiples of the basic 7- or 8-bit units.

Several different solutions emerged, but the architecture of IBM System 360 computers quickly became completely dominant. Its basic unit was the 8-bit "byte" (the term being an IBM invention), and the units could be connected for arithmetic operations in half-words of 16 bits, words of 32 bits and – in some cases – double-words of 64 bits. This architecture has since become the conventional one, especially after IBM used it also in the PC introduced in the eighties. (Note: The formally accepted term for an 8-bit unit in data processing is *octet*. In this document, however, the more common term *byte* is used throughout.)

The 8-bit principle made possible the representation of 256 different characters. Some of these were needed for control functions, like e.g. "carriage return". In the 360 series, IBM used a proprietary representation – "coding" – scheme named EBCDIC, in which the initial 64 characters are set aside for control purposes. Consequently a maximum of 192 "graphic" – i.e. printable – characters could be represented.

In the EBCDIC scheme only 114 of these were however alphabetic characters, the remaining bit-combinations used for punctuation signs, currency symbols etc. Ten combinations were also needed for the digits 0 to 9 (which in this connection are printable characters, not quantities).

As should be obvious from the section above on languages, the alphabetic characters available in EBCDIC are completely inadequate to cover even the Latin-script alphabets of Europe. IBM therefore developed a large number of alternative EBCDIC schemes, each one intended to cover a specific language (like French) or a geographical area (like Denmark/Norway and Finland/Sweden). This enabled nationally-adapted data processing, but made error-free international exchange of information quite complicated.

The need to standardise the representation of characters had been recognised already in the fifties. The rapid expansion of data processing in the sixties, and especially the adoption of the IBM products, made the matter highly urgent.

3.3 Standards for character representation

3.3.1 The ASCII scheme

Already in the fifties, work was started in US standardisation ANSI in its committee X3.4 to define a coding scheme known as the "American Standard Code for Information Interchange" (ASCII). The scheme was intended to cover the minimum character requirements for the US, and also to be acceptable to a large number of computer manufacturers**;** at the time many companies world-wide were designing and producing computers, most of them "incompatible" with each other. The ASCII was therefore designed as a minimal scheme, based on a 7-bit principle.

The original scheme used 35 bit-combinations for control characters. The graphic – i.e. printable – characters consisted of the capital letters A–Z, the ten digits, the $ symbol, and a number of punctuation signs. It left 29 combinations undefined for future use. The standard was published in 1963.

Work to complement the initial scheme continued, partly in co-operation with international standardisation. This resulted in a second ASCII edition, published in 1967. In it, the number of control characters had been reduced to 33, and the small letters a–z and some more punctuation signs added. This version was to become the definitive one.

At the time however, the 8-bit EBCDIC scheme had become dominant in data processing, and the ASCII scheme therefore covered only a subset of the characters that had become available to many computer users. The need to develop an "8-bit ASCII" was consequently obvious. Extensive work on the matter was done in co-operation between various organisations, primarily ANSI and the European Computer Manufacturers Association (ECMA), and in 1984 the latter organisation submitted a formal proposal to the International Organization for Standardization (ISO) for such a scheme, which eventually resulted in the "Latin-1" as described in the next section.

In the various EBCDIC variants, the characters appear to be distributed randomly, although some explanation for the structure of the schemes can be found in early IBM products. On the other hand, the coding of characters in ASCII is based on logical principles.

ASCII has therefore become the firm basis for practically all coding schemes developed since the sixties. Not only are all international coding standards supersets of ASCII coding-wise, up to and including Unicode (see below), but those international standards also became the basis for most computer manufacturers' implementations. Therefore, the proprietary schemes for the IBM PC are supersets of ASCII, both in the original DOS operating system and in OS/2 and Windows, as are the schemes for the Apple computers.

The main – but highly important – exception to the ASCII-based structure is EBCDIC, on which much of IBM's product line is still founded.

3.3.2 International standards for 8-bit schemes: ISO/IEC 8859

Originally a number of international standardisation bodies in addition to ISO worked on character coding, amongst them the International Electrotechnical Commission (IEC) and the International Telecommunication Union (ITU), whose standard developments part was previously named CCITT. Eventually ISO and IEC decided to establish a Joint Technical Commission, ISO/IEC JTC1, to handle standardisation work in the area of Information Technology. As regards ITU, it has now limited itself to develop standards related to telecom and not covered by the work in ISO/IEC JTC1.

As described above, the work on an "8-bit ASCII" was taken over by international standardisation, engaging mainly ISO. The result of this work was the ISO standard 8859 part 1, known as "Latin-1", and published in 1987. This scheme quickly became implemented in a number of computer systems, in some cases even before the standard

had been formally finalised. Later, it was also used as the basis for IBM OS/2 and Microsoft Windows. The corresponding Windows scheme, Code Page 1252 "Windows Western", is a superset of Latin-1, with a few characters added in coding positions left unused in Latin-1 (and not really available for use, according to the 8-bit coding framework established by ISO/IEC, which Windows is not fully conformant with).

As the Windows name implies, the coding scheme is intended to cover Western European needs in data processing (and thereby also the US needs). Actually the set of characters of Latin-1 is identical to that of the US version of EBCDIC, and information coded in that EBCDIC version can therefore always be converted to Latin-1 coding, and vice versa.

Like for EBCDIC, the 256-character limit of the ISO 8859 necessitated alternative schemes to cover more language areas than Western Europe. Today there are no less than 16 parts of ISO 8859, some of which have never become implemented by industry, although possibly forming the basis for applications developed by some user communities. (Only Part 1 is specified in References.)

3.3.3 International standards for 8-bit schemes: ISO/IEC 6937

Another ISO standard, developed in co-operation with ITU, is of interest, although not implemented by any computer manufacturer. This 8-bit standard, ISO 6937 originally published in 1983, preceded ISO/IEC 8859.

The standard is based on an excellent principle, namely to define the base letters A–Z (and the corresponding lower-case a–z) as well as European-language special-shape letters, and in addition to that a number of separate diacritics for combination with the base letters. The principle of the standard can therefore be seen as equivalent to the "type-writer method" described above for producing letters with diacritics. Excepting some small defects in its set of special-shape letters, implementation of this standard would have solved the problem of representing practically all European Latin-script letters in an 8-bit environment.

Unfortunately the principle causes problems in computerised text processing, since a letter is represented by either a single character or, in the case of base letters with diacritics, by two characters**;** making programming of editing operations complicated. The standard, therefore, has become the basis only for some user-developed applications, not for any commercially-available software.

ISO/IEC 6937 is however also of interest from another aspect, namely as an authorised documentation on the letter needs of European languages, enumerated in the standard in tabular form. Although met with opposition from some linguists, that table contains important information. Also, the standard is of interest because of its connection with keyboard standardisation (see below).

3.3.4 Other international standards for 8-bit schemes

Several other ISO/IEC standards for 8-bit environments exist. The standards ISO/IEC 2022 and ISO/IEC 4873 are of special interest.

Together these two standards define an 8-bit extension mechanism to handle more than 256 characters at a time. In the first place, a maximum of 383 graphic – i.e. printable – characters are thereby possible. This would in theory permit 8-bit representation of all Latin-script characters needed for the majority and recognised-minority languages of Europe.

These two standards form the basis for the European standard CEN EN 1923, first published in 1998, and recently updated as CEN TS 1923:2003. It covers most of the Latin-script character sets defined in the different parts of ISO/IEC 8859, and also the Cyrillic and Greek sets of its parts 5 and 7. Its coverage is not completely sufficient for European minority languages, however.

The main objection to the 2022/4873 extension mechanism is that its computer implementation is quite complicated. Also the possible increase in the number of characters is insufficient for many situations. Theoretically the framework defined by these standards permits an unlimited number of characters, but in practice the limit can be seen as 383. The computer industry has therefore not adopted the principles of these standards, and it cannot be expected that TS 1923 will become the basis for any industry-supported software.

Some other character set standards for the 8-bit environments also exist, but are not relevant to the problem of data processing of names in Sweden.

3.3.5 Standards for multi-byte schemes: Unicode and ISO/IEC 10646

The market importance of the countries using ideographic writing systems, especially Japan, was naturally recognised by the computer industry quite early. Since those writing systems cannot be accommodated in the same way as script-based systems, a number of manufacturer-specific solutions came into existence. With the general acceptance of the byte-oriented computer architecture, the natural solution to the ideographic problem became based on different two-byte representations.

An obvious development was thereafter to include in a two-byte system also all the characters covered in conventional 8-bit coding schemes, thereby achieving an all-encompassing solution. A project with this purpose was started in the eighties, with mostly computer industry participation. The project was later formalised as the "Unicode Consortium", and its work resulted in 1991–1992 in the publication of "The Unicode Standard Version 1.0".

In parallel, ISO/IEC JTC1 had started work on developing a formal coding standard covering all possible characters needed in data processing. As different from the Unicode Consortium, JTC1 considered it necessary to plan for more than the 65.536 characters that are possible in a true two-byte scheme, therefore working on a four-byte coding principle.

The obvious need of co-ordinating the efforts of the Unicode Consortium and JTC1 was recognised, and joint work resulted in changes to both the Unicode "standard" and the drafts developed within JTC1. Unicode version 2.0 published in 1996 was therefore made consistent with the corresponding ISO standard ISO/IEC 10646-1, published in 1993.

ISO/IEC 10646 now defines two forms of encoding: a complete four-byte form denoted UCS-4, and a two-byte form UCS-2. The two-byte set of characters, forming the "Basic Multilingual Plane" (BMP), is coding-wise identical to Unicode.

As regards character representation, references to Unicode and to ISO/IEC 10646 are consequently equivalent, and in the following text the writing Unicode/10646 is therefore used. It should be noted, however, that since the Unicode standard is more detailed than 10646 in some respects, not all implementations that are compliant with 10646 are necessarily fully compliant with Unicode. In the context of name-writing in Sweden such differences should however not be relevant.

Different notations are possible to indicate that a two-byte value is a code for a Unicode/10646-defined character. In this document the form U+xxyy is used, where xx is the hexadecimal value of the character's first byte and yy the value of its second.

A consequence of the two-byte scheme is that for many characters the first byte of its coding is in the range 00 to 1F (hexadecimal). When passing through communication systems, the first byte could therefore become interpreted as a control character, possibly causing corruption of data.

The “transformation format” UTF-8 specified in Unicode/10646 eliminates this problem through a specific transformation mechanism. At the same time, some data compres-

sion is achieved, in that all ASCII characters are represented as a single byte instead of two. Some data base software makes use of this feature by storing information coded in Unicode/ 10646 as UTF-8 transformation, which greatly reduces storage requirements for the Latin script.

4 Computer keyboards

4.1 General characteristics

Computer keyboard layouts have evolved from the typewriter layouts used in different countries. They are formalised, in some countries, in national standards.

The originally-used "typewriter" design principle for computer keyboards, where each key-press generated a specific code corresponding to the engraving of the key, made keyboards language-specific. This principle became replaced by a more flexible one. Most of today's keyboards are electronically identical, with each key generating a unique but character-independent code. Input software then uses a table look-up to translate the key-specific coding into a character code. Only the key engraving now differs between national variants, not their electronic design.

In addition to the keys directly corresponding to a character, "dead keys" exist for generating composite letters i.e. base letters with diacritics. On Swedish keyboards, for instance, the acute and grave accents, circumflex, diaeresis and tilde are available. The combinations of base letter and diacritic that can be generated is dependent on the look-up tables. In general only combinations that are meaningful in the respective language have been included, but this is a limitation in software, not in the design of the keyboard itself. (It could be noted that for Swedish keyboards, containing the "ready-made" letters ä and ö, those can also be constructed by typing diaeresis followed by a and o, respectively.)

4.2 The "group" concept

In many computer applications there exists a need to use the Latin script alternately with some other script, or one set of characters alternately with another set. For such situations a software mechanism for switching between different keyboard layouts was early provided, through multiple look-up tables. Generally the switching was then "locking" i.e. staying in effect until the next switching. For such alternate functions the terms "keyboard states" or "keyboard groups" are used.

This software mechanism is also used for situations with single scripts when it is desired to add more characters than the basic keyboard layout can accommodate. On Swedish keyboards, for instance, a number of additional characters – like the "at sign" @ – are available on PC-type keyboards in their uppermost key row when the "AltGr" key is pressed (on Macintosh keyboards the corresponding activating key is "Option").

This additional availability can also be considered as a "third-level shift". In this document, however, the group concept is preferred.

4.3 Keyboard standards

A number of national keyboard standards exist, generally following the layouts traditionally established – more or less formally – by typewriters in the respective country. In Sweden, the standard SIS 66 22 41 "Alfanumeriskt tangentbord" was published in 1975 by the Swedish Standards Institute (SIS). The current second edition (see References) was published in 2000.

In ISO several standards relating to keyboards have been produced, starting in the seventies. Many of them were developed for typewriters, adding machines and the like,

and therefore sometimes inconsistent in relation to each other. In the nineties ISO/IEC JTC1 tackled the problem, producing the standard ISO/IEC 9995 "Keyboard layouts for text and office systems", made up of eight parts.

The standard specifies both the concept of "third-level shift" and of "groups". The principle for groups is that the "default group 1" layout is assumed to be an existing national one. The "Common secondary group" (group 2) layout is fixed, specified in the standard in detail. Any number of other groups, from "group 3" upward, can be added in a national standard if needed.

When ISO/IEC 9995 was originally developed, the character set to be covered by the "Common secondary group" was based on ISO 6937. The basic idea was that, together with what was generally available in a national group 1, all characters defined in 6937 should be covered.

This principle caused some arbitrariness. As mentioned above, ISO 6937 was developed quite early, when the data processing situation had not reached its present comparative maturity. The standard therefore contains a number of characters not very useful in text processing, e.g. the Ohm sign and "musical note".

A more serious problem is that the Icelandic capital letter Eth is missing, the intention being that the glyph for the letter "D with stroke" could be used for it. Such a "unification" is unacceptable in the data processing of today. Further, application of the group 2 layout together with most national groups 1 will lead to some duplication of character input options..

4.4 Swedish keyboard situation

Market-driven de facto-situations for keyboard layouts have become established in most countries, generally only partly following formal national standards. This is the case for Sweden also. Although commonly available keyboards are mainly conformant with SS 66 22 41:2000 as regards to its Group 1, it can hardly be expected that its Group 2 will become implemented by industry.

A recent development is however of great interest. As mentioned above, the Sami requirements on data processing have been formalised in documents issued by the Sami Parliamentary Council. Required keyboard layouts for "Finnish/Swedish with Sami" and "Norwegian with Sami" are there specified, as are also Sami-specific layouts for Finland/Sweden and for Norway.

The "Swedish with Sami" assumes the SS 66 22 41 Group 1 layout. In an "Alternate group" (which could formally become e.g. a Swedish Group 3) the specifically Sami letters are included "ready-made". They are there allocated to key positions "intuitively natural", i.e. the letter č to the c key, the letter ŋ to the n key etc. (although this principle could not be fully followed for some of the more unusual letters).

Microsoft has decided to implement full Sami support in Windows, including the keyboard layouts specified in the Sámediggi documents on their "extended" level. Starting with the recent Service Pack 2 for Windows XP, such support is available. The characters in the "Alternate group" are accessed in the usual Microsoft fashion, i.e. by depression of the "AltGr" key.

In addition to the specifically Sami letters, a number of other non-Swedish letters can be input from this keyboard layout through the use of diacritics combined with a base letter. It appears that some further extensions of the Microsoft solution could make possible the input of all of the Latin-script letters needed for writing of names according to this document. This possibility should be investigated, in co-operation with SIS and the industry.

4.5 "Soft keyboards"

Another possibility for inputting characters is the "soft keyboard" approach. With it, a table is displayed on the computer screen, showing all available characters, and the user selects the character to input.

This function exists in many application programs, e.g. Microsoft Word. It has also been implemented in the special-purpose software systems of some Swedish Authorities.

5 Proposed guidelines for representation of names

5.1 General principles

From the previous text, the following conclusions could be drawn:

1. For Swedish Authorities, a large set of Latin-script letters is needed in data processing for representing personal, company and place names relevant to the Authorities' areas of responsibility.

2. Personal names originally written in other writing systems than the Latin script must be transliterated into that script, for which well-defined rules are necessary.

3. Data processing interchange between Authorities of information regarding personal and/or company and/or place names should be coded according to Unicode/10646, transformed according to UTF-8.

4. Information made available from Authorities' web sites for individuals and companies should in principle also use Unicode/10646 encoding. In the individual case other codings may however be necessary.

5.2 Character set: letters

As described above, the number of letters needed to correctly cover Latin-script languages is quite large. A natural starting point to decide exactly what letters should be included in the set is the document "Multilingual European Subsets in ISO/IEC 10646-1" (see References). The document is the result of work carried out in CEN Technical Committee TC304 "Information and communication technologies – European localization requirements"

In that work, the needs for characters were investigated for all European majority languages, as well as for some minority ones, and three subsets were defined. The first subset contains letters in the Latin script only, and is consequently the natural basis in the case at hand.

This subset, designated MES-1, contains the same set of letters as ISO/IEC 6937, with the addition of the Icelandic letter capital Eth. Some of those letters are however used only in older orthography or in Esperanto, and need therefore not be included in a required Swedish set of letters. These special letters are capital and small c with circumflex, i with tilde, j with circumflex and u with tilde, and small letters kra and "n preceded by apostrophe".

Further, the capital and small ligature ij, earlier used in Dutch, appears not needed in its present orthography.

As regards the German "sharp s" symbol ß, the situation for its usage is somewhat unclear, considering that the language is official not only in Germany but in Austria and Switzerland also. It should therefore be included in the required set.

The typographic construct "L with middle dot" in printed Catalan, used for the separation of double letters l in some words (like col·laboració) is generally not available in data processing, separate writing of a "middle dot" being used instead. It should therefore not need inclusion in the required Swedish letter set.

A few Welsh and Sami letters are missing in MES-1, and should be included in the required set.

Also missing are the specifically Vietnamese Latin-script letters, and possibly some other special letters used in Latin-script languages outside Europe. It appears, however, that these should not at present be included in the required set, needing further investigation.

5.3 Character set: non-alphabetic characters

For writing of personal and place names in most languages, the only non-alphabetic character needed is the hyphen. In a few languages, the apostrophe and/or a free-standing acute accent is needed also. Further, for Catalan the "middle dot" is needed, as mentioned above (although it is uncertain if it can really occur in names).

For writing of names of companies a much larger set of special characters may be desirable. It is however obviously necessary to limit the set to what can reasonably be required in the Authorities' data processing.

Such a limitation could be to only allow the non-alphabetic characters in ASCII. Its free-standing grave accent and the dollar sign (which in some coding schemes is replaced with the currency sign ¤) should not be included in the set, however.

5.4 Required character set

The complete minimum character set required according to the previous two sections is the following (only the code values are given, not the "U+" prefix):

0020–0023 0025–005F 0061–007E
00A7 00B4 00B7 00C0–00D6 00D8–00F6 00F8–00FF
0100–0107 010A–0113 0116–011B 011E–0123 0126 0127
012A 012B 012E–0131 0136 0137 013B–013E
0141–0144 0147 0148 014A–014D 0150–015B 015E–0167
016A 016B 016E–017E 018F 01B7 01E4–01E9 01EE 01EF
0259 0292 1E80–1E85 1EF2 1EF3

5.5 Character transliteration

There are several complications involved as regards transliteration.

In principle it is desirable that names originally written in other writing systems are transliterated in such a way that officials in Swedish Authorities have an indication of their correct pronunciation. This factor is behind the "newspaper transliteration" commonly used.

The first problem with this principle is that many languages differ very much from Swedish in their intonation. This applies particularly to the "tone languages" common in East Asia. The usual way of transliterating tone values is with diacritics. In the Chinese pinyin transliteration system, for instance, the acute and the grave accent, the circumflex and the diaeresis (or macron/overline) diacritics are used. In Swedish, however, these diacritics are generally interpreted not as tonal indications but in other ways, if at all.

The second problem is that the principle makes consistent transliterations across the Latin-script language area next to impossible. Most European countries' practice for "newspaper transliteration" of Russian names, for instance, differ very much from the Swedish one. Although the guidelines to be decided on at present apply to Swedish Authorities in their national work, interchange of personal and company information between countries must also be considered.

It could therefore be argued that transliteration should, as far as possible and practical, be "language-neutral". It should then also conform to established international standards. In this connection it can be noted that in the EU, where Greek is one of the working languages, transliteration of Greek names into the Latin script is done according to the international standard ISO 843 (which is the international version of an originally Greek standard).

Further work on the transliteration issue is highly needed, and comments on this issue from Swedish Authorities necessary, particularly from those with needs for close European co-ordination (like part of the Police).

References

ISO/IEC 2022:1994, *Information technology — Character code structure and extension techniques*

ISO/IEC 4873:1991, *Information technology — ISO 8-bit code for information interchange — Structure and rules for implementation*

ISO/IEC 6937:2000, *Information technology — Coded graphic character set for text communication – Latin alphabet*

ISO/IEC 8859-1:1998, *Information technology — 8-bit single-byte coded graphic character sets — Part 1: Latin alphabet No.1*

ISO/IEC 9995-1:1994, *Information technology — Keyboard layouts for text and office systems — Part 1: General principles governing keyboard layouts*

ISO/IEC 9995-2:2002, *Information technology — Keyboard layouts for text and office systems — Part 2: Alphanumeric section*

ISO/IEC 9995-3:2002, *Information technology — Keyboard layouts for text and office systems — Part 3: Complementary layouts of the alphanumeric zone of the alphanumeric section*

ISO/IEC 10646-1:2000, *Information technology — Universal Multiple-Octet Coded Character Set (UCS) —Part 1: Architecture and Basic Multiligual Plane*

CEN TS 1923:2003, *European character repertoires and their coding – 8-bit single-byte coding*

CEN CWA 13873:2000 *Information Technology - Multilingual European Subsets in ISO/IEC 10646-1*

SIS SS 66 22 41, *Informationsteknisk utrustning – Alfanumeriskt tangentbord för svenskt bruk*

The Unicode Consortium, *The Unicode Standard, Version 4.0*

Sámediggi [1] 01/850-51, *Requirements for suppport of Sami languages in data processing*

Sámediggi [1] 01/850-86, *Recommendations for Sami PC-type keyboard layouts*

Council of Europe CETS No. 148, *European Charter for Regional or Minority Languages*

LWP Consulting R 03/14-1, *Survey of state/national/official European languages*

Note [1]: Sámediggi documents are available for download at www.samediggi,no

