Reference number of working document: **ISO/IEC JTC1/SC22/WG15 N_____**
Date: 2000-07-22
Reference number of document: **ISO/IEC WD4 TR 14766**
Committee identification: ISO/IEC JTC1/SC22
Secretariat: ANSI

Information Technology - Guidelines for POSIX National Profiles and National Locales

Tecnologies de l'information - Guide de profiles nationales et locales nationales de POSIX

P1494 / D4
July 2000

This is an unapproved draft and is subject to change.
All rights reserved by the Institute of
Electrical and Electronical Engineers.
Do not specify or claim conformance to this document.

Copyright ISO/IEC

FOREWORD

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National Bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

The main task of a technical committee is to prepare International Standards but in exceptional circumstances, the publication of a Technical Report of one of the following types may be proposed:

- Type 1, when the required support cannot be obtained for the publication of an International Standard, despite repeated efforts;

- Type 2, when the subject is still under technical development or where for any other reason there is the future but not immediate possibility of an agreement on an International Standard;

- Type 3, when a technical committee has collected data of a different kind from that which is normally published as an International Standard, 'state of the art', for example.

Technical Reports of types 1 and 2 are subject to review within three years of publication, to decide whether they can be transformed into International Standards. Technical Reports of type 3 do not necessarily have to be reviewed until the date they are considered no longer valid or useful.

Technical Report ISO/IEC 14766 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information Technology*, Subcommittee 22, *Programming languages, their environments and system software interfaces*.

This Technical Report was developed in cooperation with the Institute of Electrical and Electronics Engineers, Inc. (IEEE).

Suggestions and comments for improvement of this document are welcome. They should be sent to:

Keld Simonsen
Sankt Jørgens Alle 8
DK-1615 Copenhagen V
Denmark
Email: keld@dkuug.dk

CONTENTS Page

Copyright © 2000 IEEE. All rights reserved.
This is an unapproved IEEE Standards Draft, subject to change.

Information Technology - Guidelines for POSIX National Profiles and National Locales

1 Scope

This Technical Report provides guidelines for ISO Member Bodies in the process of making POSIX National Profiles and National Locales for the ISO/IEC 9945 series of POSIX standards.

POSIX National Profiles provide requirements for making POSIX suitable for the culture, by specifying options needed of the POSIX standards and national standards to be applied. Implementers can then comply with the POSIX National Profile to make their product suited for the market, and ISO member bodies can facilitate procurement by making POSIX National Profiles that are national standards. Users can obtain products that are suited for their needs and with consistent behaviour across applications and platforms. A POSIX National Profile may include National Locale specifications.

National Locales specify options to POSIX standards in POSIX locale format, on data that varies culturally. Applications can be written in an internationally portable way by removing hard-coded culturally dependent data or functions, and using the POSIX National Locale data instead. Implementers can, using the National Locales, be relieved from specifying the often very complex internationalization data themselves and instead rely on a credible source such as the ISO Member bodies. Users can benefit from products that are suited for their cultural needs and obtain consistent behaviour across applications and platforms. ISO member bodies can facilitate this process and provide procurement specifications via national standards for National Locales.

Note: Hereafter through this document, for simplicity of wording, the word National Profile is used as synonym of the word POSIX National Profile, unless otherwise stated.

2 Conformance and testing

As this specification is a Technical Report, there cannot be any conformance claimed to this TR.

Editors note: Take something from .23, David Blackwood will provide this.

For testing a National Profile with its National Locale it is often a good idea to provide test data for some functionality, especially the collating specification. This could be done by providing an unsorted file and a correctly sorted file.

It will probably be unmanageable to provide a test suite for all of the standards referenced by a National Profile.

3 References

The following normative documents contain provisions, which, through reference in this text, constitute provisions of this Technical Report. For dated references, subsequent amendments to, or revisions of, any of these publications do not apply. However, parties to agreements based on this Technical Report are encouraged to investigate the possibility of applying the most recent editions of the normative documents indicated below. For undated references, the latest edition of the normative document referred to applies. Members of ISO and IEC maintain registers of currently valid International Standards.

ISO/IEC 9945-1:1996, *Information technology - Portable Operating System Interface (POSIX) - Part 1: System Application Program Interface (API) [C Language].*

ISO/IEC 9945-2:1993, *Information technology - Portable Operating System Interface (POSIX) - Part 2: Shell and Utilities.*

© ISO/IEC 2000

ISO/IEC 646:1991, *Information technology - ISO 7-bit coded character set for information interchange.*

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

ISO 4217:1995, *Codes for the representation of currencies and funds.*

ISO 8601:1988, *Data elements and interchange formats - Information interchange -Representation of dates and times.*

ISO/IEC 10646:1997, *Information technology - Universal Multiple-Octet Coded Character Set (UCS), including Cor.1 and AMD 1-9.*

ISO/IEC FDIS 14651, *Information technology - International string ordering - Method for comparing character strings and description of a default tailorable ordering.*

ISO/IEC 8859, *Information technology - 8-bit single-byte coded graphic character sets - Part 1, .., 10, 13, 14, 15.*

ISO/IEC Directives: 1997, *Procedures for the technical work of ISO/IEC JTC 1 on information technology*.

ISO/IEC Directives Part 2, *Methodology for the development of International Standards.*

ISO/IEC Directives Part 3:1989, *Drafting and presentation of International Standards*.

ISO/IEC 9899:1999, *Information technology - Programming language − C.*

ISO/IEC TR 14262:1995, *Information technology - Guide to the POSIX Open Systems Environment.*

IEEE P1003.18/D13 (September 1996), *Information technology - POSIX Profile.*

IEEE 1003.23-???? *User … Profiles*

ISO/IEC TR 10000-1:1998, *Information technology - Framework and taxonomy of International Standardized Profiles - Part 1: Framework.*

ISO/IEC TR 10000-2:1998, *Information technology - Framework and taxonomy of International Standardized Profiles - Part 2: Principles and Taxonomy for OSI Profiles.*

ISO/IEC TR 10000-3:1998, *Information technology - Framework and taxonomy of International Standardized Profiles - Part 3: Principles and Taxonomy for Open System Environment Profiles.*

ISO/IEC DTR 14652, *Information technology - Specification method for cultural conventions.*

ISO/IEC 15897:1999, *Information technology - Procedures for registration of cultural elements.*

ISO/IEC TR 11017:1998, *Information technology - Framework for Internationalization.*

4 Definitions and abbreviations

For the purpose of this Technical Report the following definitions and abbreviations apply.

4.1 profile: A set of one or more base standards, International Standardized Profiles (ISPs) and, where applicable, the identification of chosen classes, conforming subsets, options and parameters of those base standards, or ISPs necessary to accomplish a particular function (10000-1).

Copyright © 2000 IEEE. All rights reserved.
This is an unapproved IEEE Standards Draft, subject to change.

4.1 POSIX profile: A profile for an International Standard is a set of specifications of the parameters, the selections of the optional items and the recommendations of the implementation related matters. A POSIX Profile corresponds to the profile concept for the POSIX International Standard.

4.2 POSIX National Profile: A POSIX National Profile is a POSIX profile that is strongly related to the cultural dependent aspects of POSIX. It also contains the definitions and recommendations for the usage of national or regional standards that support the handling of the national or area specific aspects, e.g. the use of the coded character sets.

4.3 POSIX National Locale: A National Locale is a part of a National Profile, which gives profile options in the POSIX localedef format.

4.4 Conformance to a POSIX National Profile: The concept of the degree of the preciseness of the coincidence between the specifications of a realized POSIX system and the POSIX National Profile. Since the POSIX National Profile is not necessarily included in the POSIX Profile, systems that conform to the POSIX National Profile may not pass the POSIX Conformance requirements.

4.5 National Standards Profile: A National Standards Profile (NSP) is a profile of an international standard or set of international standards, possibly together with other specifications, that is adopted by an ISO member body as a national standard.

4.6 Internationalization (I18N): A process of producing an application platform or application that is easily capable of being localized for (almost) any cultural environment. (Note that an internationalized information system does not have a dependency on any specific culture, unless it is localized to that selected culture.) (TR 11017)

4.7 Localization (L10N): A process of adapting an internationalized application platform or application to a specific cultural environment. In localization, the same semantics are preserved, while the syntax may be changed. (TR 11017)

4.8 Portability (source code): the ability that an application can perform with same results on different application platforms, without changing the program source code.

4.9 Locale: The definition of the subset of the environment of a user that depends on language and culture conventions.

4.10 Charmap: a character set description file, for use with a locale.

4.11 International Standardized Profile: An internationally agreed-to, harmonized document that describes one or more profiles (10000-1).

4.12 ISP: an abbreviation for International Standardized Profile

4.13 NSP: an abbreviation for National Standards Profile

4.14 I18N: an abbreviation for internationalization

4.15 L10N: an abbreviation for localization

5 Purpose of National Profile and National Locale

5.1 Purpose of National Profiles

National Profiles for POSIX international standards define culture and language-dependent adaptation and

interpretation of POSIX for the following purposes:

- National Profiles identify the base international and national or regional standards and clarifies the relationships among them.

- National Profiles identify the base standards, together with appropriate cultural and language-specific classes, subsets, options and parameters, which are necessary to assure higher degree of portability.

- National Profiles give detailed descriptions of locale-dependent functions that are out of the scope of the base International Standard that provides frameworks for internationalization so that national bodies can define appropriate language and cultural dependent adaptation and interpretation based on it.

- National Profiles provide reference systems, on top of which cultural and language-dependent applications can be built to promote POSIX standards among users and vendors,

- National Profiles promote the development of conformance tests that produce consistent results for the systems compliant with POSIX and a given National Profile.

Various bodies throughout the world are undertaking work in the definition of National Profiles for POSIX based international standards.

These guidelines for POSIX National Profile writers has been developed by SC22/WG15 and IEEE PASC to make the development of National Profiles consistent and the harmonization of the National Profiles easier by defining the following:

- Define style, documentation scope and classification scheme for National Profiles.

- Define those items that should be included in National Profiles.

- Define those items that should not be included in National Profiles.

5.2 The purpose of the National Locale

The purpose of the National Locale is to specify values for a given culture, country and language, so that users can refer to this locale and obtain consistent behaviour across hardware and software platforms conforming to this locale. It is expected that many national standardization organizations will make national standards of their locales that can then be used also for procurement.

The National Locale will in most cases build on already existing national standards, for example on formatting and collation, but will sometimes reflect customary specifications, for example for date and time there often does not exist an adequate national standard.

6 Concept of National Profiles

POSIX is a Open System Environment (OSE) platform. An Application Environment Profile (AEP) is a set of parameters and the selection of options for the base standards included in an OSE to support the execution of application programs for a given application field. It includes the parameters and option selections for the relevant base standards such as the platform standards like POSIX and application specific standards like GKS, SQL and so on.

A National Profile for a specific cultural region or a nation is a set of parameters and option selections for several base standards like POSIX. These standards may be National Standards like JIS X0208, or they may be extensions

Copyright © 2000 IEEE. All rights reserved.
This is an unapproved IEEE Standards Draft, subject to change.

to international standards. A National Profile cannot avoid such non-international standards because it should specify the local cultural aspects.

National Profiles cannot be considered International Standardized Profiles (ISPs) in the sense of ISO/IEC TR 10000-2, as they are not international but national in nature. Thus International Standardized Profiles cannot reference POSIX National Profiles, while the referencing from National Profiles to ISPs is possible.

Application Environment Profiles and National Profiles may be based on National Standards, and therefore it is necessary to coordinate these by defining the parameters and option selections from the viewpoint of international harmonization to support international application portability and interoperability.

Given this fact, there are several levels of conformance both for a given POSIX application environment profile and a given POSIX National Profile as follows:

For Application Environment Profile:

1 Strictly Conforming POSIX Application for POSIX AEP

An application that can be executed for any parameters and options for POSIX.

2 ISO/IEC Conforming POSIX Application for POSIX AEP

An application that requires only specific POSIX related parameters and options.

3 ISO/IEC Conforming POSIX Application using Extensions for POSIX AEP

An application that requires not only specific POSIX related parameters and options but also other ISO/IEC standards and their international profiles.

For POSIX National Profile:

1 National Body Conforming POSIX Application for POSIX NP

An application that requires only the POSIX related parameters and options defined in POSIX National Profile.

2 National Body Conforming POSIX Application using Extensions for POSIX NP

An application that requires POSIX related parameters and options defined in POSIX National Profile, National Profiles for other ISO/IEC standards, and national body standards.

6.1 The relationship to base standards

Base standards specify procedures and formats that facilitate the development of internationally portable applications across many countries or regions. They may provide mechanisms for supporting language and culturally dependent (locale specific) aspects, hopefully in a locale-independent way as much as possible.

National Profiles promote the applicability of the base standards to specific countries or regions by defining how to use mechanisms specified in the base standards for a specific country or region with appropriate choices and values for options and parameters. National Profiles may also specify additional standards that are required for locale specific features support.

National Profiles shall not contradict base standards but shall make specific choices where options and ranges of values are available. The choice of the base standard options should be restricted so as to maximize the application

© ISO/IEC 2000

portability across National Profiles, consistent with achieving the objectives of the National Profiles.

6.2 The relationship to Registration Authority

Some objects specified in National Profile may be administered and registered to keep identification and to avoid conflict of values or names adopted by each of the countries.

The administration and registration of such objects may be performed by Registration Authorities, authorized by ISO/IEC JTC1, with the procedure recognized and agreed internationally. The ISO/IEC DIS 15897 registration standard provides registration mechanisms for POSIX profiles, POSIX locales and POSIX charmaps and lists of symbolic character names, 'repertoire maps'.

Note: contents of the ISO/IEC 15897/ENV 12005 registration are available at http://www.dkuug.dk/cultreg/

The following locale objects specified in a National Profile should be registered and maintained by registration authorities.

(a) Locale definitions and their names;
(b) Symbolic character names;
(c) Coded character set and their names;
(d) Character class names.

6.3 Principles of National Profile Content

6.3.1 General Principles

General principles for a profile specified in ISO/IEC TR 10000-1, sub-clause 6.3 are applied to a POSIX National Profile.

6.3.2 Principles of National Profile Content

A National Profile places a set of requirements that are useful in maximizing application portability for a specific country or region. It does not specify all of the functionalities of a system, but only that part relevant to the function being used for locale-specific operation.

The content of a National Profile shall be specified in a coded character set independent way where it is possible. When some requirements are recognized to be locale-specific but a National Profile can make no clear indication, it may include an informative guidance to implementers.

6.3.3 Main elements of a National Profile definition

The definition of a National Profile shall comprise the following elements:

(a) A definition of the scope of the country or regions for which the National Profile is defined, and of its purpose;

(b) Normative reference to base standards, including precise identification of the actual texts of the base standards being used and of any approved amendments and technical corrigenda (errata), conformance to which is identified as potentially having an impact on achieving portability using the National Profile;

(c) Normative and informative reference to any other relevant source documents, including National Body standards;

Copyright © 2000 IEEE. All rights reserved.
This is an unapproved IEEE Standards Draft, subject to change.

(d) Specification of the application or the function of each referenced base standards, covering recommendations on the choice of classes or subsets, and on the selection of options, ranges of parameter values, etc.;

(e) Specification of the locale information of each referenced base standard;

(f) A statement defining the requirements to be observed by systems claiming conformance to the National Profile.

6.4 The meaning of conformance to a National Profile

The concepts of implementation conformance and application conformance are incorporated in the concept of National Profiles. These conformances, which are defined in a National Profile, are applied only to an application platform, for interoperability and for portability of applications and data. A real system is said to exhibit conformance if it compiles with the requirements of applicable POSIX standards.

A National Profile shall address the following two topics:

(a) Implementation conformance requirements (detailed in 6.6);

(b) Application conformance requirements (detailed in 6.7);

These requirements are stated in a POSIX National Profile.

In order to conform to a National Profile, a system shall perform correctly all the capabilities defined in the base standards as mandatory and also any options of the base standards that it claims to include. Conformance to a base standard in this context is conformance to a particular identified publication of a referenced base standard.

A National Profile shall be defined in such a way that testing of its implementation can be carried out in the most complete way possible given the available testing methodologies.

6.5 Conformance requirements of POSIX National Profiles

This section duplicated 6.4 and should be removed in its entirety. This will necessitate renumbering all subsequent sections as well as any references to them.

6.6 Implementation Conformance

6.6.1 General

The choices of interfaces and functional behaviour made in a National Profile's implementation conformance requirements are specific to that National Profile and provide added facilities to the base standards.

The choices are not, therefore, arbitrary but need to be consistent with the purpose of the National Profile and consistent across the base standards referenced by it.

In order to avoid ambiguity between the National Profile and the base standards, the implementation conformance requirements of a National Profile shall be specified, where possible, by reference to the conformance requirements of the referenced base standards.

6.6.2 Requirements

All systems claiming conformance to a National Profile shall support the required interface and functionality defined in the National Profile. The system may also provide additional functions or facilities not required by the

© ISO/IEC 2000

National Profile.

6.7 POSIX Application Conformance for National Profiles

All POSIX applications claiming conformance to the National Profile shall use only language-dependent services for one or more of the Language Options defined in the National Profile and the facilities provided by the National Profile and referenced base standards, and shall fall within one of the following categories.

6.7.1 <National Body> Conforming POSIX Application

A <National Body> Conforming POSIX Application requires only the parameters and options defined in the National Profile for the said National Body. Such an application shall include a statement of conformance that documents all options and limit dependencies, and all other <National Body> standards used.

6.7.2 <National Body> Conforming POSIX Application Using Extensions

A <National Body> Conforming POSIX Application Using Extensions is an application that requires not only the parameters and options defined in the National Profile, but also other international standards and their profiles or other national standards for the said National Body. The national extensions shall only be with respect to cultural services. Such an application shall fully document its requirements for these extended facilities, in addition to the documentation required of a <National Body> Conforming POSIX Application.

7 Contents of National Profile

A POSIX National Profile shall have the following structure:

```
1 General

1.1 Scope

The scope of the National Profile shall be described. Provision of this section
is mandatory.

1.2 Normative Reference

The standards that are referred by the National Profile shall be listed.
Provision of this section is mandatory.

1.3 Objectives

The objectives of the National Profile shall be described. Provision of this
section is mandatory.

1.4 Conformance

1.4.1 Levels of conformance

If the National Body enacts some levels of conformance, the levels shall be
specified. Provision of this section is mandatory.

1.4.2 System conformance

The requirements to the National Body conforming implementation shall be
specified. Provision of this section is mandatory.

1.4.3 Application conformance

The requirements to the National Body conforming application shall be
```

Copyright © 2000 IEEE. All rights reserved.
This is an unapproved IEEE Standards Draft, subject to change.

specified. Provision of this section is mandatory.

2 Registry

The names, which must not conflict with any other National Profile, shall be listed. The names described here shall be registered with ISO, when an official registration mechanism is established. Provision of this section is mandatory.

2.1 Locale names

The name of locales that are specified in the National Profile. Provision of this section is mandatory.

2.2 Symbolic name of characters

The list of extended character’s symbolic names or the naming conventions for symbolic name of extended characters shall be specified. Provision of this section is mandatory.

2.3 Name of coded character sets

The name of coded character sets that are referred by the National Profile shall be listed. The names may be used for code conversion utilities and functions, also. Provision of this section is mandatory.

2.4 Character classes

If the National Body specifies extra character class in the LC_CTYPE category, the names and descriptions shall be specified. This section is optional.

2.5 Environment variables

If the National Body specifies environment variables that are not specified in the base standard, the names of the environment variables and their descriptions shall be specified. This section is optional.

3 Parameters

3.1 POSIX

The range of POSIX related parameters that are allowed by the National Profile should be specified. Provision of this section is mandatory.

3.1.1 Charmap

The contents of Charmaps shall be specified. Provision of this section is mandatory.

3.1.2 Locale definition

The contents of locale definitions shall be specified. Provision of this section is mandatory.

3.1.3 System parameters

The range of values of following system parameters e.g. POSIX_NO_TRANC, NAME_MAX, and NAME_MAX shall be specified. Provision of this section is mandatory.

3.2 C language

The range of C Language related parameters which are allowed by the National Profile shall be specified, e.g. CHAR_BIT. Provision of this section is mandatory.

© ISO/IEC 2000

4 Options

Options that are required to be implemented shall be specified.

4.1 POSIX

The required optional facilities of the base standards shall be listed, e.g. the charmap option of the localedef utility. Provision of this section is mandatory.

4.2 Programming language support

The facilities required with respect to programming language support, e.g. programming language C as defined in ISO/IEC 9899.

5 Error and exception handling

If the National Body specifies the error and exception handling of some functions, the methods shall be specified. This section is optional.

6 Extensions

6.1 POSIX extensions

If the National Body requires implementation of any enhanced facility, e.g. the addition of environment variables, functions, utilities and option parameters of utilities, the enhanced facilities shall be specified. Provision of this section is mandatory.

6.2 Other standards

If the National Body requires implementation of any standards other than POSIX standard to the National Body conforming systems, the standards shall be listed. Provision of this section is mandatory.

7 Data exchange

If the National Body specifies any formats or mechanism, or requires the implementation of additional standards, the facilities shall be specified. This section is optional.

7.1 Archive file format

Format of archive files, e.g. tar and cpio, shall be specified.

7.2 Identification of coded character set

The mechanism to identify coded character sets in a file shall be specified.

7.3 Protocols

Communication protocols that the National Body conforming implementation must implement shall be listed.

7.4 Profile for OSI

The profile that the National Body specified for OSI shall be referred.

7.4 Media

If the National Body has requirements for media used for data exchange, the requirements shall be specified.

Annex A Informative reference

Copyright © 2000 IEEE. All rights reserved.
This is an unapproved IEEE Standards Draft, subject to change.

If the National Body has any recommended parameters, options and extensions, these features should be listed in this section. This section is optional.

Annex B Notes and Rationale

8 Concept of POSIX National Locales

The POSIX National Locale provides information that can be applicable to each application that modifies the behaviour of the application to adapt to national and cultural preferences. In this way the same binary application can be used according to the cultural expectations of users in different cultural environments. Locales thus enable binary portability of applications to diverse cultural environments. The National Locale is logically a part of the National Profile.

The benefits of a National Locale are exemplified by the Danish locale included in ISO/IEC 9945-2.

9 Contents of National Locale

In creating a National Locale, many things must be considered. Some data may be more easily determined than others. For each locale category recommendations on its contents is given below.

9.1 Character classification and transformation

The character classification section of the locale is normally straightforward; an ’A’ is considered a letter in most Western languages and is mapped to an ’a’ when the lower case letter should be found. Normally the LC_CTYPE definition in POSIX.2 Annex G or the POSIX equivalent of the ’i18n’ FDCC-set of ISO/IEC 14652 can be used without change.

9.2 Numeric

The data here is normally easy to determine for a given language and culture. The ISO standard uses a comma (,) as the decimal punctuation, and a period (.) as the thousands delimiter.

9.3 Monetary

The monetary formats may be a bit more difficult to specify. The ISO 4217 currency code must be specified for the international format. The local specification may offer a choice, but there may be guidelines in national orthography specifications.

Some countries may have obligations to display an amount in more than one currency, for example European countries using the Euro currency and a national currency. This is currently not possible to do in an internationalized portable way with current POSIX standards. It is recommended to make a comment in the locale if this is the case.

The current POSIX standards specify that the position of the international and domestic currency symbol in relation to the monetary amount must be the same. It is recommended to make a comment in the locale if this is not in line with the national practice.

9.4 Time

There may be problems with specifying the date format, including time zone names, which may not be well defined. You could consult a number of official sources, including orthography definitions and numeric rendering standards. One thing to watch out for is if the day and month names are written with an initial small letter - many languages do this, while some proprietary sources say that the names are spelled with an initial capital letter.

© ISO/IEC 2000

9.5 Collating

The collating sequence is a major task to define. There are a number of versions of collation algorithms; each version accomplishes collation with specific requirements. For example the telephone version, with ’Mc’ the same as ’Mac’, numbers spelled out, certain words like ’the’ ignored or moved to the end, and the same entry entered several times at different places, etc. Another level is the phonetic version - soundex, which is a little less complicated. A third version is transcripted characters, as some librarians use when they see a Greek alpha and order that as a Latin ’a’.

The version that is recommended for POSIX.2 locales is the systems interface level. The collating order should be usable in POSIX systems tools like ’ls’ and ’sort’. A requirement has been that it be deterministic; if two strings are different they will also differ when compared. Another issue has been efficiency. This is also called the dictionary version.

The problem of pronunciation and transliteration has not been addressed. Instead it had been considered adequate just to look at the characters themselves - only considering characters at the systems level - and not sounds. The level provided by the example locale in the ISO/IEC 9945-2 standard is a service for comparing strings which are intended as a replacement to the standard strcmp(), etc. routines, just a little more intelligent and adhering to what is expected to be culturally acceptable.

As an example, for Danish collating, there is as much intelligence put in there as possible. The two letters <a><a> are sorted as the single letter <aa> (A WITH RING), but the <aa> single letter is before <a><a> in homonyms. The 4 level scheme of the Canadian sorting is being used, with the four levels being letter, accent, case and special character. In support of harmonization it was decided to use the reverse sorting for the accents as the Canadians do; the natural choice may have been forward sorting here too, but as most of these words would be of French origin anyway, it was decided to follow the French rules. <ss> was implemented with the German rule, as seen in several German dictionaries. <ss> is ordered as <s><s> but before it in homonyms.

As an example of specifying the collating sequence for accents, there was some rules indicated in the Danish sorting standard and in the official Danish orthography dictionary, but it was far from complete. Then the accent sequence in several ISO standards was used, where there was no clear Danish rule. About 25 accents have been ordered.

For non-Latin scripts transcription is not recommended. This allows use of the native collation order for these scripts, like ’alpha beta gamma’ for Greek and ’a be ve ghe’ for Cyrillic. Accented Greek and Cyrillic letters and ligatures should be put into the right places.

The sequence of the scripts is recommended in the ISO/IEC FDIS 14651. That should solve the question of which scripts should come before others. A national specification may then choose this order, or maybe choose to let the native script or scripts come first, and then the rest of the scripts in the order specified in 14651.

9.6 Messages

The messages category is a hook to provide real message service in the applications, and only yes/no is considered by the POSIX standard.

For the yes/no it is recommended that only the first letter of the answer in the natural language is required, and also to allow the English form ’Yes’/’No’, and the more cultural neutral 1/0 as answers. In Greek, the affirmative answer is ’ne’ written with the Greek script, so the allowing of ’n’ for negative answers could cause confusion for users of the Greek language.

Copyright © 2000 IEEE. All rights reserved.
This is an unapproved IEEE Standards Draft, subject to change.

10 Using locale templates

The ISO/IEC 9945-2 standard introduced a copy command for all sections of the locale. This is convenient for many purposes, and it ensures that two locales are equivalent for a given category. A further step in building on previous art is described here.

The collating sequences may vary a bit from country to country, but in many cases much of the collating sequence is the same. For instance the Danish sequence is quite similar to the German, English or French, but for about a dozen letters it differs. The same can be said for Swedish or Spanish. Generally the Latin collating sequence is the same, but a few characters collate differently.

With the advent of the general coded character set independent locales like the Danish example in ISO/IEC 9945-2 annex G, it would be convenient if the few differences could be specified just as changes to an existing one. The specification job could then be reduced by orders of magnitude from say about 300 Latin letters (or 30.000 characters of ISO/IEC 10646) to about 10 to 30. This would also improve the overview of what the changes really are. Therefore it is recommended that the tool to implement the 'reorder-after' construct given in Annex C be used for the LC_COLLATE section of the locale file format for producing new National Locales.

10.1 Internationalization data collections

ISO/IEC JTC1/SC22/WG15 - the ISO POSIX Working group - has been collecting POSIX locales for a number of years, and about 60 locales and 150 charmaps are available now.

Note 1: The electronic data is freely available at the address http://www.dkuug.dk/i18n/WG15-collection.

A formal registry has been established in ISO/IEC 15897 and CEN ENV 12005, with entries encompassing a number of internationalization related data, including POSIX National Profiles, POSIX locales, POSIX charmaps and lists of symbolic character names - 'repertoire maps'.

Note 2: The electronic data is freely available at the address http://www.dkuug.dk/cultreg.

11 Concept of charmap

A charmap is a file describing a coded character set. It is used together with a locale file by the localdef utility to produce a binary locale. The charmap describes the mapping between symbolic character names, as used by the locale, and the binary encoding of the characters.

One locale can be written to support a number of coded character sets or encodings, by using symbolic character names which then are mapped to actual binary encodings via a charmap for each of the coded character sets employed, thus giving a binary locale for each of the encodings. The charmaps may also be used together with different locales when these use the same symbolic character names.

WG15 - the ISO/IEC POSIX working group - has collected about 150 charmaps that then can be readily applied by the localedef utility to a locale. The collection comprises almost all of the ISO/IEC 2375 coded character set registry, and some 60 vendor specific character sets.

Note: see clause 10.1 Note 1 for availability of this data.

Thus with just one specification of a National Locale, uniform collation for many character sets is defined - the characters will always come in the same sequence regardless of which character set employed. Also there can be just one definition of date format and the other cultural items to be written, and that specification is then valid for many character sets.

© ISO/IEC 2000

12 Contents of charmap

The content of a charmap file is described in ISO/IEC 9945-2 clause 2.4.1.

A number of characters need to be present, see table 2-4 and table 2-5 for optional control characters inclusion. This is almost the same as the repertoire of ISO/IEC 646 IRV.

In the charmap file there may optionally be specified a number of keywords.

The <escape_char> and <comment_char> may specify alternate characters for the escape character and comment character, respectively. Common replacements for the default '\' and '#' characters are "/" and "%", which may lead to better portability, as '\' and '#' is known to change representation when transmitted in certain email environments.

The <code_set_name> describes the name of the character encoding, with graphic characters from ISO/IEC 646 IRV.

<mb_cur-max> and <mb_cur-min> describes the maximum and minimum number of bytes in an encoding, respectively. They default to 1 and to the value of <mb_cur_max> respectively.

Each of the lines defining the mapping between a symbolic name and an encoding may take a third argument, namely a comment. There is no need to specify a comment character before the comment, but it does no harm. Giving the ISO/IEC 10646 short identifier and the long name, for example, may enhance the readability of the charmap considerably.

Copyright © 2000 IEEE. All rights reserved.
This is an unapproved IEEE Standards Draft, subject to change.

Annex A. Locale related descriptions in POSIX

Editor's note: We have an extract in source form from the POSIX editor, with permissions to reproduce it. It is not reproduced here due to considerations for the rain forests, as it is about 70 pages. It is an extract of POSIX.2 on the first sections including 2.5 locales, and the 4.13 date format.

Annex B. Symbolic character names

Editor's note: As in POSIX.2 annex G and ISO/IEC DTR 14652 clause 6. As it is about 40 pages, it is not reproduced in this draft of the TR.

Annex C. Convenient tools for producing National Locale

The following script has been written in the 'awk' language defined in ISO/IEC 9945-2 to implement the 'reorder-after' construct.

```
BEGIN {
        comment = "%";
        back[0]= follow[0] = 0
        }

/LC_COLLATE/ { coll=1 }

/END LC_COLLATE/ {
        coll=0;
        for (lnr= 1; lnr; lnr= follow[lnr])
                print cont[lnr]
}

{ if (coll == 0) print $0 ;
 else {
        if ($1 == "copy") {
                file = $2
                while (getline < file )
                if ( $1 == "LC_COLLATE" ) copy_lc = 1
                else if ( $1 == "END"
                && $2 == "LC_COLLATE" ) copy_lc =0
                else if (copy_lc) {
                    lnr++
                    follow[lnr-1] = lnr
                    back [ lnr ] = lnr-1
                    cont[lnr] = $0
                    symb[ $1 ] = lnr
                }
                close (file )
        }
        else if ($1 == "reorder-after")
        { ra=1 ; after = symb [ $2 ] }
        else if ($1 == "reorder-end") ra = 0
        else {
                lnr++
                if (ra) follow [ lnr ] = follow [ after ]
```

© ISO/IEC 2000

```
            if (ra) back [ follow [ after ] ] = lnr
            follow[after] = lnr
            back [ lnr ] = after
            cont[lnr] = $0
            if ( ra && $1 != comment && $1 != "" ) {
                old = symb [ $1 ]
                follow [ back [ old ] ] = follow [ old ]
                back [ follow [ old ] ] = back [ old ]
                symb[ $1 ] = lnr
            }
            after = lnr
        }
    }
}
```

Copyright © 2000 IEEE. All rights reserved.
This is an unapproved IEEE Standards Draft, subject to change.

Annex D. Use of ISO/IEC 10646 in POSIX standards

D.1 Introduction and scope

For servicing the widest possible audience, POSIX standards should be able to handle the most encompassing character set, and the best candidate for this is the ISO/IEC 10646-1:2000 standard. The following gives guidance for how to accomplish this goal.

The area of application includes global organisations interested in just one character set organisation wide, European government institutions, and the eastern Asia region, among others.

ISO/IEC 10646-1:2000, the Universal Multiple-Octet Coded Character Set (UCS), provides the capability to encode multi-script text within a single coded character set.

However, because UCS is designed to use all code points available, null bytes and the code values of the other ISO/IEC 646:1991 IRV (also known as ASCII) characters, including the code value of the ISO 646 solidus ("/") character, are not protected. This makes the UCS character encoding incompatible with many existing ISO 646 based POSIX operating system implementations. The fact that UCS also uses code points used for ISO 6429 control characters introduces further problems for communication and application software. From these problems it was clear that a POSIX internal encoding was required for the ISO/IEC 10646 coded character set.

In the following, a survey of the possible coded representations of UCS and UCS-transformation formats and their respective characteristics are given. Then each of the handling areas (data storage, file names, internal processing, communications, inter-process communications) of POSIX operations is analyzed. Finally guidelines are given for POSIX standards.

A revised TR 10176 with guidelines for support of ISO/IEC 10646 has been published, and there may be further recommendations in this area of relevance to POSIX.

D.2 UCS coded representation forms and UCS transformation formats

D.2.1 POSIX internal encoding

For the POSIX internal encoding UTF-8 was considered suitable.

The objective of UTF-8 is to provide an UCS transformation format that also meets the requirement of being usable on historical POSIX operating system file systems in a non-disruptive manner.

The UTF-8 transformation format represents both UCS-2 and UCS-4 in a compatible format using multiple-octet coded characters of lengths 1, 2, 3, 4, 5, and 6 octets:

Bits	Hex Min	Hex Max	Byte Sequence in Binary
1	7	00000000	0000007F 0vvvvvvv
2	11	00000080	000007FF 110vvvvv 10vvvvvv
3	16	00000800	0000FFFF 1110vvvv 10vvvvvv 10vvvvvv
4	21	00010000	001FFFFF 11110vvv 10vvvvvv 10vvvvvv 10vvvvvv
5	26	00200000	03FFFFFF 111110vv 10vvvvvv 10vvvvvv 10vvvvvv 10vvvvvv
6	31	04000000	7FFFFFFF 1111110v 10vvvvvv 10vvvvvv 10vvvvvv 10vvvvvv 10vvvvvv

The UCS value is the concatenation of the v-bits in the multiple-octet encoding, where the v-bits are the 0's and 1's that constitute the UCS value.

© ISO/IEC 2000

Thus UTF-8 has the capability of handling existing ISO 646 files without change, and all codes in the ISO 646 range (having an octet value in the range 0-127) can be safely assumed to be representing the normal ISO 646 character.

D.2.2 Other forms of ISO/IEC 10646

ISO/IEC 10646 has two forms: UCS-2 and UCS-4, a 16-bit and 31-bit coded representation of the character set, respectively. ISO/IEC 10646 is planned to have more than 64.000 characters, so the general case of UCS-4 needs to be considered.

ISO/IEC 10646-1:1993 had a transformation format **UTF-1**, which was informative, and it has now been removed from the standard by the amendment ISO/IEC 10646-1 AM4: 1996. UTF-8 is aimed at the same purpose, and has more capability. UTF-8 has been approved as part of UCS via the amendment ISO/IEC 10646-1 AM2: 1996.

Another Transformation Format of ISO/IEC 10646, **UTF-16**, has also been approved, as ISO/IEC 10646-1 AM1: 1996, but this cannot accommodate all of ISO/IEC 10646 (it accommodates about 1 million characters) and it will employ techniques like in UTF-8 with ranges indicating how many octets are required to form one character, without the added functionality of being backwards compatible with ISO/IEC 646 and ISO/IEC 2022 encodings (which is a functionality of UTF-8).

The most general of the above encodings of ISO/IEC 10646 is the **UCS-4**. It has the property of being constant-width, which may be easier to handle than the multiple-octet UTF-8. As a file and as an interchange code it has the problematic property of using codes in conflict with ISO/IEC 646, ISO/IEC 2022 and ISO/IEC 6429, dependency on byte-ordering (little-ending vs. big-ending) of the hosting machine architecture, and also of using 4 octets per character. Here UTF-8 is clearly superior for POSIX internal encoding. UCS-4 may have advantages as an internal processing code, and as an inter-process encoding, for C language widechar-like encodings, but with the ISO/IEC C language amendment (AM1) with full support for multi-byte coded character sets that advantage may be diminishing. UTF-8 is as well defined and capable of representing all ISO/IEC 10646 characters, and given its strengths in other areas it may well be chosen also for the internal processing, and inter-process communication. Internal processing is not in the scope of POSIX interfaces, anyway.

D.2.3 UCS levelling

ISO/IEC 10646 has 3 levels of support, level 1 without combining characters, level 2 with combining characters in some scripts, and level 3 with unrestricted use of combing characters. SC22 has by resolution of the 1993 Paris plenary recommended that all SC22 standards be enabled for level 3 data, but that the semantics of combining characters not be addressed currently. Thus there is not specific SC22 request for further support of level 2 and 3, but eventually there could be a need for support of these levels. SC22 also recommended use of ISO/IEC 10646 terminology throughout SC22 standards, and this may need an alignment of current POSIX work, though it is the belief that current POSIX work is already well aligned with ISO/IEC 10646 with respect to terminology.

D.3 Problems in POSIX handling of UCS

There are several challenges presented by UCS that must be dealt with by present implementations of the POSIX operating system.

D.3.1 Data storage

The most significant of these challenges is the encoding scheme used by UCS. More precisely, the challenge is the marrying of the UCS standard with existing programming languages and existing operating systems. Prominent among the operating system UCS handling concerns is the representation of contents of data in files. An underlying assumption is that there is an absolute requirement to maintain the existing operating system software investments while at the same time taking advantage of the use the large number of characters provided by UCS.

Copyright © 2000 IEEE. All rights reserved.
This is an unapproved IEEE Standards Draft, subject to change.

For UTF-8 the representation of ISO 646 data is exactly the same, and for ISO/IEC 8859 parts, right hand side characters will need two octets for representation. For ideographic characters in the BMP, the representation will be three octets. This does not give a dramatically changed requirement for what is currently consumed for data storage.

D.3.2 File names and internal processing

The UTF-8 transformation format was originally conceived as a file system safe transformation format of UCS to allow historically ISO 646 based POSIX operating systems to cope with representation and handling in file names of the large number of characters that are possible to be encoded by UCS. In addition, from an internal operating system (kernel) viewpoint this handling of a large character set is only a problem for handling file names, which are only analyzed for the solidus ("/") delimiter to parse a name into filename components. As UTF-8 can represent the full encoding of ISO/IEC 10646 and is backwards compatible with ISO 646, UTF-8 handling is sufficient for POSIX internal encoding.

D.3.3 Communications

Current ISO POSIX standards do not address communication, but as ISO 6429 control characters are often used in communication, and the UTF-1 transformation format was originally created for avoiding control character problems in communication, UTF-1 could be the choice. As UTF-1 is being removed from UCS and UTF-8 introduced, having the same capabilities with respect to control character problem solving, UTF-8 is the recommended choice in POSIX communication interfaces.

D.3.4 Inter-process communication

Communication between POSIX processes would probably use internal data formats, for example integers should be transferred in binary form. As it could be recommended that programs internally use a C language widechar style encoding of characters, a UCS-2 or UCS-4 format could be recommended.

On the other hand inter-process communication is often across networks and between heterogeneous systems, therefore since UCS-2 and UCS-4 are dependent on machine architecture, UTF-8 may be the preferred candidate. UTF-8 would in many cases also be less space consuming, which may be a significant plus when using low-capacity network lines.

D.4 Recommendation

According to the above analysis, UTF-8 is the best candidate for POSIX internal encoding of UCS in the areas of data storage, file names and internal operating system (kernel) processing, and communication, where otherwise UCS-2 or UCS-4 would have been used for coded data. Furthermore UTF-8 is a good candidate for UCS representation in inter-process communication.

It is thus the recommendation to use the UTF-8 transformation format whenever UCS is used in POSIX interfaces.

As POSIX interfaces in principle should be coded character set independent, there is no general need to require the use of UTF-8 in POSIX standards, but guidance could be given in rationales.

A specific recommendation is that the portable archive exchange utility "pax" be revised to be able to specifically use UTF-8 for file names, and the use of UTF-8 should be clearly identified.

D.5 Consequences

The Open Group has raised a number of problems with use of ISO/IEC 10646 in POSIX in the document WG15 N621. With the preceding recommendation the problems can be addressed as follows:

- In UTF-8 the repertoire of ASCII is encoded as ASCII (ISO/IEC 646 IRV).

- We know no code sets with control characters encoded in the full single octet range 0 thru 7F, but many use 0 thru 1F hex and 7F, and some the range 80 thru 9F. UTF-8 has reserved these octet ranges for control characters.

- Zero value octets and octets equating ’/’ only appear in UTF-8 as representations of the NUL and ’/’ character respectively.

- "Combining characters" need not have special processing as per SC22 resolutions, except for possibly a width specification in a locale.

- According to the ISO/IEC 10646 standard there is no equivalences prescribed between sequences of characters with combining characters and some "pre-composed" characters, and the SC22 plenary recommendation is that there need not be special handling of this.

- It should not be necessary to process composite sequences in a special way.

Copyright © 2000 IEEE. All rights reserved.
This is an unapproved IEEE Standards Draft, subject to change.

Annex E. Registry data

The following schema is needed for registration with IS 15897/ENV 12005:

Application form for a Cultural Specification

Please specify all data relevant for the Cultural Specification type, indicating non-available data by "not available". Please fill out one form for each Cultural Specification submitted. When completed, please send it to the Registration Authority as listed in clause 4.

1. Cultural Specification type number: _____

2. Organization name of Sponsoring Authority: _____

3. Organization postal address: _____

_____

4. Name of contact person: _____

5. Electronic mail address of contact person: _____

6. Telephone number for contact person: + ___ _____

7. Fax number for contact person: + ___ _____

For Narrative Cultural Specifications and POSIX Locales (type 1 and 2):

8. Natural language, as specified in ISO 639: _____

9. Territory, as two-letter form of ISO 3166: _____

For POSIX Charmaps and POSIX Repertoire maps (type 3 and 4):

10. The proposed POSIX Charmap or POSIX Repertoire map name: _____

For all 4 types:

11. If not for general use, an intended user audience, e.g. librarians: _____

12. If for use of a special application, the short application name: _____

13. Short name for Sponsoring Authority, used in token identifier: _____

14. Version number with zero or more dots: _____

15. Revision date in ISO 8601 format: _____

The Cultural Specification identified above, and of which we hold copyright, is allowed for free distribution.

Date: _____ Authorized signature: _____

Annex F. Examples of National Profile - Japan

© ISO/IEC 2000

[It is ready to include an example of Japanese National Profile here. Since the text is so large, the example is intentionally omitted from this review version of document. Please contact Japanese National Body for the details of Japanese National Profile.]

Annex G. Examples of National Locale - Norway

[An example of Norway National Locale will be provided here.]

Bibliography

Index

Copyright © 2000 IEEE. All rights reserved.
This is an unapproved IEEE Standards Draft, subject to change.