Internet Draft
draft-hoffman-i18n-terms-00.txt
November 16, 2000
Expires in six months

Paul Hoffman
IMC & VPNC

Terminology Used in Internationalization in the IETF

Status of this memo

Abstract

This document provides a glossary of terms used in the IETF when discussing internationalization. The purpose is to help frame discussions of internationalization in the various areas of the IETF and to help introduce the main concepts to IETF participants.

1. Introduction

As [RFC2277] summarizes: "Internationalization is for humans. This means that protocols are not subject to internationalization; text strings are." Many protocols throughout the IETF use text strings that are entered by, or are visible to, humans. It should be possible to make these text strings readable to anyone, which means that the text must be able to be displayed in any human language. This is the challenge of internationalization.

1.1 About this document

Internationalization is discussed in many working groups of the IETF. However, few working groups have internationalization experts. When designing or updating protocols, the question often comes up "should we internationalize this" (or, more likely, "do we have to internationalize

this").

This document gives an overview of internationalization by covering lightly the many aspects of internationalization and the vocabulary associated with those topics. It is not meant to be a complete description of internationalization. The definitions in this document come from many earlier IETF documents and books.

As in many fields, there is disagreement in the internationalization community on definitions for many words. The topic of language brings up particularly passionate opinions for experts and non-experts alike. This document attempts to define terms to be most useful to the IETF audience.

Note that this is a very early draft of the document. Many definitions here will likely change, and some topics may be added. Discussion of this document is encouraged. Information on the mailing list for this document can be found at <http://www.imc.org/ietf-i18n-terms/>.

1.2 Foundations for internationalization

language

A language is a way that humans interact. The use of language occurs in many forms, the most common of which are writing, vocal, and visual. Each language form is independent: some languages have a close relationship between the written and vocal forms, while others have a looser relationship. [RFC1766] and [RFC1766bis] discuss languages in more detail.

script

A script is the written form of a language. It can be considered a writing system which has many attributes, such as the number of written characters and the combining rules for written characters. Most IETF protocols that deal with languages deal only with scripts, not spoken or visual forms. [RFC2277] discusses scripts in more detail.

It is common for internationalization novices to mix up the terms "language" and "script". This can be a problem in protocols that differentiate the two, such as mail content protocols. Almost all internationalized protocols deal with scripts (the written systems), while fewer deal with languages. Many languages can be expressed using different scripts.

grapheme, phoneme, alphabet

A grapheme is an abstract, atomic written entity of a script. A phoneme is an abstract, atomic spoken entity of a spoken language. An alphabet is a script that maps between graphemes and phonemes.

internationalization

In the IETF, the verb "internationalize" means to add or improve the handling of international information in a protocol. Many protocols that handle text only handle one script, the one that contains the letters used in English text. Internationalizing such a protocol means to make the protocol able to handle more scripts, hopefully all of the ones useful to anyone in the world.

localization

Internationalized applications can handle a wide variety of languages. Typical users only understand a small number of languages, so the program must be tailored to interact with users in just the languages they know. Localization involves not only changing the language interaction, but also other relevant changes such as display of numbers, dates, currency, and so on.

i18n, l10n

These are abbreviations for "internationalization" and "localization". "18" is the number of characters between the "i" and the "n" in "internationalization", and "10" is the number of characters between the "l" and the "n" in "localization".

2. Fundamental Terms

This section covers basic topics that are needed for almost anyone who is involved with internationalization of IETF protocols. Many terms in this section are based on [IDN-REQ].

characters

A character is a member of a set of elements used for organization, control, or representation of data. In written form, a language is expressed in characters. The same set of characters can often be used in many languages.

coded character

A coded character is a character with its coded representation. A coded character set (CCS) is a set of unambiguous rules that establishes a character set and the relationship between the characters of the set and their coded representation. The set of characters in a CCS is often called the "repertoire".

character encoding scheme

A character encoding scheme (CES) is a mapping from one or more coded character sets to a set of octets. Some CESs are associated with a

single CCS; for example, UTF-8 [RFC2279] applies only to ISO 10646. Other CESs, such as ISO 2022, are associated with many CCSs.

charset

A charset is a method of mapping a sequence of octets to a sequence of abstract characters. A charset is, in effect, a combination of one or more CCS with a CES. Charset names are registered by the IANA according to procedures documented in [RFC2278]. A particular charset may have different glyphs depending on the language being used.

transfer encoding syntax

A transfer encoding syntax (TES) is a reversible transform of already-encoded data represented in one or more character encoding schemes. TESs are useful for encoding type of character data into an another format, usually for allowing new types of data to be transmitted over legacy protocols.

3. Standards Bodies and Standards

This section describes some of the standards bodies and standards that appear in discussions of internationalization in the IETF. This is an incomplete and possibly over-full list; listing too few bodies or standards can be just as politically dangerous as listing too many.

ISO

The International Organization for Standardization has been involved with standards for scripts since before the IETF was started. ISO is a non-governmental group made up of national bodies. ISO has many diverse standards in the international scripts area; the one that is most used in the IETF is commonly referred to as "ISO 10646" [ISO10646]. ISO 10646 describes a CCS that covers almost all known written characters in use today.

ISO 10646 is controlled by the group known as "ISO/IEC JTC 1/SC 2 WG2", often called "WG2" for short. ISO standards go through many steps before being finished, and years often go by between changes to ISO 10646. Information on WG2, and its work products, can be found at <http://anubis.dkuug.dk/JTC1/SC2/WG2/>.

Unicode Consortium

The second important group for international character standards is the Unicode Consortium. The Unicode Consortium is a trade association of companies and governments interested in promoting the Unicode Standard [Unicode3]. The Unicode Standard is a CCS whose repertoire is identical to ISO 10646. The Unicode Consortium has added features to the base CCS

which make it more useful in protocols, such as defining attributes for each character.

The Unicode Consortium publishes addenda to the Unicode Standard as Unicode Technical Reports. There are many types of technical reports at various stages of maturity. The Unicode Standard and affiliated technical reports can be found at <http://www.unicode.org/>

encodings and transformations of ISO 10646

Characters in the ISO 10646 CCS can be expressed in many ways. Encoding forms are direct addressing methods, while transformation formats are methods for expressing encoding forms as bits on the wire. Two encoding forms are defined for ISO 10646:

UCS-2 addresses the first 2^15 characters as 16-bit values. This range is also called the "Basic Multilingual Plane" (BMP), and is also called "plane 0".

UCS-4 addresses the entire range of 2^31 characters as 32-bit values.

There are many transformation formats of the CCS that are used in IETF standards. The two most common are:

UTF-8, defined in [RFC2279], is the preferred encoding for IETF protocols. Characters in the BMP are encoded as one, two, or three octets.

UTF-16 (BE & LE), defined in [RFC2781], is used much less often than UTF-8. Characters in the BMP are always encoded as two octets, and many characters outside the BMP as four octets.

native CCSs and charsets

Before ISO 10646 was developed, many countries developed their own CCSs and charsets. Many dozen of these are in common use on the Internet today. Examples include the ISO 2022 series, the ISO 8859 series, and Shift-JIS. The official list of the registered charset tags is maintained by IANA.

ASCII

Probably the most well-known native CCS is ASCII [US-ASCII]. This CCS is used in numerous IETF protocols that have not yet been internationalized.

local and regional standards organizations

Just as there are many native CCSs and charsets, there are many local and regional standards organizations to create and support them. Common examples of these are ANSI (United States), JIS (Japan), GB (China), and

ETSI (Europe).

4. Linguistic Issues

This section contains terms and topics that are commonly used in linguistics and therefore are of concern to people internationalizing protocols. These topics are standardized outside the IETF.

composition and decomposition

In some CCSs, some characters consist of combinations of other characters. For example, the letter "a with acute" might be a combination of the two characters "a" and "combining acute". The rules for combining two or more characters are called "composition", and the rules for taking apart a character into other characters is called "decomposition".

normalization and canonicalization

These two terms are often used interchangeably in internationalization. Generally, they both mean to convert a string of one or more characters into another string based on standardized rules. In internationalized text, these rules are usually based on decomposing combined characters or composing characters with combining characters. [UTR15] describes the process and many forms of normalization in detail.

locale and region

Because languages differ from country to country (and even region to region within a country), the locale of the user of internationalized text can often be an important factor. Typically, the locale information for a user includes the language(s) used. Locale issues go beyond character use, and can include things such as the display format for currency, dates, and times.

case

In many scripts, particularly those from or based on European alphabets, there are two forms for letters: uppercase and lowercase. There is usually (but not always) a one-to-one mapping between the same letter in the two cases. However, there are many examples of characters which exist in one case but for which there is no corresponding character in the other case. Case conversion can even be dependant on locale. Converting between the two cases is sometimes called "folding".

sorting

The characters in a CCS each have a code point, and the code points can be sorted, but the result of sorting characters based on code points is often not what a native reader of the language would expect. Therefore,

there are many different rules for sorting that depend on the CCS, on the language, and on the locale.

glyph

A graphic character or glyph is a character, other than a control function, that has a visual representation normally handwritten, printed, or displayed. There are many types of glyphs, including alphabetic characters, digits, punctuation, diacritics, and symbols.

types of characters
    alphabetic
    ideographic
    symbol
    spacing character
    punctuation
    diacritic
    combining character
    control character
    formatting character

5. User interface for text

Although the IETF does not standardize user interfaces, many protocols make assumptions about how a user will enter or see text that is used in the protocol. Many protocols make inherent assumptions such as that text will be typed on a standard (that is, US-centric) keyboard, or that text will be displayed on a character-based monitor. Internationalization challenges assumptions like these, and it is therefore useful to consider how users typically interact with internationalized text.

input methods
display methods
rendering characters
bidirectional scripts
undisplayable text

6. Text in current IETF protocols

Many IETF protocols started off being fully internationalized, while others have been internationalized as they were revised. In this process, IETF members have seen patters in the way that many protocols use text. This section describes some specific protocol interactions with text.

protocol elements
on-the-wire encoding
name spaces
identifiers
charset tagging

language tagging
MIME
base64
quoted printable
ASN.1 text formats
ASCII-compatible encoding

7. Other Common Terms In Internationalization

This is a hodge-podge of other terms that have appeared in internationalization discussions in the IETF. It is likely that additional terms will be added as this document matures.

Latin
romanization
CJK and Han
compression
translation
regular expressions
private use characters

8. Security Considerations

Security is not discussed in this document.

9. References

[IDN-REQ] "Requirements of Internationalized Domain Names", work in progress (draft-ietf-idn-requirements), Z. Wenzel and J. Seng.

[ISO10646] ISO/IEC 10646-1:1993. International Standard -- Information technology -- Universal Multiple-Octet Coded Character Set (UCS) -- Part 1: Architecture and Basic Multilingual Plane. Five amendments and a technical corrigendum have been published up to now. UTF-16 is described in Annex Q, published as Amendment 1. 17 other amendments are currently at various stages of standardization.

[RFC1766] "Tags for the Identification of Languages", RFC 1766, H. Alvestrand.

[RFC1766bis] "Tags for the Identification of Languages", work in progress (draft-alvestrand-lang-tag-v2), H. Alvestrand.

[RFC2277] "IETF Policy on Character Sets and Languages", RFC 2277, H. Alvestrand.

[RFC2279] "UTF-8, a transformation format of ISO 10646", RFC 2279, F. Yergeau.

[RFC2781] "UTF-16, an encoding of ISO 10646", RFC 2781, P. Hoffman and F. Yergeau.

[Unicode3] The Unicode Consortium, "The Unicode Standard -- Version 3.0", ISBN 0-201-61633-5. Described at <http://www.unicode.org/unicode/standard/versions/Unicode3.0.html>.

[US-ASCII] Coded Character Set -- 7-bit American Standard Code for Information Interchange, ANSI X3.4-1986.

[UTR15] "Unicode Normalization Forms", Unicode Technical Report #15, M. Davis & M. Duerst.

10. Additional Interesting Reading

ALA-LC Romanization Tables, Randall Barry (ed.), ISBN 0844409405

Blackwell Encyclopedia of Writing Systems, Florian Coulmas, ISBN 063121481X

MultiLingual Computing & Technology magazine, ISSN 1098-7665

Unicode Standard version 3.0, Unicode Consortium, ISBN 0201616335

Writing Systems of the World, Akira Nakanishi, ISBN 0804816549

A. Acknowledgements

The definitions in this document come from many sources, including a wide variety of IETF documents.

James Seng contributed to the initial outline of this document.

B. Author Contact Information

Paul Hoffman
Internet Mail Consortium and VPN Consortium
127 Segre Place
Santa Cruz, CA 95060 USA
paul.hoffman@imc.org and paul.hoffman@vpnc.org