Network Working Group
Request for Comments: 2978
BCP: 19
Obsoletes: 2278
Category: Best Current Practice

N. Freed
Innosoft
J. Postel
ISI

October 2000

IANA Charset Registration Procedures

Status of this Memo

Copyright Notice

Abstract

Multipurpose Internet Mail Extensions (MIME) (RFC-2045, RFC-2046, RFC-2047, RFC-2184) and various other Internet protocols are capable of using many different charsets. This in turn means that the ability to label different charsets is essential.

Note: The charset registration procedure exists solely to associate a specific name or names with a given charset and to give an indication of whether or not a given charset can be used in MIME text objects. In particular, the general applicability and appropriateness of a given registered charset to a particular application is a protocol issue, not a registration issue, and is not dealt with by this registration procedure.

1. Definitions and Notation

The following sections define terms used in this document.

1.1. Requirements Notation

This document occasionally uses terms that appear in capital letters. When the terms "MUST", "SHOULD", "MUST NOT", "SHOULD NOT", and "MAY" appear capitalized, they are being used to indicate particular requirements of this specification. A discussion of the meanings of these terms appears in [RFC-2119].

1.2. Character

A member of a set of elements used for the organization, control, or representation of data.

1.3. Charset

The term "charset" (referred to as a "character set" in previous versions of this document) is used here to refer to a method of converting a sequence of octets into a sequence of characters. This conversion may also optionally produce additional control information such as directionality indicators.

Note that unconditional and unambiguous conversion in the other direction is not required, in that not all characters may be representable by a given charset and a charset may provide more than one sequence of octets to represent a particular sequence of characters.

This definition is intended to allow charsets to be defined in a variety of different ways, from simple single-table mappings such as US-ASCII to complex table switching methods such as those that use ISO 2022's techniques. However, the definition associated with a charset name must fully specify the mapping to be performed. In particular, use of external profiling information to determine the exact mapping is not permitted.

HISTORICAL NOTE: The term "character set" was originally used in MIME to describe such straightforward schemes as US-ASCII and ISO-8859-1 which consist of a small set of characters and a simple one-to-one mapping from single octets to single characters. Multi-octet character encoding schemes and switching techniques make the situation much more complex. As such, the definition of this term was revised to emphasize both the conversion aspect of the process, and the term itself has been changed to "charset" to emphasize that it is not, after all, just a set of characters. A discussion of these issues as well as specification of standard terminology for use in the IETF appears in RFC 2130.

1.4. Coded Character Set

A Coded Character Set (CCS) is a one-to-one mapping from a set of abstract characters to a set of integers. Examples of coded character sets are ISO 10646 [ISO-10646], US-ASCII [US-ASCII], and the ISO-8859 series [ISO-8859].

1.5. Character Encoding Scheme

A Character Encoding Scheme (CES) is a mapping from a Coded Character Set or several coded character sets to a set of octet sequences. A given CES is sometimes associated with a single CCS; for example, UTF-8 applies only to ISO 10646.

2. Charset Registration Requirements

Registered charsets are expected to conform to a number of requirements as described below.

2.1. Required Characteristics

Registered charsets MUST conform to the definition of a "charset" given above. In addition, charsets intended for use in MIME content types under the "text" top-level type MUST conform to the restrictions on that type described in RFC 2045. All registered charsets MUST note whether or not they are suitable for use in MIME text.

All charsets which are constructed as a composition of one or more CCS's and a CES MUST either include the CCS's and CES they are based on in their registration or else cite a definition of their CCS's and CES that appears elsewhere.

All registered charsets MUST be specified in a stable, openly available specification. Registration of charsets whose specifications aren't stable and openly available is forbidden.

2.2. New Charsets

This registration mechanism is not intended to be a vehicle for the design and definition of entirely new charsets. This is due to the fact that the registration process does NOT contain adequate review mechanisms for such undertakings.

As such, only charsets defined by other processes and standards bodies, or specific profiles or combinations of such charsets, are eligible for registration.

2.3. Naming Requirements

One or more names MUST be assigned to all registered charsets. Multiple names for the same charset are permitted, but if multiple names are assigned a single primary name for the charset MUST be

identified. All other names are considered to be aliases for the primary name and use of the primary name is preferred over use of any of the aliases.

Each assigned name MUST uniquely identify a single charset. All charset names MUST be suitable for use as the value of a MIME content type charset parameter and hence MUST conform to MIME parameter value syntax. This applies even if the specific charset being registered is not suitable for use with the "text" media type.

All charsets MUST be assigned a name that provides a display string for the associated "MIBenum" value defined below. These "MIBenum" values are defined by and used in the Printer MIB [RFC-1759]. Such names MUST begin with the letters "cs" and MUST contain no more than 40 characters (including the "cs" prefix) chosen from from the printable subset of US-ASCII. Only one name beginning with "cs" may be assigned to a single charset. If no name of this form is explicitly defined IANA will assign an alias consisting of "cs" prepended to the primary charset name.

Finally, charsets being registered for use with the "text" media type MUST have a primary name that conforms to the more restrictive syntax of the charset field in MIME encoded-words [RFC-2047, RFC-2184] and MIME extended parameter values [RFC-2184]. A combined ABNF definition for such names is as follows:

```
mime-charset = 1*mime-charset-chars
mime-charset-chars = ALPHA / DIGIT /
           "!" / "#" / "$" / "%" / "&" /
           "'" / "+" / "-" / "^" / "_" /
           "`" / "{" / "}" / "~"
ALPHA        = "A".."Z"    ; Case insensitive ASCII Letter
DIGIT        = "0".."9"    ; Numeric digit
```

2.4. Functionality Requirement

Charsets MUST function as actual charsets: Registration of things that are better thought of as a transfer encoding, as a media type, or as a collection of separate entities of another type, is not allowed. For example, although HTML could theoretically be thought of as a charset, it is really better thought of as a media type and as such it cannot be registered as a charset.

2.5. Usage and Implementation Requirements

Use of a large number of charsets in a given protocol may hamper interoperability. However, the use of a large number of undocumented and/or unlabeled charsets hampers interoperability even more.

A charset should therefore be registered ONLY if it adds significant functionality that is valuable to a large community, OR if it documents existing practice in a large community. Note that charsets registered for the second reason should be explicitly marked as being of limited or specialized use and should only be used in Internet messages with prior bilateral agreement.

2.6. Publication Requirements

Charset registrations MAY be published in RFCs, however, RFC publication is not required to register a new charset.

The registration of a charset does not imply endorsement, approval, or recommendation by the IANA, IESG, or IETF, or even certification that the specification is adequate. It is expected that applicability statements for particular applications will be published from time to time that recommend implementation of, and support for, charsets that have proven particularly useful in those contexts.

Charset registrations SHOULD include a specification of mapping from the charset into ISO 10646 if specification of such a mapping is feasible.

2.7. MIBenum Requirements

Each registered charset MUST also be assigned a unique enumerated integer value. These "MIBenum" values are defined by and used in the Printer MIB [RFC-1759].

A MIBenum value for each charset will be assigned by IANA at the time of registration. MIBenum values are not assigned by the person registering the charset.

3. Charset Registration Procedure

The following procedure has been implemented by the IANA for review and approval of new charsets. This is not a formal standards process, but rather an administrative procedure intended to allow community comment and sanity checking without excessive time delay.

3.1. Present the Charset to the Community

Send the proposed charset registration to the "ietf-charsets@iana.org" mailing list. (Information about joining this list is available on the IANA Website, http://www.iana.org.) This mailing list has been established for the sole purpose of reviewing

proposed charset registrations. Proposed charsets are not formally registered and must not be used; the "x-" prefix specified in RFC 2045 can be used until registration is complete.

The posting of a charset to the list initiates a two week public review process.

The intent of the public posting is to solicit comments and feedback on the definition of the charset and the name chosen for it.

3.2. Charset Reviewer

When the two week period has passed and the registration proposer is convinced that consensus has been achieved, the registration application should be submitted to IANA and the charset reviewer. The charset reviewer, who is appointed by the IETF Applications Area Director(s), either approves the request for registration or rejects it. Rejection may occur because of significant objections raised on the list or objections raised externally. If the charset reviewer considers the registration sufficiently important and controversial, a last call for comments may be issued to the full IETF. The charset reviewer may also recommend standards track processing (before or after registration) when that appears appropriate and the level of specification of the charset is adequate.

The charset reviewer must reach a decision and post it to the ietf-charsets mailing list within two weeks. Decisions made by the reviewer may be appealed to the IESG.

3.3. IANA Registration

Provided that the charset registration has either passed review or has been successfully appealed to the IESG, the IANA will register the charset, assign a MIBenum value, and make its registration available to the community.

4. Location of Registered Charset List

Charset registrations will be posted in the anonymous FTP file "ftp://ftp.isi.edu/in-notes/iana/assignments/character-sets" and all registered charsets will be listed in the periodically issued "Assigned Numbers" RFC [currently RFC-1700]. The description of the charset MAY also be published as an Informational RFC by sending it to "rfc-editor@isi.edu" (please follow the instructions to RFC authors [RFC-1543]).

5. Charset Registration Template

To: ietf-charsets@iana.org
Subject: Registration of new charset [names]

Charset name:

(All names must be suitable for use as the value of a
MIME content-type parameter.)

Charset aliases:

(All aliases must also be suitable for use as the value of
a MIME content-type parameter.)

Suitability for use in MIME text:

Published specification(s):

(A specification for the charset MUST be
openly available that accurately describes what
is being registered. If a charset is defined as
a composition of one or more CCS's and a CES then these
definitions MUST either be included or referenced.)

ISO 10646 equivalency table:

(A URI to a specification of how to translate from
this charset to ISO 10646 and vice versa SHOULD be
provided.)

Additional information:

Person & email address to contact for further information:

Intended usage:

(One of COMMON, LIMITED USE or OBSOLETE)

6. Security Considerations

This registration procedure is not known to raise any sort of security considerations that are appreciably different from those already existing in the protocols that employ registered charsets.

7. Changes made since RFC 2278

Inclusion of a mapping to ISO 10646 is now recommended for all registered charsets. The registration template has been updated to include this as well as a place to indicate whether or not the charset is suitable for use in MIME text.

8. References

[ISO-2022]
International Standard -- Information Processing -- Character Code Structure and Extension Techniques, ISO/IEC 2022:1994, 4th ed.

[ISO-8859]
International Standard -- Information Processing -- 8-bit Single-Byte Coded Graphic Character Sets
- Part 1: Latin Alphabet No. 1, ISO 8859-1:1998, 1st ed.
- Part 2: Latin Alphabet No. 2, ISO 8859-2:1999, 1st ed.
- Part 3: Latin Alphabet No. 3, ISO 8859-3:1999, 1st ed.
- Part 4: Latin Alphabet No. 4, ISO 8859-4:1998, 1st ed.
- Part 5: Latin/Cyrillic Alphabet, ISO 8859-5:1999, 2nd ed.
- Part 6: Latin/Arabic Alphabet, ISO 8859-6:1999, 1st ed.
- Part 7: Latin/Greek Alphabet, ISO 8859-7:1987, 1st ed.
- Part 8: Latin/Hebrew Alphabet, ISO 8859-8:1999, 1st ed.
- Part 9: Latin Alphabet No. 5, ISO/IEC 8859-9:1999, 2nd ed.
International Standard -- Information Technology -- 8-bit Single-Byte Coded Graphic Character Sets
- Part 10: Latin Alphabet No. 6, ISO/IEC 8859-10:1998, 2nd ed.
International Standard -- Information Technology -- 8-bit Single-Byte Coded Graphic Character Sets
- Part 13: Latin Alphabet No. 7, ISO/IEC 8859-10:1998, 1st ed.
International Standard -- Information Technology -- 8-bit Single-Byte Coded Graphic Character Sets
- Part 14: Latin Alphabet No. 8 (Celtic), ISO/IEC 8859-10:1998, 1st ed.
International Standard -- Information Technology -- 8-bit Single-Byte Coded Graphic Character Sets
- Part 15: Latin Alphabet No. 9, ISO/IEC 8859-10:1999, 1st ed.

[ISO-10646]
ISO/IEC 10646-1:1993(E), "Information technology -- Universal Multiple-Octet Coded Character Set (UCS) -- Part 1: Architecture and Basic Multilingual Plane", JTC1/SC2, 1993.

[RFC-1590] Postel, J., "Media Type Registration Procedure", RFC 1590,March 1994.

[RFC-1700] Reynolds, J. and J. Postel, "Assigned Numbers", STD 2, RFC 1700, October 1994.

[RFC-1759] Smith, R., Wright, F., Hastings, T., Zilles, S. and J. Gyllenskog, "Printer MIB", RFC 1759, March 1995.

[RFC-2045] Freed, N. and N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies", RFC 2045, November 1996.

[RFC-2046] Freed, N. and N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types", RFC 2046, November 1996.

[RFC-2047] Moore, K., "Multipurpose Internet Mail Extensions (MIME) Part Three: Representation of Non-Ascii Text in Internet Message Headers", RFC 2047, November 1996.

[RFC-2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC-2130] Weider, C., Preston, C., Simonsen, K., Alvestrand, H., Atkinson, R., Crispin, M. and P. Svanberg, "Report from the IAB Character Set Workshop", RFC 2130, April 1997.

[RFC-2184] Freed, N. and K. Moore, "MIME Parameter Value and Encoded Word Extensions: Character Sets, Languages, and Continuations", RFC 2184, August 1997.

[RFC-2468] Cerf, V., "I Remember IANA", RFC 2468, October 1998.

[RFC-2278] Freed, N. and J. Postel, "IANA Charset Registration Procedures", BCP 19, RFC 2278, January 1998.

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

10. Authors' Addresses

Ned Freed
Innosoft International, Inc.
1050 Lakes Drive
West Covina, CA 91790 USA

Phone: +1 626 919 3600
Fax: +1 626 919 3614
EMail: ned.freed@innosoft.com

Jon Postel

Sadly, Jon Postel, the co-author of this document, passed away on October 16, 1998 [RFC-2468]. Any omissions or errors are solely the responsibility of the remaining co-author.

11. Full Copyright Statement

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.