Network Working Group
Internet-Draft
Obsoletes: 2978 (if approved)
Intended status: Best Current Practice
Expires: October 26, 2015

M. Mcfadden
IANA
A. Melnikov, Ed.
Isode Ltd

April 24, 2015

IANA Charset Registration Procedures
draft-iana-charset-reg-procedure-01

Abstract

Multipurpose Internet Mail Extensions (MIME) (RFC-2045, RFC-2046, RFC-2047, RFC-2231) 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.

This document obsoletes the IANA Charset Registration Procedures originally defined in [RFC2978]. Specifically, this document completely revises the registration procedures and the charset registries. The charset registry is now divided into three parts with separate registration procedures for each.

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.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on October 26, 2015.

Table of Contents

1. Definitions and Notation

The following sections define terms used in this document.

1.1. Requirements Notation

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

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 [RFC0020] to complex table switching methods such as those that use ISO 2022's [ISO-2022] 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 [ISO-8859] 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 [RFC2130].

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 [RFC0020], 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 [RFC3629] 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 media type MUST conform to the restrictions on that type described in [RFC2045]. 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 (see Section 5.1 of RFC 2045). 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 [RFC1759]. [[RFC 1759 got obsoleted by RFC 3805 and MIBEnum is no longer there. Should we point to http://www.iana.org/assignments/ianacharset-mib instead?]] 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 [RFC2047] [RFC2231] and MIME extended parameter values [RFC2231]. A combined ABNF [RFC5234] 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 [RFC2046], 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 (Unicode) [Unicode7.0] 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 [RFC1759]."

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. The Charset Registry

The following procedure has been implemented by the IANA for review and approval of new charsets. In [RFC2978] an Expert Review process was used to add new charsets into the registry. This document changes that model by creating a new charset registry with three new subregistries. For each of the new registries, the registration procedures and initial registrations are provided.

3.1. The Recommended charset registry

The first sub-registry of the full charset registry is the "recommended" charset registry.

New registrations in the "recommended" charset registry require "Standards Action" as defined by [RFC5226]. Specifically, the charset MUST have a standards track RFC that defines the charset itself and MUST ALSO have a standards track RFC recommending its use.

In the RFC that defines the charset, the document MUST have a single recommended MIME charset label following the "mime-charset" syntax defined in Section 2.3. It MUST also state whether it is suitable for MIME text and have a reference to a formal specification or translation table to Unicode [Unicode7.0].

There is one, initial entry in the Recommended charset registry: UTF-8 [RFC3629].

3.2. The Widely-used Open Standard charset registry

The second sub-registry of the full charset registry is the "Widely-used Open Standard" charset registry.

New registrations in the "Widely-used Open Standard" charset registry require "Expert Review" as defined by [RFC5226]. In Section 3.2.2 of this document a template is provided that allows proposals for new charsets in this subregistry.

In the template that describes the charset, the template MUST provide a single recommended MIME charset label following the "mime-charset" syntax defined in Section 2.3. It MUST ALSO state whether it is suitable for MIME text and have a reference to a formal specification or translation table to Unicode.

The following charsets are to be moved from the historic charset registry into the new "Widely-used Open Standard" subregistry: INSERT A LIST OF CHARSET NAMES HERE. [[GUIDANCE IS REQUIRED FOR THIS ENTRY]]

3.2.1. Submitting "Widely-used Open Standard" charset Proposals to the IETF Community

Send the proposed "Widely-used Open Standard" charset proposal 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 [RFC2045] 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.

3.2.2. IANA 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, see Section 5.1 of RFC 2045.)

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)

3.2.3. 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.

3.2.4. IANA Registration of "Widely-used Open Standard" charsets

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.

3.3. The Other charset subregistry

The third subregistry is for all other charsets. Registration of charsets in the "other" charset subregistry is done on a "First Come, First Served" basis as defined by [RFC5226].

4. IANA Considerations

This document requests that IANA completely revise the existing charset registry. The new registry shold be divided into three subregistries. These subregistries are: "Recommended charsets", "Widely-used Open Standard charsets" and "Other charsets".

The registration procedure for the "Recommended charset" subregistry is Standards Action required. IANA is directed to move the following

entries from the [RFC2978] legacy registry to this subregistry: UTF-8 [RFC3629].

The registration procedure for the "Widely-used Open Standard charset" subregistry is Expert Review. IANA is directed to move the following entries from the [RFC2978] legacy registry to this subregistry: INSERT A LIST OF CHARSET NAMES HERE. [[GUIDANCE IS REQUIRED FOR THIS ENTRY]]

The registration procedure for the "Other charset" subregistry is First Come First Served. IANA is directed to move the following entries from the [RFC2978] legacy registry to this subregistry: INSERT A LIST OF CHARSET NAMES HERE. [[GUIDANCE IS REQUIRED FOR THIS ENTRY]]

In all cases the registration template specified in Section 3.2.2 must be used.

4.1. Publication of Registered Charset List

This document directs IANA to create a new XML-based registry for charset registrations. This registry will be divided into three subregistries as specified in Section 3 of this document."

New charset registrations will be published in the new, XML-based registry. The proposed charset will use the approval process appropriate for the indended, designated subregistry.

Legacy charset registrations will be converted to the new XML registry. The instructions for converting the legacy registrations into entries in the new subregistries are documented in Section 4 of this document.

HISTORICAL NOTE: Previously, charset registrations were posted in the anonymous FTP file "ftp://ftp.isi.edu/in-notes/iana/assignments/character-sets" and all registered charsets were listed in the periodically issued "Assigned Numbers" RFC.

5. Security Considerations

The conversion of this IANA registry - and the changes made to the registration procedures for the new subregistries - introduces no known security considerations. Security issues that relate to charsets are dealt with in the RFCs that describe the protocols that use those charsets.

6. Acknowledgements

This document is a revision of RFC 2978 by Ned Freed and Jon Postel and is largely based on their original text.

7. References

7.1. Normative References

[RFC0020] Cerf, V., "ASCII format for network interchange", RFC 20, October 1969.

[RFC1759] Smith, R., Wright, F., Hastings, T., Zilles, S., and J. Gyllenskog, "Printer MIB", RFC 1759, March 1995.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2045] Freed, N. and N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies", RFC 2045, November 1996.

[RFC2046] Freed, N. and N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types", RFC 2046, November 1996.

[RFC2047] Moore, K., "MIME (Multipurpose Internet Mail Extensions) Part Three: Message Header Extensions for Non-ASCII Text", RFC 2047, November 1996.

[RFC2231] Freed, N. and K. Moore, "MIME Parameter Value and Encoded Word Extensions:
Character Sets, Languages, and Continuations", RFC 2231, November 1997.

[RFC3629] Yergeau, F., "UTF-8, a transformation format of ISO 10646", STD 63, RFC 3629, November 2003.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

[RFC5234] Crocker, D. and P. Overell, "Augmented BNF for Syntax Specifications: ABNF", STD 68, RFC 5234, January 2008.

[Unicode7.0]
The Unicode Consortium, "The Unicode Standard, Version 7.0.0", 2014, <http://www.unicode.org/versions/Unicode7.0.0/>.

7.2. Informative References

[RFC2978] Freed, N. and J. Postel, "IANA Charset Registration Procedures", BCP 19, RFC 2978, October 2000.

[RFC2130] Weider, C., Preston, C., Simonsen, K., Alvestrand, H., Atkinson, R., Crispin, M., and P. Svanberg, "The Report of the IAB Character Set Workshop held 29 February - 1 March, 1996", RFC 2130, April 1997.

[ISO-2022]
International Organization for Standardization, "Information technology - Character code structure and extension techniques", ISO Standard 2022, 1994.

[ISO-10646]
International Organization for Standardization, "Information Technology - Universal Multiple-octet coded Character Set (UCS) - Part 1: Architecture and Basic Multilingual Plane", ISO Standard 10646-1, May 1993.

[ISO-8859]
International Organization for Standardization, "Information processing - 8-bit single-byte coded graphic character sets - Part 1: Latin alphabet No. 1 (1987) - Part 2: Latin alphabet No. 2 (1987) - Part 3: Latin alphabet No. 3 (1988) - Part 4: Latin alphabet No. 4 (1988) - Part 5: Latin/Cyrillic alphabet (1988) - Part 6: Latin/Arabic alphabet (1987) - Part 7: Latin/Greek alphabet (1987) - Part 8: Latin/Hebrew alphabet (1988) - Part 9: Latin alphabet No. 5 (1989) - Part 10: Latin alphabet No. 6 (1992)", ISO Standard 8859, 1992.

Appendix A. Changes Since RFC 2978

Created 3 new subregistries with different IANA registration procedures instead of a single existing one.

Updated references, split them into Normative and Informative. Erratum 357.

Disallow single quotes in charset names (as per RFC 2231). Erratum 1912. Note that vertical bar and backslash characters were prohibited in RFC 2978 (a change from RFC 2278), but the change was never noted in RFC 2978.

Authors' Addresses

Mark Mcfadden
IANA

EMail: mark.mcfadden@icann.org

Alexey Melnikov (editor)
Isode Ltd
14 Castle Mews
Hampton, Middlesex TW12 2NP
UK

EMail: Alexey.Melnikov@isode.com