Internet Engineering Task Force (IETF)
Request for Comments: 6067
Category: Informational
ISSN: 2070-1721

M. Davis
Google
A. Phillips
Lab126
Y. Umaoka
IBM

December 2010

BCP 47 Extension U

Abstract

This document specifies an Extension to BCP 47 that provides subtags that specify language and/or locale-based behavior or refinements to language tags, according to work done by the Unicode Consortium.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc6067.

Table of Contents

1. Introduction 2
1.1. Requirements Language 2
2. BCP 47 Required Information 2
2.1. Summary 4
2.1.1. Canonicalization 5
2.2. Registration Form 6
3. Acknowledgements 6
4. IANA Considerations 6
5. Security Considerations 7
6. References 7
6.1. Normative References 7
6.2. Informative References 7

1. Introduction

[BCP47] permits the definition and registration of language tag extensions "that contain a language component and are compatible with applications that understand language tags". This document defines an extension for identifying Unicode locale-based variations using language tags. The "singleton" identifier for this extension is 'u'.

1.1. Requirements Language

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 RFC 2119 [RFC2119].

2. BCP 47 Required Information

Language tags, as defined by [BCP47], are useful for identifying the language of content. They are also used as locale identifiers (or can be mapped to locales) in many operating environments and APIs. However, many locale identifiers also require additional "tailorings" or options for specific values within a language, culture, region, or other variation. This extension provides a mechanism for using these additional tailorings within language tags for general interchange.

The Unicode Consortium defines a standardized, structured set of locale data and identifiers for locale data in the "Common Locale Data Repository" or "CLDR". The maintaining authority for the extension defined by this document is the Unicode Consortium:

Item	Value
Name	Unicode Consortium
Contact Email	cldr-contact@unicode.org
Discussion List Email	cldr-users@unicode.org
URL Location	cldr.unicode.org
Specification	Unicode Technical Standard #35 Unicode Locale Data Markup Language (LDML), http://unicode.org/reports/tr35/
Section	Section 3 Unicode Language and Locale Identifiers

The specification of extension subtags is provided by Section 3, Key Type Definitions of Unicode Technical Standard #35: Unicode Locale Data Markup Language [UTS35]. As required by BCP 47, subtags follow the language tag ABNF and other rules for the formation of language tags and subtags, are restricted to the ASCII letters and digits, are not case sensitive, and do not exceed eight characters in length. Note that any "well-formed" language tag (see RFC 5646, Section 2.2.9 [BCP47]) is also a well-formed locale identifier.

LDML [UTS35] specifies a canonical representation. LDML is available over the Internet and at no cost, and is available via a royalty-free license at http://unicode.org/copyright.html. LDML is versioned, and each version of LDML is numbered, dated, and stable. Extension subtags, once defined by LDML, are never retracted and never change in meaning in a substantial way.

The structure of the Unicode locale extension is determined by the Unicode CLDR Technical Committee, in accordance with the policies and procedures in http://www.unicode.org/consortium/tc-procedures.html, and subject to the Unicode Consortium Policies on http://www.unicode.org/policies/policies.html.

Changes that can be made by successive versions of LDML [UTS35] by the Unicode Consortium without requiring a new RFC include: the allocation of new attributes, keywords, and types; clarifications or non-material changes to an existing attribute, keyword, or type; and compatible extensions to the overall syntactic structure of attributes, keywords, and types. A new RFC would be required for material changes to an existing attribute, keyword, or type, or an incompatible change to the overall syntactic structure of attributes, keywords, and types; however, such a change would be contrary to the policies of the Unicode Consortium, and thus is not anticipated.

2.1. Summary

The subtags available for use in the 'u' extension consist of a set of attributes, keys, and types. Attributes, keys, types, and their respective meanings are defined in Section 3 (Unicode Language and Locale Identifiers) of [UTS35]. The following is a summary of that definition:

- An 'attribute' is a subtag with a length of three to eight characters following the singleton and preceding any 'keyword' sequences. No attributes were defined at the time of this document's publication.

- A 'keyword' is a sequence of subtags consisting of a 'key' subtag, followed by zero or more 'type' subtags (so a 'key' might appear alone and not be accompanied by a 'type' subtag). A 'key' MUST NOT appear more than once in a language tag's extension string. The order of the 'type' subtags within a 'keyword' is sometimes significant to their interpretation.

  A. A 'key' is a subtag with a length of exactly two characters. Each 'key' is followed by zero or more 'type' subtags.

  B. A 'type' is a subtag with a length of three to eight characters following a 'key'. 'Type' subtags are specific to a particular 'key' and the order of the 'type' subtags MAY be significant to the interpretation of the 'keyword'.

For example, the language tag "de-DE-u-attr-co-phonebk" consists of:

- The base language tag "de-DE" (German as used in Germany), exactly as defined by [BCP47] using subtags from the IANA Language Subtag Registry.

- The singleton 'u', identifying this extension.

- The attribute 'attr', which is an example for illustration (no attributes were defined at the time this document was published).

- The keyword 'co-phonebk', consisting to the key 'co' (Collation) and the type 'phonebk' (Phonebook collation order).

Only the first occurrence of an attribute or key conveys meaning in a language tag. When interpreting tags containing the Unicode locale extension, duplicate attributes or keywords are ignored in the following way: ignore any attribute that has already appeared in the tag and ignore any keyword whose key has already occurred in the tag.

Successive versions of [UTS35] could define additional attributes, keys, and types. Once defined, attributes, keys, and types will never be removed.

Beginning with CLDR version 1.7.2, machine-readable files are available listing the valid attributes, keys, and types for each successive version of [UTS35]. These releases are listed on http://cldr.unicode.org/index/downloads. Each release has an associated data directory of the form "http://unicode.org/Public/cldr/<version>", where "<version>" is replaced by the release number. For example, for version 1.7.2, the "core.zip" file is located at http://unicode.org/Public/cldr/1.7.2/core.zip. Inside the "core.zip" file, the path "common/bcp47" contains the data files defining the valid attributes, keys, and types. The most recent version is always identified by the version "latest" and can be accessed by the URL in Section 2.2.

To get the version information in XML when working with the data files, the XML parser must be validating. When the 'core.zip' file is unzipped, the 'dtd' directory will be at the same level as the 'bcp47' directory; this is required for correct validation. For each release after CLDR 1.8, types introduced in that release are also marked in the data files by the XML attribute "since", such as in the following example:
<type name="adp" since="1.9"/>

The data is also currently maintained in a source code repository, with each release tagged, for viewing directly without unzipping. For example, see:

- http://unicode.org/repos/cldr/tags/release-1-7-2/common/bcp47/

- http://unicode.org/repos/cldr/tags/release-1-8/common/bcp47/

Some data in the CLDR data files might require reference to LDML [UTS35]. For specific information, see Appendix Q in that document. For example, LDML reserves the type 'codepoints' to define specific code point ranges in Unicode for specific purposes.

2.1.1. Canonicalization

As required by [BCP47], the use of uppercase or lowercase letters is not significant in the subtags used in this extension. The canonical form for all subtags in the extension is lowercase. The canonical order of attributes is in [US-ASCII] order (that is, numbers before letters, with letters sorted as lowercase US-ASCII code points). The canonical order of keywords is in [US-ASCII] order by key. The order

of subtags within a keyword is significant; the meaning of this extension is altered if those subtags are rearranged. Thus, the canonical form of the extension never reorders the subtags within a keyword.

2.2. Registration Form

Per RFC 5646, Section 3.7 [BCP47]:

```
%%
Identifier: u
Description: Unicode Locale
Comments: Subtags for the identification of language and cultural
   variations.  Used to set behavior in locale APIs.  Data is
   located in the "common/bcp47" directory inside the referenced
   URL.  Unicode Technical Standard #35 (LDML) provides additional
   reference material defining the keys and values.
   For more details please see
   <http://cldr.unicode.org/index/bcp47-extension>.
Added: 2010-09-02
RFC: RFC 6067
Authority:     Unicode Consortium
Contact_Email: cldr-contact@unicode.org
Mailing_List:  cldr-users@unicode.org
URL: http://www.unicode.org/Public/cldr/latest/core.zip
%%
```

3. Acknowledgements

Thanks to John Emmons and the rest of the Unicode CLDR Technical Committee for their work in developing the BCP 47 subtags for LDML.

Thanks also to Doug Ewell, for his many suggestions for improvements to this document.

4. IANA Considerations

According to this document, IANA has inserted the record in Section 2.2 into the Language Extensions Registry, according to Section 3.7 (Extensions and the Extensions Registry) of [BCP47], "Tags for Identifying Languages". Per Section 5.2 of [BCP47], there might be occasional (rare) requests by the Unicode Consortium (the "Authority" listed in the record) for maintenance of this record. Changes that can be submitted to IANA without the publication of a new RFC are limited to modification of the Comments, Contact_Email, Mailing_List, and URL fields. Any such requested changes MUST use the domain 'unicode.org' in any new addresses or URIs, MUST explicitly cite this document (so that IANA can reference these

requirements), and MUST originate from the 'unicode.org' domain. The domain or authority can only be changed via a new RFC.

5. Security Considerations

The security considerations for this extension are the same as those for [BCP47]. See RFC 5646, Section 6, Security Considerations [BCP47].

6. References

6.1. Normative References

[BCP47] Phillips, A., Ed. and M. Davis, Ed., "Tags for Identifying Languages", BCP 47, RFC 5646, September 2009.

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

[US-ASCII] International Organization for Standardization, "ISO/IEC 646:1991, Information technology -- ISO 7-bit coded character set for information interchange.", 1991.

[UTS35] Davis, M., "Unicode Technical Standard #35: Locale Data Markup Language (LDML)", December 2007, <http://www.unicode.org/reports/tr35/>.

Section 3: http://unicode.org/reports/tr35/#Unicode_Language_and_Locale_Identifiers

Appendix Q: http://unicode.org/reports/tr35/#Locale_Extension_Key_and_Type_Data

6.2. Informative References

[ldml-registry] "Registry for Common Locale Data Repository tag elements", September 2009.

Authors' Addresses

Mark Davis
Google

EMail: mark@macchiato.com

Addison Phillips
Lab126

EMail: addison@lab126.com

Yoshito Umaoka
IBM

EMail: yoshito_umaoka@us.ibm.com