Internet Engineering Task Force	M.E. Davis
Internet-Draft	Google
Intended status: Informational	A. Phillips
Expires: December 19, 2011	Lab126
	Y. Umaoka
	IBM
	June 17, 2011

BCP 47 Extension T
draft-davis-t-langtag-ext-00

Abstract

This document specifies an Extension to BCP 47 which provides subtags for specifying the source language or script of transformed text, including text that has been transliterated, transcribed, or translated. It also provides for additional information used for identification.

Status of this Memo

Copyright Notice

Table of Contents

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 specifying the source of a text transformation, including text that has been transliterated, transcribed, or translated. The "singleton" identifier for this extension is 't'.

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.

2. BCP47 Required Information

Language tags, as defined by [BCP47], are useful for identifying the language of content. There are mechanisms for specifying variant subtags for special purposes. However, these variants are insufficient for specifying text transformations, including text that has been transliterated, transcribed, or translated. That is, for fully specifying such text, it is important to specify the source language and/or script. In addition, it may also be important to specify a particular specification for the transformation.
For example, if one is transcribing the names of Italian or Russian cities on a map for Japanese users, each name will need to be transliterated into katakana using rules appropriate for the source

language and target languages. When tagging such data, it is important to be able to indicate not only the resulting content language ("ja" in this case), but also the source language.
Transforms such as transliteration may vary depending not only on the basis of the source and target script, but also language. Thus the Russian <U+041F U+0443 U+0442 U+0438 U+043D> (which corresponds to the Cyrillic <PE, U, TE, I, EN>) transliterates into "Putin" in English but "Poutine" in French. The identifier may need to indicate a desired mechanical transformation in an API, or may need to tag data that has been converted (mechanically or by hand) according to a transliteration method.
Such identification is accomplished by using the 't' extension defined in this document. This extension is formed by the 't' singleton followed by a sequence of subtags that would form a language tag defined by [BCP47]. This allows for the source language or script to be specified to the degree of precision required. There are restrictions on the sequence of subtags. They MUST form a regular, valid, canonical language tag, and MUST neither include extensions nor private use sequences introduced by the singleton 'x'. Where only the script is relevant (such as identifying a script-script transliteration) then 'und' is used for the primary language subtag.
For example:

Language Tag	Description
ja-t-it	The content is Japanese, transformed from Italian.
ja-Kana-t-it	The content is Japanese Katakana, transformed from Italian.
und-Latn-t-und-cyrl	The content is in the Latin script, transformed from the Cyrillic script.

Note that the sequence of subtags governed by 't' cannot contain a singleton (a single-character subtag), because that would start a new extension. For example, the tag "ja-t-i-ami" does not indicate that the source is in "i-ami", because "i-ami" is not a regular language tag in [BCP47]. That tag would express an empty 't' extension followed by an 'i' extension.
In addition, it is sometimes necessary to indicate additional information, such as the mechanism used to do the transformation, optionally including the version of the mechanism. The mechanism can be supplied by using the 'm0' separator. The format of such a 't' extension is thus:
"t-<language-tag>-m0-<mechanism>".
(The full format reserves some additional syntax for future expansion, as described below.)
The transform <mechanism> is a series of subtags that indicate the specification used for the transformation, such as "UNGEGN" for the the United Nations Group of Experts on Geographical Names transliterations and transcriptions.

For example:

Language Tag	Description
und-Cyrl-t-und-latn-m0-ungegn-2007	the content is in Cyrillic, transformed from Latn, according to a UNGEGN specification dated 2007.

The separator subtags such as 'm0' were chosen because they are short, visually distinctive, and cannot occur in a language subtag (outside of an extension and after 'x'), thus eliminating the potential for collision or confusion with the source language tag.

The subtags that are valid after in the 't' extension are provided by Section 3 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.

EDITORIAL NOTE: This new facility has been accepted by the Unicode CLDR committee for incorporation into the next version of Unicode CLDR, parallel with the structure of the 'u' extension [RFC6067], for which it is already the maintaining authority. The data and specification will be available by the time this internet draft has been approved. 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 or change in meaning in a substantial way.

The structure of 't' subtags 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 subtags for use after the 't' extension. A new RFC would be required for material changes to an existing 't' subtag, or an incompatible change to the overall syntactic structure of the 't' extension; however, such a change would be contrary to the policies of the Unicode Consortium, and thus is not anticipated.

The maintaining authority for the 't' extension is the Unicode Consortium:

Item	Value
Name	Unicode Consortium
Contact Email	cldr-contact@unicode.org
Discussion List Email	cldr-users@unicode.org

Item	Value
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

2.1. Summary

The following is a summary of the definition for the 't' subtags defined by Section 3 of Unicode Technical Standard #35: Unicode Locale Data Markup Language *[UTS35]*, which is relevant for this specification. The subtags in the 't' extension are of the following form:

Label	ABNF	Comment
t_ext=	"t-"	Extension
	[lang]	Source
	*("-" field)	Optional information
lang=	language	[BCP47], with restrictions
	["-" script]	
	["-" region]	
	*("-" variant)	
field=	sep 1*("-" 3*8alphanum)	With restrictions
sep=	1ALPHA 1DIGIT	Subtag separators

Description and restrictions:

a. The 't' extension MUST have at least one subtag.

b. The 't' extension normally starts with a source language tag, which MUST be a regular, canonical language tag as specified by [BCP47]. Tags described by the 'irregular' production in BCP 47 MUST NOT be used to form the language tag. The source language tag MAY be omitted: some field values do not require it.

c. There is optionally a sequence of fields, where each field is a separator followed by a sequence of subtags. Two identical separators MUST NOT be present.

d. One field is initially specified in [UTS35]: the transform mechanism.

    a. The transform mechanism consists of a sequence of subtags starting with the 'm0' separator followed by one or more mechanism subtags. Each mechanism subtag has a length of 3

to 8 alphanumeric characters. The sequence as a whole provides an identification of the specification for the transform, such as the mechanism subtag 'UNGEGN' in "und-Cyrl-t-und-latn-m0-ungegn". In many cases, only one mechanism subtag is necessary, but multiple subtags MAY be defined in [UTS35] where necessary.

b. Any purely numeric subtag is a representation of a date in the Gregorian calendar. It MAY occur in any mechanism field. If it does occur:

   *it MUST occur as the final subtag in the field,

   *it MUST NOT be the only subtag in the field, and

   *it MUST consist of a sequence of digits of the form YYYY, YYYYMM, or YYYYMMDD.

   For example, 20110623 represents June 23th, 2011. A date subtag SHOULD only be used where necessary, and then SHOULD be as short as possible. For example, suppose that the BGN transliteration specification for Cyrillic to Latin had three versions, dated June 11th, 1999; Dec 30th, 1999; and May 1st, 2011. In that case, the corresponding first two DATE subtags would require months to be distinctive (199906 and 199912), but the last subtag would only require the year (2011).

c. Some mechanisms may use a versioning system that is not distinguished by date, or not by date alone. In the latter case, the version will be of a form specified by [UTS35] for that mechanism. For example, if the mechanism XXX uses versions of the form v21a, then a tag could look like "ja-t-it-m0-xxx-v21a". If there are multiple subversions distinguished by date, then a tag could look like "ja-t-it-m0-xxx-v21a-2007".

e. Successive versions of [UTS35] could define additional separator subtags, and additional subtags for those separators. Once defined, those subtags will never be removed.

f. The order of the subtags is significant (see Section 2.1.1 Canonicalization).

EDITORIAL NOTE: The following parallels the structure used for the 'u' extension [RFC6067], for which the Unicode Consortium is the maintaining authority. The data and specification will be available by the time this internet draft has been approved.
Beginning with CLDR version 1.7.2, machine-readable files are available listing the data defined for BCP47 extensions 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 data defined for BCP47 extensions. The most recent version is always identified by the version "latest" and can be accessed by the URL in Section 2.2.

<type name="adp" since="1.9"/>

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; that 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:
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/

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, with the fields ordered by the separators, alphabetically.

2.2. Registration Form

Per RFC 5646, Section 3.7 *[BCP47]* :

```
%%
Identifier: t
Description: Transform Specification
Comments: Subtags for the identification of text transforms,
    including transliteration, transcription, and translation.
Added: 2010-mm-dd
RFC: [TBD]
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.

4. IANA Considerations

This document will require IANA to insert the record in Section 2.2 into the Language Extensions Registry, according to Section 3.7. Extensions and the Extensions Registry of "Tags for Identifying Languages" in [BCP47]. 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.
This document does not require IANA to create or maintain a new registry or otherwise impact IANA.

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

[UTS35]	Davis, M, "Unicode Technical Standard #35: Locale Data Markup Language (LDML) ", December 2007. Section 3: http://unicode.org/reports/tr35/#Unicode_Language_and_Locale_Identifiers

[BCP47]	Davis, M.E., "Tags for the Identification of Language (BCP47)", September 2009.
[RFC6067]	Davis, M.E., "BCP 47 Extension U", September 2010.
[US-ASCII]	International Organization for Standardization, "ISO/IEC 646:1991, Information technology -- ISO 7-bit coded character set for information interchange. ", 1991.

6.2. Informative References

, "

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

Authors' Addresses

Mark Davis Davis Google EMail: mark@macchiato.com

Addison Phillips Phillips Lab126 EMail: addison@lab126.com

Yoshito Umaoka Umaoka IBM EMail: yoshito_umaoka@us.ibm.com