Network Working Group
Internet Draft
Intended status: Informational
Expires: December 2011

C. Falk
Infinite Automata

June 13, 2011

Tags for the Identification of Transliterated Text
draft-falk-transliteration-tags-01.txt

Status of this Memo

This Internet-Draft will expire on December 13, 2011.

Copyright Notice

Abstract

This document describes the structure, content, creation, and semantics of language tags for use in describing text that was transliterated from one orthographic system to another.

Table of Contents

1. Introduction

1.1. Problems Concerning Language Tags

Language tags are a common tool used in the Internet. Such tags are useful in content localization and machine translation. Many different standards exist for how to represent language information in machine-readable formats.

Existing language tags all suffer from the same problem in that they represent only the language and not the orthography used in writing said language. Many languages such as Russian, Chinese, and Arabic have multiple orthographies for written content. A few languages,

including Serbian, are digraphic, which means they are natively written in two or more different scripts.

A further complication arises when including the practice of transliteration, or changing orthographies. Most often this is seen when languages written in non-Latin orthographies are rewritten using Latin characters. These orthographies are not mutually intelligible. So to say that two different pieces of text are, "Chinese written in Latin script," is not useful if one is transliterated using the Wade-Giles system while the other is using the Pinyin system.

The problems a complete language tag must address are:

1. Identify the content's language.
2. Identify the language's current orthography.
3. Identify the original orthography used if the content was subject to transliteration.
4. Identify the system used in the transliteration, if the current content differs from the original.

To date no single language tag standard can address all these problems.

1.2. Tags for Identifying Languages

While there are several existing language tag standards only a handful of these standards advance us toward the goal of a complete language tag system. Chief among these is the RFC 5646 document as edited by Phillips and Davis. RFC 5646 satisfies the first two criteria of the proposed complete language tag.

First, RFC 5646 it represents the content's language. This is the very first portion of a BCP 47 language tag. If an alpha-2 code belonging to the ISO 639-1 standard is available then that code is used. If no alpha-2 code is available then the longer alpha-3 code belonging to the ISO 639-3 standard is used.

Second, RFC 5646 represents the languages current orthography. This is an optional portion of the BCP 47 tag. Language orthography representation is handled by the alpha-4 tags defined in the ISO 15924 standard.

What RFC 5646 doesn't address is the last two transliteration-related criteria for a complete language tag.

2. Transliteration Tags

While RFC 5646 does have its shortcomings, it provides for future growth and expansion through extension sub-tags. By using these extension sub-tags we can add a second layer of analysis upon the existing RFC 5646 tags to satisfy our transliteration tag criteria.

As discussed in section 1.1. , the transliteration tag needs to define two additional pieces of data:

1. Original orthography.
2. The transliteration system used.

There will be a new extension tag for each of these pieces of data:

1. The original source orthography will be denoted by the singleton "s" followed by the ISO 15924 for the source script.
2. The transliteration system will be denoted by the singleton "t" followed by a 2-8 character alphanumeric string abbreviation of the transliteration system.

3. Security Considerations

The transliteration tag described in this document includes information about the transliteration system used. Some transliteration standards are proprietary, and the information of their use in a public exchange might constitute a breach of privacy.

4. IANA Considerations

There are no IANA considerations for this document.

5. Conclusions

This document shows how, using the extension mechanisms built into the language tag standard of RFC 5646, a more complete way of representing written languages is achieved to include any transliteration performed upon the text.

6. References

6.1. Normative References

[1] Phillips, A. and Davis M. (Editors), "Tags for Identifying Languages", BCP 47, RFC 5646, September 2009.

[2] International Organization for Standardization, "ISO 639-1:2002. Codes for the representation of names of languages - Part 1: Alpha-2 code", July 2002.

[3] International Organization for Standardization, "ISO 639-3:2007. Codes for the representation of names of languages - Part 3: Alpha-3 code for comprehensive coverage of languages", February 2007.

[4] International Organization for Standardization, "ISO 15924:2004. Information and documentation -- Codes for the representation of names of scripts", January 2004.

6.2. Informative References

[5] Dale, I.R.H., "Digraphia", International Journal of the Sociology of Language 26 (1980) pp. 5-13.

[6] Buckwalter, T., "Buckwalter Arabic Transliteration", Qamus, 2002.

[7] International Organization of Standardization, "ISO 9:1995. Transliteration of Cyrillic characters into Latin characters - Slavic and non-Slavic languages", 1995.

7. Acknowledgments

Thanks to Tim Buckwalter of the University of Maryland for patiently answering questions about his Arabic transliteration system.

This document was prepared using 2-Word-v2.0.template.dot.

Appendix A. Examples of Transliteration Tags (Informative)

ar-Latn-s-Arab-t-buckwalt (Arabic-language text transliterated from the Arabic script into the Latin script via the Buckwalter transliteration system)

ru-Latn-s-Cyrl-t-iso9 (Russian-language text transliterated from the Cyrillic script into the Latin script via the ISO 9 transliteration system)

zh-Latn-s-Hans-t-pinyin (Mandarin Chinese-language text transliterated from the simplified Han script into the Latin script via the Pinyin transliteration system)

Authors' Addresses

Courtney Falk
Infinite Automata

Email: court@infiauto.com