fhiso

CFPS 27
(Call for Papers Submission number 27)

Requirement for Representing Uncertain Characters

Submitted by: Proctor, Tony

Type: Functional requirements

Created: 2013-03-27

Last updated: 2013-04-20

URL: Most recent version: http://fhiso.org/files/cfp/cfps27.pdf
This version: http://fhiso.org/files/cfp/cfps27_v1-1.pdf

Description: Requirement for representing uncertain characters in transcribed data

Keywords: Transcription, Mark-up, UCF

Contents

1. Abstract

This requirement is to be able to represent uncertain characters in transcribed data. This is often referred to as Uncertain Character Format, or UCF.

2. Requirement

There are several schemes already in use for representing UCF. Most are based on the RegEx syntax.

Whether the representation involves special escape characters in the text, or some tagging syntax (as with XML), it must cater for one-or-more characters that are unreadable, and one-or-more characters that may be possibilities from a specified set.

3. Not Covered or Not Required

This requirement only describes support for uncertain characters but there are other types of transcription anomaly that must be supported in order to record evidence correctly. For instance, stuck-out characters (possible using the XHTML <s> element), marginalia, and interlinear/intralinear notes. This must be addressed in a separate proposal.

4. Illustration

From the FreeBMD Web site:

_ (Underscore)	A single uncertain character. It could be anything but is definitely one character. It can be repeated for each uncertain character.
* (Asterisk)	Several adjacent uncertain characters. A single * is used when there are 1 or more adjacent uncertain characters. It is not used immediately before or after a _ or another *. Note: If it is clear there is a space, then * * is used to represent 2 words, neither of which can be read.
[*abc*]	A single character that could be any one of the contained characters and only those characters. There must be at least two characters between the brackets. For example, [79] would mean either a 7 or a 9, whereas [C_] would mean a C or some other character.
{*min*,*max*}	Repeat count - the preceding character occurs somewhere between *min* and *max* times. *max* may be omitted, meaning there is no upper limit. So _{1,} would be equivalent to *, and _{0,1} means that it is unclear if there is any character.
? (Question mark)	Only used where it is unambiguous that there are no characters in the field, e.g a missing Volume. The question mark must be the only character in the field. Note: If it is unclear whether the field is empty or not _{0,1} is used.

Technical note: Although this UCF format has many similarities to regular expressions (e.g. Perl, Unix) it is not identical and in particular there is no escape mechanism.

Another scheme may be found at http://igenie.org, under Transcriptions, although this includes support for more types of transcription anomaly.

5. Use Cases

It is obviously important to be able to record evidence “as is”, without changes. However, it is common during transcriptions to be uncertain about a sequence of characters. While is possible to make an educated guess, the transcription must reflect that since it could result in later failures such as misdirected searches or name mismatches.

6. References

FreeBMD notes on UCF. http://freebmd.rootsweb.com/Format.shtml#UCF.