UTC/L2/05- 277
Date: October 6, 2005

Title:	**Problems with a language-based security approach**
Authors:	Michel Suignard and Mark Davis
Action:	Input to UTR#36 (AI 104-A24)

Problems with a language-based security approach

It is very hard to determine exactly which characters are used by a language. For example, English is commonly thought of as having letters A-Z, but in customary practice many other letters appear as well. For examples, consider proper names such as "Zoë", words from the Oxford English Dictionary such as "coöperate", and many foreign words, proper or not, that are in common use: "René", 'naïve', 'déjà vu', 'résumé', etc… Thus the problem with restricting identifiers by language is the difficulty in defining exactly what that implies. The problem with using language identifier in a security approach derives from the complexity to define what a language is. See the following definitions:

> **Language**: Communication of thoughts and feelings through a system of arbitrary signals, such as voice sounds, gestures, or written symbols. Such a system including its rules for combining its components, such as words. Such a system as used by a nation, people, or other distinct community; often contrasted with dialect.
> *(From American Heritage, Web search)*

> **Language**: The systematic, conventional use of sounds, signs, or written symbols in a human society for communication and self-expression. Within this broad definition, it is possible to distinguish several uses, operating at different levels of abstraction. In particular, linguists distinguish between language viewed as an act of speaking, writing, or signing, in a given situation […], the linguistic system underlying an individual's use of speech, writing, or sign […], and the abstract system underlying the spoken, written, or signed behaviour of a whole community.
> *(David Crystal, An Encyclopedia of Language and Languages)*

> **Language** is a finite system of arbitrary symbols combined according to rules of grammar for the purpose of communication. Individual languages use sounds, gestures, and other symbols to represent objects, concepts, emotions, ideas, and thoughts.
> …
> Making a principled distinction between one language and another is usually impossible. For example, the boundaries between named language groups are in effect arbitrary due to blending between populations (the dialect continuum). For instance, there are dialects of German very similar to Dutch which are not mutually intelligible with other dialects of (what Germans call) German.
> Some like to make parallels with biology, where it is not always possible to make a well-defined distinction between one species and the next. In either case, the ultimate difficulty may stem from the interactions between languages and populations.
> *http://en.wikipedia.org/wiki/Language, September 2005*

In contrast, the definitions of writing systems and scripts are much simpler:

Writing system: A determined collection of characters or signs together with an associated conventional spelling of texts, and the principle therefore.
(extrapolated from Daniels/Bright: The World's Writing Systems)

Script: A collection of symbols used to represent textual information in one or more writing systems
(Unicode 4.1.0 UAX #24)

The simplification originates from the fact that writing systems and scripts only relate to the written form of the language and do not require judgment call concerning language boundary. Therefore security considerations that relates to written form of languages are better served by using the concept of writing system and or script.

> Note: A writing system uses one or more scripts, plus additional symbols such as punctuation. For example, the Japanese writing system uses the scripts Hiragana, Katakana, Kanji (Han ideographs), and sometimes Latin.

Nevertheless, language identifiers are extremely useful in other contexts. They allow cultural tailoring for all sorts of processing such as sorting, line breaking, and text formatting. There are just a poor predicate to qualify a finite set of characters. For example, the Unicode Common Locale Data Repository (CLDR) supplies a set of exemplar characters per language, the characters used to write that language. Originally, there was a single set per language. However, it became clear that a single set per language was far too restrictive, and the structure was revised to provide auxiliary characters, other characters that are in more or less common use in newspapers, product and company names, etc. For example, auxiliary set provided for English is: [áà éè íì óò úù âêîôû æœ äëïöüÿ āēīōū ăĕĭŏŭ åø çñß]. As this set makes clear, (a) the frequency of occurrence of a given character may depend greatly on the domain of discourse, and (b) it is difficult to draw a precise line; instead there is a trailing off of frequency of occurrence.

> Note: As mentioned below, some sorts of language identifiers, called language tags, may contain information beyond the language itself, such as country and scripts, and can help to determine an appropriate script.

As explained in the section 6.1 Writing Systems of the Unicode Standard 4.0, scripts can be classified in various groups: Alphabets, Abjads, Abugidas, Logosyllabaries, Simple or Feature Syllabaries. That classification, in addition to historic evidence, makes reasonably easy to arrange encoded characters into script classes.

The set of characters sharing the same script value determines a script set. The script value can be easily determined by using the information available in the Unicode Standard Annex UAX#24 (Script Names). No such concept exists for languages. It is generally not possible to attach a single language property value to a given character. Similarly, it is not possible to determine the exact repertoire of characters used for the written expression of most common languages. Languages tend to be fluid; words are added or disappear, foreign words using new characters from the original script may be borrowed.

> Note: A well known example is English itself which is commonly considered to only use the Latin letters A to Z, while in fact the large borrowing from the French language has introduced words or expressions such as ‘naïve’, ‘déjà vu’, ‘résumé’, etc…

> Note: There are few cases where script and languages are tightly connected, like Armenian, Lao, etc…However, using scripts in these cases preserves the general model.

Creating ‘safe character sets’ is an important goal in a security context. The benefit is to create a collection of characters that are deemed familiar for a given cultural environment. Incorporating all characters necessary to express the written language associated with the culture is the obvious choice. However, because of the indeterminate set of characters used for a language, it is much more effective to move to the higher level, the script, which can be determinately specified and tested.

Customarily, languages are written in a small number of scripts. This is reflected in the structure of language tags, as defined by RFC 3066 "Tags for the Identification of Languages", which are the industry standard for the identification of languages. Languages that require more than one script are given separate language tags. Examples can be found in http://www.iana.org/assignments/language-tags.

The proposed successor to RFC3066, which has just finished IETF last call, makes this relationship with scripts more explicit, and provides information as to which scripts are implicit for which languages. CLDR also provides a mapping from languages to scripts which is being extended over time to more languages. The following table below provides examples of the association between language tags and scripts.

Language tag	Script(s)	Comment
en	Latin	Content in ‘en’ is presumed to be in Latin script, unless where explicitly marked
az- Cyrl-AZ	Cyrillic	Azeri in Cyrillic script used in Azerbaijan
az-Latn-AZ	Latin	Azeri in Latin script used in Azerbaijan
az	Latin, Cyrillic	Azeri as used generically, can be Latin or Cyrillic
ja or ja-JP	Han, Hiragana, Katakana	Japanese as used in Japan or elsewhere

The strategy of using scripts works extremely well for most of the encoded scripts because users are either familiar with the entirety of the script content, or the outlying characters are not very confusable. There are however few important exceptions, such as the Latin and Han scripts. In those cases, it is recommended to exclude certain technical and historic characters except where there is a clear requirement for them in a language, as is done in UTR #36.

Lastly, text confusability is an inherent attribute of many writing systems. However, if the character collection is restricted to the set familiar to a culture, it is expected by the user, and he or she can therefore weight the accuracy of the written or displayed text. The key is to (normally) restrict identifiers to a single script, thus vastly reducing the problems with confusability.

> Example: In Devanagari, the letter aa: आ can be confused with the sequence consisting of the letter a अ followed by the vowel sign aa ा. But this is a confusability a Hindi speaking user may be familiar as it relates to the structure of the Devanagari script.

In contrast, text confusability that crosses script boundary is completely unexpected by users within a culture, and unless some mitigation is in place, it will create significant security risk.

> Example: The Cyrillic small letter pe п is undistinguishable from the Greek letter pi π (at least with some fonts), and the confusion is likely to be unknown to users in cultural context using either script. Restricting the set to either Greek or Cyrillic will eliminate this issue.

Conclusion: Although a language identifier can uniquely determine a safe set of characters in some rare cases, it is preferable to use the script property as predicate on a given culture to determine the safe character set.