L2/08-407

To: UTC

From: Peter Edberg (Apple), Markus Scherer (Google)

Subject: Proposal to provide new UTC text discussing encoding conversion security issues

An important class of computer security problems is related to conversion of text from one encoding to another. Differences in encoding converters between servers and web browsers can be used to pass potentially dangerous text through a server-side security check to a web browser: The server may convert text to something that is not seen as problematic, while the browser converts text to something different which in fact does cause problems.

One way to avoid these sort of problems is to handle text with malformed or unmappable sequences defensively, if the conversion continues past the problematic sequences. To facilitate this it would be useful to have some official UTC text that describes the relevant encoding conversion security issues and provides a practical set of "best practice" guidelines (with pseudocode) for implementing text encoding converters to make them less vulnerable to such issues.

Why should the Unicode Consortium do this? These days, most text encoding conversion is used to convert text to or from some form of Unicode (or to convert from one form of Unicode to another). Furthermore, UTC members involved in developing and enhancing the ICU encoding converters have gained a significant amount of experience in addressing security issues, and the ICU encoding converters can provide a model implementation of encoding converters that adhere to the suggested guidelines.

Where or how should this text be provided? Mark Davis and Markus Scherer favor adding it to UTR #36 (Unicode Security Considerations) or possibly UTS #22 (Character Mapping Markup Language) [What about UTS #39 (Unicode Security Mechanisms)?]. On the other hand, at the last Editorial Committee meeting (which Mark and Markus were not able to attend), the committee (including Peter Edberg) favored creating a separate, new UTR, for the following reasons:

- These issues relate to encoding conversion, not specifically to Unicode (many of the same issues apply, for example, when converting between ISO-2022-JP and EUC-JP).
- These issues are also not primarily related to the description of the mapping between two encodings (the subject of UTS #22), but rather to the implementation of the encoding conversion process.

We would like to get general approval from the UTC to proceed with drafting such text, and also a decision from the UTC as to where this text should end up.