JTC 1/SC2/WG2/IRG N 1335

Date: 2007-6-6

(Draft) Response to WG2 on regarding IICORE safe character

According to the request from WG2, IRG reviewed IICORE and its safe characters for security with reference to UTR36 and RFC3743.

The goal for encoding ideographs in ISO/IEC 10646 is to be as complete as possible because of the standard's "universal" nature so that different applications can be developed within the framework of ISO/IEC 10646. However, because a functional set for day-to-day use is only a few thousand characters in size, users need to confront a large list of generally unfamiliar characters and there are also a large number of non-unifiable but potentially synonymous variants. The introduction of IICORE solves this issue.

'The idea of creating 'safe character set' is to create "a collection of characters that are deemed familiar for a given cultural environment" (UTR36) and in that sense, IICORE is designed to be such a commonly used collection of characters that gives sufficient coverage for everyday needs and are familiar to most users. However because of its 'international' nature, users of IICORE must be aware that some of IICORE characters common to a given country/region may be unfamiliar for people in other countries/regions.

As to 'variants' considered as a potential security risk for the use of IDN in RFC3743, IICORE does contain 'variant' form of CJK Ideographs.
UTR36, discussed about many cases of spoofing by visual confusion of non-ideographic character shapes. Those situations will be possible on ideographs as well as non-ideographic characters. However, this problem might be more serious with ideographs than that of non-ideographic characters because there are cases that two or more ideographs having the same meaning and the choice is sometimes dependent on preference or other considerations.

For example, even Japanese users may be confused on some Japanese ideographs by different reasons:
(1) Possible confusion by shapes. Characters used in the specific region (e.g. in Japan) therefore be confused

(凜凜), (靭靱), (高髙), (俱俱), (胄胃胄), (采采), (日曰), (壷壹壺壼), etc.

(2) Easily recognized shape differences but users may be spoofed by their meanings

(2-1) Glyph variants of one character

(学學), (島嶋嶌), (斉齊), (鴬鶯), etc.

(2-2) Different characters but used as a same word

(新撰組 vs 新選組), (寿司 vs 鮨 vs 鮓), (皐月 vs 五月), (苺 vs 莓), etc.

(2-3) Different characters used in the different word but same readings. Some of them have similar meanings also.

(科学 vs 化学), (共同 vs 協同 vs 協働), (回答 vs 解答), (保険 vs 保健), etc.

(2-4) Different expressions of names for people, place or other specific individuals

(太田 vs 大田), (渡部 vs 渡辺 vs 渡邉 vs 渡邊), (斉藤 vs 斎藤 vs 齊藤 vs 齋藤 vs 齎藤), etc.

However it is difficult to make the exhaustive list that could eliminate the opportunity to exploit 'variants' to divert users and even if it were possible, excluding all such characters will make the resulting set totally unuseable. So IRG would need more clarification on the request from WG2, preferably with a reference to a formal request from IETF community to understand the background and the requirements.

(END OF DOCUMENT)