

An OOXML extension proposal for checking and normalizing characters

Japanese member body for SC34

9 February 2009

Summary: An amendment to ISO/IEC 29500 for introducing two optional attributes for checking and normalizing characters is proposed. This amendment has political advantages as well as technical advantages.

1. Outline of the amendment

This amendment should introduce two optional attributes wherever document designers would like to impose constraints on permissible characters.

The first attribute selects one of the four Unicode normalizations provided by the Unicode standard. Implementations may use this attribute for normalizing characters typed by end users.

The second attribute specifies a URI as a reference to a character repertoire description in DSDL Part 7 (Character Repertoire Description Language). Implementations may use this attribute for checking characters typed by end users and provide diagnostic message.

The amendment is expected to be very short and can be standardized in 2009, if everything goes very well.

2. Background

Web documents or office documents are sometimes used as "forms". Such "forms" are meant to collect data such as strings, integers, and date from end users.

However, end users sometimes type inappropriate characters. These are classified into three groups as below:

- Foreign characters
- Too difficult kanji characters (e.g., those character which are not taught in elementary or junior high schools)
- Half-width characters, when full-width ones are preferred (and vice versa)

Note: I do not know requirements of non-Japanese users very well, but I am sure that similar requirements exist.

Web forms do not provide any mechanisms for normalization and checking. At present, AJAX or server-side programming is used for normalization and checking.

Note: Microsoft Office provides some support for normalization and checking but the support is ad-

hoc and slightly inconsistent.

3. Advantages

- Useful for those who are annoyed by too many ideographic characters
- Useful for those who do not want to receive foreign characters
- Useful for those who do not want to handle variations (e.g., half-width and full-width)

4. Impacts on OOXML implementations

Conformant implementations have to read and write these two optional attributes.

Implementations can be conformant without using these attributes for normalization and checking as long as this omission is clearly stated. It would be nice if future implementations provide normalization and checking, though.

5. Remaining problems

The idea of being able to specify a Unicode normalization form might be problematic, since Unicode normalization forms (even conservative ones) cause surprising changes.

References

ISO/IEC FCD 19757-7, Information technology -- Document Schema Definition Languages (DSDL) -- Part 7: Character Repertoire Description Language (CRDL) , available at <http://www.itscj.ipsj.or.jp/sc34/open/0978c.htm>

Martin J. Du''rst, A Notation for Character Collections for the WWW, 2000. Available at <http://www.w3.org/TR/charcol/>