

Forward compatibility of `text_encoding` with additional encoding registries

Document No. **P2498 R0**

Reply To Peter Brett pbrett@cadence.com

Date 2021-12-01

Audience: SG16, LEWG

Introduction

As currently proposed [1], `text_encoding` refers only to the Internet Assigned Numbers Authority (IANA) Character Sets database [2]. This registry is known to be incomplete and, in some respects, does not provide a perfect match to the requirements of C++ [2]. It is possible that future enhancements to `text_encoding` may wish to refer to additional/alternative registries.

Alter the names of `text_encoding` facilities that directly map IANA database data to explicitly reference `iana`.

Design

Do not rename `aliases()`

The `text_encoding::aliases()` member function currently returns a range of alternative names for a particular `text_encoding`. Although this range is required to include the aliases registered with IANA, it may also include additional, implementation-defined aliases.

This means that there is no need to rename this to `iana_aliases()`; the contract is already sufficiently wide to accommodate aliases from other registries.

Normative guidance for future compatibility

Currently, the exposition-only member variables of `text_encoding` contain only a `text_encoding::id` without scope for disambiguation of IDs or the capacity for representing non-IANA IDs, if required in the future.

This is adequate for now. However, we should provide normative guidance that implementors should consider the possibility of additional/alternative text encoding registries being used in the future and make accommodations in the layout of `text_encoding`. This will prevent any such evolution from being blocked due to ABI concerns.

Proposed wording

Editing notes

All wording is relative to P1885R8 [1].

[`text.encoding`]

Throughout this section:

1. Rename `text_encoding::id` to `text_encoding::iana_id`, and update all references to this type accordingly.
2. Rename `text_encoding::mib()` to `text_encoding::iana_mib()`, and update all references to this member function accordingly.

Add an additional item of 'recommended practice':

- When defining the member variables and layout of `text_encoding`, implementations should consider the possibility that future revisions of this standard may reference additional or alternative text encoding registries.

Acknowledgements

Thank you to Tom Honermann for suggesting making the IANA link explicit in identifiers, and to Jens Maurer for highlighting the downsides of the way that IANA registry text encodings are specified.

References

- [1] C. Jabot and P. Brett, "P1885R8 Naming Text Encodings to Demystify Them," 11 Nov 2021. [Online]. Available: <http://www.open-std.org/jtc1/sc22/wg21/docs/papers/2021/p1885r8.pdf>.
- [2] N. Freed and M. Dürst, "Character Sets," Internet Assigned Numbers Authority, 2021.
- [3] J. Maurer, "P2491R0 Text encodings follow-up," 15 Nov 2021. [Online]. Available: <http://www.open-std.org/jtc1/sc22/wg21/docs/papers/2021/p2491r0.html>.