L2/02-165

Ken Whistler

A Unicode Conformance Model

April 29, 2002

I. Introduction

The Unicode Standard is a very large and complex standard. Because of this, and because of the nature of the material in the standard, it is often rather difficult to determine, in any particular case, just exactly what conformance to the Unicode Standard means. People have raised issues regarding this difficulty, both from a theoretical point of view, and from the practical standpoint of determining what products "support" Unicode, and what such claims of support actually mean.

In an effort to fill this gap, this Unicode Conformance Model has been developed. It aims at explaining what conformance means for the Unicode Standard. It defines terminology regarding the topic of conformance, specifies different areas and levels of conformance, and describes what it means to make a claim of conformance or "support" of the standard.

This model is not, in itself, a framework for compliance testing, although it could be used to develop such a framework, should that prove desireable.

II. Terminology

This section gives a basic introduction to the terminology that will be discussed in more detail in sections below.

Conformance

In the context of formal standard, conformance refers to a set of rules or criteria whereby a relevant entity (element of information interchange, device, application, piece of hardware, etc., etc.) can be determined to either be meeting or not meeting the specification in the standard.

In general, a formal standard will have a conformance clause or clauses, which will be stated in terms of conditionals ("X is in conformance with Y specification of this standard if Z") or modals ("An X that conforms with Y specification of this standard SHALL Z"). The modal verbs that standards language generally associates with such statements may themselves be carefully defined, and typically involve specialized usage of "SHALL" and "MUST", to avoid any ambiguities of interpretation.

If a standard is complex, the conformance clause or clauses themselves may also be complex. But on occasion, a conformance clause may simply be stated along the lines of "X is in conformance with this standard if it follows the specification in section W", where section W may consist of hundreds of pages and constitute most of the rest of the standard.

Normative/Informative

Formal standards often distinguish between normative and informative content. This distinction may be highly conventionalized, or even be

subject to rules specified in other standards, as for ISO standards, or the distinction may be much less formally maintained.

Normative content of a standard is that which is required for all of the conformance requirements to be meaningful. Typically a standard will have normative definitions for terms used in the rest of the specification, will have normative references to other standards or sources whose content is referred to indirectly, and will have normative clauses, specifications, or sections, which actually define the content of the standard itself -- that which the conformance clauses apply to.

Informative content of a standard is that material which has been added for clarification, but which, in the judgement of the standard's maintainers, could in principle be omitted without materially affecting the specification which the conformance clauses refer to.

If a standard is changed over time, the status of some particular content could change from informative to normative, or vice versa, depending on whether it was newly required for conformance or became unrequired for conformance.

Compliance

The term compliance is often used synonymously with the term conformance. However, it is possible to draw a meaningful distinction.

In the context of the Unicode Conformance Model, compliance is used to mean an external determination that a particular relevant entity actually does meet one or more conditions of the conformance clauses of the standard. Thus while conformance is merely a logical statement of requirements, compliance is a state met when entity X is actually determined, under some specified set of circumstances, to meet the logical statement of requirements. As such, conformance clauses exist in the standard on their own, but compliance determination implies the existence of compliance tests, applied to entities to make such determinations.

A conformance claim can simply be stated. It is an assertion that entity X meets a requirement of the standard.

A compliance claim, on the other hand, is the result of the specific application of a test designed to determine the validity of a conformance claim. Such tests are called compliance tests.

Conformance Tests and Compliance Tests

A standard may include tests or "benchmarks" as part of the text of the standard, or as external documents associated with the standard. Once again, while there is some overlap in general usage of the terms "conformance test" and "compliance test", in the Unicode Conformance Model a systematic distinction is drawn between the two.

A conformance test for the Unicode Standard is a list of data certified by the UTC to be "correct" in regard to some particular requirement for conformance to the standard. In some instances, as for example, the implementation of the bidirectional algorithm, producing a definitive list of correct results is difficult or impossible, and in such cases, a conformance test may itself consist

of an implemented algorithm certified by the UTC to produce correct results for any pertinent input data. Conformance tests for the Unicode Standard are essentially benchmarks that someone can use to determine if their algorithm, API, etc., claiming to conform to some requirement of the standard, does in fact match the data that the UTC claims defines such conformance.

A compliance test for the Unicode Standard, on the other hand, is a test, usually designed and implemented by a third party not associated with the Unicode Standard or the UTC, intended to test a product which claims conformance to one or more aspects of the Unicode Standard, for actual compliance to the standard. Thus a compliance test is a test *of a product*. A compliance test, may, of course, make use of one or more of the Unicode conformance tests in order to determine the results of its test of compliance.

Support

The term support, in the context of the Unicode Conformance Model, refers to a more generalized claim of intent to conform to one or another requirement of the standard. A claim of Unicode support may in fact be difficult to verify, since it can be and often is vague in detail. But in principle, at least, it indicates that the developer or user of an entity intends conformance.

More specifically, support often refers to a claim of particular repertoire coverage. For example, an application may claim support for Unicode Greek. That should be interpreted as meaning that Unicode Greek characters will be handled conformantly with the standard, and furthermore that all other relevant aspects of processing of those characters which that particular application is concerned with, will also be done in such a way as not to violate conformance clauses of the standard.

Stability and Invariance

Some formal standards are developed once and then are essentially frozen and stable forever. For such standards, stability of content and the corresponding stability of conformance claims is not an issue.

For a large, complex standard aimed at the universal encoding of characters, such as the Unicode Standard, such stability is not possible. The standard is necessarily evolving and expanding over time, to extend its coverage of all the writing systems of the world. And as experience in its implementation accumulates, further aspects of character processing also accrue to the formal content of the standard. This fundamentally dynamic quality of the Unicode Standard complicates issues of conformance, since the content to which conformance requirements pertain continually expands, both horizontally to more characters and scripts, and vertically to more aspects of character processing.

Invariance refers to those aspects of the content of the Unicode Standard that have been determined to be unchangeable, even as the standard continues its dynamic development. A fairly trivial example can be seen in the guarantee of the stability of the formal Unicode character names. While in principle such names *could* be changed, and in very early versions of the standard were changed (between Version 1.0 and Version 1.1, for example), the UTC has determined that such changes are too disruptive and have too little benefit to be tolerated. Accordingly, the stability of character names

has been promoted to the status of an invariant in the standard.

Conformance claims need to be distinguished in terms of their relationship to invariants and non-invariants in the standard, because of their different risk levels for stability.

Versions

The Unicode Standard is regularly versioned, as new characters are added. A formal system of versioning is in place, involving major, minor, and update versions, all with carefully controlled rules for the type of documentation required, handling of the associated data files, and allowable types of change between versions. For more information about the details of Unicode versioning see [link].

Conformance claims clearly must be specific to versions of the Unicode Standard, but the level of specificity needed for a claim may vary according to the nature of the particular conformance claim being made.

[The following content is just sketched out in outline form.]

III. Structure of Unicode Conformance

This section will serve as a guide to unravelling the particular way that the Unicode Standard expresses conformance requirements, both in terms of where they are located and how they are expressed.

It also explores the peculiar aspects of conformance related to the synchronized status of the Unicode Standard and the independent but closely aligned International Standard ISO/IEC 10646, which has its own conformance clauses expressed using ISO conventions.

Definitions
Conformance Clauses
Unicode Standard Annexes
Identification of Normative Content
Relation to 10646 Conformance

IV. Areas of Conformance

[Borrowing from Asmus' suggestions:]

1) representation

Representation would cover being able to express and transmit Unicode data, it would be a requirement applicable to certain protocols (e.g. XML), but might apply to the storage aspects of databases as well.

This would also apply to correct use of encoding forms and encoding schemes.

2) transcoding

Transcoding between Unicode and legacy (all other) character encodings.

3) string processing

String processing would
generically cover all operations on Unicode texts that can be carried out without considering layout and specifically not considering fonts.

4) text layout, including display and selection

Layout would comprise all operations that go from backing store to displayed text (and the reverse, for selection). These operations are dependent on font data.

5) fonts

Primarily refers to CMAP's for fonts, and to claims of "coverage" of Unicode repertoire by fonts.

6) input

Issues of coverage of Unicode repertoire, conversion of input to Unicode character values for storage, and consistency with the text models required for particular scripts and text layout. The entities here are mostly IME's and keyboards (drivers).

V. Levels of Conformance

This section will provide both a typology for levels of conformance (i.e., an alternative to the notion that all aspects of Unicode conformance are either/or issues), and specific lists of levels of conformance and support where they can be pulled out of the standard. For example, the standard explicitly talks about levels of surrogate support -- that should be abstracted, along with others, to provide the basis for determining how to make various claims of conformance.

Repertoire coverage
Full conformance (in an area)
Partial conformance (in an area)
  - levels of support defined
Best practices

VI. Interoperability

Matching areas and levels of conformance between implementations and components.

Repertoire matching.

Downrev and uprev compatibility issues.