JTC1/SC2/WG2 N2733

L2/04-126

ISO/IEC JTC 1 N7409

2004-03-29

Replaces:

ISO/IEC JTC 1
Information Technology

Document Type: Other document (Defined)

Document Title: Comments Received on JTC 1 N 7335, Responses on Sophia Resolution #39-Development of a Solution for the Unambiguous Identifiation and Interworking of Codes Representing Countries, Langauage and Currenciies (prepared on behalf of SC 32/WG 1)

Document Source: JTC 1 Secretariat

Project Number:

Document Status: This document is circulated to JTC 1 National Bodies for information. This document is forwarded to the National Body of Canada for review and consideration. If the National Body of Canada wishes to respond and recommend further action, they are asked to submit this to the JTC 1 Secretariat.

Action ID: FYI

Due Date:

Distribution:

Medium:

Disk Serial No:

No. of Pages: 9

Secretariat, ISO/IEC JTC 1, American National Standards Institute, 25 West 43rd Street, New York, NY 10036; Telephone: 1 212 642 4932; Facsimile: 1 212 840 2298; Email: lrajchel@ansi.org

Comments Received on JTC 1 N 7335, Responses to JTC 1 Sophia Resolution #39- Development of a Solution for the Unambiguous Identification and Interworking of Codes Representing Countries, Languages and Currencies (prepared on behalf of SC 32/WG 1)

Italy:

"Italy recognizes the need to develop a Solution for the Unambiguous Identification and Interworking of Codes Representing Countries, Languages and Currencies.

The solution to be adopted needs to have a wide consensus, therefore, Italy recommends the distribution of the document for comments to all the JTC 1/SCs and related ISO/TCs mentioned in the document."

United States:

US National Body Comments in Response to JTC 1 N 7335 - Response to JTC 1 Sophia Antipolis Resolution #39: Development of a Solution for the Unambiguous Identification and Interworking of Codes Representing Countries, Lanaguages, and Currencies (prepared on behalf of SC 32/WG 1)

1. COMMENTS ON THE TRANSMITTAL

In the Foreword of N7335, it states:

> This contribution on behalf of SC32/WG1 responds to JTC1 Sophia Resolution #39 ...

We don't believe this document has been authorized by SC32/WG1 because:

- There are no resolutions authorizing this document to be transmitted to JTC1 in the SC32/WG1 resolutions of the 2003-01 and 2003-06 meetings.
- There are no resolutions authorizing this document to be transmitted to JTC1 in the SC32 resolutions of the 2003-01 SC32 Plenary.
- The document was not submitted by the SC32/WG1 Convener, SC32 Chair, or SC32 Secretariat.

More properly, the document should be regarded as a proposal from Canada NB without regard to SC32/WG1.

2. GENERAL COMMENTS

The following are general comments.

2.1 Document should be referred back to SC32

Considering the potential conflicts in N7335 with existing standards work in SC32 (including new conflicting terminology), we suggest that the discussion be referred back to SC32 and its WGs to present a consistent SC32 perspective.

2.2 Not JTC1's role to interpret TC37, TC46, and TC68 standards

Considering that the document concerns the interpretation of standards developed and coordinated by TC37, TC46, and TC68, we suggest that these TCs provide the official interpretation of their standards. In other words, if there are concerns about "semantic interoperability" (and we believe there are no interoperability concerns), then TC37, TC46, and TC68 should provide the appropriate interpretation. We recommend that the document be re-scoped to either (1) ask a question of interpretation, or (2) express Canada's implementation concerns.

2.3 Significant implementation problems with recommendations

Considering that the document proposes changes in the structure and syntax of languages codes, etc. that would greatly affect existing implementations, *the N7335 document provides no implementation rationale for a change the would break millions of existing implementations*.

2.4 Lack of implementation experience with ISO/IEC 15944

It appears that ISO/IEC 15944 has been used as a basis for identifying "e-business" needs that exceed the requirements for IT interoperability, but a Google search of "ISO 15944 conform OR conformity OR conforming OR implementation" only produced hits that refer to SC32/WG1 documents and copies of the SC32/WG1 documents -- we found no implementations of 15944 which would help us understand particular needs identified in N7335. It would be helpful if implementations of 15944 (and their issues) were referenced in N7335.

3. SPECIFIC COMMENTS

The document JTC1/N7335 is a 44-page presentation that may be best summarized by the following:

3.1 Concepts and terms that lack harmony

Both N7335 and 15944 introduce new concepts and terms that lack harmony with existing IT terminology, e.g., terms like "semantic interoperability", "commitment exchange", "coded domain", and "pivot code". We are concerned that these kind of concepts and terms blur the existing JTC1 standards work.

In the context of N7335 (and 15944), "semantic interoperability" is no different than "IT interoperability". Quoting from ISO/IEC 2382-01, Information Technology Vocabulary, Fundamental Terms:

> interoperability: The capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units.

N7335 implies that "semantic interoperability" implies agreement upon meaning, but so does 2382-01 "interoperability": "requires ... little or no knowledge of the unique characteristics [but requires knowledge (i.e., agreement upon "meaning") of the general characteristics".

In N7335, "commitment exchange" does not concern the exchange of commitments, per se, but agreement upon a "technical specification by the partners of some data exchange. The term "technical specification" is defined in ISO/IEC Guide 2:

> technical specification: document that prescribes technical requirements to be fulfilled by a product, process or service
>
> NOTE 1: A technical specification should indicate, whenever appropriate, the procedure(s) by means of which it may be determined whether the requirements given are fulfilled.
>
> NOTE 2: A technical specification may be a standard, a part of a standard or independent of a standard.

The terms "conformity" and "supplier's declaration" are also terms of ISO/IEC Guide 2. It is unclear why the concept and the term "commitment exchange" is created when the Guide 2 terms *technical specification* and a *supplier's declaration* of *conformity* state the same notion and an used the broadly agreed terminology of ISO/IEC Guide 2. In short, a supplier's declaration of conformity to a technical specification is very close to the intent of the term "commitment exchange".

In N7335, the term "coded domain" is used. This term conflicts with the well-established term of "value domain" in ISO/IEC 11179 (an SC32 standard) that is widely implemented internationally.

In N7335, the term "pivot" is used suggesting that, in the case of codes sets that use multiple representations, one of the codes sets is preferable (i.e., designated the "pivot code"). From an ISO/IEC 11179, there may be metadata that includes a unique registration identifier that may be everlasting, but the registration identifier is not the code. As one can see from this history of 3166-1 and a careful reading of the 3166-1 standard, there can be no everlasting "pivot" or "pivot code"; and 3166-1 provides no special distinction of any of the 7 columns: 3-numeric, 3-alpha, 2-alpha, short English,

long English, short French, long French. Thus, the notion of "pivot" is an artificial and inaccurate construction of N7335. Again, this concept of "pivot" conflicts with the ISO/IEC 11179 notion that the registration identifier is separate from the code itself.

These four concepts and terms are not the only areas of conflict, but it is important for JTC1 to be careful to avoid using terminology that is overlapping and conflicting with existing standards work that is widely adopted internationally.

3.2 Recommendation that JTC1 adopt proposed solution on codes and delimiters

In N7335, the following recommendations are made:

> Requested Actions of JTC1:
>
> one pertaining to a proposed solution for the unambiguous identification and interworking of codes representing "countries", "languages", and "currencies" in terms of their ordering, identification, referencing, and use of specific code sets of those standards (in order to support semantic interoperability requirements of e-business); and
>
> one pertaining to the need to develop a "common" set of delimiters for locale identifiers and
> language identifiers.

The ISO 639-*, ISO 3166-1, and ISO 4217 standards already provide "unambiguous identification and interworking of codes". For example, since ISO 3166-1 provides a one-to-one mapping among 3-numeric, 2-alpha, and 3-alpha codes, there no differences among any of these choices, e.g., the choice of 3-numeric vs. 2-alpha is merely a choice of implementation convenience, not a choice of different "semantics".

N7335 implies that the numeric codes have longer longevity but, this is simply not true. When countries split or merge, the code set changes for all sets. Additionally, it should be noted that the UPU (Universal Postal Union) has chosen the 2-alpha country codes (over, say, 3-numeric) because of there usability.

N7335, states the following as a justification of a "semantic interoperability" concern:

> However, the development of the 3-alpha code sets of ISO 639-2 was done in response to real world requirements (and inadequacies of ISO 639-1 which currently contains only 42% of the languages for which codes are now available via ISO 639-2).

There has been a long history in ISO of overlapping and more expansive code sets. For example, there is ISO/IEC 646, ASCII, ISO/IEC 8859-1, and ISO/IEC 10646-1. The ISO/IEC 646 (and its national variant ASCII) only represent a small subset of codes within ISO/IEC 10646-1. However, there are no "semantic interoperability" concerns:

the meaning of the character code "LATIN CAPITAL LETTER A" of ISO/IEC 646 is the same as its corresponding code in ISO/IEC 10646. Likewise, the meaning of "en" in ISO 639-1 is the same as "eng" in 639-2/T.

Thus, more expansive code sets do not imply a lack of semantic interoperability, but imply a broader coverage of the problem space. Industry is still likely to continue ongoing use of both narrow (e.g., ASCII) and broad (ISO/IEC 10646-1) code sets.

Regarding the shortcomings of ISO 639, Enthologue has prepared a document titled "An Analysis of ISO 639: Preparing the way for advancements in language identification standards" by Peter Constable and Gary Simons that is available at:

http://www.ethnologue.com/iso639/An_analysis_of_ISO_639.pdf

The following excerpt summarizes the issues:

> 4. Possible implications for new or enhanced standards
>
> Momentum is building for an effort to enhance the ISO 639 family of standards toward comprehensive coverage of the world's languages. Specifically, there has been a proposal to develop a "Part 3" of the standard that would offer a set of code elements (possibly using four letters) that encompass all the languages of the world. The lessons learned from the experience of mapping the existing ISO 639 code elements onto the languages identified in the Ethnologue provide some insights that could inform that effort. The key insights are:
>
> • An operational definition for every category of language code should be part of the standard.
> • The type of category denoted by each code element should be clearly identified.
> • The denotation of every language-level code element needs to be defined with much more information than just a name.
> • The denotation of every collective code element should be defined by enumerating the more specific collective code elements and the language-level code elements it encompasses.
>
> *An operational definition for every category of language code element should be part of the standard.* In the current code elements of ISO 639-2, we can distinguish at least three categories of language: ancient languages, modern languages (including recently extinct languages) and constructed (or artificial) languages. Among the collective language codes there are similarly at least four categories: collections corresponding to a single genetic subgroup, collections associated with a particular region, collections based on a shared name, and other collections. There are, of course, other ways to categorize languages and collections and a future standard could end up with more or fewer categories. The main point is that such a standard needs to name the categories for which it provides codes and then provide an operational definition for each category. Such

standardised definitions are needed to ensure that all the codes in the new standard consistently meet criteria for what deserves to have a code and for the category of thing a code should represent. In Constable and Simons 2000, we have discussed the need for clarification of categories and operational definitions in more detail.

The type of category denoted by each language code element should be clearly identified. This could be done by means of naming conventions or other signals in the documentation about the code elements, but our proposal is that this be done in the code elements themselves. We recommend that in a four-character code, ranges of letters for the first character could be reserved to denote the category of the code. For instance, the category of ancient languages might be coded using initial characters of a–c.27 The first letter would thus be analogous to a namespace designator.28 The different namespaces could even be assigned to different registration authorities for the purpose of managing the last three letters of the code. This could be advantageous in that different institutions might prove best qualified to manage the different categories of codes. In Constable and Simons 2000, we have discussed the notion of alternate namespaces for language codes in more detail.

The denotation of every language-level code element needs to be defined with much more information than just a name. Our experience in mapping the current ISO 639 code elements to Ethnologue languages demonstrates that a simple name is woefully inadequate for defining the denotation of a language identifier. A basic description such as that given in an Ethnologue entry is needed. This implies that the kind of standards document used to define ISO 639 Parts 1 and 2 would not serve to define a comprehensive Part 3. It would not be feasible to produce and maintain a printed volume including the codes and descriptions for 6000+ languages. Rather, a web site operated by the designated registration authority would list the current codes and definitions; the standard would give the operational definitions of categories and describe all other practices and constraints that the registration authority would be required to follow.

The denotation of every collective code element should be defined by enumerating the more specific collective code elements and the language-level code elements it encompasses. If collective code elements represent, by definition, collections of languages, then the web site that documents the code elements should make their denotation explicit by enumerating the code elements found elsewhere in the standard that are part of the collection. These could be language-level code elements, or they could be code elements for more specific collections that are included within the scope of the collection.

Thus, it appears that TC37 is working with industry to incorporate the latest research and recommendations. JTC1 should let TC37 provide this expertise and the new or revised standards. If JTC1 has concerns about interpretation, defects, or implementation issues,

then concerns should simply be referred via liaison requests. JTC1 should not "second-guess" the experts of TC37 (or TC46 or TC68).

3.3 Several misunderstandings about the nature of language codes

There are several misunderstandings in N7335 about language codes and their use.

N7335 suggests that ISO/IEC 10646 contributes to these so-called problems of "semantic interoperability":

> The number of ISO 639-2 3-alpha codes continues to grow. A major contributing factor here is the introduction and increasingly widespread use of ISO/IEC 10646-1:2000

The 10646-1 standard has little to do with language codes because 10646-1 specifies scripts (and symbols), not languages. Of course a script is used in written language, but the same script may be used for multiple languages (e.g., latin script used for English, French, Spanish) and a single language may use multiple scripts (e.g., Japanese may be presented in kanji, kana, and latin (romaji) scripts). So it is difficult to see how 10646-1 contributes to any of the problems cited in N7335.

N7335 makes the claim:

> ISO/IEC 9945-1 (POSIX) and IETF RFC 3066 contain examples of combinations of "language code" and "country code" as well as "administrative subdivisions" within a country, i.e., states, provinces, etc. {See further 2.1 above}. These examples are no longer representative of current user requirements in a global context let alone from an e-business needs perspective.

N7335 confuses the issue by assuming that 9945-1 and 3066 are specifying the same thing. As pointed out in JTC1/N6866 (an SC36 contribution inquiring about this issue):

> In the [SC22/]WG20 E-mail discussion, it appears that the subtle distinction is that the locale identifier in 9945-1 describes the "user's" environment, while the locale identifier in 3066 describes the language. In the E-mail discussion (message number SC22WG20.4131), it appears that one can write a locale identifier such as "en-GB_US", which might mean "the language is British English, but the user is operating within the US environment".

In appears that N7335 believes that all uses of country codes serve the same purpose and the problem, simply, it to agree upon a common delimiter. This is wrong, as illustrated in the excerpt above. Given the position-based syntax in "xx-YY_ZZ", "xx" is the language, "YY" specializes that language according to a particular national or local variant (e.g., British English, US English, Australian English, etc.), "ZZ" identifies that national or local user environment. So in the example "en-GB_US", the language, date, and time

would be presented in a British English format, but the default set of time zones would be US-based (e.g., Eastern Time, Central Time, etc.).

Furthermore, N7335 suggests that the ordering be changed from language+country to country+language. This misunderstands the nature of the requirements. In IETF RFC 3066, the requirement is to specify a language code. The 3066 specification allows for further specialization by including the country code, e.g., "en", "en-GB", and "en-US" are all valid values with different meanings (the first, "en", provides no country-specific specialization). It is incorrect to think of the country code in the primary (first) position since it is only a secondary attribute to the language code, i.e., the purpose of 3066 is to identify languages, not countries. *This proposed change in ordering would break millions of existing applications*.

Likewise, 3166-1 country codes can be used in a variety of applications, such as for identifying national boundaries based on land, sea, air, citizenship. A person at a given point on Earth may be in different countries depending upon whether the question is asked concerning land, sea, and air, e.g., a person may be over Mexican land but still in US airspace. Asking the question "Where are all the citizens of country Q?" would produce a geography that is different than the political boundaries of familiar international maps. There is no "semantic" problem here, it is merely different applications of the use of the 3166-1 codes. The 3166-1 standard makes it clear that it makes not claim on status (e.g., some country codes might not be countries) or boundary of a country code; and 3166-1 makes it clear it has a broad range of applicability.

4. CONCLUSIONS

We recommend the following:

1. JTC1 should not accept any of the recommendations in N7335.
2. JTC1 should request that SC32 to provide harmonized terminology across its standards, based upon existing JTC1 terminology and existing implementation experience.
3. JTC1 should not interpret TC37, TC46, and TC68 standards, but JTC1 should facilitate the appropriate liaison communications so that requests for interpretation, defect reports, and implementation experience may be communicated.