

**in070278**

February 14, 2007

**Subject: Stephen Anderson Contribution on Behalf of the Linguistic Society of America (LSA) - Comment on ECMA 376 OOXML Proposed Standard**

Sara Desautels  
ANSI Program Manager  
ISO/IEC JTC1/SC 34 Information technology—Document description and processing languages

Dear Ms. Desautels,

I am writing you as President of the Linguistic Society of America (LSA), on behalf of the Executive Committee of the Society and its members. The LSA understands that the ECMA 376 Office Open XML (OOXML) standard is being proposed for adoption as an ISO/IEC standard by JTC1/SC34. The LSA has reviewed the OOXML standard in relation to use of language identifiers and requests that any ISO/IEC standard for OOXML incorporate revisions to consistently specify the use of the recommendations in IETF BCP 47 for language tags in OOXML documents. A detailed explanation follows.

The LSA has reviewed the ECMA 376 Office Open XML standard in relation to internationalization and, specifically, metadata elements for language identification. As observed in §4.2 of SC34/N0809, WordprocessingML and DrawingML use language identifiers for each paragraph and run. The specifications for these in §4:2.18.51 and §4:5.1.12.72, however, are vague, unnecessarily inconsistent, and underrepresent the world's languages. To be specific:

- WordprocessingML uses the simple type, ST\_Lang, defined in §4:2.18.51, while DrawingML uses a different simple type, ST\_TextLanguageID, defined in §4:5.1.12.72. There is no reason why these should be different: the same conventions should be used in every place for all of the OOXML document formats where language identification metadata elements are used.
- The specification of ST\_Lang in §4:2.18.51 requires values to consist of an “ISO 639-1 letter code plus a dash plus an ISO 3166-1 alpha-2 letter code”. This roughly corresponds to IETF specification RFC 1766 (superseded by RFC 4646), though it is more restrictive. This specification has undesirable qualities:
  - It requires an ISO 3166-1 country identifier even if one is unnecessary or even inappropriate.
  - It does not permit important distinctions related to written form that are essential

for linguistic processing, such as script or orthography conventions (except where these coincidentally correlate with country distinctions, which very frequently is not the case).

- It allows reference to only the small portion of the world's languages that are supported in ISO 639-1 rather than the more comprehensive set supported in ISO 639-3. Use is, therefore, limited to roughly 200 out of some 7000 known languages.

In the specification of ST\_TextLanguageID in §4:5.1.12.72, the type is said to be a restriction of the XML Schema string data type, yet no restriction of any sort is, in fact, described. The ST\_TextLanguageID allows any string, and no convention for the form and semantics of these strings is specified.

- These issues can best be remedied by having a single data type that is used throughout the document formats wherever language identification metadata elements are used and by having the specification for that data type normatively reference the IETF specification BCP 47. Even if a change is not made to unify data types, at least the specification for the two existing data types in §4:2.18.51 and §4:5.1.12.72 should be revised to reference IETF BCP 47.
- The IETF BCP 47 specification is consistent with the examples given in both §4:2.18.51 and §4:5.1.12.72. It would relax the specification in §4:2.18.51, and it would constrain the specification in §4:5.1.12.72, as well as giving formal and semantic definition to the metadata elements used in DrawingML; these are both desirable changes. BCP 47 has specifically been designed to account for all languages and all relevant language distinctions and is also being maintained in accordance with this design.

In summary, then, if ECMA 376 is considered as a proposed ISO/IEC standard, then the LSA requests that it be revised to unambiguously specify the use of the recommendations in IETF BCP 47 for language tags per the specific changes described above.

Sincerely yours,



Stephen R. Anderson  
Dorothy R. Diebold Professor of Linguistics  
Yale University  
President (2007), Linguistic Society of America