Need for Revision of XML 1.0 to account for Localization Issues with Particular Emphasis on Ethiopic Script and Writing System

January 2002

This Version: http://www.digitaladdis.com/sk/ETXMLFinal.pdf
Latest Version: http://www.digitaladdis.com/sk/ETXMLFinal.pdf

Authors/Editors:

Samuel Kinde Kassegne (UC San Diego, Engg Ext.) <bikila_97@yahoo.com>
Bibi Ephraim (Hewlett Packard) <thisbibi@yahoo.com>

Abstract

This document contains XML example codes, surveys and recommendations that support the need for Ethiopic script-based XML element type and attribute names in Amharic, Afaan Oromo, Tigrigna, Agewgna, Guragina, Hadiya, Harari, Sidama, Kembatigna and other Ethiopian languages that use Ethiopic script. The document also contains recommendations for the inclusion of U+1361 (Hulet Netib) as valid name character and U+1372 – U+137C as valid numeric characters that can be used in XML tags.

Status of this Document

This document is a proposal that addresses and in some cases incorporates recommendations received during the comment period.

Table of Contents

1. Scope

The work reported in this document contains XML example codes, surveys and recommendations that support the need for a revision of current XML standards to allow Ethiopic-based XML element type and attribute names in Amharic, Afaan Oromo, Tigrigna, Agewgna, Guragina, Hadiya, Harari, Sidama, Kembatigna and other Ethiopian languages that use Ethiopic script. The document also contains recommendations for the inclusion of U+1361 (Hulet Netib) as a valid name character and U+1372 – U+137C as valid numeric characters that can be used in XML tags. This work is supported by the Ethiopic XML interest group (See Appendix A).

2. Introduction

The current version of XML (V 1.0, Edition 2.0) is based on character standards of an older version of Unicode (v 2.0). Over the years, however, Unicode has evolved to newer versions that have added a number of character sets from various scripts around the world. As a result, not all character sets defined and standardized in Unicode 3.1 can be used as XML names such as element type names, attribute names, processing instruction targets, and the like. This has prevented a fully-native XML markup ability in Ethiopian languages such as Amharic, Afaan Oromo, Tigrigna, Agewgna, Guragina, Hadiya, Harari, Sidama, Kembatigna, etc which use Ethiopic script.

The so-called Blueberry revision of XML that addresses this issue affecting Ethiopic, Syriac, Myanmar, Ogham, Runic, Sinhala, Thaana, Yi and other scripts was recently proposed by the World Wide Web consortium (W3C) XML Core working group. The latest version of the Blueberry draft document is published at http://www.w3.org/TR/xml-blueberry-req.

This work, therefore, is in response to the Blueberry requirement and deals with the demonstration and proposal for standardization of Ethiopic-based element type and attribute names in Amharic, Afaan Oromo, and Tigrigna languages.

3. Objectives

Through sample XML codes and DTDs (Document Type Definitions) for various applications (eCommerce, telemedicine, and calendar), we demonstrate that the use of non-Ethiopic script or transliteration is inadequate, inconvenient and in some cases erroneous for XML documents marked-up for use by speakers of one or more of the major Ethiopian languages. The examples codes include schemas based on transliteration and translations through the Latin script.

We also demonstrate that through the use of example XML codes and DTDs, hypothetically supported by Unicode 3.1, correct Ethiopic XML documents can be marked-up.

Further, we include an example (telemedicine) that demonstrates a real need for support in XML for Ethiopic name tags.

A review of the status of Romanization (the use of the Latin alphabet to represent Ethiopic-script languages) in both HTML and XML languages, how widespread it is, what standardization exist, if any, and how readily it is understood is included as a starting point.

4. Review of Status of HTML and XML Applications in Ethiopic

4.1. Evolution of Ethiopic Content in HTML

A review of the history of development of HTML markup in Ethiopian languages, we feel, will provide an appropriate and helpful insight into how XML markup for Ethiopian languages will evolve. To aid this review process, we had carried out a survey of major Ethiopian web sites.

Ethiopian web sites started with HTML markups around 1994 and 1995 and were naturally limited to, in almost all the cases, to contents written in the English language. Whenever the need arose for writing in Ethiopic script, either image files were used or in the more sophisticated but awkward cases, partial Ethiopic fonts that fit into ASCII table were used. The introduction of transliteration by D. M. Yaqob, et al provided a number of news web sites with a clever option to markup their pages in Ethiopic [5-6]. Further, the introduction of an Ethiopic character set and character code conversion algorithms and utilities by the same authors has also helped a number of web sites develop web contents in Ethiopic script. The Ethiopian Headlines News web site is an early and consistent user of such **transliteration** and "**conversion"** solution. EthioIndex, a News search site is a recent addition to the pool of transliteration and conversion users.

Over the past 18-24 months, encoding using Ethiopic Unicode character sets has slowly been introduced. Abass Alamenehe developed the so-called "**Ethiopia Jiret**" Unicode TrueType font [12] that consists of Ethiopic and Latin-1 blocks[1]. The Ethiopic block confirms to the Unicode 3.0 standard. GF Zemen is another Unicode font for Ethiopic content. Examples of sites that use Unicode encoded Ethiopic character sets are Senamirmir and Ethiopian Headlines News.

While transliteration, conversion and now Unicode encoding of Ethiopic character sets had introduced an interesting and helpful solution in HTML marked-up pages for contents in Ethiopic, the majority of the most active Ethiopian web sites still continue to employ awkward image files to display Ethiopic characters. In the case of Seleda, a popular monthly art and literature site targeting Ethiopians in the Diaspora, transliteration has been limited to depicting Ethiopian language words in their Latin transliteration equivalence. MediaEthiopia, another popular site with readership in both inside and outside Ethiopia uses image files for displaying documents written in Amharic and Afaan Oromo. Again, it needs to be mentioned that encoding using Ethiopic Unicode character sets is at its infancy and we expect it to be widely used in the next 3-5 years.

We feel various factors contribute to the rather non-uniform solution to the problem of HTML rendering in Ethiopic. The major problems are, however, the lack of any easy to use Ethiopic-based HTML software (including Unicode-based software) and the presence of numerous fonts and character encoding procedures. Despite the introduction of standards by bodies like CEC (Committee for Ethiopic Computing)

[1] the Ethiopia Jiret font adds two Ethiopian quotation marks that are not featured in the Unicode character set.

and ECoSA (Ethiopian Computer Standards Association), the problem still persists at a large scale.

4.2. Ethiopic and XML

Coming back to XML, to date, there is no publicly available data on the use of XML for marking-up web pages using any of Ethiopia's major languages. The survey in this report is the first such survey. This survey of major Ethiopian web-sites carried out by the authors of this report suggest that, at this point, HTML remains the markup language of choice in almost 100% of the sites with XML usage limited to one or two experimental sites. Our survey indicated that with the possible exception of EthioIndex which uses some elements of XML markup, none of these major web sites have implemented XML markups for content and structured data in Ethiopic script.

However, the proliferation of well-architectured Ethiopian eCommerce sites like AnythingONETHIOPIA and EthioTrade and the slow but steady adoption of both consumer and B2B eCommerce practice suggests that the demand for XML markups for content and data in Ethiopic script will reach a critical mass in the near future. Further, we believe that Microsoft's .NET strategy will ultimately not only encourage but also drive the introduction of Ethiopic software solutions based on XML.

With regard to the current limitation of XML 1.0, using the history of HTML markup language in Ethiopic as a guide, one may be tempted to extrapolate that transliteration and "conversion" at the server side may provide a way-around for hybrid XML markups in Ethiopic. However, XML documents are not merely documents but contain valuable data that are and should be amenable to manipulations of any sort the user wishes. Therefore the issue of pragmatism takes precedence over a simple rendering in native Ethiopic script. In other words, unlike HTML where it is adequate, transliteration, in the context of XML, faces the additional burden of providing meaningful and structured data sets in addition to correct rendering in Ethiopic script. Therefore, the use of transliteration in hybrid XML document markups in Ethiopic, we feel, will be very much resisted by content providers, eCommerce sites and Ethiopic researchers and linguists for the simple but critical reason that it fails to address the major requirement of XML marked-up documents, i.e., providing a mechanism to manipulate data in a structured, accurate and non-ambiguous fashion.

Looking forward, the two areas where we see XML finding immediate use among Ethiopic script users are eCommerce and telemedicine. ECommerce application in Ethiopic, we propose, will follow the overall industry's trend as the need for presentation of structured data in eCommerce is universal. On the other hand, the promise of telemedicine is more relevant in least developed nations such as Ethiopia, where there is single physician for 100,000+ people. It is now well accepted that telemedicine holds a pragmatic promise to alleviate some of the most common infrastructure problems in medicine in developing countries. One can imagine remote villages connected to medical centers via wide area networks where appropriate medical personnel could remotely diagnose and prescribe remedies as

recently done in Ethiopia [9]. Further, one can imagine the development of systems and tools that transmit medical symptoms and conditions from remote locations to medical centers for advice and resolution. XML has the potential to be a tool of choice for exchange of such information among the various fragmented medical institutions.

Table 1 contains a list of major Ethiopian news, information, research and community sites that were reviewed for this work. While there are hundreds of thousands of web sites related to Ethiopia [Google.com returned about 1.54 million pages when searched under the keyword "**Ethiopia**"], only the most important and representative web-sites were included in this survey.

The term "**transliteration**" as used here to describe support for Ethiopic content applies to transliteration and "**conversion**" schemes applied both at the server-level [a good example being **Ethiopian Headlines News**] and at the individual page level. The determination on the use of XML was based on the issue of availability of structured data and content in Ethiopic at the sites. With the possible exception of EthioIndex, our survey indicated that almost none of the sites covered in the survey use XML markup to present documents with structured data.

Table 1. Use of HTML and XML in Ethiopic Script in Major Ethiopian web-sites.

Name of Site	Location/URL	Nature of Content	Support for Ethiopic Content	XML
1. **Abyssinica**	http://www.abyssinica.com	Arts and Literature	None	no
2. **Addis Chamber of Commerce**	http://www.addischamber.com	News, limited eCommerce	Bitmaps	no
3. **AddisTribune**	http://www.addistribune.com	News	Bitmaps	no
4. **Amharic.com**	http://www.amharic.com	software	Bitmaps	no
5. **Anything-onEthiopia**	http://www.anythingonethiopia.com	News and eCommerce	Bitmaps	no
6. **CyberEthiopia**	http://www.cynberethiopia.com	News, Search and community	Bitmaps	no
7. **DebreHayk**	http://www.the3rdman.com/ethiopia nart	Arts	No	no
8. **ECS**	http://www.ethiopic.com	Software vendor	Bitmap and pdf	no
9. **Ethio.com**	http://www.ethio.com	News	None	no
10. **EthioExpo**	http://www.ethioexpo.com	ECommerce	None	no
11. **EthioGift**	http://www.ethiogift.com	ECommerce	None	no
12. **EthioGuide**	http://www.ethioguide.com	News, tourist info	None	no
13. **EthioIndex**	http://www.ethioindex.com	News	Unicode, transliteration, conversion	yes
14. **EthiopiaFirst**	http://www.ethiopiafirst.com	News	None	no
15. **Ethiopian Headline News**	http://www.ethiozena.net	News	Unicode, transliteration, conversion, pdf	no
16. **Ethiopian Yellow Pages**	http://ethiopiayellowpages.com	Yellow pages, eCommerce	Bitmaps	no
17. **EthiopianReporter**	http://www.ethiopianreporter.com	News and limited eCommerce	WEFT embedded font	no
18. **EthioSearch**	http://www.ethiosearch.com	Search engine	Bitmaps	no
19. **EthioSports**	http://www.ethiosports.com	Sports news	Bitmap and pdf	no
20. **EthioSystems**	http://www.neosoft.com/~ethiosys	Software vendor	Unicode, bitmap, pdf	no
21. **EthioTrade**	http://www.ethiotrade.com	ECommerce	None	no
22. **Harar**	http://www.harar.com	Community, news	Bitmaps	no
23. **MediaEthiopia**	http://www.mediaethiopia.com	News, Community	Bitmaps	no
24. **Mesob**	http://www.mesob.org http://www.mesob.net	Community, news, eCommerce	Bitmaps	no
25. **PAHA**	http://paha.ooi.net/main.phtml	Health issues (HIV)	None	no
26. **Seleda**	http://www.seleda.com	Arts and Literature, Webzine.	Bitmaps and Latin Script	no
27. **Senamirmir**	http://www.senamirmir.com	Research in Ethiopic, Webzine	Unicode + pdf	no
28. **Walta**	http://www.waltainfo.com	News	Bitmaps + WEFT embedded font	no

5. Hybrid XML Document Markups in Ethiopic

With the current XML 1.0 recommendations, Ethiopic script users are restricted to using either Latin script with English translation or Latin script with transliteration for defining their XML tags and attributes. Theoretically, it could be argued that translation could offer a quick and dirty short cut for helping native markups. Along the same lines, a case could be made for, at least hypothetically, the use of transliteration to help native markup.

5.1. Examples:

The following examples (A-E) demonstrate how the use of Latin script in translation and transliteration to define tags and attributes for an Ethiopic based XML document could be employed at least hypothetically. Here, we have used documents written in some of the major Ethiopian languages such as Amharic, Afaan Oromo, and Tigrigna. Example D, demonstrates the use of a transliteration based on the proposals of Yakob, et al [5-6].

I. **Example A - Ethiopian Calendar Data in Amharic using Amharic-English translation**

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE HybridEXMLETCalendar SYSTEM
"HybridEXMLETCalendar.dtd">
<yabeshakalender>
        <qen> 25 </qen>
        <ilet> እሁድ </ilet>
        <beal_qen> ስላሴ </beal_qen>
        <wer> ግንቦት </wer>
        <yeAbeshaamet>1955</yeAbeshaamet>
        <yezemensim> ዘመነ ዮሀንስ </yezemensim>
</yabeshakalender >
```

The DTD file, HybridEXMLETCalendar.dtd, will look like this:

```
<!DOCTYPE yabeshakalender [
<!ELEMENT yabeshakalender (qen,ilet,beal_qen,wer,
yeAbeshaamet, yezemensim*)>
<!ELEMENT qen (#PCDATA)>
<!ELEMENT ilet (#PCDATA)>
<!ELEMENT beal_qen (#PCDATA)>
<!ELEMENT wer (#PCDATA)>
<!ELEMENT yeAbeshaamet (#PCDATA)>
<!ELEMENT yezemensim (#PCDATA)>
]>
```

Note: The name of a "**tabot**" could interchangeably be used with "**beal_qen**"; even though "**beal_qen**" is more generic and includes all holidays[2].

Further, the Ethiopian calendar data set could be expanded to contain attributes with the equivalent Gregorian calendar system as shown below.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE  HybridEXMLETCalendar SYSTEM
"HybridEXMLETCalendar.dtd">
<yabesha_ina_ferenj_kalender>
        <yabesha_qen> 25 </yabesha_qen>
        <yeferenj_qen> 2 </yeferenj_qen>
        <ilet> እሁድ </ilet>
        <beal_qen> ስላሴ </beal_qen>
        <yabesha_wer> ግንቦት </yabesha_wer>
        <yeferenj_wer> ሰኔ </yeferenj_wer>
        <yeAbeshaamet>1955</yeAbeshaamet>
        <yeFerenjochamet>1963 </yeFerenjochamet>
        <yezemen sim> > ዘመነ ዮሀንስ </yezemensim>
</yabesha_ina_ferenj_kalender >
```

Again, the DTD file, HybridEXMLETCalendar.dtd, will look like this:

```
<!DOCTYPE yabesha_ina_ferenj_kalender [
<!ELEMENT yabesha_ina_ferenj_kalender (yabesha_qen,
yeferenj_qen,
                ,ilet,beal_qen, yabesha_wer, yeferenj_wer,
yeAbeshaamet,
                yeFerenjochamet , yezemensim*)>
<!ELEMENT yabesha_qen (#PCDATA)>
<!ELEMENT yeferenj_qen (#PCDATA)>
<!ELEMENT ilet (#PCDATA)>
<!ELEMENT beal_qen (#PCDATA)>
<!ELEMENT yabesha_wer (#PCDATA)>
<!ELEMENT yeferenj _wer (#PCDATA)>
<!ELEMENT yeAbeshaamet (#PCDATA)>
<!ELEMENT yeFerenjochamet (#PCDATA)>
<!ELEMENT yezemensim (#PCDATA)>
] >
```

[2] Some argue that the term "bealat" (literal translation, holidays) is a more generic and therefore better description than "tabot". However, "tabot" is a term that is deeply entranced in the Ethiopian social, and cultural tradition and we have opted for its use here.

II. Example B - Ethiopian Calendar Data in Afaan Oromo Language using Afaan Oromo-English translation

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE  HybridEXMLETCalendar SYSTEM
"HybridEXMLETCalendar.dtd">
<kaaleendaraabesha>
        <qeni> 25 </qeni>
        <guya> ዲልባታ </guya>
        <guya_senbeta> ስላሴ </guya_senbeta>
        <weri> ኤብላ </weri>
        <waga_abesha>1955</waga_abesha>
        <zemenimeqa> ዘመኒ ዮሀንስ </zemenimeqa>
</kaaleendaraabesha>
```

III. Example C - Ethiopian Calendar Data in Tigrignna Language using Tigrignna -English translation

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE  HybridEXMLETCalendar SYSTEM
"HybridEXMLETCalendar.dtd">
<yabeshakalenderti>
        <mealti> 25 </mealti>
        <ilet> ሰንበት</ilet>
        <beal> ስላሴ </beal>
        <werhi> ግንቦት </werhi>
        <ametbeEthiopiaAkostasira>1955</
ametbeEthiopiaAkostasira >
        <yezemen sim> ዘመነ ዮሀንስ </yezemensim>
</yabeshakalender >
```

IV. **Example D - Address Data in Amharic using transliteration**

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE  HybridEXML SYSTEM
"HybridEXML.dtd">
<adraxa>
        <kfle hager> ሸዋ </kflehager>
        <wereda>16</wereda>
        <qebelE>07</qebelE>
        <bEt>475</bEt>
        <ketema> አዲስ አበባ </ketema>
</adraxa>
```

V. **Example E – Application of XML in Telemedicine using Romanization**

In the Ethiopian context, given the particular unique and ambiguous ways of describing medical conditions/symptoms, translations, transliteration and Romanization could insert a level of misrepresentation. Example E presents an example of such a case.

Let us look at an example of a stomach ache being reported through a telemedicine system. An example DTD for transmission of type of ailment could look like:

```
<!DOCTYPE Beshita [
<!ELEMENT Beshita
(ayenet,mekekit,lesentgeze,yebetesebtarik*)>
<!ELEMENT ayenet (#PCDATA)>
<!ELEMENT meleket (#PCDATA)>
<!ELEMENT lesentgize (#PCDATA)>
<!ELEMENT yebetesebtarik (#PCDATA)>
]>
```

Almost all the terms used in this particular data set could potentially be written in many different ways depending on the particular style of the author. For example, "**beshita**" [translation: ailment] could be spelled also as "**beshta**". Similarly, "**ayenet**" as "**aynet**", "**melekt**" as "**milikit**" or "**mlkt**", and "**lesentgize**" as "**lesntgze**".

5.2. Discussions

While the schemas shown above demonstrate that a hybrid (English and Ethiopic) XML document could be written with some effort using translation and transliteration, the following handicaps make the usefulness of such "work-around" very difficult, cumbersome and impractical.

5.2.1. **Switching between two scripts**. The Ethiopic XML document author will have to switch between Ethiopic and Latin scripts for each and every attribute of the document.

5.2.2. **Translation skills**. With regard to translation, the usefulness of the data set depends on the translation skills of the author. For example, some unique Ethiopian concepts like "**tabot**" **[ታቦት]** do not have an equivalent English (or any foreign language) words. Here, we had used the word "**beal qen**" [**በአል ቀን**] as an equivalent translation; but this word just means a "**holy day**" and does not necessarily establishes the distinction between a regular holiday like Ethiopian New Year (**Meskerem 1**) and a "**tabot**" like **Lideta/Kidist Mariam** which also falls on the first of each month. Therefore, this attribute could potentially be carrying a wrong data.

5.2.3. **Romanization styles**. As shown through Example E (telemedicine example), the Romanization style of the author has a significant influence on how the attributes are presented. There are simply many styles of writing Amharic, Afaan Oromo or Tigrigna words in Latin script; none of them being more correct than the other.

5.2.4. **Cultural sensitivity.** The powerful Ethiopian church and its army of linguists have historically been very much opposed to any attempt to replace Ethiopic script usage by even traces of Latin script. The Ethiopian Church has regarded itself as the keeper of the Ethiopic (also called Geez) script since the early days of its foundation in the late 6th and 7th Century. Over the years, all attempts to modernize the script even with the blessing of powerful political leaders have failed[3]. We see transliteration and facing the same fate. A detailed account of

[3] There are a number of proposals put forward by leading intellectuals, linguists, and authors. These include, Haddis Alemayyehu (various works) and Getachew Bekele (Ye Agelgaye Mestawat, published 1954 Eth. Calendar) among others. For a more complete history of efforts by authors to change Ethiopic script, refer to Reidulf K. Molvaer's, "Black Lions", a look at the creative lives of modern Ethiopia's Literary Giants and Pioneers [13].

such prior attempts and the church's sensitivity to script issues is given by a classic publication of Abraham Demoz [11].

5.2.5. **Lack of standard and acceptance**. Currently, apart from the work of a single group of researchers, transliteration is not a widely accepted practice. As a result, there simply isn't even agreement on what transliteration scheme to adopt. A case in point is, for example, the different ways the Ethiopian name **ያቆብ** is spelled and transliterated. Variations of this include, Yacob, Yaqob and Yakob. It is interesting to note while the letters "**c**", "**q**", and "**k**" have been used to represent the Ethiopic explosive sound, **"ቆ"** ("**qwo**"), none of them actually is more accurate than the other.

Another example that demonstrates the lack of agreement and conventions is the Amharic word for address (i.e., **አድራሻ**). The "**sha**" sound in Amharic is usually transliterated, as seen in the at least 50 years tradition of Ethiopian English newspapers like the Ethiopian Herald, with the equivalent "**sha**" in English. However, as Example D taken from Reference [6] shows, there is at least one transliteration scheme where "**x**" has been used to transliterate the Amharic sound, "**sha**". Further, the transliteration for this particular sounds is identical to what the Somali language has adopted in its Latin-based writing system. This, in our opinion, fails to acknowledge the significance of the long history of Ethiopian script by proposing a solution that basically equates it to languages with little or no written history[4]. This fact alone can guarantee, in our opinion, resistance from Ethiopian linguists and Church officials, two important bodies traditionally regarded as the stewards of the Ethiopic (**Geez**) script.

[4] An indigenous Somali script was proposed but later abandoned around 1972 in favor of Arabic. The script was briefly used in government and public school systems. A proposal for encoding the script is found at: http://www.dkuug.dk/jtcl/sc2/wg2/docs/n2361r.pdf

6. Fully-Native XML Document Markups in Ethiopic

The Ethiopic script was formally adopted by the ISO in Unicode 3.0. Here, we use the Unicode (3.0) representations of Ethiopic characters to demonstrate the application of a potentially fully-native markup for Ethiopic XML documents.

6.1. Examples

We will use the same example used in Section 5 (i.e., Example A) to demonstrate this.

I. Ethiopian Calendar Data in Amharic in Fully-Native XML markup

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE HybridEXMLETCalendar SYSTEM
"HybridEXMLETCalendar.dtd">
<የአበሻካሌነደር>
        <ቀን> 25 </ቀን>
        <እለት> እሁድ </እለት >
        <በአል_ቀን> ስላሴ </በአል_ቀን>
        <ወር> ግንቦት </ወር >
        <የአበሻ_አመት>1955</የአበሻ_አመት>
        <የዘመን_ስም> ዘመነ ዮሀንስ </የዘመን_ስም>
</የአበሻካሌነደር >
```

The DTD file, HybridEXMLETCalendar.dtd, will look like this:

```
<!DOCTYPE የአበሻካሌነደር [
<!ELEMENTየአበሻካሌነደር
(ቀን,እለት,በአል_ቀን,ወር,የአበሻ_አመት,የዘመን_ስም*)>
<!ELEMENT ቀን (#PCDATA)>
<!ELEMENT እለት (#PCDATA)>
<!ELEMENT በአል_ቀን (#PCDATA)>
<!ELEMENT ወር (#PCDATA)>
<!ELEMENT የአበሻ_አመት (#PCDATA)>
<!ELEMENT የዘመን_ስም (#PCDATA)>
]>
```

6.2. Discussions

The example in Section 6.1 demonstrates that a fully-native XML markup in Ethiopic script provides a cleaner solution.

Particular advantages of such a fully-native XML markup in Ethiopic script include:

- Clarity of data definition. With fully-native markup, there will not be any confusion on what the author uses to represent a certain data.
- Better readability of XML codes by humans. This is indeed one of the requirements in XML 1.0 recommendations.
- Ease in writing a good and clean XML code with no frequent switching of scripts.

7. Recommendations for Valid Character List for Ethiopic XML

In name tags made up of two or more words, the under score character (_) has been widely used to declare them as valid name tags. In Ethiopic writing system, U+137 (Hulet Netib) that is traditionally used to separate words could serve as a more natural and logical alternative to connect two or more words of a valid name tag. Its inclusion in upcoming revision of XML 1.0 will be a welcome development to Ethiopic script users[5].

Further, we propose that U+1372 - U+137C are valid numeric characters. The rationale here is that the numbers are digital with respect to the Ethiopic numeral system as per the rule that the characters can be concatenated together to form a decimal value. The Unicode designation of U+1369 - U+1371 as digits was a matter of convenience to allow software to map the characters onto the ASCII 1-9 digital range. This convenience mapping should not otherwise negatively impact Ethiopic digits beyond the value of 9.

Appendix B gives a list of proposed valid character classes.

[5] These sets of recommendations were incorporated as a result of feedback obtained from Daniel Yacob during discussions at the Ethiopic XML Interest Group.

Conclusions and Recommendations

The example schemas in Section 5 demonstrated that a hybrid (English and Ethiopic) XML document could be written with some effort using translation and transliteration. However, we have demonstrated that, its usefulness is seriously hampered by the following handicaps described in depth in Section 5:

I. Inconvenience of switching between two scripts.

II. The translation skills of the author and user of XML pages have a large influence on the usefulness of the data and document.

III. The numerous personal preferences and styles of writing Ethiopic words in Latin script also have a large influence on the usefulness of the data and document

IV. Cultural sensitivity has historically been a major issue in Ethiopia [10]

V. Lack of standard and acceptance of transliteration poses a problem

VI. Need for equal representation of scripts. The notion of XML not being able to support complete native markup for certain scripts is very hard to justify as it seems to carry a negative stigma of unfairness and lack of equal representation of writing systems around the world. Discriminating against languages simply because their scripts were not encoded in Unicode 2.0 is inherently unjustifiable. The ramification is not only to drive certain cultural heritages to extinction but also to deprive future generations from understanding and using unique social, cultural and scientific concepts outside of the mainstream.

VII. One of the specific goals mentioned in XML 1.0 specification states that "XML documents should be human-legible and reasonably clear". For users of Ethiopic, Syriac, Myanmar, Ogham, Runic, Sinhala, Thaana and Yi scripts, this requirement is violated. Certainly, such contradictions are extremely hard to justify.

Further, in Section 6, we have demonstrated that a fully-native markup for Ethiopic will encourage the use of XML markup for Ethiopic use in education, media, eCommerce and the like.

In Section 7, we forwarded a recommendations for the inclusion of U+1361 (Hulet Netib) as valid name character and U+1372 – U+137C as valid numeric characters that can be used in XML tags.

In the final analysis, we propose that a hybrid markup will in fact stifle the use and proliferation of Ethiopic script use in XML documents. A completely native XML markup in Ethiopic, we propose, is the justifiable solution to the proliferation of useful Ethiopic content and data in XML.

Acknowledgements

The authors would like to give special thanks to Abass Alamenehe for the numerous online and offline discussions with regard to this topic and other numerous topics in Computational Ethiopic. Many thanks go to Daniel Yacob who had commented on our draft proposal and recommended the inclusion of U+137 (Hulet Netib) and U+1372-U+137C as valid name and numeric characters.

We also would like to thank John Cowan for his valuable suggestions.

Appendix A

The Ethiopic XML Interest Group (formally called EthiopicXML Specifications Working Group) was established in July 2001 and aims to:

1. Support the Blueberry draft proposal that addresses the need for a revision of current XML standards to allow XML element type and attribute names in scripts such as Ethiopic that had their character sets included in Unicode only in version 3.0 and later.

2. Create and consolidate working groups and organizations with interest in Ethiopic and XML.

3. Educate the public on the need for Ethiopic extensions to XML for eCommerce, literature, education, and research applications. Also encourage the development of native XML grammar and vocabulary for a variety of applications in Ethiopic. This will include work in the major languages of the country, namely, Amharic, Afaan Oromo, Tigrigna and others.

4. Bring the topic for wider discussion that will include Ethiopic linguists, church personalities, academics and students of Ethiopian languages.

5. Finally, submit the draft proposal to the W3C XML Core working group.

Appendix B Character Classes

Following the characteristics defined in the Unicode standard, characters are classed as base characters (among others, these contain the syllographic characters of the Ethiopic syllabary), digits and non-digital integer characters.

BaseChar ::= [#x1200-#x1206] | [#x1208-#x1246] | #x1248 |
[#x124A-#x124D] | [#x1250-#x1256]
| #x1258 | [#x125A-#x125D] | [#x1260-#x1286] | #x1288 |
[#x128A-#x128D]
| [#x1290-#x12AE] | #x12B0 | [#x12B2-#x12B5] |
[#x12B8-#x12BE] | #x12C0
| [#x12C2-#x12C5] | [#x12C8-#x12CE] | [#x12D0-#x12D6] |
[#x12D8-#x12EE] | [#x12F0-#x130E]
| #x1310 | [#x1312-#x1315] | [#x1318-#x131E] |
[#x1320-#x1346] | [#x1348-#x135A]

Digit ::= [#x1369-#x1371]
Number Other ::= [#x1372-#x137C]

The character classes defined here can be derived from the Unicode 3.0 character database as follows:

- [#1372-#137C] are allowed as Ethiopic usage does not distinguish between the Nd and No designations assigned by Unicode.

- Character #x1361 is allowed as a name character.

- Characters [#x1362-#x1368] are excluded (in accordance with Unicode 3.0).

References

1. XML Blueberry Requirements -W3C Working Draft 21 September 2001 - http://www.w3.org/TR/xml-blueberry-req.html, John Cowan, Editor.

2. Ethiopic XML Interest Group - http://groups.yahoo.com/group/EthiopicsXML also, see Appendix A in the present document.

3. The Unicode Standard - http://www.unicode.org/unicode/standard/standard.html

4. Ethiopic character code for The Unicode Standard, Version 3.0 - http://www.unicode.org/charts/PDF/U1200.pdf

5. System For Ethiopic Representation In ASCII (SERA), Daniel Yacob and Yitna Firdyiwek - http://www.mediaethiopia.com/yitna.html

6. Introduction to SERA in SERA FAQ – Daniel Yacob- http://www.abyssiniacybergateway.net/fidel/sera-faq_1.html

7. Committee for Ethiopic Computing (CEC) - http://www.neosoft.com/~abassa/ytna/projects.htm

8. The Ethiopian Computer Standards Association, (ECSA) http://ecosa.ethiopiaonline.net/

9. Samuel Kinde Kassegne, "Internet in Ethiopia Revisited – A Mixed Bag of Progress and Opportunities on-Hold", MediaETHIOPIA, April 2002.

10. Assefa Fesseha, University of Utrecht, Netherlands, Personal Correspondences for translation to Tigrigna, November 2001.

11. Abraham Demoz, ``Amharic Script Reform Efforts'', `Ethiopian Studies: Dedicated to Wolf Leslau on the Occasion of his 75th Birthday November 14th 1981, Otto Harrassowitz, 1983.

12. Ethiopian Jiret Font, http://www.senamirmir.com/download/jiret.zip.

13. "Black Lions: A look at the creative lives of modern Ethiopia's Literary Giants and Pioneers". Reidulf K. Molvaer.