Translation Journal

Volume 4, No. 2
April 2000

Dr. Melby was born in 1948, earned his B.S. majoring in mathematics *Magna Com Laude* and his Ph.D. in Computational Linguistics from the Brigham Young University. He is a full-time professor of Linguistics at BYU and has worked extensively with different U.S. and international organizations in the areas of machine translation, machine-aided translation, Machine-Readable Terminology Interchange Formate (MARTIF), Standard Graphic Markup Language (SGML), and Extended Markup Language (XML). He is listed in International Who's Who in the World and International Who's Who in Translation and Terminology.

Dr. Melby can be reached at akm@byu.edu.

Web page: http://www.ttt.org.

- **Front Page**
- January ‘00 Issue
- October ‘99 Issue
- July ‘99 Issue
- April ‘99 Issue

XML and the Translator

by Alan K. Melby, Ph.D.

Why should you be interested in XML? Well, if you are not interested in HTML, then you probably won't be interested in XML. On the other hand, if you have translated a Web page or are even thinking about doing it someday, then you had better learn about XML.

As you know, a Web page is basically tags and text. Here is an extremely simple Web page:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 3.2
Final//EN">

<HTML>

<HEAD>

<TITLE>ASTM Workshop Information</TITLE>

</HEAD>

<BODY BGCOLOR="WHITE">

<H1 ALIGN=CENTER>ASTM Workshop</H1>

<H2>Your Invitation to the March 2000
Workshop</H2>

<P>ASTM Subcommittee F15.48 on Language
Translation cordially

invites you to actively participate in our
committee work

to leverage common interests in the field of
translation

and ensure premium quality service. We hope you
```

```
will attend.</P>

</BODY>

</HTML>
```

Here are some of the tags:

1. <BODY>

2. <H1 ALIGN=CENTER>

3. </BODY>

Everything that is not a tag is text.

So far, so good. Now, what kinds of tags are there?

Tag number 1 (<BODY>) is called a start tag, since it comes at the beginning, and tag number 3 (</BODY>) is called an end tag, since it terminates the "element" which is everything between the start tag and its end tag. And, as in this case, the stuff in between can include other tags.

Tag number 2 is a start tag that has an "attribute" (ALIGN), and that attribute has a value (CENTER).

Another important concept is an "empty" tag. One empty tag is <BR>, which forces a line break. It is empty because it has no end tag.

Now what if we deleted the </P> tag in the sample Web page? Everything would still work, since the end tag is optional for a <P>, but would that make <P> an empty tag? Not at all! It just makes the end tag implicit. There is no implicit end tag for <BR>.

Now we have some terms under our belt (start tag, end tag, empty tag, implicit end tag, and attribute), and we can go on to XML.

HTML and XML are both members of the SGML family, but HTML is an "application" of SGML while XML is a "subset" of SGML. What is the difference? A big one. First of all, don't worry about exactly what SGML is. It is an ISO standard (number 8879), but that doesn't help you understand it. I have studied SGML for over ten years now, and I don't know all the ins and outs of it. It is a very abstract system for defining systems of tags. SGML is used to express the fact that <BODY> is an HTML tag, but <SHOULDER> and <EAR> are not. Someone (Tim Berners-Lee, to be exact) used SGML to define an application called HTML, and that changed the world because it was a key element in bringing the Internet to the average person. So it is important to make a good selection of tag names and how they fit together.

Han Yang, Ph.D.

Science & Technology

- A Translator’s Guide to Organic Chemical Nomenclature XIX
by Chester E. Claff, Jr., Ph.D.

Caught in the Web

- Web Surfing for Fun and Profit
by Cathy Flick, Ph.D.
- Translators’ On-Line Resources
by Gabe Bokor

Translators’ Tools

- More Translation Memory Tools
by Suzanne Assénat-Falcone
- Translators’ Emporium

- **Translators’ Events**

- **Call for Papers and Editorial Policies**

HTML has a fixed list of tag names. You can't just start using a new tag name, like <SUITCASE> and expect a browser to understand it.

The single biggest difference between HTML and XML, the difference that hits you in the face, is that in XML there is *no fixed list of tag names*! Instead, XML is a simplified form of SGML, and you can use it to define your own XML application that has its own tag names. In an XML application there could very well be a <SHOULDER> tag or even a <SUITCASE> tag. The list of tag names, along with other technical information about the tags, such as what attributes they can have and how they fit together, is found in the an awful looking thing called a DTD or a schema. May you never have to touch or even see a DTD or schema. Just the other day, I spent over twelve hours working a two persnickity DTDs. Just punishment for being ornery, some would say.

At at rate, now we can get the to the meat of the whole question of XML. Why would a translator even want to know about XML? I hinted at the answer at the beginning, when I said that those who translate Web pages will need to know XML. The reason is that XML is being combined with HTML in sophisticated Web pages already, and this trend should continue rapidly.

There are several ways to combine HTML and XML. One way is insert XML inside an HTML page. This is called an XML island.

Another way is to use XML *instead* of HTML for a Web page. This can be done already in Internet Explorer 5. The XML describes the content and structure of the information to be presented; then, in order to display the Web page nicely, you attach to the Web page a set of instructions called an XSL stylesheet. An XSL stylesheet is kind of like a cascading style sheet in HTML 4, but much more powerful. It is actually a computer program written in XSL, which is a programming language designed specifically for use with XML.

So now you might be wondering how to deal with an XML page if you are asked to translate it.

Well, first of all, you will probably not translate the tag names. Just as you would not translate the tag names in an HTML page. This should be verified with whoever is requesting the translation.

Secondly, you need to get a list of the tag names for the particular XML application you are working with, so you will be able to spot typos and data corruption. Every XML document for that application must select from those tag names, unless they use what is called an "open" model, which allows creative tag names that cannot be found even in the DTD/schema. This is kind of wild, but you might run across it. Another exception is what is called a namespace, an advanced topic of another tutorial. You can spot a namespace by the prefix followed by a colon and a tag name.

Thirdly, there are a few key differences between XML and HTML you need to keep in mind, besides the basic difference (fixed set of tag names vs. different set of tag names for each XML application):

a. Every empty tag has a slash at the end.

   This will look strange at first. In the case of HTML, a break would be <BR/> to show that there is no end tag to look for.

b. There are no implicit end tags.

   There will never be a <P> with no </P>, for example. This is a hard and fast rule for XML, and sets it apart from HTML in a big way. Of course, things change. There is a kind of HTML that is XML compliant, and it, of course, has no implicit end tags. This is called XHTML, and is not yet very well known.

c. Quotes are required on attribute values.

   This is an easy one to mess up on. In HTML, you can leave off the quotes in some cases, such as ALIGN=CENTER, but in XML, it must be ALIGN="CENTER" or ALIGN='CENTER'.

There is much, much more to XML, and new things are happening in the XML world every day, but the few principles just explained should help a translator deal with the differences between HTML and XML and be able to ask questions instead of being completely lost in XML land.