

Six Things You Should Know About XML

R. Jason Weiss

Development Dimensions International

Here is a one-item test on emerging Internet technologies:

1. What do the letters “XML” stand for?
 - (a) eXtensible Markup Language
 - (b) eXciting Mathematical Logic
 - (c) eXtreme Motor-scooter League

Bonus question: Does XML hold important implications for I-O psychology?

- (a) Yes
- (b) No

The answer to both questions is “a.” eXtensible Markup Language (XML) is an Internet technology that promises to improve the way we work with data. For those interested in learning more, I present the six things you should know about XML:

- XML is a language for organizing and storing data.
- XML ensures that data definitions remain with the data.
- XML permits straightforward communication of data across diverse systems.
- XML can be used to map industry-wide data vocabularies.
- Efforts are underway to create XML data vocabularies for HR.
- The input of I-O psychologists is necessary for these efforts to succeed.

XML is a language for organizing and storing data. Let’s start our discussion of XML by contrasting it with its more familiar cousin, HyperText Markup Language (HTML). HTML communicates the format and linkage of a document’s content to the Web browser. HTML accomplishes this task by using tags, which are commands enclosed in less-than and greater-than (< and >) symbols, embedded within the document. For example, the tag signals the Web browser to use a boldface font for the text that follows. Tags are typically used in pairs. The second tag in a pair indicates the end of the command’s scope, and is indicated by a forward slash before the command, as in . The HTML fragment “Add salt and pepper to taste.” is therefore rendered “Add **salt** and **pepper** to taste.”

XML looks like HTML, but has a somewhat different focus. HTML tags are commands that define how to format and link the information they contain. XML tags are descriptors for the data they contain. For example, in the XML fragment “<TestScore>50</TestScore>,” <TestScore> is the tag, and 50 is the datum contained by the tag. The tag name makes it clear what the enclosed datum represents.

Another key difference between HTML and XML is that HTML tags are predefined, while XML tags are user-defined. This ability to create your own tags is what makes XML the eXtensible Markup Language. If you don't like <TestScore>, you are free to use <Score> or whatever you consider the most accurate descriptor when you create the XML data file. You can extend the range of tag names as far as necessary.

There are two important similarities between HTML and XML that warrant discussion. First, like HTML, attributes can be used in XML to communicate additional information related to a tag. For example, we can enhance the <TestScore> tag by including as attributes information describing the test name and test form. Attributes take the form VariableName="value", and are located in the opening tag of a pair. Therefore, the <TestScore> example above with TestName and TestForm attributes added would look like the following: <TestScore TestName="Mental Rotation" TestForm="B">50</TestScore>.

A second similarity to HTML is that XML tags can be "nested," or arranged hierarchically so that groups of tags can be organized under higher-order tags. Exhibit 1 illustrates an example in which each participant's data from a fictional goal-setting study are nested within opening and closing <ParticipantData> tags.

Taken together, these features suggest that XML is a powerful language for modeling data. However, there are more benefits to describe, so let's move on.

XML ensures that data definitions remain with the data. Imagine this scenario: you would like to do additional analyses on data you collected some time ago. After some searching, you are able to locate the raw data file, but each participant's data are just a string of numbers and words separated by spaces. Exhibit 2 illustrates a typical raw data file using more of the fictional data shown in Exhibit 1. Given time and a good filing system, it is possible to locate the definitional information and put the data to use. On the other hand, I'm sure there are people who have been through this process and determined that it would have been faster just to collect the data over again.

Contrast this example with the data in Exhibit 1. In XML, data and definitions are stored together by default. As described above, tag attributes and higher-order organizing tags further enhance the "readability" of the data file. Together, these features ensure that XML data files can include all of the information necessary to minimize reliance on external supporting documentation for the data.

XML permits straightforward communication of data across diverse systems. A common problem for both academics and practitioners is that data are frequently stored on different systems, in a number of different, and potentially incompatible formats. Incompatible formats can arise due to different computer operating systems and in the software used to store and work with the data, among other things. The problem typically rears its head

Exhibit 1. XML data file.

```
<?xml version="1.0"?>
<ParticipantData>
  <Date>1/2/01</Date>
  <StartTime>9:14</StartTime>
  <Administration>Computer</Administration>
  <ParticipantID>123456789</ParticipantID>
  <Gender>Male</Gender>
  <Age>21</Age>
  <Difficulty>30</Difficulty>
  <Score>73</Score>
  <CompletionTime>15:03</CompletionTime>
</ParticipantData>
<ParticipantData>
  <Date>1/2/01</Date>
  <StartTime>10:23</StartTime>
  <Administration>Paper</Administration>
  <ParticipantID>987654321</ParticipantID>
  <Gender>Female</Gender>
  <Age>20</Age>
  <Difficulty>50</Difficulty>
  <Score>80</Score>
  <CompletionTime>15:10</CompletionTime>
</ParticipantData>
```

Exhibit 2. Raw data file.

```
1/2/01 9:14 Computer 123456789 Male 21 30 73 15:03
1/2/01 10:23 Paper 987654321 Female 20 50 80 15:10
1/2/01 11:11 Paper 111111111 Female 21 70 83 18:00
1/3/01 15:43 Computer 222222222 Male 23 50 45 12:32
1/3/01 16:01 Computer 333333333 Male 19 70 79 14:59
```

when you are migrating data from one system to another, or combining data from multiple systems. It can be mitigated somewhat by specialized translation routines that allow you to save data from one software application into a format compatible with your target application, although such translation routines are not always completely reliable.

XML addresses this problem in several ways. First, XML data are stored in plain text files, which are the lowest common denominator of data file.

Text files are readable by a wide variety of software and can be edited using simple text editors, such as the Windows Notepad. Second, software developers can avoid having to translate their data to and from different formats by using XML for importing and exporting data. The next point begins to address the power of XML in this context.

XML can be used to map industry-wide data vocabularies. Efforts are underway to create XML data specifications for all data within entire industries. These specifications are commonly known as XML vocabularies. While mapping out all data within a given industry is an ambitious goal, the promised benefits are highly motivating. For software makers within specific industries, creating translation routines for an endless array of target applications will no longer be necessary. An accepted, public standard ensures that all developers will understand how they must format data for export and what format they can expect imported data to take. The key benefit that plays out at the software user level is that these diverse systems will be able to “speak” to each other seamlessly, and permit data collected in one system to be used by another as required.

Efforts are underway to create XML data vocabularies for HR. There are two groups working on modeling HR data in XML. These are the HR-XML Consortium (www.hr-xml.org) and the Object Management Group (OMG; www.omg.org). Both groups are nonprofit corporations dedicated to establishing vendor-neutral standards, and both extend invitations for new members to join. HR-XML is dedicated to mapping out the HR space exclusively. OMG casts a wider net, establishing cross-industry standards in addition to developing specifications for vertical markets such as healthcare and finance.

Currently, there are a number of groups within HR-XML in different phases of standards generation. These include Benefits Enrollment, Competencies, Payroll, Recruiting and Staffing, and Time Reporting, among others. Within the Consortium’s process guidelines, additional groups can be formed to develop standards in other areas. Standards created by the groups are approved by membership vote. So far, HR-XML has published an established standard for posting information about job opportunities on job boards and retrieving information about job/position seekers in return.

The OMG operates according to a different principle. Where HR-XML takes an active role in organizing open groups to develop standards, the OMG issues Request for Proposals (RFPs) for the creation of specifications. The standards submitted in response to the RFPs are then evaluated and approved by task forces. As a result, several different submitting groups may work simultaneously on a given standard.

The input of I-O psychologists is necessary for these efforts to succeed. The move to establish standards for HR data is well underway. For any set of standards to succeed, it is necessary for a critical mass of stakeholders to accept and adopt it. Given the innumerable ways in which I-O psychologists

use this data, we can ensure that our needs are met by giving our input into the standards-setting process.

By getting actively involved with HR-XML and/or OMG, we can ultimately achieve three goals. First, we can monitor the developing standards and give input as to how they influence the ways in which we use specific data. Second, we can offer a critical theoretical perspective on a number of fronts, such as how data may be represented for particular applications, or why certain types of data may or may not be compatible. Third, we can locate improvements in the way we currently use data and help get them implemented for the benefit of all.

XML is an exciting technology, and we are fortunate enough to have the opportunity to shape its implementation in the software that we use. If you are interested in contributing to the standards-setting process, both HR-XML and OMG encourage you to take part. Contact the HR-XML Consortium by e-mail at info@hr-xml.org. OMG can be reached at info@omg.org. If you have any questions or comments for me, please e-mail me at jweiss@ddiworld.com.