MatML: XML for Materials Property Data

E.F. Begley and C.P. Sturrock
National Institute of Standards and Technology, Gaithersburg, MD

Introduction

In the August 2000 issue of Advanced Materials & Processes (AM&P), ASM International published a guest editorial[1] concerning materials databases. Accompanying this issue of AM&P was a special supplement entitled "Directory of Materials Property Databases[2]." Both the editorial and the directory are pertinent to the effort to develop MatML, an extensible markup language for the exchange of materials property data on the World Wide Web (Web). This effort is being coordinated by the National Institute of Standards and Technology, and is driven by the MatML Working Group, whose members include several ASM International Fellows, members, and staff.

The editorial described the following scenario for materials data systems of the future:

> "Wide and easy access to data is achieved by creating a system that is Web enabled with links to CAE software via either direct data transfer or by export routines for the most common CAE applications."

This scenario implies that Web documents and computer-aided engineering (CAE and also called computer assisted engineering) applications are interoperable, which means that the former may be integrated seamlessly and automatically into the latter, without human intervention or checking. However, anyone who has recently tried to download materials property data from the Web for subsequent processing has found that interoperability is poor at best, and in most cases nonexistent.

Considering the ASM Directory of Materials Property Databases, one is immediately struck by the quantity and diversity of materials databases listed therein. Perhaps even

more striking, however, is the fact that very few of these databases are accessible via the Web, which is by far the dominant medium for the exchange of digital information today. What is impeding the transition from printed handbooks and proprietary databases to Web accessibility? Economics are partially responsible – currently it is much easier to forestall piracy of intellectual property via the distribution of printed or magnetic media. However, with e-commerce becoming increasingly ubiquitous, economic impediments to this transition are gradually vanishing – only technical impediments remain.

Both of these problems, namely (1) the lack of interoperability between Web documents and CAE applications, and (2) technical impediments to the transition of materials data from print and magnetic media to the Web, arise as a result of the lack of a common data exchange format for materials data, which in turn motivated the conception and development of MatML.

The Web and HTML

To understand the significance of establishing a Web-based exchange format for materials data, consider how most of these data are encoded in Web documents today. The current *lingua franca* of the Web is hypertext markup language (HTML). Figure 1 contains a very small HTML code fragment for part of a table from a document in the

```
<table>
  <tr>
    <td align="center"><b>Magnetic Field (T)</b></td>
    <td align="center"><b>Temperature (K)</b></td>
    <td align="center"><b>Critical Current Density (kA/cm<sup>2</sup>)</b></td>
  </tr>
  <tr>
    <td align="center">0</td>
    <td align="center">3</td>
    <td align="center">3040</td>
  </tr>
</table>
```

Figure 1. Code fragment from the NIST Ceramics WebBook

Magnetic Field (T)	Temperature (K)	Critical Current Density (kA/cm^2)
0	3	3040

Figure 2. How code fragment in Fig. 1 would appear in a web browser

NIST Ceramics WebBook[3] while Fig. 2 illustrates how the code would be displayed in a browser.

Fully describing HTML (and, later, XML) is beyond the scope of this article but one does not need too much information to understand Fig. 1. In HTML, <table> and </table> are the tags used to start and end a table, respectively. Similarly, <tr> and </tr> begin and end a row within the table. Finally, <td> and </td> begin and end a cell within a row. So, the table in Fig. 1 is composed of 2 rows, each containing 3 cells. The content inside each cell is centered and the entries in the first row of cells are displayed in bold font (turns bold on and turns bold off). Also, ^{and} turn superscripting on and off.

Notice that the tags such as <table>, <tr>, and <td> do not describe the content, i.e., Critical Current Density is not identified as a property and Magnetic Field and Temperature are not identified as measurement parameters. The tags merely tell the browser how to display the content, as seen in Fig. 2. That the tags do not help identify the content, in this case, materials property data, is a serious drawback for those wishing to use the data in computer applications such as simulation and CAE software. One would need detailed knowledge about the data as well as the containing document's structure in order to write a computer application that could extract the data for subsequent processing.

Worse yet, even if data interpretation as described were not an issue for one source of

data, it is very likely that interoperability issues would arise if one were to apply the same data interpretation rules to a different source of materials property data.

The MatML effort is addressing these problems of interpretation and interoperability through the development of an extensible markup language (XML)[4] for materials property data that will allow a person or a computer application to interpret and use the data regardless of the source.

What is Extensible Markup Language?

Fig. 3 represents the markup for the data in Fig. 2 if one were to use the current working draft of MatML[5]. Unlike the code fragment in Fig. 1, in Fig. 3 it is clearly easier to understand what each datum is and one can almost see how each datum is related to

```
<Property>
  <PropertyDescription>
    <PropertyName>Critical Current Density</PropertyName>
    <PropertyUnits> kA/cm<sup>2</sup></PropertyUnits>
  </PropertyDescription>
  <PropertyValue type="integer">3040</PropertyValue>
  <Parameter>
    <ParameterName>Magnetic Field</ParameterName>
    <ParameterValue type="integer">0</ParameterValue>
    <ParameterUnits>T</ParameterUnits>
  </Parameter>
  <Parameter>
    <ParameterName>Temperature</ParameterName>
    <ParameterValue type="integer">3</ParameterValue>
    <ParameterUnits>K</ParameterUnits>
  </Parameter>
</Property>
```

Figure 3. Possible markup of Fig. 1 data using MatML

others without any other assistance. Part of the power of XML is that it permits the definition of domain-specific tags that allow the data content to become self-describing. As a consequence, interpretation of the data is streamlined and amenable to automation. So, what is XML? For an extensive introduction, the interested reader is referred to

"XML, Element Types, DTD's, and All That[6]" on the MatML web site. In short, like HTML, XML is derived from the Standard Generalized Markup Language[7] (SGML) that was approved in 1986 (ISO 8879:1986) by the International Organization for Standardization as a standard way of marking up electronic documents for subsequent processing by applications such as automated typesetting.

HTML is the small, fixed set of tags that in large part contributed to the explosion of the World Wide Web. Humans are naturally gregarious and HTML provides a simple, easy-to-learn language that allows anyone to establish a web presence. Yet, the very strength of HTML, a small easily learned tagset, can also be viewed as a weakness.

HTML is maintained centrally by the World Wide Web Consortium (W3C) which, in and of itself, is good because it allows millions of users worldwide to "converse" using a "universal language." However, HTML is a very specific and limited application of SGML. It does not support flexibility for those in a given community to establish their own language for their own purposes, i.e., HTML is not extensible. Furthermore, HTML is used for describing a document's form rather than its content. This is important to note since communities of web users rapidly reached the point where they wanted to move beyond the *dissemination* of information into the *exchange* of information.

Recognizing the shortcomings of HTML and the needs of the user community for extensibility, the W3C embarked on an effort to bring an abridged implementation of SGML to the Web. SGML is a metalanguage, meaning it is a language for describing languages and is well suited for allowing one to design one's own markup language. The result of the W3C's effort was the 10 February 1998 release of the version 1.0 specification of XML, the e**X**tensible **M**arkup **L**anguage.

XML brings the three principal features of SGML to the Web.

1. ***Extensibility***: users can define their own tags and attributes used in their documents

2. ***Structure***: users can define their own document type definition (DTD) which is the information model of a document describing how the tags and attributes are combined

3. ***Validation***: users can test the conformance of their documents to the structure defined by the DTD.

Additionally, XML intentionally does not include some of the complex functionality of SGML. It was designed to be easy to learn, easy to write, easy to interpret, and easy to implement; characteristics perfectly suited for use on the Web.

There are now many applications of XML in use by different communities. Two well-known examples in the sciences are MathML[8] and CML[9, 10]. MathML was developed by mathematicians to bring mathematical notation to the Web and CML, the Chemical Markup Language, was designed to manage chemical information. MatML (Fig. 4) was launched in October 1999 to develop a markup language for materials property data.

Strategy

The strategy for MatML's development encompasses 6 steps:

1. Establishment of a working group

2. Delineation of the scope and specifications for MatML

3. Development of the formal MatML document type definition

4. Development of a catalog of examples

5. Application development and acceptance testing

6. Dissemination

The MatML Working Group has been established and represents a cross section of the

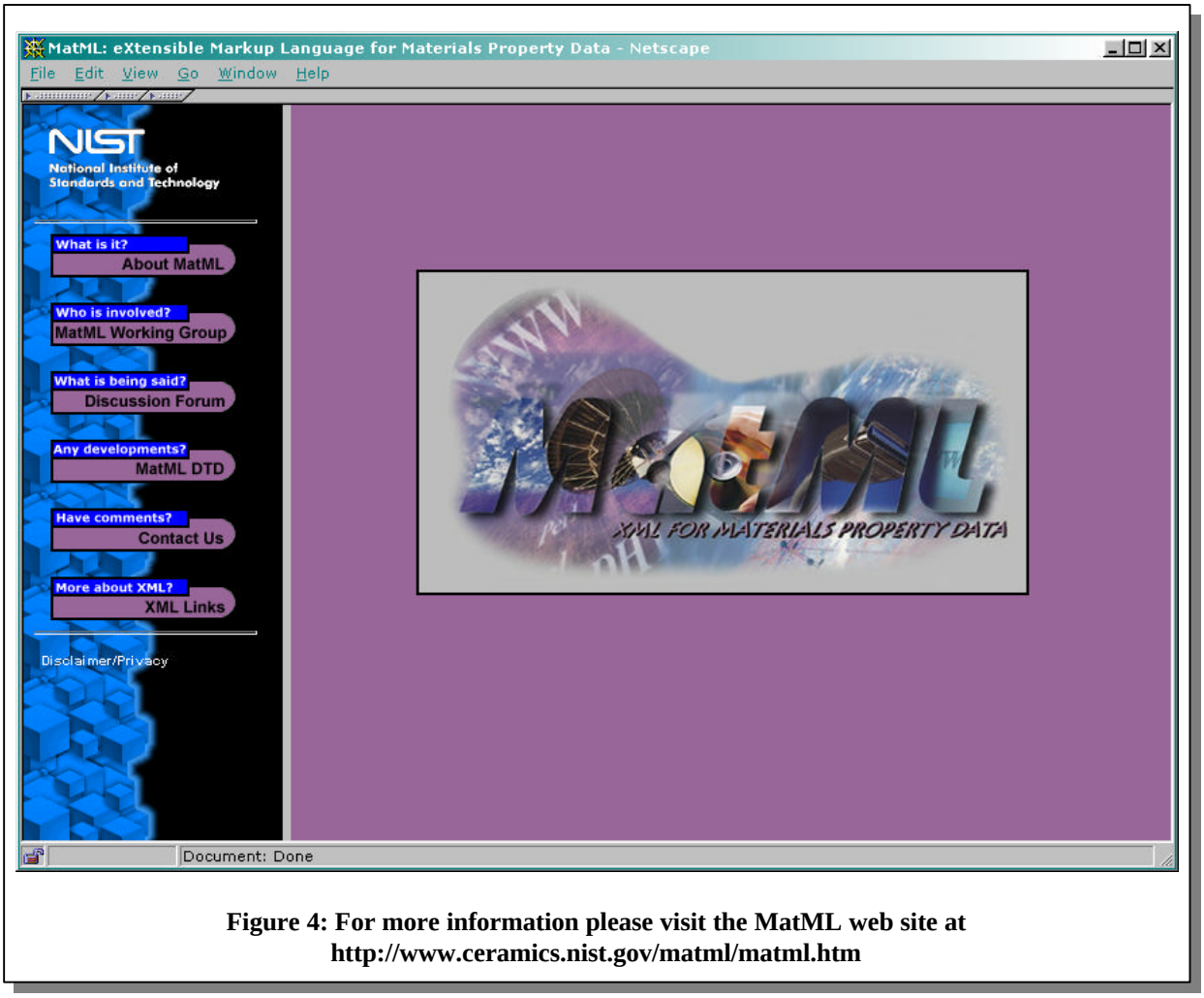

Figure 4: For more information please visit the MatML web site at http://www.ceramics.nist.gov/matml/matml.htm

materials community with members from private industry, government laboratories, universities, standards organizations, and professional societies. The Working Group uses an online forum for discussing issues such as the scope of and specifications for MatML and has recently produced a working draft of the document type definition for the markup language. The MatML DTD contains structures for transferring information concerning the material and its properties, terms that may help with the interpretation of the transferred data, and graphs. The DTD is the XML semantic and syntactic formalism that software will need to parse, interpret, and use the data contained in MatML documents. The next task after the DTD is prepared will be to apply MatML to materials property

data and to build a catalog containing these examples. The catalog will demonstrate how to prepare MatML documents and will also reveal opportunities for revision of the DTD. Once the MatML DTD moves beyond the draft stage, it will enter the next critical phase of application development and acceptance testing. This phase will be crucial for demonstrating that MatML works, testing its robustness, and serving as a guide for others interested in using MatML. Lastly, the strategy for MatML's development includes its dissemination through a variety of standards and professional organizations.

---

[1] "Databases Rule!," by C. Grant, Advanced Materials & Processes, Vol. 158, No. 2, 2000, p. 6
[2] "Directory of Materials Property Databases," Special Supplement to Advanced Materials & Processes, 2000
[3] For more information: NIST Ceramics WebBook, www.ceramics.nist.gov/webbook/webbook.htm
[4] For more information: World Wide Web Consortium XML Pages, www.w3.org/XML/
[5] For more information: MatML DTD Initial Working Draft, www.ceramics.nist.gov/matml/matmldtd.htm
[6] "XML, Element Types, DTD's, and All That," www.ceramics.nist.gov/matml/allthat.htm, 2000
[7] *GCA Standard 101-1983 Document Markup Metalanguage: GENCODE™ And The Standard Generalized Markup Language (SGML)*, Graphics Communications Association, Arlington, VA, 1983, ISBN:0-89740-224-8.
[8] For more information: MathML, www.w3.org/Math
[9] "Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles," by P. Murray-Rust and H. Rzepa, J. Chem. Inf. Comput. Sci., Vol. 39, No. 6, 1999, p. 928-942
[10] For more information: CML Pages, www.xml-cml.org