The Background and Development of MatML, a Markup Language for Materials Property Data

E.F. Begley and C.P. Sturrock
National Institute of Standards and Technology
Gaithersburg, Maryland 20899
begley@nist.gov; sturrock@nist.gov

1. Abstract

MatML is an extensible markup language (XML) for the management and exchange of materials property data. Launched in October 1999 and coordinated by the National Institute of Standards and Technology, an agency of the U.S. Department of Commerce, the MatML project has drawn upon the expertise of an international group of materials scientists and engineers from private industry, government laboratories, universities, standards organizations, and professional societies. The background and development of MatML is described including a discussion of its features and its relationship to other scientific markup languages.

2. Background and Development

Introduction

On 01 October 1999, the National Institute of Standards and Technology (NIST) initiated the development of MatML[1], a markup language for the management and exchange of materials property data. At NIST, the need for such a markup language became evident in May 1997 during the development of the Ceramics WebBook[2], which includes databases of evaluated data for advanced ceramics, a guide to materials data centers and sources, and a collection of materials-related tools and resources. It was during the May 1999 ASTM-NIST Workshop on Materials Data in the Internet Age, however, when it became clear there were many independent realizations that a materials data markup language was needed. A few months later, NIST addressed this need by launching the MatML development project. The first order of business was the assembly of an international working group composed of materials data experts. The MatML Working Group's central activities were to determine the scope of MatML's application and to formulate its vocabulary and syntax. There was an especially strong emphasis during the working group discussions that: MatML ought to be flexible in the scope of its application, i.e., that it could be used to markup a wide variety of materials data from a broad range of sources; MatML's vocabulary should be materials-oriented and easy to learn; and MatML's syntax must be extensible to accommodate both alternative and unanticipated forms of materials data and their representation.

---

[1] See: http://matml.nist.gov
[2] See: http://www.ceramics.nist.gov/webbook/webbook.htm

Perhaps the best way to describe MatML is by example and in contrast to HTML. Table 1 contains an excerpt of data drawn from the NIST High-Temperature Superconducting Materials Database (WebHTS)[3] as it would appear when rendered by a web browser. In this excerpt, 3040 is the value of the measured property, Critical Current Density, whose units are given in kA/cm^2. The measurement parameters are the Magnetic Field, value of 0, and Temperature, value of 3, with units T and K, respectively.

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

Table 1: Tabular Display of Data Excerpt from WebHTS

Figure 1 contains the HTML code for this table, with the tags shown in italics and the data shown in normal text. Note that the HTML tags specify only how the data are to be formatted for display and convey no description of the data themselves. While this aspect of HTML has enabled the publication of billions of web documents, it, nonetheless, represents a serious drawback to those who wish to automate the processing of the data contained in those documents. To address this problem, the Extensible Markup Language[4] (XML) was developed and released by the World Wide Web Consortium (W3C)[5], an international group of approximately 500 member organizations dedicated to developing common protocols that promote the evolution of the Web and ensure its interoperability.

```
<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>
```

Fig. 1: HTML code for Table 1

XML enables communities, such as materials scientists and engineers in the case of MatML, to define their own domain-specific tags and document structure, i.e., they can create their own languages for data management and exchange that, in turn, permit efficient parsing and interpretation of those data via software. As an example, Figure 2 contains the excerpt for Table 1 in MatML markup with the tags shown in italics and the data shown in normal text.

[3] See: http://www.ceramics.nist.gov/srd/hts/htsquery.htm
[4] See: http://www.w3.org/XML/
[5] See: http://www.w3.org/Consortium/

The descriptive nature of the MatML tags, such as <*Properties*>, <*Name*>, and <*Units*> is plainly evident. Note also that the MatML code for Table 1 includes other information about the pedigree of the data, as indicated by <*DataSource*> and <*DataType*>. The descriptive nature and extensibility of the language renders it far more intelligible than non-descriptive fixed tagsets such as HTML. The complete MatML tagset is shown in Table 2.

```
<Properties>
   <PropertyDetails>
      <Name>Critical Current Density</Name>
      <Units>kA/cm<sup>2</sup></Units>
      <DataSource>Journal</DataSource>
      <DataType>Evaluated</DataType>
   </PropertyDetails>
   <Value>3040</Value>
   <Parameters>
      <Name>Magnetic Field</Name>
      <Value type="integer">0</Value>
      <Units>T</Units>
      <Name>Temperature</Name>
      <Value type="integer">3</Value>
      <Units>K</Units>
   </Parameters>
</Properties>
```

Fig. 2: MatML code for Table 1

At the same time, MatML defines a coherent and consistent document structure for its tags, which ensures that any programming language can be used to parse and process the data in whatever manner required. The formal specification for MatML, which covers syntax and semantics, is described in the Annotated MatML Document Type Definition (DTD) Version 2.0[6,7]. The DTD includes extensive descriptive information as well as sample markup.

Associate	Elements	Processing
Associations	Form	Properties
BulkDetails	Formula	PropertyDetails
Characterization	Geometry	Qualifier
ChemicalComposition	Graphs	Relationship
Class	Material	Result
ComponentDetails	MatML_Doc	Shape
Compound	MeasurementTechnique	Source
Concentration	Name	Specification
DataSource	Notes	Subclass
DataType	Orientation	Terms
DimensionDetails	Parameters	Units
Dimensions	PhaseComposition	Value

Table 2: MatML DTD Version 2.0 Tagset

MatML Workshop

After nearly two years of development activity, a workshop[8] was convened at NIST during June 2001 for members of the materials data community to assess openly the

[6] Important note: As this paper is being prepared, the data content model for MatML is being revised and the representation of the language is being ported from XML DTD to XML Schema format. The revised specification, MatML Schema Version 1.0, will be available on the MatML website in early autumn 2002.

[7] See: ftp://www.ceramics.nist.gov/matml/MatMLv20.pdf

[8] Sturrock, C.P., Begley, E.F., and Kaufman, J.G., "MatML - Materials Markup Language Workshop Report," NISTIR 6785, National Institute of Standards and Technology, Gaithersburg, MD, August 2001 [ftp://www.ceramics.nist.gov/matml/workshop.pdf]

progress-to-date as well as to provide valuable input on the technical and strategic future of MatML. In brief, the main findings of the workshop were as follows:

- The development of MatML has been met with great interest, with frequent visits to the MatML website, downloads of the DTD, and inquiries to NIST and other parties that have contributed to the development of the language;

- The extensibility of MatML makes it possible to address the full range of materials information needed by materials researchers, specialists, designers, and quality control analysts, as well as the full range of materials-related applications used by these professionals;

- MatML markup is relatively easy for materials scientists and engineers to understand, as the tags consist largely of materials and engineering terminology (see example above), which has the key advantage of enabling materials specialists to review the markup in detail to ensure that all of the required parameters involving test methods, quality, reliability, and applicability of the data have been addressed;

- A good set of illustrations demonstrating the power and flexibility of MatML should be established in order to advance the widespread adoption of MatML;

- MatML must be tested thoroughly, involving individuals at various stages in the material development/application cycle ranging from materials specialists who select materials to designers responsible for addressing load-carrying capacity and failure limits. In addition, the needs of materials scientists exploring the fundamental structure and properties of materials, and of journal publishers disseminating materials information, must also be addressed;

- At the appropriate time MatML should be recognized officially and formally via registry with XML repositories such as xml.org and/or by standards organizations such as the Object Management Group, ASTM, or ANSI;

- The overall effort should be linked with key industries, that might, via some of their applications, be in position to provide good tests for MatML and, assuming success, encourage or perhaps even require users of their information systems to use MatML as the data exchange medium.

To address some of the key strategic needs among these findings, the MatML Steering Committee (Table 3) was formed during the autumn of 2001. The principal activities of the committee have been promotion of MatML including information dissemination, coordination of acceptance testing activities, and exploration of opportunities for MatML's long-term stewardship.

J.G. Kaufman, Chairman, the Aluminum Association (ret.)
E.F. Begley, Lead Technical Expert, NIST
F. Cverna, ASM International
C. Grethlein, AMPTIAC
S. McCormick, ESM Software
D. Mies, MSC Software
C. Nunez, CenTOR Software Corporation
C. Sturrock, NIST
M. Sullentrup, Boeing Company

Table 3: MatML Steering Committee (2002)

The MatML Steering Committee represents the interests of a much larger community of materials and computing professionals interested in MatML and its objective of automating the exchange of materials property data over the Web. This larger community is in turn kept abreast of developments via the MatML website and dedicated mailing list (for information contact begley@nist.gov).

Issues and Activities

At the time of this writing, the data content model and the XML representation of MatML are being revised. The revisions will be available on the MatML website as MatML Schema Version 1.0 during early autumn 2002. Several aspects of the content model are being refined to resolve technical issues concerning composite materials, data redundancy, and data quality and reliability. In addition, the MatML specification is being reformatted using XML Schema[9] rather than XML Document Type Definition. The reasons for the port to XML Schema are many but primarily include obviating notable shortcomings of DTD's and, in anticipation of MatML's acceptance testing phase, providing an improved language for software development. XML Schema, for example, provides robust support for data typing, including user-defined data types, whereas DTD's treat all data content essentially as text. Support for data types makes it easier to define permissible data values and other restrictions, to validate data, and to convert data between different data types, all of which are, among other applications, essential for working with data from databases. XML Schemas, furthermore, are written in XML whereas DTD's are not. This is especially significant since it means that: XML editors and parsers can be used to edit and parse Schema files; Schemas can be manipulated using the XML Document Object Model (DOM) (The DOM views the document as a collection of objects that can be manipulated, i.e., items can be added, deleted, and modified programmatically[10].); and Schemas can be transformed using XML stylesheet language transformations (XSLT) (XSLT is used to transform and format documents for display[11].)

With respect to acceptance testing, there are two principal kinds of XML applications: document publishing and document processing. Document publishing, which is the manipulation of information for human consumption, by necessity, comes first and will

---

[9] See: http://www.w3.org/XML/Schema
[10] See: http://www.w3.org/DOM/
[11] See: http://www.w3.org/TR/xslt

exhibit the strengths and weaknesses of MatML. In a sense, acceptance testing is already underway with the development of a compendium of markup examples, but wider participation is needed in order to assess MatML's application to a broader and deeper range of materials property data. Future NIST testing might include markup of NIST WebSCD[12], NIST WebHTS, and NIST WebPDS[13], which are online compilations of structural ceramic, high temperature superconductor and property data summaries, respectively.

Document processing, which is the manipulation of information intended for software consumption, represents the second phase of MatML's acceptance testing. While no specific plans have yet been identified, several projects are already exploring the use of MatML, including:

- Green's Function Research and Education Enhancement Network[14] (Kent State University, Massachusetts Institute of Technology, NIST);

- femML[15] – the finite element modeling markup language (J. Michopoulos, U.S. Naval Research Laboratory).

Another important issue is technical education. It is not sufficient to publish MatML simply as a language specification. It is necessary to provide additional resources to promote understanding and adoption of the new language. This philosophy was behind the publication of MatML in an annotated format on the MatML website, where, in addition to the syntactic and semantic formalism of the language, English language explanations are provided as well as sample markup. Examples of other supporting resources for MatML found on the MatML website include links to sources of information about XML, and related expositions such as "XML, Element Types, DTDs and All That[16]." Finally, the compendium of markup examples will be extremely valuable for assisting users with learning how to apply MatML to their own data.

Finally, in addition to technical matters there are two main strategic issues for MatML: promotion and outreach, and MatML registration. Direct communications, journal publications, the MatML website, conference presentations, and the MatML workshop proceedings can be used to reach data providers and consumers such as journal editors, handbook and database publishers, materials' suppliers (data sheets), instrument manufacturers, and researchers. Registering MatML with the www.xml.org registry as well as the proposed NIST registry/repository for XML also should provide broader exposure and access to MatML.

---

12 See: http://www.ceramics.nist.gov/srd/scd/scdquery.htm
13 See: http://www.ceramics.nist.gov/srd/summary/advmatdb.htm
14 See: http://shreve.mcls.kent.edu/NSDLGreen/GREENProject.htm
15 See: http://www.istos.org/femML/
16 See: http://matml.nist.gov/allthat.htm

MatML's Relationship to Other Scientific Markup Languages

As an extensible markup language, MatML permits the use of non-MatML tagsets within a document, satisfying the requirement put forward by the MatML Working Group that the language must be extensible. Accessing non-native tagsets is accomplished through the use of XML namespaces.[17] In brief, non-technical terms, an XML namespace can be thought of as a set of uniquely named tags that are associated with an owner. MatML, in fact, incorporates a namespace, SVG[18], the Scalable Vector Graphics markup language, within its specification for marking up graphs. It was recognized early during MatML's development that graphs are commonly used to represent materials property relationships and tags would be needed to mark up the information contained in those graphs. Rather than expending valuable resources redefining a set of tags for marking up graphs, the MatML Working Group decided to leverage the energy, time, and standard of the SVG development community and formally included SVG as a namespace within MatML. The Annotated MatML DTD Version 2.0 contains sample markup for the phase diagram of sulfur and illustrates the use of the SVG namespace for inclusion of 2D graphical images within a MatML document.

Namespace usage within MatML is, of course, not limited to SVG. There are now many extensible markup languages available[19] including those for scientific disciplines. Table 4 shows some markup languages that could serve as namespaces for MatML. It is also important to recognize that MatML itself can serve as a namespace for other markup languages; the femML effort, for example, is exploring the use of MatML for marking up composite materials data that are processed subsequently by femML software.

Markup Language	Description	Developer(s)
MathML[20]	Mathematical Markup Language for encoding mathematical notation and content	W3C
CML[21]	Chemical Markup Language for the management of chemical information	P. Murray-Rust (Virtual School of Molecular Sciences, UK) H. Rzepa (Imperial College, UK)
UnitsML[22]	Units Markup Language for encoding measurement units (development in progress)	F. Olken and J. McCarthy (Lawrence Berkeley National Laboratory) R. Dragoset, B. Taylor, and M. McLay (NIST)
SpectroML[23]	Molecular Spectrometry Markup Language for the interchange of analytical chemistry spectroscopy data (development in progress)	M. Rühl (University of Applied Sciences Wiesbaden, Germany) M. Peschke and G. Kramer (NIST)

Table 4: Some extensible markup languages that could be used as namespaces for MatML

[17] See: http://www.w3.org/TR/REC-xml-names/
[18] See: http://www.w3.org/Graphics/SVG/Overview.htm8
[19] See: http://www.oasis-open.org/cover/siteIndex.html#toc-applications
[20] See: http://www.w3.org/TR/2001/REC-MathML2-20010221/
[21] See: http://www.xml-cml.org/
[22] See: http://physics.nist.gov/TechAct/Div840/ecsed.html
[23] See: http://www.mel.nist.gov/div826/msid/sima/02_instrmnt_chemrefdata.html

3. Conclusion

At present, the materials data "marketplace" can be very chaotic and difficult for both providers and consumers. Until the arrival of MatML, there was no common exchange format but instead hundreds of proprietary formats that resulted in wasteful duplication of effort and poor scale-up. Also plaguing the materials data community was (is) the absence of software interoperability, whereby one computer program, such as a finite element code, can automatically input material properties from another computer program or database without the necessity of human intervention. The present situation overall yields very inefficient data processing.

MatML will serve the materials data marketplace as a common, public-domain materials data exchange format - a non-proprietary and generic language for materials property data. With considerable built-in flexibility and extensibility as attributes of the language, MatML will provide for direct program-to-program interoperability, efficient data processing, and rapid, reliable, and useful response to searches for materials data over the Web. Its widespread adoption will provide for powerful and malleable connection of the various materials data sources found on the Web today, as well as those to be added to the Web tomorrow.