Metadata Structure for Geotechnical Physical Models (and Simulations?)

Bruce L. Kutter
University of California, Davis

Daniel W. Wilson
University of California, Davis

J.P. Bardet
University of Southern California

Note: this paper is related to one that the above authors submitted to the International Conference on Physical Modeling in Geotechnics (Kutter et al. 2002). This paper was updated based on a breakout group discussion at the International Workshop on Earthquake Simulation in Geotechnical Engineering held at Case Western Reserve University, November 8-10, 2002.

ABSTRACT

Advances in data volume associated with sensor and data acquisition technology and the increased complexity and value of model test data make it clear that there is a real need to implement better, flexible, and more general standards for archiving data. The new standards should make it easy to create, edit, share, query, and search through databases and they should facilitate comparison of numerical simulations with experimental data. This paper presents a strawman data structure for physical model test results; perhaps a similar structure would be useful for numerical simulation results. It is anticipated that that this structure will be used to establish a Model Test Markup Language (MTML). XML (eXtensible Markup Language) provides an ideal syntax for MTML. The goal of this paper is to stimulate convergence toward an accepted structure of metadata.

INTRODUCTION

An important issue facing the geotechnical modeling community is how to properly document and archive data sets for use by other researchers. Sufficient *metadata*, defined as *data about data* must be archived along with the data to make it useful to others. This data archiving issue has come to the forefront of the National Science Foundation NEES (George E. Brown Jr., Network for Earthquake Engineering Simulation) program. A goal of the NEES program is to link several large scale experimental facilities and researchers for real time interaction through a high performance internet and to provide general access to curated data archives.

NEESgrid (www.neesgrid.org), the system integrator of the NEES program is planning to characterize the Earthquake Engineering community use of data and metadata, and future requirements in January, 2002, distribute preliminary metadata standards, data models and representations for review in May 2002, and publish recommended standards for data and metadata models and representations by September 2002 (Prudhomme and Mish, 2001). To have an influence on these standards, the community should begin to organize its thoughts now. This workshop has presented an opportunity to initiate discussions on metadata structure and

data archives. The geotechnical modeling community in general may benefit from, and have a positive influence on these standards if we begin to organize and express our metadata needs. With advances in data acquisition and sensor technology, it is becoming possible to include more and more sensors in experiments. Sound and image data from photographs and audio-video recordings are also increasing exponentially. The increase in data volume is encouraged by new hardware that makes it possible to store and process larger quantities of data. At UC Davis, we are routinely monitoring 100 channels of data in one model test. Each test series usually involves several different types of events: simulated earthquakes, spin-up, penetration tests, consolidation, cyclic loading with an actuator. From each event we may have data from transducers (pore pressure, acceleration, displacement, strain gauges), film, digital and video cameras, and hand-written data in a laboratory notebooks.

Data from several series of experiments at Davis is archived with metadata and can be freely downloaded (CGM 2001). The data archives at this site are quite complete and they include text descriptions of the tests and results, sketches of model configurations, spreadsheet tables of sensor numbers, channel numbers, amplifier gains, tables that show the sequence of testing, many pages of plots of data from several sensors in several events, and ASCII files of raw and processed data from the experiments. Despite our attempts to standardize our own data report formats, we have found that our internal standards are continuously evolving.

Figure 1 depicts the flow of metadata and data from sensors to an archiver. This indicates that we anticipate automatic formatting of much of the metadata by to-be-defined software represented by the Metadata Generator at the top left of Figure 1.

A strawman metadata hierarchy is proposed to be implemented in the framework of XML (Extensible Markup Language) (e.g., O'Reilly 2001). Progress toward metadata standards also depends upon a change in the culture of academia; credit must be given for establishing archival databases, just as credit is given for publication of archival journal articles.

BACKGROUND STORY

Graduate student #1 wants a Master's degree and is recruited by University X because Professor A offers financial support to work on a research project involving centrifuge model tests. Student #1, a good student, performs the tests, coauthors a conference paper, and then leaves a notebook, a set of electronic data files, and a stack of photographs with Professor A. Professor A is a junior faculty member in need of journal publications to achieve promotion. The merit review committee puts almost no value on the establishment of a detailed archive of the data and metadata. Thus Professor A may refuse to release the data to Professor B before he has a chance to publish the data himself.

Student #1 then disappears to a job in industry, perhaps with great intentions to coauthor a journal article. Two years pass and the paper is not submitted. Then Professor A decides to work on the paper himself or to ask a new student (Student #2) to re-evaluate Student #1's data or to analyze it by a new procedure. The data files were written using a different program and different operating system. One of the files is missing, and he cannot figure out if Accelerometer

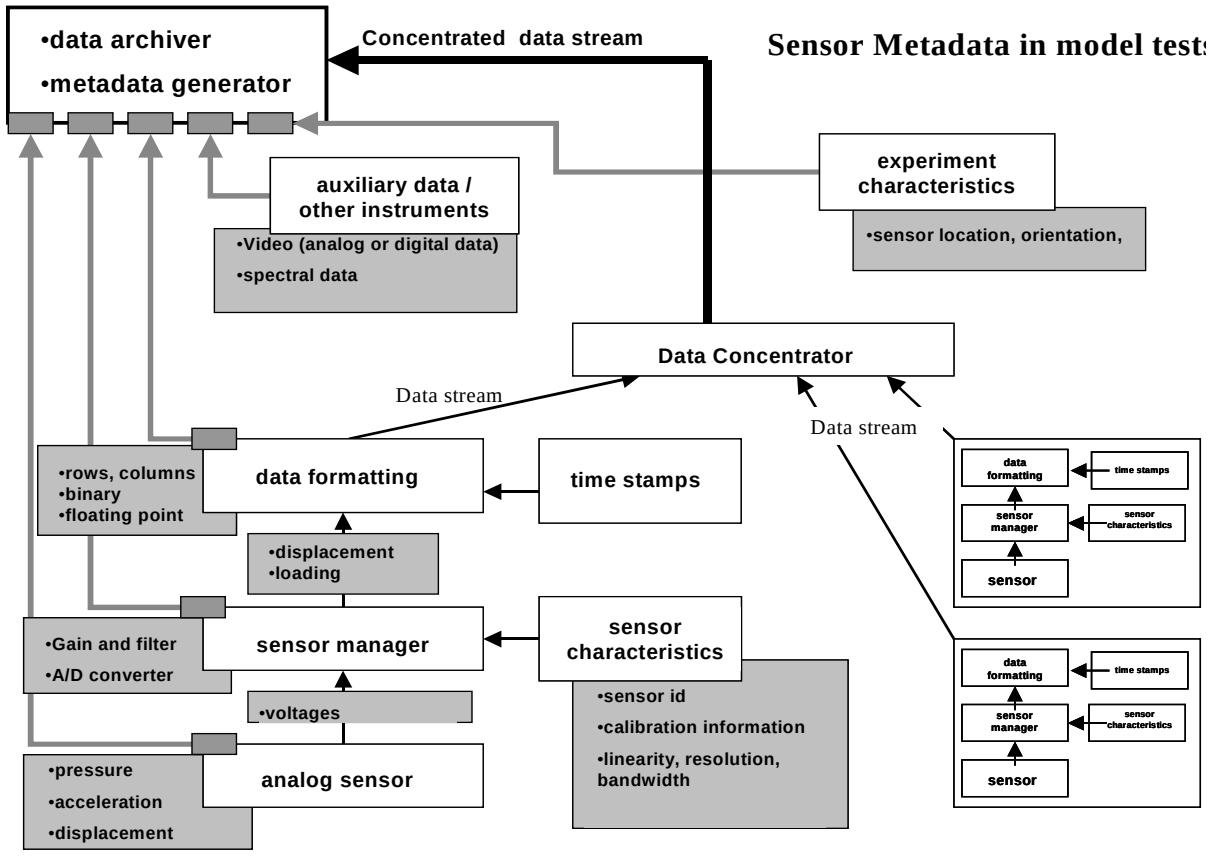

Figure 1. Sensor metadata and sensor data flow.

46 was plugged into Amplifier Channel 14 or 19 because the 4 looks like a 9 in Student #1's hand written data book. Professor A gives up and never publishes the data in an archival journal.

Furthermore, it is often the case that a model test or a series of numerical simulations will generate far more data than can be archived in a conventional paper. Unpublished data that is not properly archived is doomed to disappear.

The above examples illustrate the importance for more formal data archives that help us maintain valuable databases.

SUMMARY OF ANTICIPATED USES OF DATA AND METADATA

Before proceeding to develop a metatdata standard, it is useful to summarize the potential uses of the data and metadata. Some of these are:

1. Document procedure to permit duplication of the work
    a. by the same researcher
    b. by other researchers
2. Real time interaction with experiment by remote researchers
3. Numerical simulation of experimental data

    a. interactive decision making during experiment
    b. years after the test
4. Automated control (feedback control) of the experiment
5. Visualization
    a. research, education, sponsors
6. Data search/query filter
7. Artificial Intelligence, inverse/system identification
8. Facilitation of software sharing by common interface (e.g., opensees)

Many of the above uses of data and metadata require much of the same metadata.

STRAWMAN METADATA STRUCTURE

In order to take a step toward the creation of metadata standards, a "strawman" of metadata structure and content is proposed in outline form in Table 1. An electronic copy of this outline can be found at http://cgm.engr.ucdavis.edu/NEES/mtml.

Section 1 of the outline in Table 1 contains metadata associated with the research project.

Section 2 is a catalog of physical objects used to construct or test the model. This includes: apparatus used to test the model, passive materials and markers that are placed in the model, and sensors that are used in the model tests.

Section 3 describes sequencing of events. A sequence can be the measurement of the location of an object, or an event involving activation of an actuator or a penetrometer sounding.

Section 4 includes the sensor-channel-gain lists; this documents which sensors are plugged into which amplifier channels, and also includes the sequence in which the sensor data was recorded, and parameters that define gains and filters.

Section 5 describes image data. This could include photographs, video camera data, and/or engineering drawings of configuration.

Section 6 describes the data required to control the experiment. This could, for example, determine the location of a CPT sounding, the rate of penetration of a penetrometer, or command files to control a shaker.

Some of the items in Table 1 are expanded into greater detail than others. The community should review this outline for completeness, lack of redundancy, logical groupings. Certain of these categories of metadata may not be applicable to all metadata sets and therefore, should be considered optional in a metadata record. Other outlined metadata types may have multiple entries. For example in Table 2, at 2.2.1, we have indicated "soil deposit (1)". The (1) in parentheses is to emphasize that there could be an array of soil deposits. The numerical order of the items is not meant to imply any ranking of importance.

Table 1. Strawman metadata structure for geotechnical model tests

Modified 9/26/01, 11/07/01, 11/10/01 - BLK

Model Test

1. **Project Identifiers**
    - 1.1. report title
    - 1.2. authors, publishers
        - 1.2.1. address book link
    - 1.3. date of report
    - 1.4. acknowledgements:
        - 1.4.1. sponsors
        - 1.4.2. others
    - 1.5. conditions and limitations
    - 1.6. purpose of project
    - 1.7. purpose of model test
2. **Catalog of Materials, Objects, Sensors and Apparatus**
    - **2.1. apparatus used to test model**
        - 2.1.1. centrifuge ID and metadata
        - 2.1.2. container ID and metadata
        - 2.1.3. actuators ID and metadata
        - 2.1.4. shaking apparatus ID and metadata
        - 2.1.5. cone penetrometer ID and metadata
    - **2.2. materials placed in the model**
        - 2.2.1. soil deposit (1)
            - 2.2.1.1. deposit ID
            - 2.2.1.2. material source or supplier,,USCS classification
            - 2.2.1.3. method of preparation
                - 2.2.1.3.1. density and water content during preparation
                - 2.2.1.3.2. method of saturation
            - 2.2.1.4. shear strength parameters
            - 2.2.1.5. index tests: emax, emin, grain size, specific gravity
        - 2.2.2. structure (1)
            - 2.2.2.1. structure ID
            - 2.2.2.2. mass, dimensions, location of CG, orientation, mass moment of inertia
        - 2.2.3. markers (1)
            - 2.2.3.1. colored sand layers
            - 2.2.3.2. lead shot markers
            - 2.2.3.3. fiducial marks
    - **2.3. sensors used in the model test (1)**

2.3.1. manufacturer's serial number

2.3.1.1. sensor type

2.3.1.2. manufacturer

2.3.1.3. manufacturers model or part number

2.3.1.4. calibration information for sensor

2.3.1.4.1. date of calibration, sensitivity coefficients (with units), bandwidth, resolution, range

2.3.1.4.2. scale factor exponent to be used to convert sensor data to prototype data.

2.3.1.4.3. Pointer to more metadata about this sensor

2.4. data acquisition system

2.4.1. amplifiers

2.4.2. filters

2.4.3. A/D converters

3. Sequence of Model Test Events and Measurements

3.1. Event(1)

3.1.1. device ID (e.g., centrifuge, penetrometer, actuator)

3.1.1.1. control data file for this device

3.1.1.2. parameters to scale control file

3.1.2. output data (raw digital data)

3.1.2.1. name of record, *output file name*

3.1.2.2. format (e.g., XML version, 12 bit offset binary)

3.1.2.3. file size, date modified, other file attributes

3.1.2.4. sensor-channel-gain-list (SCGL)

3.1.2.5. sampling frequency

3.1.2.6. time data acquisition begins

3.1.2.7. number of samples (data points per channel)

3.1.2.8. unprocessed data {data 1, data 2, …data n} stored at *output file name*

3.1.3. processed output data record

3.1.3.1. name of unprocessed data file, *output file name*

3.1.3.2. format, file size, date modified, other file attributes

3.1.3.3. processed data record file name, *processed file name*

3.1.3.3.1. processing parameters

3.1.3.3.1.1. apply calibration specified from SCGL

3.1.3.3.1.2. filter algorithms and parameters

3.1.3.3.1.3. scale factors applied

3.1.3.3.2. processed data {t, data 1, data 2, …data n} to be stored in *processed file name*

3.2. location measurement of an object

3.2.1. device used to measure location

3.2.1.1. name of object or site (this could be anything identified in Catalog Section)

3.2.1.1.1. metadata associated with device

3.2.1.1.2. coordinate system used (local or global)

3.2.1.1.2.1. {t,x1,x2,x3, x1', x2', x3'} (time, location and orientation)

3.3. centrifuge speed data

3.3.1.{t, angular velocity}

3.4. earthquake loading data

3.5. shear wave velocity measurement

3.6. dissecting

4. **Sensor Channel Gain Lists(1)**

4.1. sensor ID(1)

4.1.1.cable number

4.1.2.first amplifier number

4.1.2.1. channel number, gain1, filter

4.1.3.second amplifier number

4.1.4.A/D converter

4.1.4.1. channel number

4.1.4.2. gain

4.1.5.intermittent sensor behavior

4.1.5.1. time span of intermittent behavior

4.1.5.2. who noticed the intermittent behavior

4.1.5.3. when they reported it

5. **Image Data**

5.1. photographic data

5.1.1.raw data, compression algorithms, time date stamps, audio descriptor

5.2. video data

5.3. engineering drawings of configuration

6. **Control Data Files**

6.1. Penetrometer

6.2. Robot manipulation

6.3. shaker

METADATA AND AN XML SOLUTION

XML (Extensible Markup Language) is a promising language to document data with appropriate metadata. XML documents look similar to HTML documents. There are two key differences between HTML and XML. Firstly, HTML is used to specify how data is displayed while XML describes the information content of the data. Secondly, XML is extensible (it allows users to create new tags); HTML does not. The XML extensibility will allow us to create our own language, which we could call MTML (Model Test Markup Language) for now; a language we could specially design to help us archive the metadata involved in our experiments. MTML standards could be created by establishing *Document Type Definition* (dtd) files that define the tags and specify the structure of the metadata. Excerpts of an example MTML dtd are provided

by Kutter et al. (2002). Examples of a Mathematics Markup Language and Chemical Markup Language are available at W3C (2001) and Jirat (2001).

XML documents contain ASCII data organized in a tree like structure with categories, sub-categories, and sub-sub-categories, etc. Table 2 shows excerpts of an example of XML document for storing the information about centrifuge sensors. XML documents can have their own structures, or they can follow the rules and syntax defined in external (DTD) files. DTDs can be reused and shared between many XML documents, and can be called through hyperlinks from remote DTD repositories. Graphical User Interfaces to XML data, freely available on the internet, make it possible to view, query, enter, and edit XML data conveniently. Fig.2 presents one such interface, called XML Notepad, which reads data from XML code (such as that shown in Table 2, and allows one to display and edit the XML file. Once there is an established standard for the metadata structure, it is expected that metadata can be automatically or conveniently generated and edited on user interfaces specific to model tests.

New technology such as TEDS (Transducer Electronic Data Sheets) and SCEDS (Signal Conditioner Electronic Data Sheets) will facilitate automatic metadata generation. For example a Transducer with TEDS can tell the data archiver its serial number, calibration factor, etc. A computer can provide the time stamps to the metadata generator according to a specified clock.

Table 2. Excerpts from a XML document, documenting sensor related data. Four segments of code describe the cataloging of the sensor, the location of the sensor, the signal path to the digitizer (SCGL), and the recording of the sensor data.

```
<?xml version="1.0"?>
<ModelTest>
……
   <Catalog>
      <Sensors>
         <Sensor SN="PCB3245">
            <Type>Piezoelectric Accelerometer</Type>
            <Manufacturer>PCB</Manufacturer>
            <Model>352</Model>
            <CalibrationDate>092899</CalibrationDate>
            <Sensitivity Unit="mV/g">100</Sensitivity>
            <Range>50g</Range>
            <SensorData> http://www.pcb.com/pcb3245 </SensorData>
         </Sensor>
      </Sensors>
   <Catalog>


   <Sequence>
      <LocationMeasurement>
         <MeasuringTool> Caliper44
            <Sensor>PCB3245
               <CoordinateSystem1>092890_2.37pm, 45mm, 103 mm, 37mm,
               1,0,0</CoordinateSystem1>
            </Sensor>
```

```
         <Sensor>PCB3246
            <CoordinateSystem1>092890_2.43pm, 48mm, 223 mm, 41mm,
            0,-1,0</CoordinateSystem1>
         </Sensor>
      </Measuring Tool>
      <MeasuringTool> MeterStick45
         <Structure>Structure#1
            <CoordinateSystem1>092890_2.49pm, 101mm, 129 mm, 98mm,
            1,0,0</CoordinateSystem1>
         </Structure>
      </MeasuringTool>
   </LocationMeasurement>
</Sequence>

<SCGL> EQ1
   <Sensor>PCB3246
      <Cable> C492 </Cable>
      <Amplifier1>PVL23
         <Channel>32</Channel>
         <Gain>20</Gain>
            <FilterParameters>100 Hz, 5th order, butterworth
            </FilterParameters>
      </Amplifier1>
      <ADC>DT2839
         <Channel>83</Channel>
         <Gain>4</Gain>
      </ADC>
      <IntermittentBehavior>
         <TimeSpan> T1, T2 </TimeSpan>
      </IntermittentBehavior>
   </Sensor>
</SCGL>


<Sequence>
   <Event> Shake1
      <Shaker>
         <ControlFile>C:/shaker/motions/PortIsland.txt
         </ControlFile>
         <Parameters> 0.4,2.2,39 </Parameters>
      <Output>
         <FileName>C:/shaker/Shake1.out</FileName>
         <DateCreated>11042002 11:34:59</DateCreated>
         <SCGL>EQ1</SCGL>
         <SamplingFrequency> 2073.58</SamplingFrequency>
         <NumberOfSamples> NSamples </NumberOfSamples>
         <BeginRecording> time digitizer starts
         </BeginRecording>
      </Output>
   </Event>
</Sequence>
```

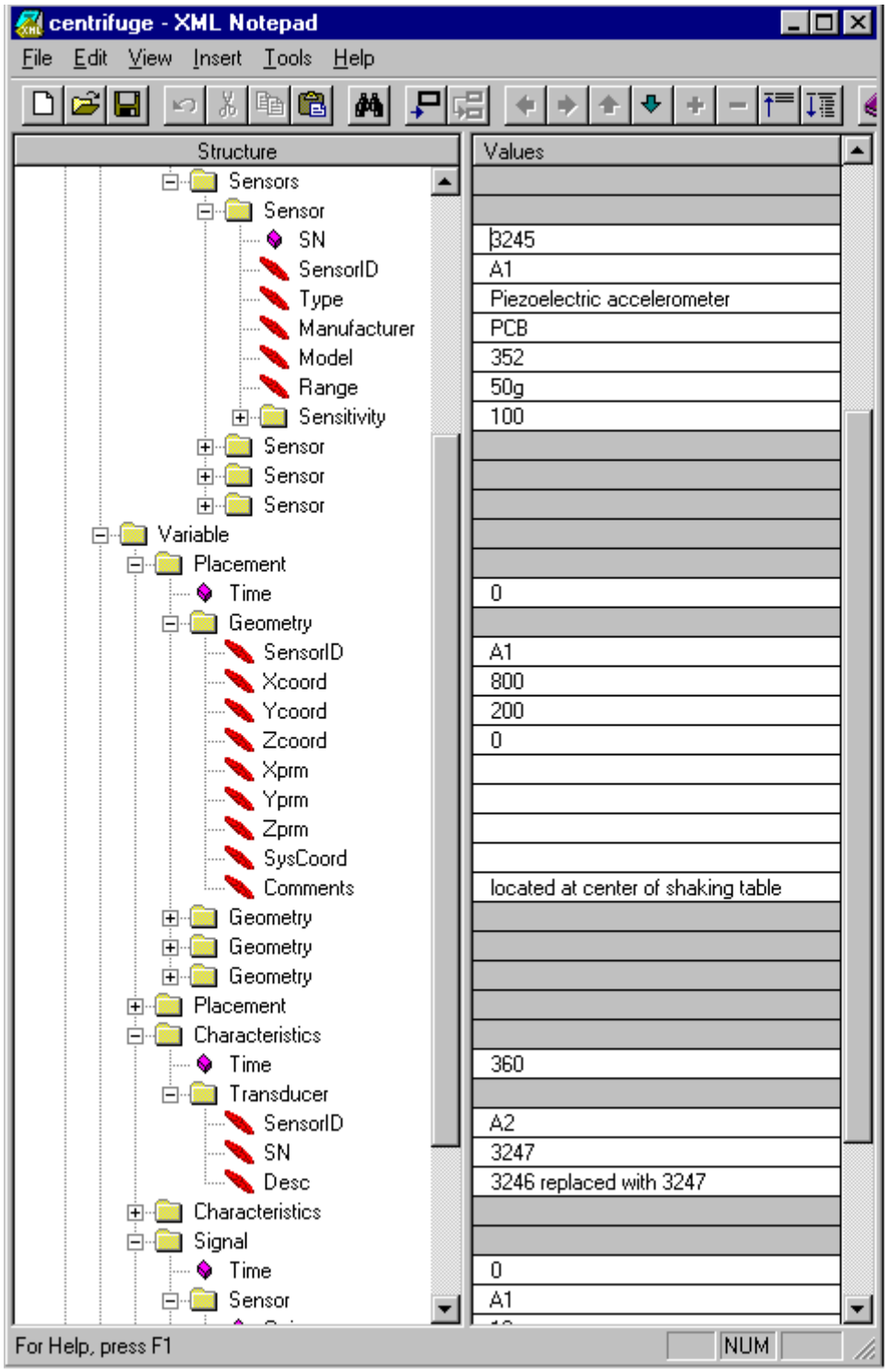

Figure 2. . Example of XML data displayed for entry or editing by *XML Notepad*.

WHICH COMES FIRST, THE CHICKEN OR THE EGG?

On one hand, one may argue that establishment of metadata standards will facilitate sharing of software that enables exploration and sharing of data. On the other hand it may be argued that development of a fantastic software package (the so-called *killer app*), that assumes a metadata standard, would provide incentive for adoption of a metadata standard. If the software application is wonderful enough, people will be willing to invest the time to learn the standard. There are several types of software applications that could serve as incentive for acceptance of a standard and/or develop as a consequence of the adoption of a standard:

1. Convenient portals and user interfaces to test data. These interfaces could facilitate entering, querying, editing and sharing of data and metadata.
2. Automatic real-time compilation of metadata will be facilitated by use of smart sensors and hardware that communicate their configuration and sensitivity to a metadata generator. A progression from paper laboratory books toward electronic laboratory books could also streamline metadata generation.
3. Numerical analysis data, stored in a similar format as the experimental data, could be easily compared to experimental data using portable visualization tools. Portability will be enhanced by standardization.
4. Data published in standard format can be archived with confidence that it will be readable by unknown individuals in the future
5. Software tools that facilitate effective teleparticipation in a remote experiment could also necessitate the adoption of metadata standards that are available to both the local and remote participants. Conversely, the existence of a standard would lead to community development of new teleparticipation tools.

POSSIBLE PITFALLS

Model test procedures and experiment characteristics may be too different from one test series to another or from one research center to another. Future tests may require a new type of metadata that was not considered in the previous version of MTML. The researcher would then need to either create a new subset of dtds. The revised MTML dtd file will need to be archived with the data. If the initial standard is not general enough, frequent MTML revisions could become bothersome and confusing.

There could be a struggle between alternative proposed metadata standards. Different professional societies and organizations may each propose different standards. As long as the standards are open, it is likely that, over time, competing standards may evolve toward each other. Procedures need to be developed for establishment and maintenance of the MTML dtd's.

There will also be a need to establish archival data repositories for the data and metadata. Who will ensure the security of the data, perform quality control checks on the data, and provide access to the archived data? Who will pay for maintaining the data?

CONCLUSIONS

On the path toward design and development of metadata standards we must attempt to think ahead to future uses of the data. The standard must be flexible to facilitate evolution of standards without encouraging proliferation of competing standards. Visualization of large data sets is demanding; making visualization tools convenient and efficient may impose severe constraints on the metadata. One philosophical trade-off in the design of metadata standards is a decision as to how much of the metadata needs to be archived with the data and what subset of the metadata needs to be re-archived every time the data is re-used. If the metadata is copied from its original source there is a potential for one of the copies of the data could be "corrected" or corrupted. Contradictory data archives may evolve.

In the future, advanced testing facilities will generate larger and more comprehensive data sets due to new sensor and data acquisition technologies. The expansion of data volume will require the use of modern data management techniques with query capabilities and secure archival storage. We recommend the development of a "standard" version(s) of MTML tailored to organize the data and metadata generated in geotechnical model tests. This paper presents a strawman outline of a data structure that needs to be reviewed and improved by the community of potential users. The NEES program of NSF will be forcing the issue of formalizing data archives in the US. Computer scientists working with the NEES system integrator (www.neesgrid.org) and the NEES consortium developer www.nees.org, will organize efforts to establish metadata standards and archiving protocols. It will be beneficial to the geotechnical modeling community to begin to contemplate the issues associated with metadata and data archives and to make our collective needs known to the NEES organizations.

ACKNOWLEDGEMENTS

The authors are supported by funding from the NSF George E. Brown, Jr. Network for Earthquake Engineering Simulation (NEES) Program. Kutter and Wilson are funded by award number CMS-0086566 and Bardet through award number CMS-0084529. The authors thank Jianping Hu for his help in generating XML examples. Special thanks also go to the participants in the CWRU workshop breakout discussion group on this topic, in particular Michael Stokes and the session reporter, Stein Sture helped organize and develop the ideas presented herein.

REFERENCES

CGM. 2001. *Center for Geotechnical Modeling Publications,* http://cgm.engr.ucdavis.edu/pubs/pubs.html.

Jirat, J. 2001. *Chemical Markup Language 1.0 reference with examples,* http://www.zvon.org/xxl/CML1.0/Output/.

Kutter, B.L., Wilson, D.W., and Bardet, J.P. *Metadata and Data Archives for Geotechnical Model Tests,* Submitted for publication in the International Conference on Physical Modeling in Geotechnics, to be in St. Johns, Newfoundland, Canada, 2002.

NEES. 2001. *George E. Brown, Jr. Network for Earthquake Engineering Simulation home page,* http://www.eng.nsf.gov/nees/.

O'Reilly. 2001. *XML web page*, http://www.xml.com/.

Prudhomme and Mish, 2001. NEESgrid: A Distributed Virtual Laboratory for Advanced Earthquake Experimentation and Simulation: Project Execution Plan for NEES Systems Integration Component, submitted to NSF, June 2001.

W3C. 2001. *Mathematical Markup Language (MathML) Version 2.0, W3C Recommendation 21 February 2001*, http://www.w3.org/TR/MathML2/.