Development of Geotechnical Data Schema

Revised: 4/6/2005

Purpose: To develop a standard XML schema and data dictionary for geotechnical data. This will be accomplished through a survey of stakeholders to identify their specific geotechnical data needs (at a dictionary description level). The survey results will be used to develop a consensus for a national standard geotechnical data dictionary and to define the national standard XML (GML compliant) data interchange format schema. The achievement of a consensus on the data dictionary is the majority of the effort in creating a standard. Existing standards and efforts will be used as the basis for this effort.

Background: A number of state and federal agencies are developing Geotechnical Databases which may be queried for information used for maintenance of existing projects as well as design, construction, etc. of nearby projects. Unfortunately through the lack of a standard data definition for Geotechnical data, there exists significant difficulty in archiving, reusing and sharing data. This problem has been exacerbated with the dependence on computer software as the standard for design, construction, and maintenance of new and existing infrastructure projects. For instance, numerous computer programs have been developed to electronically collect and present geotechnical in-situ data by hardware manufacturers (i.e. CPT, PMT, etc.), as well as borehole/boring logs or fence diagrams. However, each software suite has developed its own data needs and formats. Some commonality exists in the definition of data through the use of standards such as ASTM, AASHTO, ANSI and others. However, some state DOTs, federal agencies as well as software vendors may have adopted variations of the standards. In order to collect, view, and share geotechnical data there is a great need to establish a Geotechnical Data Dictionary from which a national standard XML (GML compliant) data interchange format schema may be established. The establishment of standards for the development of geotechnical management systems will provide the means for state DOTs to efficiently capture, store, retrieve, and share geotechnical data and information internally as well as with external agencies and user groups. The standards will also provide the means for IT departments and software developers to build components or modules for geotechnical management systems that would meet each state DOT's needs and be compatible with other modules developed by other software providers. These standards will reduce each State's time and cost involving software development, maintenance, and updates. In addition, the standards, if adopted by other local, state and federal agencies, would provide the means to electronically share geotechnical data obtained by other sources near DOT highway projects.

Objectives: A first step in this process is to survey state and federal agencies and their consultants to define their geotechnical field and laboratory testing practices and the types of geotechnical data that they collect, archive and reuse through a web based survey method. Specifically of interest is the type of field and laboratory tests that are routinely preformed, associated data collected, as well as metadata (data describing data: type of equipment, etc). Also of concern is the uniformity of testing practices (i.e. ASTM, AASHTO, etc.), and description of the data (e.g. soil classification, strengths, etc.). The survey will cover data at the dictionary level and will require very detailed and specific

information. From the survey information, develop a consensus of data definitions to be accepted in the standard schema. The data dictionary specifies the meaning of the terms used in the data base.

The second objective involves the development of an open and flexible XML (GML compliant) based data structure and data dictionary geotechnical management systems. The data structure will define the form and content (alpha or numeric) of the data, the precision, the units, the field size, the type of data acquired, other data attributes, and the relationships between the attributes.

.

Scope: The survey and schema development will be a collaborative effort by a **Core Team** consisting of technical representatives from the University of Florida, Department of Civil Engineering (UF), Association of Geotechnical and Geoenvironmental Specialists in the United Kingdom (AGS), Consortium of Organizations for Strong-Motion Observation Systems (COSMOS). Oversight of development by the Core Team will be provided by the **Geotechnical Data Coalition** with representation from UF, AGS, COSMOS, Construction Industry Research and Information Association (CIRIA), Federal Highway Administration (FHWA) and the Ohio Department of Transportation (ODOT). A **Geotechnical Management System Group** (GMS group) composed of representatives from 12 State DOTs, FHWA, US EPA, US Army Corps of Engineers, and the US Geological Survey has been formed to govern the development of the standards for all geotechnical data and to provide all final decisions for this project.

The UF group has developed a data dictionary, database and XML exchange format for the Florida DOT that allows web based sharing of geotechnical laboratory test data, classification data, in-situ test data and as-built construction data. The AGS membership comprises UK organizations and individuals having a common interest in the business of site investigation, geotechnics, geoenvironmental engineering, engineering geology, geochemistry, hydrogeology, and other related disciplines. AGS has a flat file exchange format that has been used for 14 years in the UK, Europe and Asia which handles geotechnical field data, lab data, chemical and hydrological data. They also have a draft standard for an XML version that is GML compliant. COSMOS has developed a data dictionary and a virtual data center for sharing borehole data on the internet, as well as an ongoing project on geotechnical lab data.

A survey of information needs will be developed by the core team in close cooperation with GMS group. The survey will cover basic demographic data (business type, use of data, etc), methods of collection, and specific data needs (data needed, current data collected, and priority of needed data). The detailed content of the survey is described in the "Survey Format" section. The survey will be sent to a wide group of stakeholders that use Geotechnical data including state and federal agencies, civil software developer/vendors, consulting and design firms as well as others specified with the help of the GMS group.

The Geotechnical Data Field and Laboratory Data survey is the most difficult and time consuming part of the standards development. The survey will identify the types of tests, hardware, as well as data description necessary for the development of a data dictionary. Consequently, it is envisioned that over half the effort (~60%) will occur in this phase compared to the development of a final data dictionary and XML (GML compliant) schema. The results of the survey will be used as the data requirements and definitions for defining a standard schema.

1. **Survey Format:** The contents of the survey will be developed by a small core team from the University of Florida, AGS, CIRA and COSMOS. The principal players will be: Dr. Marc Hoit (UF), Dr. Michael McVay (UF), Mr. Roger Chandler (AGS), Mr. Tim Spinks (CIRA), and TBD (COSMOS). A draft version will be created by the core team by merging all the data definitions found in the UF system, AGS, COSMOS and the Army Corps of Engineers and COSMOS survey results. The draft will be a prototype data dictionary merging the existing dictionaries and any determined missing data and not start with a blank slate. The draft version of the survey will be presented to the GMS group and refined based on their input. Once a final version is approved, the survey will be electronically administered. The GMS group will help define the stakeholders to be included in the survey and will be responsible for collecting the required input for each data section of the survey.

The survey will also ask questions about metadata for such as equipment, location, contractor/person etc. The core team will also use questions and results from the COSMOS survey where appropriate.

The survey will also include general questions about Geohazard and Geotechnical asset inventory and condition data to prepare for a further phase in defining a dictionary for that data. Questions such as: What data do you currently collect on Geohazards? Do you have a database collecting with Geohazard or condition information? What format is the data stored in? And others will be asked that will help determine the existing data and information to use as a starting point for development of a Geohazard and asset condition schema. The final data dictionary delivered from this project may or may not include Geohazard and asset condition information. The determination will be based upon the amount of time required to develop them which will largely depend upon whether sufficiently refined data definitions exist for this data.

A multiple choice web based survey questionnaire will be developed that includes each of the Geotechnical Tests identified above. For example, for SPT field tests, the survey would identify equipment: 1) method of creating borehole (auger, drilling mud, casing, etc), and size; 2) SPT Hammer calibration; 3) SPT blow counting; 3) Split spoon Sampling; 4) as well as soil field classification, and sampling for laboratory: a) Soil-Shelby tubes, Piston Samplers and b) Rock – Single, double and triple wall samplers. And, request the testing method utilized in performing the test.

Concerns: The proposed survey will require a significant amount of time and thoroughness by each person filling it out. The survey will be a detailed prototype data dictionary at a very detailed level. While the electronic format will make it easier to handle the mechanics of the survey, there is significant effort required by the responders to research how their data is used, what is collected by different groups etc. ***In order to collect this data effectively, it is expected that a minimum of a person week of effort will be required by each respondent.*** This effort should be distributed to the people most familiar with the section of data to be discussed (e.g. Lab testing, In-situ testing etc). Generally, surveys are not taken this seriously and return percentages are also often low.

Note: The working group members are responsible for their survey and the survey of any assigned agencies. This effort will be considered as in-kind contributions.

Data Reduction: While all raw data will be given to the working group, it will be the reduced data that is most effective for future work. Key to the reduction effort is to understand that there are two types of data: 1) Generic/common/fundamental data and 2) Derived/internal/specialized data. Generic data is that data which all stakeholders will want to share and is the bases for their work. An example of basic information is data like location, standard test data results, basic properties etc. Derived data is data that is developed during a process (design, engineering, software intermediate values, etc) and most all of the time is internal to or specialized to the group doing the work. The derived data is not expected to be part of any future dictionary or schema since it is specialized to particular processes, assumptions and methods and should be contained within the local application information data.

In reducing the data, it will require the working group to help decide on the relative importance of survey results in making decisions on what to include in the final report on required data. Much of the data will be common and easy to agree upon. The difficulty arises because often there are differing standards, test methods and equipment used. Decisions by the GMS group will need to be made on how to select the most common and important information that needs to be saved and shared in the future.

Note: The AGS dictionary will remain a sub-set of the final result. Additions to cover the more diverse data needs, larger scope of areas considered and differences in testing procedures and standards will be included. AGS will consider adopting theses additions in future releases of their standard.

Development of Schema: The resulting data dictionary from the above process will be used to develop a final XML data hierarchy and XML schema. This will be done through a meeting of the core team to develop a straw version of the schema and hierarchy in a 1 ½ day meeting. Existing schemas will be used as the basis (AGS, COSMOS and Florida) in order to speed the process and reduce the new development. A facilitated 1 ½ days meeting with the working group will be used to finalize a draft hierarchy and schema. The draft will be posted for comment and be sent to all groups participating in the data

dictionary survey. Depending on the comments, the final schema will be modified and submitted to the working group for adoption. If there are a sufficiently large number of concerns and comments, an additional meeting with the working group will be scheduled.

The resulting accepted standard dictionary, hierarchy and schema will be submitted to international standards bodies for acceptance.

Process: The University of Florida will lead the core team effort with AGS, CIRIA, and COSMOS as its partners. The following steps will be used to reach a consensus:

NOTE: The timeline is only a suggested one. If possible, the working group and core team will try to accelerate the process. Timely completion of the work outlined herein is dependent upon the effort provided by the participants, core team and the working group.

1. The initial straw survey will be developed by the core team from UF, AGS and COSMOS with input from the DOTs, FHWA, U.S. Army Corps of Engineers and the U.S.G.S. This effort will be accomplished through a physical meeting of the core team. Target meeting is May (Week of 16th or 23rd) (Completed August 05)
2. The draft will be distributed to the GMS group for comment, changes and additions. A 1 ½ days meeting will take place with the GMS group to refine and finalize the survey. (Oct 05)
3. The final survey will be distributed to the selected stakeholders (depending on the options chosen). (Dec 05 –March 06, depending on options)
4. The facilitated meeting(s) will be scheduled inviting groups to reduce the survey results into a report on required common geotechnical data. The facilitated meetings will be 1 ½ days long with the working group and others as decided. (Completed August 06)
5. The draft version of the required common geotechnical data will be distributed to the GMS group for comment. A physical meeting is proposed. After approval, the draft will be sent to a large stakeholder group for comment. A web distribution with a moderated forum will be used for the large comment phase. (Completed October 06)
6. The final version of the required common geotechnical data dictionary will be delivered to the GMS group for approval and publication. (December 06)
7. The core team will meet to develop a draft data hierarchy and schema. A draft will be distributed to the GMS group for comment and discussion. (Jan 07)
8. A 1 ½ days facilitated meeting with the GMS group will finalize the hierarchy and schema. (Feb 07)
9. The draft data dictionary and schema will be posted to the web for a comment before final acceptance. All stakeholders will be contacted and asked to review and comment on the draft. (Feb –March 07)
10. A final version of the schema will be presented to the GMS group. If sufficient changes are required, a 1 day meeting will be scheduled. If comments are easily handled through electronic discussions, then the final dictionary and schema will be posted and distributed. (June 07)

Tasks: The following are the tasks defined for this proposal.

Task	Time frame	Format	Participants
Develop straw survey	May 05 – July 05	2 day meeting of core team	Core team & GMS group
Finalize Survey Questions	July 05 – Oct 05	1 ½ day facilitated meeting	Core team & GMS group
Distribute Survey	Dec 05 – March 05	Electronic	GMS Group and external stakeholders
Reduce Survey Data	May 06 – August 06	1 ½ day facilitated meeting	Core team & GMS group
Draft for comments	Aug 06 – Oct 06	Electronic	FHWA
Final dictionary delivered	Dec 06	Electronic	GMS group
Develop straw hierarchy and schema	Jan 07	2 day meeting of core team	Core team & GMS group
Finalize draft schema	Feb 07	1 ½ days facilitated meeting	Core team & GMS group
Post and distribute draft for comment	Feb – May 07	Electronic	All stakeholders
Finalize schema and hierarchy	June 07	Electronic or physical meeting (depends on level of changes by public comment)	Core team & GMS group

Budget: Below is the proposed budget for the survey. The budget is developed using the following assumptions:

1) Team participants (UF, AGS, and COSMOS) will not receive salary compensation. Only their travel expenses will be covered.
2) The GMS group will be expected to provide effort as in-kind contributions in addition to any funding they contribute. The members are expected to participate by acting as the lead for their state in providing the effort in answering the survey. This includes getting in-depth reviews of the survey by their state experts in the areas of interest. They will also be expected to be the contact to other state DOTs and agencies not in the working group in order to find an equivalent lead responsible for answering the survey and following up with them to receive results. The GMS group needs to consider other groups they wish to involve in answering the survey (consultants, etc) and again act as lead contact.
3) The below budget is for the entire process including: a 1 ½ days working group meeting to refine and finalize the survey, a 1 ½ days meeting with the working group to reduce the results of the survey, and 1 ½ days meeting for working group

to finalize draft of schema. Finally, travel to the meetings by the working group members will be covered by the pooled fund study.

4) Additional meetings to reduce the data by working with other groups selected by the working group are given as an option.
5) The budget includes a part time graduate student at UF (1/2 time for 18 months) to help with the mechanics of developing the survey, entering the questions into the survey system, and collecting the results into a readable form and reducing the results and meeting results, helping to merge the schemas etc.

Budget Description:

The following table represents total project costs including in-kind service.

Task	Hrs	Personnel	Travel	Total
Develop straw survey	860	$92,060.00	$13,046.00	$105,106.00
Finalize survey questions	508	$51,920.00	$31,046.00	$82,966.00
Implement survey	290	$30,500.00	$0.00	$30,500.00
Reduce Survey data	804	$85,060.00	$31,046.00	$116,106.00
Draft for comments	300	$27,850.00	$0.00	$27,850.00
Final dictionary delivered	280	$27,280.00	$0.00	$27,280.00
Draft Schema	860	$92,060.00	$13,046.00	$105,106.00
Final Draft Schema	828	$88,060.00	$13,046.00	$101,106.00
Post draft schema for comments	300	$27,850.00	$0.00	$27,850.00
Final schema	280	$27,280.00	$0.00	$27,280.00
	5310	**$549,920.00**	**$101,230.00**	**$651,150.00**

The majority of costs for the project are being provided as in-kind service from the core team participants. For the pool fund contribution, participates will cover the costs for meetings, travel, and miscellaneous items like printing, meeting materials, etc.

.**Final Estimated Budget:**

The estimated cost to participants of the Pooled-Fund Study:

Travel	$101,230.00
Graduate Student, part-time	$31,920.00
Miscellaneous Expenses	$4,560.00
Optional Survey Reduction meeting	$35,970.00
Optional 2nd Graduate Student	$57,000.00
10% Contingency	$65,115.00
Total	**$295,795.00**

With participation by 12 state DOTs, the total cost per state would be approximately $25,000. With funding limitations and previous commitments, state DOT contributions are expected to be variable, especially with contributions for the first year.

In addition, several options are listed to provide some flexibility to the project as the project develops. The 10% contingency amount was included to cover costs associated with work group meetings, additional travel, technical support, and other miscellaneous items.