6th Annual Technical Forum
GEOHAZARDS IN TRANSPORTATION IN THE APPALACHIAN REGION
August 2006

PRACTICAL CONSIDERATIONS IN WORKING WITH DATA INTERCHANGE STANDARDS

Salvatore Caronna
gINT Software
Windsor, California

ABSTRACT: The upcoming DIGGS data interchange standard raises many questions concerning practical application of how data are stored and exchanged. In dealing with interchange standards, all too often the tendency is to try to mimic the interchange format as closely as possible. This approach can lead to improper work specifications, awkward data entry procedures, difficulty in data validation, and reduced querying capabilities. This paper explores some of the major considerations in dealing with DIGGS so as to achieve the maximum benefit to the user while still maintaining the ability to read and write interchange files.

1. INTRODUCTION

Data interchange standards have only one purpose: to house data for shipment in an accurate and readily accessible manner, much like the exchange standard for the freight industry, the freight container. Data interchange standards are not work specifications or standards for the design of databases that are used to work with the data.

The design work specifications, data interchange standards, databases, and reports each have their own unique considerations. Data interchange standards can be designed to meet the needs of a large segment of the user community but the other three items must be designed to meet the needs of each user. However, each of these four pieces need to work together and therefore each affect the design of the other to some degree.

Many interchange standards have been used, and continue to be used, in the geotechnical industry. The upcoming DIGGS standard is the most encompassing, flexible, and extensible of any that have existed to this time and presents unique challenges in its usage.

2. WORK SPECIFICATIONS

2.1. The interchange standard is not a substitute for a work specification.

The AGS standard of the United Kingdom has been in use since 1992. A complaint among the site investigation contractors is that some clients use the AGS standard as a work specification, that is, all the fields of all the relevant tables must be supplied with data. For example, in the sample data section of the standard are fields for the sampling date and time. The contractor may not as general practice collect that information and the client may never have asked for that information in the past but is now demanding it after the work has been done because it is part of the AGS standard. The collection of unnecessary data could result in significant cost increases to the client.

This implicit use of an interchange standard as a work specification leads to misunderstandings and can result in some data not being collected that are important to the client. The best way to specify work deliverables is explicitly in a traditional scope of work specification. The standard can be used in the specification by specifying which fields in the specification are expected to be supplied with data.

2.2. Specifying Code Values

Another issue important issue with a data interchange standard as part of a work specification is guidance by the client as to codes that are to be used for certain data items. For example:

- Hole Types
- Sample Types
- Layer Legend Codes
- Triaxial Test Types

The AGS and DIGGS standards have lists of recommended codes for many such fields. It is important that the client specify which lists are to be used with which fields, using either the published standards lists, custom lists of the client's design, or some combination. These need to be part of the work specification. Without this guidance, the client could get back valid data which his reporting and querying tools cannot use. This would result in a time-consuming process of reentering the data manually.

2.3. Description Classes

An additional, important consideration applies to the DIGGS standard. Unlike the AGS standard, which has only one field for layer descriptions and one for detail descriptions, the DIGGS standard allows multiple description "classes" so that any components can be captured in the data interchange. Therefore, it is crucial that the client specify which types of data are to be captured and what class names are to be used.

2.4. Use the Standard as much as Possible

Most data interchange standards allow for expansion of the specification to allow for additional data items not covered in the base standard. This is true for both the AGS and DIGGS standards. Never expand the specification if a construct exists in the specification that can accommodate the data. This avoids additional work by all involved in the transfer process to deal with a nonstandard specification.

This may seem obvious but there can be great temptation to add a new table of data to store, let's say slope inclinometer or piezometer data, when generic structures exist in the standard (MONP/MONR groups in AGS and the MonitoringPoint features in DIGGS) to deal with time-dependent data.

Another example is the Texas Cone test. This can be handled by the DrivenPenetrationTest features in DIGGS, even though at first inspection of the standard it would appear to not be supported.

2.5. Don't Overload Standard Fields

If you have data that needs to be collected and transferred and there is no structure in the specification to store it, don't overload a field you are not using, or a remarks field, with this new information. For example, let's say you want the site investigation consultant to record push pressure on Shelby tube samples. There is no such field currently in DIGGS. You could direct the consultant to log that information in the Remarks field. This makes the data difficult to validate and difficult to extract in reports and queries if the consultant is using that field for other information as well. Instead, expand the specification to include the new field.

2.6. Summary

The interchange standard is an important part of the overall work specification and care must be taken to properly use the standard appropriately and to make the work specification as clear as possible.

3. DATABASE DESIGN

3.1. Data Interchange Standards Are Not Working Databases Structures.

Much of the DIGGS standard can be adopted as a usage standard, but there are many areas of the standard that act quite well for the purpose of data interchange but are less than optimal for the purpose of data collection, validation, manipulation, and reporting. Usage databases need to be set up to best meet the requirements of the users of the data. The interchange medium should only affect the design of the usage database minimally. Even table and field names in the working database do not need to match the standard. Care must only be taken to ensure that the usage database can be mapped to and from the transfer standard.

3.2. Examples

3.2.1 MonitoringPoint

This group of five objects and features is based on the AGS MONR/MONP groups. It is designed to handle any type of depth- and time-related data. This avoids creating special objects for each type of test. This makes the interchange standard robust and easily extensible without changing the structure. However, this structure is probably not one that either the data producer or consumer would wish to work with because there are fields to cover every type of monitoring device. Many, if not most, are irrelevant for any particular test type. Further, this structure mixes all the different monitoring tests together.

For a working structure, it is much easier to work with separate structures for each of the different monitoring tests. Therefore, there would be tables specific for piezometer information and their fields would only apply to piezometers. The same for slope inclinometer, settlement gauges, extensometers, strain gauges, and so on.

3.2.2 Soil and Rock Descriptions

If your requirements for descriptions entail just a single field, as below, then your database structure could match the DIGGS structure quite closely:

Depth	Bottom	Description
0	6	Silty SAND: very loose, fine to medium, moist, green.
6	12	Silty SAND: loose, fine to medium, dry to moist, bluish red.
12	18	Silty SAND: medium dense, fine to medium, moist, brown.
18	24	Sandy CLAY: very stiff, moist, brown.
24	27	Sandy FAT CLAY: hard, wet, gray.

If you have specific needs for component descriptions to better capture and query the data, then the interchange structure and the probable working structure are vastly different. Following is a sample of a simple working component description structure that is set up the way most people would like to work with these models:

Depth	Bottom	Main	Qualifier	Consistency	Grain Size	Moisture	Color
0	6	sand	silty	very loose	fine to medium	moist	green
6	12	sand	silty	loose	fine to medium	dry to moist	bluish red
12	18	sand	silty	medium dense	fine to medium	moist	brown
18	24	clay	sandy	very stiff		moist	brown
24	27	fat clay	sandy	hard		wet	gray

Following is how one might set up a working structure to mimic the DIGGS structure.

Depth	Bottom	Component	Value
0	6	main	sand
0	6	qualifier	silty
0	6	consistency	very loose
0	6	grain size	fine to medium
0	6	moisture	moist
0	6	color	green
6	12	main	sand
6	12	qualifier	silty
6	12	consistency	loose
6	12	grain size	fine to medium
6	12	moisture	dry to moist
6	12	color	bluish red
12	18	main	sand
12	18	qualifier	silty
12	18	consistency	medium
12	18	grain size	fine to medium
12	18	moisture	moist
12	18	color	brown
18	24	main	clay
18	24	qualifier	sandy
18	24	consistency	very stiff
18	24	grain size	
18	24	moisture	moist
18	24	color	brown
24	27	main	fat clay
24	27	qualifier	sandy
24	27	consistency	hard
24	27	grain size	
24	27	moisture	wet
24	27	color	gray

Like the DIGGS structure, the above table design is quite easy to add new components without having to alter the structure. All that is needed is to add more components to the lookup list. Since this structure mimics DIGGS the import and export process becomes easier as well. However, it is doubtful that many people (if any) would want to work with such a structure.

3.2.3 Laboratory and Field Test Results

The DIGGS standard does not contain raw data fields. Therefore, the data producer always needs many more fields and tables than the interchange standard provides to properly record the raw data for laboratory and field testing.

DIGGS has a multitude of objects and features devoted to laboratory and field test results. The data consumer may wish to have just one laboratory testing table that contains just the results of all laboratory tests and another table for field tests.

In both these scenarios the database needs to be set up differently than the interchange structure to better meet the needs of the data producer and consumer.

3.3. How the Database Structure needs to change to conform to the Interchange Standard

The interchange standard cannot be ignored when designing a database. Certain fundamental structural elements must be in place. Without some basic structural alignment with the standard, it is possible that invalid data could be exported from the database and data could be lost on import into the database.

3.3.1 Table Key Fields

The key fields in a table uniquely define each data record. For example, in most tables in the U.S. storing depth-related data, the keys are generally the Hole ID and Depth, that is, within one hole there will always be a maximum of one record with any particular depth.

DIGGS generally defines more key fields for uniqueness in their tables than most organizations in North America are used to dealing with. For example, each record in the SAMPLE feature is defined by the Hole ID, Top Depth, Sample Type, and Sample Number. This allows for multiple records at the same depth in a hole, but the combination of all four key fields must be unique for each record.

Note that DIGGS does not explicitly define key fields for each table. Rather, the keys are implied via the nesting of the tables within the XML structure.

In exporting data to DIGGS, having fewer keys will result in valid records in DIGGS, as long as mandatory key fields are included. Importing data from DIGGS could result in loss of data if there are fewer keys in the database table than the feature in DIGGS. For example, let's say a DIGGS file has two different sample types at the same depth in the same hole. If the target database requires uniqueness of the Hole ID and Depth, only one of the records will be imported.

3.3.2 Table Relationships

It is common for the lab testing parent table to be a direct child of the HOLE table in North America. In DIGGS the corresponding feature (SPECIMEN) is a child of SAMPLE (which is a child of HOLE). With this non-conforming relationship, on export the mapping procedures must somehow associate each specimen with a sample. If this is not done the lab testing results will be "orphans" and this will invalidate the DIGGS file.

3.3.3 Data Types

It is common practice for database designers to specify the data type of fields that should be storing just numeric or date/time data as text fields. This allows input of comments when the data are not recorded. For example, a field that holds water depth could then have a note like "not encountered". Allowing such an entry is not good database design. With the DIGGS standard there is one more reason not to do this. The fields in DIGGS are properly typed, that is, a field holding water depth would be a numeric field and there would be an associate note field for comments. Exporting text data to a numeric field in DIGGS will invalidate the file.

3.3.4 Code Lists

A big part of the DIGGS standard (and many others) is associating lists of valid values (called "code lists" in DIGGS) with certain fields. These lists can be dictated by the standard, a national body, or by a client letting contracts. These lists need to be in line with those that are required, either by using the lists in the design of the database or by mapping your custom lists to the required lists in the import and/or export processes.

This is an important consideration in data usability. For example, if analyses are set up to correlate characteristics of soil based on certain types of drive sample results, and the data generator used different codes than those expected by the analyses, the analyses will not work or will be inaccurate.

3.4. Summary

The interchange standard's purpose is strictly to get from data from one place to another. Database design needs to reflect the requirements of the end user of the database. Database design needs to accommodate the requirements of the interchange standard but not be dominated by them. The tail should not wag the dog.

4. VALIDATION

DIGGS files are self-validating:

- the schema can be checked for accuracy
- data are checked against the specified data types
- values in fields associated with code list can be validated
- minimum and maximum values can be assigned to fields

However, there are no automated methods inherent in the standard for performing dependent validations. For example:

- RQD must be greater than or equal to Total Recovery.
- If an SPT penetration is 1.5 feet, there must be three blows.
- Specimen depths must be in the range of the corresponding parent sample depth range.
- Layers within the same description classification must not overlap.
- If the liquid limit is 35, a plastic limit of 52 is unreasonable.

With time, budget, and enough people, these dependent validation rules can be written. In the meantime, do not assume that because DIGGS is self-validating that the data are all reasonable.

5. DIGGS IS NOT HUMAN

Some data interchange standards can be edited by real people using text editors or spreadsheets. DIGGS is not one of those standards. Following is a snippet of a DIGGS file:

```
- <subsurface>
  - <Hole gml:id="D6DD2E0C-7BFF-4ebc-83E4-4F69BC1D71A1">
      <gml:name codeSpace="http://www.ags.org.uk/id">TS150</gml:name>
    - <geometry>
      - <gml:LineString gml:id="72A638B0-731A-4474-BA1F-BC4CF48DC052" srsName="urn:ogc:crs:epsg:6.9:27700">
          <gml:pos dimension="3">97488.580 103170.658 54.894</gml:pos>
          <gml:pos dimension="3">97488.580 103170.658 54.894</gml:pos>
        </gml:LineString>
      </geometry>
    - <gml:engineeringCRSRef>
      - <gml:EngineeringCRS gml:id="43AF3014-9E65-4faf-B93C-E78A2A938559">
          <gml:srsName>TS150 CRS</gml:srsName>
        - <gml:usesCS>
          - <gml:LinearCS>
            - <gml:usesAxis>
              - <CoordinateSystemAxis gml:id="E1A43C6C-E9B4-4593-9D10-472C12F809E4" gml:uom="units.xml#m">
                  <gml:axisName>Depth</gml:axisName>
                  <axisDirection xlink:href="72A638B0-731A-4474-BA1F-BC4CF48DC052" />
                </CoordinateSystemAxis>
              </gml:usesAxis>
            </gml:LinearCS>
          </gml:usesCS>
        - <gml:usesEngineeringDatum>
          - <EngineeringDatum gml:id="374AD02E-4CEE-4d16-ADC3-260B799DAD69">
            - <origin>
              - <gml:Point srsName="urn:ogc:def:crs:epsg:6.9:27700">
                  <gml:pos>97488.580 103170.658 54.894</gml:pos>
                </gml:Point>
              </origin>
            </EngineeringDatum>
          </gml:usesEngineeringDatum>
        </gml:EngineeringCRS>
      </gml:engineeringCRSRef>
      <type codespace="AGSHoleTypeCodeList.xml">INST</type>
    - <remark>
        <comment>Tunnel progress 0</comment>
        <dateTime>2001-05-02T12:00:00</dateTime>
      </remark>
    - <remark>
        <comment>Tunnel progress 0</comment>
        <dateTime>2001-05-07T12:00:00</dateTime>
      </remark>
    - <remark>
        <comment>Tunnel progress 0</comment>
        <dateTime>2001-05-11T12:00:00</dateTime>
      </remark>
    - <remark>
        <comment>Tunnel progress 0</comment>
```

DIGGS is based on XML (Extensible Markup Language) and GML (Geographic Markup Language) which makes it self-describing and self-validating. It also makes DIGGS understandable to some degree by many GIS software packages without special translation. Finally, a host of tools are available for validation, coordinate transforms, and unit conversions. The downside is that no human can hope to properly create or edit a DIGGS file in any reasonable period of time.

Specialized software will be needed to read and write DIGGS files. The experience in the UK with the AGS is that the inability to manually manipulate the interchange files is probably a good thing. AGS files can be edited with text editors and spreadsheets and many people have done this and continue to do so. Unfortunately, the results of such human manipulation are generally not satisfactory.

6. SUMMARY

The holy grail of easily interchangeable data is within our reach. However, this brave new world will require significant changes in the way we deal with data and software. New structures need to be put in place, and work must be put into the exchange process to ensure that it be routine and accurate.

7. REFERENCES

The Association of Geotechnical and Geoenvironmental Specialists (2005), "Electronic Transfer of Geotechnical and Geoenvironmental Data, (Edition 3.1) including addendum May 2005"

DIGGS Usage Guide and Data Dictionary. Draft of 1 July 2006 (unpublished).

Caronna, S and Wade, P (2005), “Problems with Using the AGS Format As a Working Database Structure“, *Geotechnical and Geoenvironmental Data in Electronic Format Production, Management, and Application*, Birmingham, United Kingdom, 19 October 2005.

Caronna, S. (2005). "Data Granularity in the Storage and Reporting of Soil Exploration Information", *The Second Annual Geotechnical, Geophysical, and Geoenvironmental Technology Transfer Conference and Expo*, Charlotte, North Carolina, 14-15 April 2005.

Caronna, S. (2005). "Geotechnical Data Management Issues for Transportation Authorities", *6th Transportation Specialty Conference*, Toronto, Ontario, 2-4 June 2005.