Consolidation, Automation, and Visualization: CHS Atlantic's GPS & Tidal Data Holdings

Mike MCMAHON, Canada

SUMMARY

The primary goal of the CHS's ongoing National Continuous Vertical Datum Project is to enable hydrographic surveys to collect data referenced directly to the ellipsoid. An important initial requirement is to link our existing tide station-based vertical datum holdings with the appropriate ellipsoidal and geodetic reference frames. To establish these links, the CHS has been collecting precise GPS data at those tidal stations with the best quality tidal records and vertical datums. Once collected, these GPS datasets are submitted to NRCAN's Precise Point Positioning (PPP) post processing service which returns the station's best possible coordinates. These coordinates are then used to establish the link between the station's vertical datums and the appropriate reference frames. With over 350 occupation datasets in the Atlantic region alone, it became a major task to keep track of the data.

Transcribing the salient elements of PPP's output files into a spreadsheet would have consolidated the data, but would become increasingly difficult to maintain over time. Instead, CHS Atlantic has implemented a database solution, not only avoiding maintenance problems, but also facilitating the automation of data loading, initial data analysis, and the generation of kml-based data summaries. These summaries provide geodetic offsets, available datums, and links to benchmark and constituent information for every station. CHS Atlantic now has the most current version of its data easily constantly available and can easily determine which stations need to be reoccupied, which have "suspicious data", and can quickly identify "holes" in the data set requiring further work.

KEY WORDS: PPP, CGVD28, CD, Visualization, Data Management

1. INTRODUCTION

CHS Atlantic's Tidal Section is actively collecting GPS data, and required a mechanism for storing and accessing the most current information. Spreadsheets were an obvious option, since they offer a quick and easy way to store information. However, spreadsheets tend to reproduce and change, and it quickly becomes unclear which version is the most recent, and whether or not any version contains all of the available data. In short, they can result in a data management nightmare.

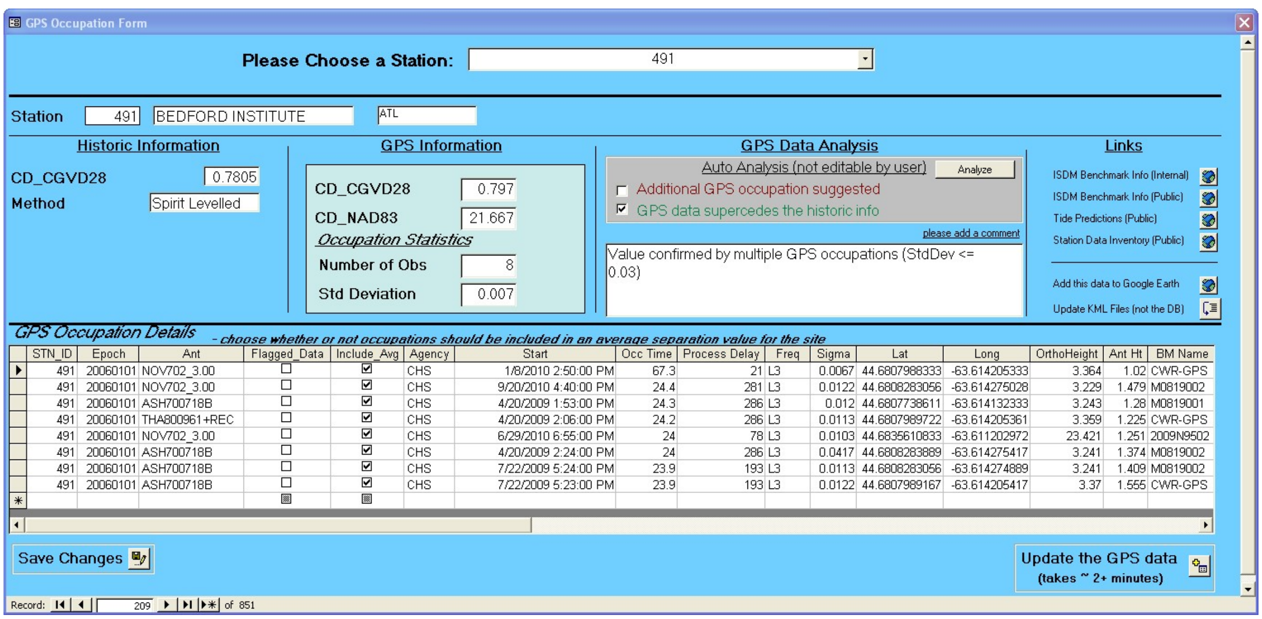
It became apparent that a database made the most sense - multiple people could look at it simultaneously, and the input could be sanitized to prevent blunders. Several other advantages were quickly realized - including the ability to automate updates, the ability to automate the initial validation of the data, and the ability to automate the generation of Google earth files of the data. These benefits have greatly streamlined our ability to collect and process data. For example, we can now more easily identify data gaps and better focus our data collection efforts. Additionally, within moments of receiving new data from the field, we can not only process it, but also have the results rolled out to the entire organization.

I will first describe the application's interface, its outputs, and how they have benefitted CHS Atlantic. Next, I will describe the actual steps taken by a user to update the data. Finally, I will conclude with the more technical aspects of the application – including the database design, and how VB code is leveraged to automate tasks such as updating, and analyzing the data, as well as reformatting it into user-friendly kml files.

2. OVERVIEW OF THE APPLICATION

2.1 THE INTERFACE

The application interface consists mainly of the following form within MS Access. The primary functions of the form are to provide 1) a mechanism for viewing existing data, and 2) a means for updating the form with new data. Central to the form is a listing of all of the GPS occupations that have occurred at each station and the salient information from each occupation.



STN_ID	Epoch	Ant	Flagged_Data	Include_Avg	Agency	Start	Occ Time	Process Delay	Freq	Sigma	Lat	Long	OrthoHeight	Ant Ht	BM Name
491	20060101	NOV702_3.00	☐	☑	CHS	1/8/2010 2:50:00 PM	67.3	21	L3	0.0067	44.6807988333	-63.614205333	3.364	1.02	CWR-GPS
491	20060101	NOV702_3.00	☐	☑	CHS	9/20/2010 4:40:00 PM	24.4	281	L3	0.0122	44.6808283056	-63.614275028	3.229	1.479	M0819002
491	20060101	ASH700718B	☐	☑	CHS	4/20/2009 1:53:00 PM	24.3	286	L3	0.012	44.6807738611	-63.614132333	3.243	1.28	M0819001
491	20060101	THA800961+REC	☐	☑	CHS	4/20/2009 2:06:00 PM	24.2	286	L3	0.0113	44.6807989722	-63.614205361	3.359	1.225	CWR-GPS
491	20060101	NOV702_3.00	☐	☑	CHS	6/29/2010 6:55:00 PM	24	78	L3	0.0103	44.6835610833	-63.611202972	23.421	1.251	2009N9502
491	20060101	ASH700718B	☐	☑	CHS	4/20/2009 2:24:00 PM	24	286	L3	0.0417	44.6808283889	-63.614275417	3.241	1.374	M0819002
491	20060101	ASH700718B	☐	☑	CHS	7/22/2009 5:24:00 PM	23.9	193	L3	0.0113	44.6808283056	-63.614274889	3.241	1.409	M0819002
491	20060101	ASH700718B	☐	☑	CHS	7/22/2009 5:23:00 PM	23.9	193	L3	0.0122	44.6807989167	-63.614205417	3.37	1.555	CWR-GPS

Figure 1 The Interface

All of the data values in the form are read-only. This eliminates the possibility of introducing blunders while reviewing data. The only way to alter the displayed values is to alter the source data files themselves.

The only items a user can modify are:
- comments about the station itself;
- comments about individual occupations; and
- checkboxes flagging an occupation as being questionable and/or acceptable.

For each station, values are shown for CD_CGVD28 and for CD_NAD83. These values are the average for all of the occupations at that station that have been manually accepted via a checkbox.

Also shown is an analysis window, which indicates whether or not the data for a station has been deemed acceptable by the application. A station will be flagged for an additional occupation if:
- the GPS-derived geodetic value is significantly different from a historic, spirit levelled value;
- several GPS-derived values differ from each other;
- only a single GPS occupation exists

2.2 THE OUTPUTS

While the intent was for the database form to be the focus of the application, it has become secondary to the kml files that it generates. Each time the database is updated, three types of kml files are generated, and stored on our local network. These types can be categorized as:
- All of the tidal stations for which we have data available.
- All of the GPS occupations for every station.
- A summary file showing a single data point for every station for which we have at least one geodetic value.

For each of the "types" of kml files described above, 2 files are available – 1 with data shown on the surface of the earth, and the second with the data extruded vertically proportionate to the CD_CGVD28 value. Each of these files is discussed in more detail below.

2.2.1 <u>All Tide Stations – Available to all CHS</u>

This file is available to all of CHS, and shows every tide station with all of the available tidal datums. The symbology facilitates the identification of those stations for which datums and constituents are readily available, the criteria essential to deciding if a station should be occupied with a GPS. The text balloon shows all of the tidal datums for each station.

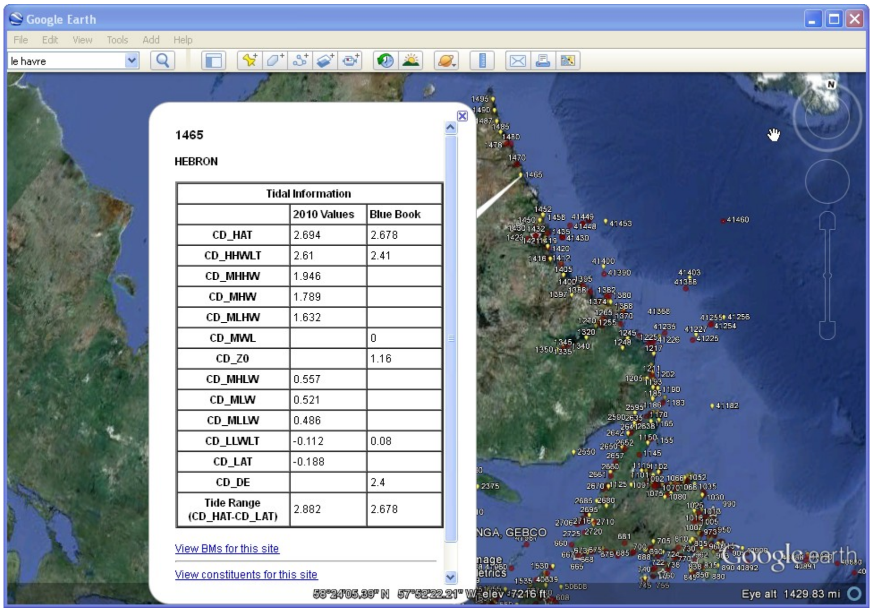


Figure 2 Tidal Information for all Stations with Links to BM and Constituent Information

2.2.2 All Occupations – Available to Tidal Section only

This file is available to tidal section staff only, as it includes more information than most people want, but is very valuable internally. The file displays all of the occupations for every station. Within the text balloon for each occupation is all of the information about the occupation (e.g. date and length of occupation, antenna height, benchmark used, etc).

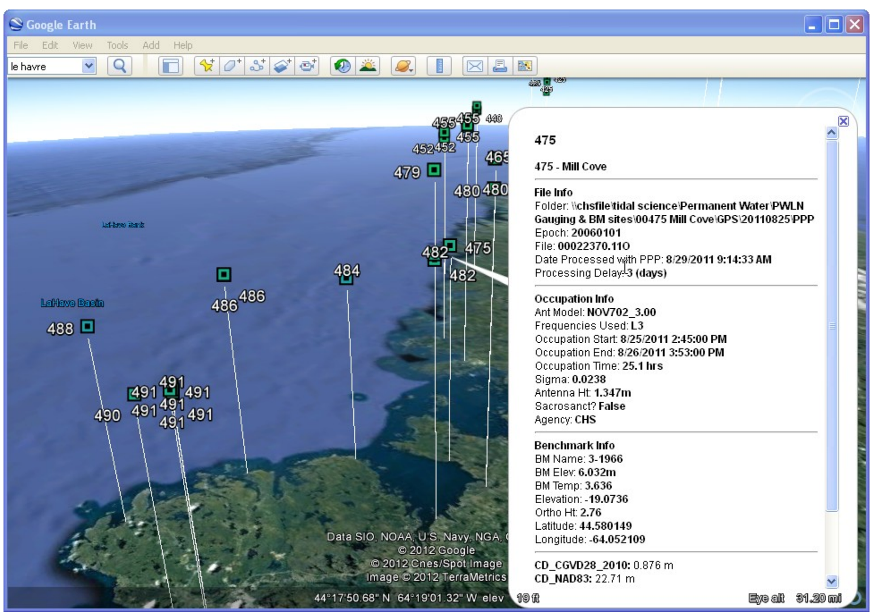


Figure 3 All Details For Every Occupation Available

2.2.3 <u>Summarized Geodetic Data – Available to all CHS</u>

This file is available to all of CHS Atlantic, and is probably the most heavily used by tidal section staff. It distills all of our geodetic data for each station into a single point. For these stations, the text balloon is populated with the mean CD_CGVD28 and CD_NAD83 values, all of the available datums for that station, and links to the constituents and benchmark information for that station.

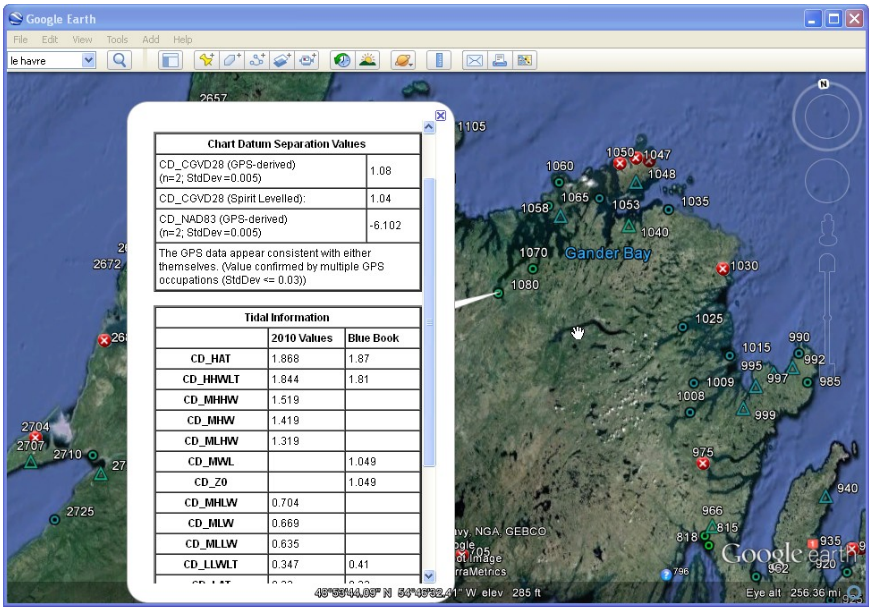


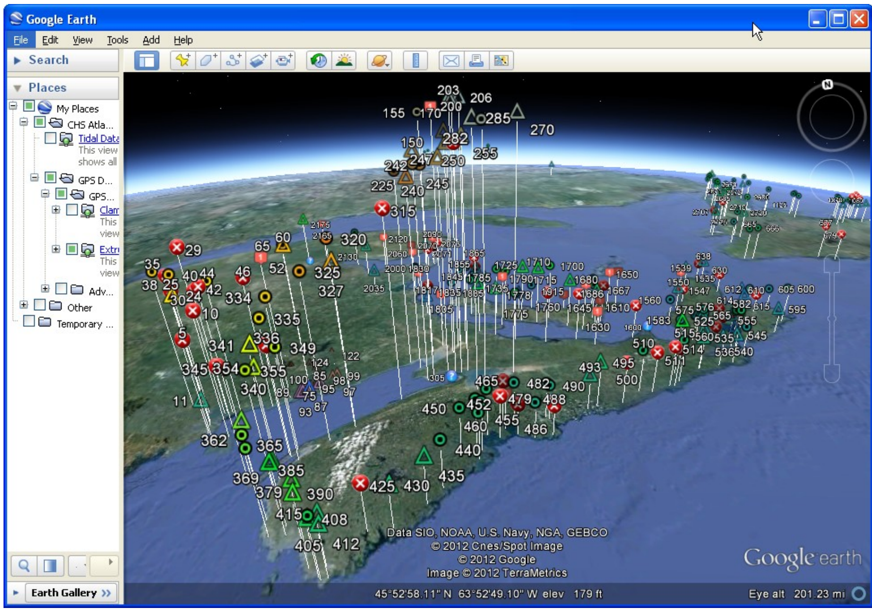
Figure 4 Summarized Geodetic and Tidal Datum Information and for All Possible Stations

2.3 THE BENEFITS

The database has done what it was supposed to do – it has become the one and only source of our geodetic information. It has reduced blunders and facilitated data comparisons. Moreover, it has allowed Tidal Section to simplify and automate aspects of our data management that we did not anticipate. Dumping the data to kml started out as a small side project, but has become one of the largest benefits of the entire application.

The KML output, particularly the summarized geodetic data, is especially useful when dealing with external clients, because it makes all of the datums and geodetic data available for every station visible simultaneously. Most of the requests received consist of people asking for information such as a tide range, a datum, or the Chart Datum-Geodetic offset for a given location. Because the data is viewed spatially, relationships between nearby stations are readily apparent, and it has become much easier to interpolate values for locations where we have never had a station.

Internally, we can quickly identify those stations for which we need additional occupations, since the symbology reflects the confidence we have in the data. Additionally, the CD_CGVD28 offsets can be viewed in 3D, helping us recognize outlier stations with dramatically different values than their neighbours.



k

Figure 5 The CD_CGVD28 Relationship can be Easily Explored and Understood in 3D

3.   WORKING WITH THE APPLICATION

The application was designed to import data from PPP files with minimal user interaction – the philosophy being that the more automated the process, the less potential for introducing blunders. All that is required for the application to work is 1) processed PPP data, and 2) a text file for each occupation containing information about the name and elevation of the benchmark that was occupied. The following steps describe exactly what is required of the user to update the database with new information. Note that these steps assume that the GPS data has already been processed by NRCAN's PPP service.

<u>Step 1 – Creation of Benchmark.mmm File</u>
The PPP data is received, and stored in a PPP folder under the appropriate station. Since the PPP results have no knowledge of which benchmark was used, or its elevation, the user must provide information for linking the geodetic and chart datum information. To do this, a text file called

"Benchmark.mmm" is created within the PPP folder. The format of the file is simple (as shown below):

<u>Example contents of Benchmark.mmm</u>

Benchmark:	69N261
Elevation:	12.853
Temp BM:	12.767
Agency:	CHS

Of those values shown above, only values for the benchmark name and elevation are normally required. In cases where the GPS is set up over a temporary pin that has been levelled from an existing BM, a value for Temp BM is also required.

<u>Step 2 – Initial Update</u>
The user opens up the database and clicks a button entitled "Update the GPS data". The database will then iterate through all of the folders within a particular directory for PPP data, and grab information out of those files and use it to populate fields within the database.

<u>Step 3 – User QC</u>
Within the form, the user then browses to each of the new datasets, and does some straightforward QC. They will verify that all of the necessary information was carried over, that the session was long enough, that both phase centres of the antenna were used, and that the sigma of the occupation was not extraordinarily high. If everything looks fine, the dataset is flagged as being acceptable. Comments can be added for each occupation.

<u>Step 4 – Final Update</u>
Once all of the new data is QC'd, the user clicks the "Analyze" button. This makes the database compares data within each station to ensure consistency. If the datasets for a given station aren't consistent, or if new GPS data is too different from older, spirit-levelled data, the station gets flagged as being potentially problematic. Simultaneously, the database generates KML files for use by Google Earth of the updated data. The KML symbology reflects the analysis of the data, so a quick glance quickly reveals which sites need to be reoccupied.

With the completion of Step 4, the new data is available to the whole organization. Users who have the application open during an update will see new data appear without having to refresh anything.

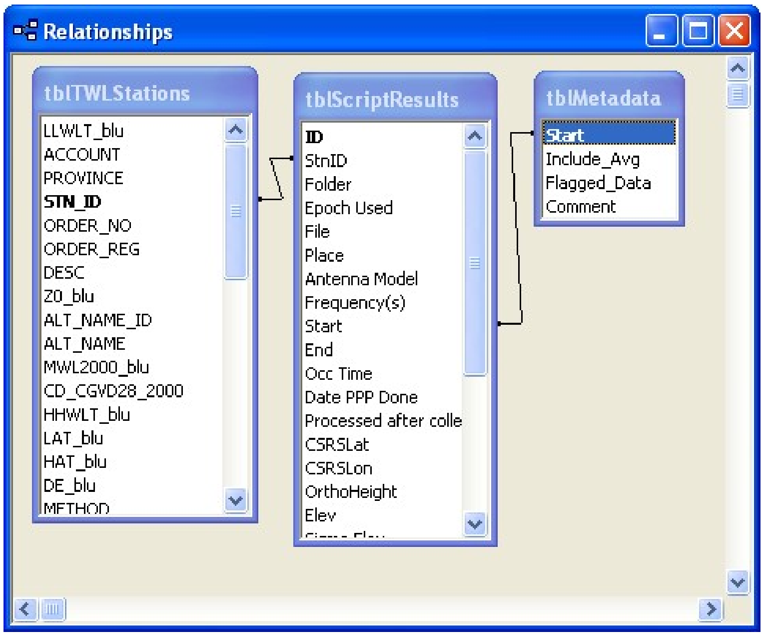
4. UNDER THE HOOD

The database is MS Access, and all of the automation (e.g. auto-updating; auto-analysis; and auto generation of KML files) relies on the ability to execute snippets of VB code called modules. These modules will be discussed following a brief description of the database itself.

4.1 DATABASE DESIGN

The database is very simple, consisting primarily of 3 tables. One table (tblTWLStations) holds information about each station, another table (tblScriptResults) holds information about individual GPS occupations, and a third table (tblMetadata) holds metadata about each occupation. tblTWLStations is joined in a 1 to many relationship to tblScriptresults, and tblScriptResults is joined to tblMetadata in a one to one relationship (using a date field). Simply put, any given station can have numerous GPS occupations, and every GPS occupation can have some user-generated information associated with it (e.g. comments).

The reason for keeping the metadata separate from the occupation data is that the occupation data gets refreshed with every update, and we don't want to lose all of the comments and flags every time new data is added.



4.2 UPDATING THE DATABASE

Clicking the "Update the GPS" button initiates a series of operations, all of which serve to delete the contents of tblScriptResults, and fill it back up with updated values. The specific operations are:

- Delete all of the existing occupation data;
- Copy over any "sacrosanct" data - that which has been deemed as correct (and should not be overwritten);
- Iterate through a directory, find PPP data, and pull the relevant information into tblScriptResults;

— Alert the user that new data is available and QC must be performed.

The VB procedure for updating the data is called suckTheData, and it works by looking within a particular directory for PPP folders. Once found, it looks for the *.sum file, which contains all of the relevant PPP data. Within that file, it looks for 20 uniquely identifiable snippets of text, such as the phrase “Antenna Model”. Depending on which text is found, the script knows where to find the actual data we want. In the case of the antenna model, this would be a value like “NOV702_3.00”. In VB, this would look something like the following.

```
If InStr(1, Line, "Antenna Model", 1) > 0 Then
    sumData(5) = Mid(Line, 24, 16)
```

In this case, the actual value for the antenna model is temporarily stored as “sumData(5)”. The script continues scanning the sum file until it has values for all 20 desired pieces of information. At that point, it looks for the presence of a Benchmark.mmm file in the same folder, and then scans that file for up to 4 other pieces of information, such as the benchmark’s name and elevation.

Once the script has all of the values it wants from both files, it calculates several other values (such as the duration of the occupation, and CD_CGVD28). Then, it runs a SQL INSERT statement that drops all of the captured information into a single row in tblScriptResults. The script continues iterating through the directory looking for more PPP data and each time it finds some, repeats the steps above.

4.3 ANALYZING THE DATA

The purpose of the “Analyze” button is to look at all of the user-approved (or “QC’d”) data for each station and ensure that the data does not have any obvious problems. A VB procedure, called AnalyzeGPS, performs these operations. For any station to have ”acceptable data”, one of the following scenarios must occur.

- If 2 or more QC’d GPS occupations are available, the standard deviation of all occupations must be within 0.03.
- If no other QC’d GPS occupations exist, the GPS data is compared to existing spirit-levelled data, and it should be within 4cm of the historic value.

In short, every station must have at least 2 consistent values to have an acceptable value. Whether or not a station is deemed acceptable or not is specific to each station, and thus, the result is stored in master tide station table (tblTWLStations). Comments are also generated about the station describing any problems that have been detected, and this information is visible to the user both via the application interface as well as via the kml files. The status of each station is reflected in the symbology used in the kml summary file.

4.4 GENERATING KML FILES

This application generates 3 kml files and these are stored in 2 different locations on the network, depending on the intended audience. CHS staff only ever load a single file, and this file is simply a pointer to all of the data that Tidal Section wants to make available. The advantage of this method is that the data can be updated while people are looking at it, and changes are visible immediately without having to refresh anything. The summarized GPS data and the generic tide station data are accessible to all of CHS, and the files containing specific details of every occupation is only accessible to Tidal Section.

The process of generating each of the kml files is very simple and is controlled by another VB procedure (i.e. writeKMLFiles), and consists of the following steps:

- Open a file, and make it available for editing.
- Run a SQL SELECT statement against a database table to get all of the data needed for each station. At a minimum, this consists of a Latitude and Longitude, and a name, but in practice, much more information is desired.
- Determine how each station/occupation will be symbolized.
- Wrap the data in the appropriate tags to create a correctly formed KML placemark.
- Repeat for all stations/occupations.
- Close the file.

To facilitate the visualization of the data, each kml file is actually generated in 2 different ways – once with the data clamped to the earth's surface (elevation of 0), and once with the data extruded up and out of the earths surface (elevation > 0). Kml files expect coordinates to be in the form of *<coordinates>-66.817323, 44.606119, 0 </coordinates>*, where the values are longitude, latitude, and elevation, respectively. By assigning the elevation value for each station as a value proportional to the value for CD_CGVD28, the placemarks are shown as being at the top of different sized "sticks" coming out of the earth's surface. For the scale of data displayed by this application, the CD_CGVD28 difference was multiplied by 30,000.

5. FUTURE DIRECTION – WHAT'S NEXT?

At the present, this application meets the needs of CHS Atlantic's tidal section, and there is no pressing reason to modify it. However, tidal sections in other regions are collecting similar data, and have expressed interest in this work. There is no reason why this or a similar application could not be implemented nationwide. Ideally, implementation would be centralized, with a single application accessible to all regions. There is no good reason for each region to reinvent the wheel, and many good reasons why they shouldn't.

If this does grow into something larger, and is required to server a wider audience, it will probably have to move from a simple MS Access application into an enterprise level database (e.g. perhaps Oracle?).

BIOGRAPHICAL NOTES

Mike joined CHS as an MDH in 2007, and is probably best known for his development of GooDir – a Google Earth Interface for CHSDir. Prior to his work with CHS, Mike worked as an ArcIMS Administrator at the Hawaii Biodiversity and Mapping Program, as well as a Project Manager at Latitude Geographics. Mike received his B.S. in Zoology from the University of Calgary in 1998, and a Master of Marine Management from Dalhousie University in 2000.

CONTACTS

Mike McMahon
Canadian Hydrographic Service (Atlantic)
Bedford Institute of Oceanography
PO Box 1006
Dartmouth, NS
CANADA
Tel. (902) 426-2389
Fax (902) 426-1839
Email: Mike.McMahon@dfo-mpo.gc.ca