Award Number 08HQGR0119
FDEP/FGS Metadata Project (Part I)
Florida Geological Survey
Final Technical Report

Principal Investigator
Clint Kromhout
Co-Principal Investigator
Sarah Ramdeen
903 West Tennessee St.
Tallahassee, Fl 32304-7716
Phone: (850) 488 – 4191
Fax: (850) 488-8086
Clint.Kromhout@dep.state.fl.us
Sarah.Ramdeen@dep.state.fl.us

Award Dates: 9/1/2008 – 8/31/2009
Report Submission Date:
Report Written and Compiled by: **Brian G. Clark, Lisa Fulton, Clint Kromhout, Sarah Ramdeen, Amber Raynsford, and Keith Wood**

Proposal Abstract

The purpose of this project is to substantially enhance metadata for up to 10,000 well records in the Florida Department of Environmental Protection/Florida Geological Survey (FDEP/FGS) sample repository database during the one-year grant period. Two tasks will be completed to attain this goal: 1) a GIS-based location refinement of up to 10,000 well entries and 2) metadata creation and refinement in XML format for these wells. In order to complete metadata for the FDEP/FGS collections, well record validation and refinement of locational accuracy of well cuttings, cores, and core chips are required. Currently all three sample types (collections) have up to five different recorded locations which may, or may not, all agree. Combined, these three collections represent approximately 19,000 wells.

The FDEP/FGS has been working for a considerable number of years to develop a digital database of its geological collections. Impediments to progress include insufficient staffing, funding, and bureaucratic obstacles. As a result, the preservation of and accessibility to the FDEP/FGS data has suffered. End users of the data are faced with quality assurance and quality control (QA/QC) issues within the metadata.

Rejuvenated interest in the last four years has lead to some success in database development; however, metadata population challenges exist due to locational uncertainties and limited staffing to migrate relevant data from paper records to the database following appropriate QA/QC procedures. The granting of the requested funds will greatly benefit both the Nation and State by not only improving and preserving the data in a nationally consistent format, but by opening the collection to serve a wider audience through the NGGDPP National Catalog portal.

Technical Report

Overview

In 2008, the Florida Geological Survey (FDEP/FGS) was awarded funds by the USGS through the National Geological and Geophysical Data Preservation Program (NGGDPP) to provide the USGS with metadata from our well cuttings and cores collections. Funding from the earlier phase in 2007 allowed FDEP/FGS to inventory our sample collections, determine their current standing, assess future needs and conduct a number of improvements. These improvements helped to ready FDEP/FGS for undertaking the second phase of the grant period which involved validation and refinement of individual records within the collections.

At the end of the first year of the NGGDPP granting program (phase one), two major goals were established for the collection needs. The first goal was to address physical preservation of the collection data that is in danger of loss through the decline of historic data sources such as fading ink and deteriorating boxes in the physical sample collections. There is still a need to collect this information and find a way to better preserve the physical samples. The second goal was to clean up paper and electronic records for the sample collections and create corresponding metadata for an internal database. Work on this second goal comprised a majority of the effort expended during the second phase of funding (phase two). This need was expanded to our own internal database to include submission to the National Catalog. Developing plans of action for these goals as well as the work done in the two phases of the NGGDPP granting program have help to conceptualize a first draft of a long range data-preservation plan.

Funding was available through the phase two of the NGGDPP granting program to send a FDEP/FGS representative to the 2009 USGS Data Preservation Workshop in Indiana. The FDEP/FGS representative at this workshop was Sarah Ramdeen, one of the Co-Principle Investigators. During the workshop she gave a presentation entitled *Qualitative Study of Preservation Practices at State Geological Surveys*. The presentation included information from a study Ms. Ramdeen is conducting as part of her Doctoral research at the University of North Carolina in the field of digital preservation and access as well as knowledge gained from working with the sample collections at the FDEP/FGS during the first two phases of the NGGDPP granting program.

Process

In order to clear up differences between various paper and electronic sources for each sample set, a GIS data refinement process was completed by verifying a digital database with recorded locations for each sample in the collections against historic hardcopy data sources. FDEP/FGS houses hardcopy sample data in several formats, including card indexes, hanging file folders, and three-ring binders. Each data format correlates to a single sample record by FDEP/FGS well number and well location. The data verification process involved the analysis of sample records across each format, allowing the QA/QC team (consisting of two reviewers) to capture information from historic resources and enter it into the digital tables and GIS datasets from which metadata could be generated. The purpose of the data verification process was to determine the most accurate information for each record, including digital point location and sample characteristics. The verification process was conducted using many resources as detailed above. These collections include information from driller's logs, lithostratigraphic reports, district geologist cards, and sample description cards. Additional sources that were consulted as needed included well check-in logs, water management district records, and oil and gas information (Appendix A).

The basic steps in the verification process began by creating a dataset for each county from a base shapefile containing the total number of well locations statewide. Reviewers analyzed the data sets one county at a time. During the verification process, the reviewer scrutinized the location of each point

within the county dataset by comparing historic and other hardcopy documentation to determine the point's best location. Variances in the documented location were remedied by selecting the information with the most sources in agreement. Once the best location was determined, the database record and point data were corrected, if necessary, and the correct information was noted in the hardcopy documentation. Once each record in a county dataset was completed, the dataset was passed to a second reviewer who repeated the verification process, making changes or corrections if needed.

Following completion of several county datasets, the project team became aware of additional resources containing recently verified reliable location data for many samples. It was also determined that the exhaustive updating of each of the historic resources which contained redundancies proved to be time consuming. Considering these findings, the project team re-evaluated the data verification methods in order to streamline the process and incorporate the additional resources. To address these issues, the project team decided to create a new master reference dataset and to only update a select set of the historical resources.

The primary purpose of overhauling the master reference dataset used by the QA/QC team was to vastly increase the speed of the location and sample data verification process while at the same time enhancing accuracy and eliminating redundancy. These goals required a shift in focus for a period of time from the QA/QC process to researching the accuracy of locational sources. A series of meetings between the project team and key FDEP/FGS staff with high levels of institutional knowledge resulted in the consensus that digital versions of two FDEP/FGS resources, known as the Blackbooks and Graybooks, contained the most accurate internally held data for the well samples. Since the Blackbooks represented a conglomeration of previous historical data, it was determined that their use would cancel the need for reviewers to reference a majority of the historic resources. However these electronic records have never undergone a QA/QC process and as such they contain typos and other minimal errors. The locations taken from these two sources would be the skeleton of the new master dataset but the physical copies and the other historical resources would still be used as a reference to verify the accuracy of their digital conterparts.

While the locations from both sources are generally considered to be accurate, there was variance between the data they contained. For example, the Blackbooks contain only information for wells up to number 17,000, while the Graybooks contain location information for wells after 17,000. In addition, the Blackbook locations lack the precision of the Graybooks, as the Blackbook locations were recorded using Township-Range-Section (TRS). The TRS locations included in the Blackbooks are precise to four quarter sections at best, while most Graybook locations were recorded as latitude and longitude coordinates, often calculated using GPS, assuring a high-level of positional accuracy. During additional planning meetings, an extensive list of recent project-specific data sources was compiled to supplement the new reference dataset. These data sources identified include well locations that had been previously verified for specific projects at the FDEP/FGS, Florida's water management districts, and other agencies. Cross-referencing these with the FDEP/FGS well unique identifier, or 'WNUMBER', produced a set of highly accurate well locations to be compared with recorded locations in the FDEP/FGS master reference dataset.

Previously, QA/QC personnel spent much of their time manually entering required information during the verification process. This information consisted of lengthy text fields that were blank and needed to be populated according to what was listed on paper sources. Due to the changes made to the master dataset, this part of the verification process has been streamlined greatly reduced decreasing the effort required to verify each record. The final step in overhauling the master dataset was to create a series of 'Yes/No' columns indicating whether another electronic resource existed for a particular well's location as part of one or more external projects. This was accomplished with a series of table joins. The 'Yes/No' value informs reviewers whether another electronic resource for the well location is available in

a dataset from the column referenced dataset. Since these auxiliary datasets are also digital, the need to review paper records has been vastly reduced.

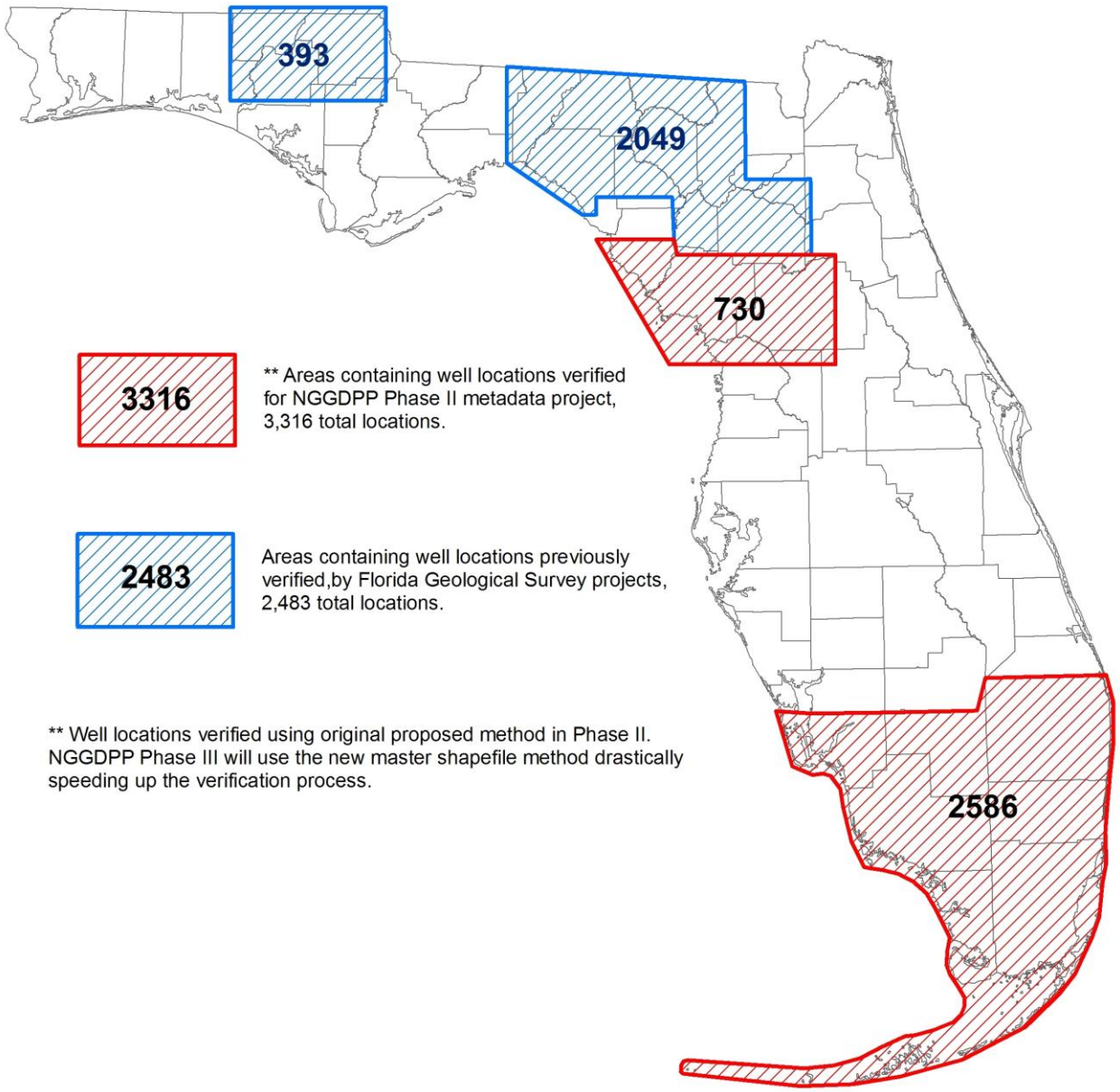
Outside project sources that were used in this endeavor include datasets cleaned for prior research publications, driller notes kept by FDEP/FGS Drilling Operations staff, wells drilled by the FDEP/FGS Coastal section for research projects, highly scrutinized wells providing the basis for the STATEMAP project, wells documented through permits tracked by FDEP's Oil & Gas Section, and well collections maintained by water management districts. These newly inventoried auxiliary sources will provide a solid comparative basis for QA/QC staff in the process of selecting the most accurate and precise location for each FDEP/FGS well.

During this project 3,316 records were verified for the FDEP/FGS cuttings and core sample collections. An additional 2,483 records wore previously verified from prior FDEP/FGS projects. In total, metadata for 5,799 records have been uploaded to the USGS National Catalog (Appendix A). These records contain the six fields required for the USGS National Catalog, as well as 35 additional fields that were determined to be critical to the FDEP/FGS. As the National Catalog will act as a portal leading users from the USGS site to the FDEP/FGS for more detailed records, it was determined that the inclusion of these fields was essential to completing the data set. As part of this process we have also created archive sets of shapefiles for each county that have undergone the QA/QC process and a data dictionary was created which contains definitions of all of the columns contained in the master files (Appendix C).

In accordance with the proposal, the completed dataset was to be uploaded to the National Catalog in XML format; however, after submission of the proposal, the primary database manager for the project left FDEP/FGS. Unfortunately, due to budget issues, there was a delay in the replacing that member of the project team. Given the shortened time window available for testing and uploading files to the National Catalog, the determination was made to begin the upload in CSV format (Appendix B), the format most familiar to the new database manager and appropriate for use with the National Catalog. This determination was made to ensure ample time to overcome any obstacles that may arise with the upload process and meet the targeted deadline. During the next project phase, the project team will determine whether to continue with the CSV process or to develop a procedure for XML uploading.

While the initial goals to clean up to 10,000 records has not been met in this phase, the FDEP/FGS has made significant progress towards one of the two major goals established in phase one of the NGGDPP funding program. At the start of this phase, there was a steep learning curve which required a redesign of the QA/QC process. This has lead to the creation of a better base metadata set for our sample collections which will only improve the end results of the QA/QC process. This also ultimately benefits the FDEP/FGS as there is now documentation about historical resources and the quality of their data which was not previously recorded and was at risk of being lost as staff members retire. Funding for phase three has already been authorized and will continue on the work established during the course of phase two. This process is now streamlined and will proceed at a much faster pace.

Appendix A


393
2049
730
2586
3316
** Areas containing well locations verified for NGGDPP Phase II metadata project, 3,316 total locations.
2483
Areas containing well locations previously verified,by Florida Geological Survey projects, 2,483 total locations.
** Well locations verified using original proposed method in Phase II. NGGDPP Phase III will use the new master shapefile method drastically speeding up the verification process.

Paper data sources from left to right: 3 sets of index cards cataloged and cross referenced by PLSS, well number, and county and 158 "Greenbooks" binders cataloged by county containing driller's logs, lithologic descriptions, and miscellaneous information

Appendix B

Details for the CSV conversion and upload process:

1) Go to the folder U:\Staff\OPS\Wood_K\dbfs_to_XML-CSV
2) Run the batch file: run_USGS > except_mm_dd_yyyy.txt (where mm=month number, dd=day number, yyyy=millennium
3) After the run ends (a few minutes) check the output files created exceptions_01.csv and except_mm_dd_yyyy.txt. Check over the files listed (if any) and they will need correcting later.
4) The translated files will all be named with prefix P584 or P582. These will be the ones to upload to USGS.
5) Load each output file into a text editor like Notepad++ (not the usual windows notepad!!). Look for extraneous line feeds (line breaks) and backspace over them to get rid of them.
6) Go to the web page: http://my.usgs.gov/csc/nggdpp/upload and log in there.
7) Follow the on screen instructions for each file for upload and verify. If a file fails verification read the diagnostics at the bottom of the web page and figure out what needs to be done to satisfy the verification process they use at USGS. Usually something simple.
8) If the file passes verification then you can click the button to publish it on USGS.
9) Continue with step #7 - #8 on each file till you are done.
   Take the exception files and see what needs to be done to fix the records for upload later.

Appendix C

FGSWells / NGGDPP Phase II – III Master Reference Shapefile and Spreadsheet Data Dictionary

Field properties are shown in this font and color.

WNUMBER – FGS-assigned master catalog number for samples received from borehole locations. *Long integer, 0 decimal places.*

COUNTYCODE – 2-character label that corresponds to an FGS-only county labeling convention. *Text, 2 characters.*

TOWNSHIP – The number of the township in which a borehole is located. *Long integer, 0 decimal places.*

TWP_CARD – The cardinality (N/S direction) of the township. *Text, 1 character.*

RANGE – The number of the range in which a borehole is located. *Long integer, 0 decimal places.*

RANGE_CARD – The cardinality (E/W direction) of the range. *Text, 1 character.*

SECTION – The number of the section within which a borehole is located. *Long integer, 0 decimal places.*

QUARTERS – The quarter-section location of a borehole, based on FGS' internal style of labeling: NW¼ =A, NE¼=B, SW¼=C, SE¼=D. Quarters increase in size when reading this field from left to right. *Text, 4 characters.*

LAT_DEG – Degrees Latitude of a borehole's location. Always a whole number. *Long integer, 0 decimal places.*

LAT_MIN – Minutes of latitude. Always a whole number. *Long integer, 0 decimal places.*

LAT_SEC – Seconds of latitude. Generally has no more than 4 decimal places. *Double, variable number of decimals.*

LAT_DIR – Cardinality of a well's latitude. Auto populated and always 'N' or 'S'. *Text, 1 character.*

LAT_DD – Calculated latitude in Decimal Degrees (DD) format. Will be auto-calculated prior to datasets being uploaded to the National Catalog. *Double, variable number of decimal places. ArcGIS truncates at 6 decimals, generally.*
LONG_DEG – Degrees Longitude of a borehole's location. These values appear as positive numbers, so their values must be multiplied by -1 to get the true longitude. Always a whole number. *Long integer, 0 decimal places.*

LONG_MIN – Minutes of longitude. Always a whole number. *Long integer, 0 decimal places.*

LONG_SEC – Seconds of longitude. Generally has no more than 4 decimal places. *Double, variable number of decimal places.*

LONG_DIR – Cardinality of a well's longitude. Auto populated and always 'E' or 'W'. *Text, 1 character.*

LONG_DD – Calculated longitude in Decimal Degrees (DD) format. Will be auto-calculated prior to datasets being uploaded to the National Catalog. *Double, variable number of decimal places. ArcGIS truncates at 6 decimals, generally.*

COUNTYNAME – Full name of the county in which the borehole is located. *Text, 50 characters.*

WELL_NAME – The documented name associated with a borehole. *Text, 255 characters.*

WELL_USE – Purpose for which a well is used, if available. One of several standardized categories. *Text, 50 characters.*

DOC_ELEV – Documented elevation at the land's surface at the top of the borehole, usually in vertical datum NGVD29. May or may not be furnished with other documentation. *Double, 2 decimal places.*

DEM_ELEV – The elevation at the surface of the borehole according to the statewide 15meter Digital Elevation Model (DEM). This will be auto-calculated prior to the next stage of QA. *Double, 2 decimal places.*

SMP_DELETE – Auto-populated Yes/No column. Whether a well's samples are currently listed as being previously removed from the FGS collection. Will be updated pending Drew Butler's finalization of Warehouse inventory. *Text, 3 characters.*

DELCONFIRM – Whether FGS Warehouse staff have confirmed the removal (deletion) of a well's samples. *Text, 3 characters.*

OWNER_DRLR – Name of the owner, or owner and driller, of the well. Empty fields should be updated whenever possible. *Text, 255 characters.*
ISMOREINFO – Yes/No column: Is there more information of public interest available for a borehole, which is listed in well index cards but not recorded in this file? In most cases this will be 'Yes'. *Text, 3 characters.*

TOTALDEPTH – Total depth in feet below land surface (BLS) at the bottom of a drilled well. May or may not be included in available documentation. *Double, 2 decimal places.*

SAMP_TYPE1 – Type of sample listed first in the database's Sample Table. Either Cuttings or Core. Is auto-populated but needs cross-checking with other sources. *Text, 50 characters.*

TOP_SAMP1 – Depth in feet at which the upper limit of sample type 1 is located. Is auto-populated but needs cross-checking with other sources. *Double, 2 decimal places.*

BTM_SAMP1 – Depth in feet at which the lower limit of sample type 1 is located. Is auto-populated but needs cross-checking with other sources. *Double, 2 decimal places.*

SAMP1_INT – Interval in feet at which sample type 1 was collected (i.e. 5, 10, 100). Is auto-populated but needs cross-checking with other sources. Note that according to Sarah this information is not always recorded, so don't spend too much time trying to find it.
Long integer, no decimal places.

SAMP_TYPE2 – Type of sample listed second in the database's Sample Table. Either Cuttings or Core. Is auto-populated but needs cross-checking with other sources. *Text, 50 characters.*

TOP_SAMP2 – Depth in feet at which the upper limit of sample type 2 is located. Is auto-populated but needs cross-checking with other sources. *Double, 2 decimal places.*

BTM_SAMP2 – Depth in feet at which the lower limit of sample type 2 is located. Is auto-populated but needs cross-checking with other sources. *Double, 2 decimal places.*

SAMP2_INT – Interval in feet at which sample type 2 was collected (i.e. 5, 10, 100). Is auto-populated but needs cross-checking with other sources. Note that according to Sarah this information is not always recorded, so don't spend too much time trying to find it.
Long integer, no decimal places.

VERIFIED – Initials, and brief notes if necessary, corresponding to the first and second verifications of a record. *Text, 50 characters.*

VERIFYDATE – Date of the final (second or cross-check) verification of a record. In MM/DD/YYYY format. *Date of type MM/DD/YYYY.*

DRILL_DATE – Date on which the drilling of a borehole was completed, in MM/DD/YYYY format. Not always available, but the date on which samples were received should *not* be substituted for this. *Date of type MM/DD/YYYY.*
COMMENTS – Any pertinent information describing something of interest encountered during the well QA/QC process. *Text, 255 characters.*

LOC_SOURCE – The name of the source whose coordinates are ultimately recorded as verified and more accurate than those from any other source (see post-Blackbook sources below). This field allows for some additional notes to be entered, possibly to indicate why a particular source was chosen, or other pertinent information. If more accurate recent locational data cannot be found for a well, this should either be Blackbooks or Graybooks. On occasion, the use of a location from a paper archive source such as a district geologist's card or a record from the Greenbooks may be warranted, and should be recorded here as such. *Text, 255 characters.*

Reference Sources

ALLWELLS06 – Yes/No. Whether a well's origin is from the 'allwells_withBB_2006' shapefile created by Alex Wood in 2006, and later joined with the digital Blackbooks. NOTE: This source contains many locations known to have flaws, but also has some locations that were re-plotted and corrected. This should *only* replace an existing Blackbooks or Graybooks location if it correlates closely with one of these locations. In cases where other sources list locations that also appear in this source, the other sources should be weighted more heavily. Auto-populated prior to verification. *Text, 3 characters.*

BULLETIN68 – Yes/No. Whether a well was QA'd for inclusion in FGS Bulletin 68. Auto-populated prior to verification. *Text, 3 characters.*

IS_COASTAL – Yes/No. Whether a well has been QA'd for prior use in a FGS Coastal Section project. Collection maintained and updated by Michelle Ladle. Auto-populated prior to verification. *Text, 3 characters.*

HAS_LITHO – Yes/No. Whether a well has a lithologic log (.DAT file) associated with it for possible additional info. These don't include new, unverified lithologs. Auto-populated prior to verification. *Text, 3 characters.*

IS_NWFWMD – Yes/No. Whether an FGS well can be cross-referenced with the NWF Inventoried Wells collection. Auto-populated prior to verification. *Text, 3 characters.*

IS_OG_PRE – Yes/No. Whether a well has a WNUMBER (but no permit number), and can be cross-referenced using Oil & Gas pre-permitted locations. Auto-populated prior to verification. *Text, 3 characters.*

IS_OG_POST – Yes/No. Whether a well has a WNUMBER and an Oil & Gas permit number (and also an associated location) for cross-referencing. Auto-populated prior to verification. *Text, 3 characters.*
IS_ROMP – Yes/No. Whether a well has been included – and therefore extensively QA'd – for the ROMP project. Auto-populated prior to verification. *Text, 3 characters.*

IS_SFWMD – Yes/No. Whether a well's location has been previously verified by SFWMD and pulled from their online database. Auto-populated prior to verification. *Text, 3 characters.*

IS_SJRWMD – Yes/No. Whether a well's location has been previously verified by SJRWMD staff and provided to us by Don Boniol (SJRWMD hydrologist). Auto-populated prior to verification. *Text, 3 characters.*

IS_SRWMD – Yes/No. Whether a well's location has been previously verified by SRWMD staff and provided to us for cross-referencing. Auto-populated prior to verification. *Text, 3 characters.*

ISSTATEMAP – Yes/No. Whether a well has been previously scrutinized by Dave Paul as part of a STATEMAP project. Auto-populated prior to verification. *Text, 3 characters.*

ISDRILLOPS – Yes/No. Whether a borehole was drilled by FGS Drill Ops staff and recorded in Ken Campbell's spreadsheet of completed wells. Auto-populated prior to verification. *Text, 3 characters.*