Toxics Redo Process Steps

1. Pull data from WQP. See 01\_WQPQuery.R for exacts. Basically used REST service to query WQP. Built out WQPNameMatch\_05142014.csv to relate CharacteristicName from WQP to criteria names. This was a manual process searching the EPA Substance Registry System (<http://ofmpub.epa.gov/sor_internet/registry/substreg/searchandretrieve/substancesearch/search.do>) and checking CAS numbers to Table 30 and Table 40 values and against the CharacteristicName to make sure we were getting the same compounds. There more than 100 and maybe more than 200 individual characteristicNames we checked the WQP for and putting them all into a single REST query took WAY too long. To get around this I wrote a for loop to query each parameter individually and kept track using the success table which logged whether the query was successful and how many stations and rows of data were returned. Note that in this processing the WQP would only allow a certain number of successive queries before it would return an error. I put in the ability to break the loop so you could restart at the last iterator after a few minutes when this happens. Once data was retrieved, whole row duplicates were removed and the success table updated accordingly. The results of the query were saved to the WQAssessment database with the names WQPData\_05022014, WQPStations\_05022014 and WQPQuery\_Status (the success table).
2. Then, QC had to be completed on the returned data. This is captured in 02\_WQPData\_Preprocessing.R. In order to complete this, we first consulted with Scott Hoatson and Sarah Rockwell of the QA section at the Lab to determine what minimum QC requirements we had. We determined that to give a complete grade required complete QC information which was not available, however for the purposes of the Integrated Report we utilize Grade B and Grade A data so we clarified what was required to make the Use/Not Use determination. Basically, if there was any indication of contamination or complete lack of indication of QC processes used then we wouldn’t use the data. Things like hold time violations only lead to a grade of B and so were still included in the usable dataset. Once this was determined I took a two phase approach to QC the data. The first was high level and only checked the dataset to see if ANY QC comments were present within a single organization’s dataset. If no comments were present or there were inconsistencies in the reported fields (e.g. Bureau of Reclamation – Composite versus Grab) then all the data were excluded. If there were QC comments present within an organization’s dataset then the data were retained. The second phase was to process the QC comments for the included data. If the QC comment indicated contamination or major QC problem, then the data were excluded.
3. Then LASAR was queried for all toxics parameters from Tables 30 and 40. This was done using 01a\_LASARQuery.R. The relationship between LASAR names and Criteria names was completed in 2010 but the criteria have changed enough that a new effort was completed and this is captured in the R script (should probably be output to a table…..). The LASAR query specified A+, A and B data only. The only other processing in this file involved creating a numeric result field.
4. Next, we started in on locating Monitoring locations. This starts with 03\_USGS\_Station\_table\_build.R. Several USGS stations were utilized and located in the first 2012 IR effort, however, it was discovered that there were inconsistencies between the Toxics RM assignment and the RM assignment for the DO analysis. Because of this, all previously located USGS stations will be included in the locating process again to determine correctness and to resolve differences in the data. There were a couple stations that we got from the USGS Oregon Water Science Center (from the first 2012 IR attempt) that aren’t in NWIS so those we made sure to include in the list of stations to locate.
5. Well actually, the first station location attempt started with EPA only stations but we found that several of those stations only had latitude and longitude to one or two decimal places which resulted in several false spatial duplicates. We abandoned that first attempt and worked through the EPA dataset to determine what should be where. This is in the file 03a\_EPA\_LASAR\_Stations\_Resolution.R. Knowing that ODEQ were the ones to collect the EMAP data I thought maybe folks at the lab would have better station locations. I communicated with Greg Coffeen who conducted most of the CEMAP monitoring and compiled that data into an Access database. That db is the CoastalEMAP1999-2006\_FinalDatabase\_AIIMASTERTables.mdb. It didn’t have a complete lookup between EMAP station ID and LASAR ID. He then sent an excel file that I converted to a .csv for ease of import, OR\_CEMAP\_FinalStationList99-06.csv. This showed that all the CEMAP stations (From OrganizationFormalName ‘Environmental Monitoring and Assessment Program’) we had in our dataset had been given LASAR Station ID’s. Because we queried LASAR for the same parameters I checked for overlap in the data as well. There was some overlap but only in the Nitrate data. These duplicates were removed from the WQP data set and kept in the LASAR data set. It does not appear that any of the Superfund sites are in LASAR nor are any of the NARS/NRSA sites. There is a subset of the NRSA sites that have similar lat/lon truncation issues. I spoke with Shannon Hubler at the lab who informed me that NRSA sites were NOT entered in LASAR. He did however have better latitude and longitude than what was in STORET for 139 of the 175 stations. This was considered good enough and the remaining 36 with poor lat/lon were excluded. The list of stations then to be located from the EPA includes all the other organizations except for Environmental Monitoring and Assessment Program & ‘National Aquatic Resource Survey Data’. This file also identifies lasar stations NOT in the geodatabase that will also need to be located.
6. Moving onto to preparing the data and conducting the evaluations. In order to calculate the evaluations several decision rules from the Assessment Methodology had to be implemented. These are captured in 04\_Data\_Preparations.R. The first step was combining the datasets and pulling in the criteria. These processing steps were adapted from an earlier effort used to compare Toxics Monitoring Program (TMP) data to all available criteria or benchmarks. As such a criteria table was created for that effort which was used for this effort subbing out the DEQ specific standards. In working through these criteria and implementing Assessment Methodology steps it became apparent that there were Parameters/Pollutants that were not included in the original queries in addition to the condition specific parameters required to calculated hardness based and ammonia criteria. In order to include these parameters in the original datasets the file 01b\_Missed\_Parameters\_Query was run in many times as it was realized that another parameter had not yet been included. This file cannot be run from top to bottom. Instead the additional data were included in updated tables in the WQAssessment database. Handling detects and non-detects had to occur on each individual source dataset in order to accommodate the different methods that each system used to indicate detection status. Once all the data were compiled units were made consistent. One critical decision rule that took some time to sort out conceptually was the application and handling of sample fractions. The criteria can be for dissolved or total fractions so I initially thought it made sense to make the criteria specific to fraction which worked for the TMP data since we were collecting both fractions for the whole project. It did not work for the Integrated Report for two reasons. First, the Assessment database stores pollutants by elemental name only and makes no distinction for fraction even if it was considered in past assessments. Second, the data we queried is inconsistent for each sample. Some have dissolved only, some have total only and only a few have both. So it became necessary to track pollutant names in several different forms. There are the pollutant names alone from the original data source, the fraction associated with those pollutant names, the criteria table names alone and their fractions and the assessment database pollutant names alone. We also had to deal with duplicate samples which occurred in a couple different forms. First there are field duplicates, also lab duplicates and further duplicates where a single sample is reported twice with two different MRLs. Station type or ‘matrix’ are pulled in from Mike’s station locating efforts. Per the Assessment Methodology in order to evaluate some of the criteria we have to total different speciations. This is the case for Endosulfan, Nitrosamines, Hexachlorocyclohexane-Technical, Chlordane and PCBs/Aroclors. DDT is also a sum for Aquatic life but the criteria for the total is always less stringent than the criteria for the individual speciations. This part of the script also calculates hardness where we have calcium and magnesium ions. Then we merge each sample with the criteria and calculate the condition specific criteria. It turns out that like for the DDT the more complicated way of calculating the standard for pentachlorophenol is never more stringent than the static value standard. Then we pick the minimum criteria for each sample. I tried this using ddply but took WAY too long. Then I tried dplyr but it didn’t allow me to apply a custom function. Through stackoverflow searching I found a data table solution which actually runs VERY fast. A package worth considering when working with other large datasets although I prefer the attempts Hadley Wickham is making to merge human readability with performance. Next, we also decided that we would convert sample fractions when we didn’t have the appropriate fraction available. Then we actually determine the exceedances (which is only 1 line! A lot of work for something so very simple!). Following are the validity determinations dealing with MRLs, appropriate fractions and resolving duplications with total and dissolved.
7. Then we move onto the Station summaries in the file 05\_Station\_Summary\_and\_Determination.R which I implement using dplyr for its readability. Then we apply the methodology for determining Category assignment at each station accounting for only assigning 3B status to Alkalinity and Phosphate phosphorus. These summaries and the data evaluations were output to an access database (2012\_WorkingTables.mdb) where they were linked as subtables so that the data evaluations would be immediately available from the station summary in the table. I wanted to this from the SQL tables but the subdatasheet function in Access only works on local tables. Then we had to relate the names the WQAssessment pollutant names (Thanks Karla!). The last part of the file pulls in the ars in its most current edited form and checks each of the stations in the redo to see if it’s falls on an existing assessment unit and separates them out into separate data objects. One thing that came up during our review was that I had assumed that the LLID\_Stream\_Lake was the field we should be using to use for the look back for identifying existing segments and using for creating new ones. It turns out this field is really just for display purposes and facilitating searching in the web database and conceptually we really only care about the LLID\_Stream for matching purposes. So in cases where no LLID\_Stream is identified we look for a LLID\_Lake but we don’t want the assessment units to inherently account for reservoirs based on the current model we use for determining assessment units.
8. The next script 06\_Exist\_Recs.R works through updating the ARS entries for the assessment units and updated them with the new analyses.
9. The next script 07\_New\_Recs.R identifies those LLIDs where that pollutant has been analyzed before but the assessment unit does not capture the new stations and those LLIDs that have not been analyzed before for each pollutant. The LLIDs that have been analyzed before required some pretty manual editing which is handled in the file 07a\_2012\_IR\_New\_Recs\_Exceptions.R. For some reason in this set of data we didn’t really end up with mixed category stations on a new LLID. Maybe this suggests that when there is a problem with a pollutant it is truly persistent. That or it just reflects that we really only have a very small number of stations with enough data to draw our conclusions. Both the exist\_recs and new\_recs scripts were adapted from the first draft edit. They have a lot less specific edits to the ARS and so are probably a better starting point to go from when trying to make these more general.
10. The next script 08\_Final\_Table\_Edits.R make some changes to the final tables to make sure they have consistent formatting. There is also some draft code trying to think through how to put the Record\_ID assigned to the grouped stations for new records back into the station so that it can be used as a lookup to link the station to the assessment unit beyond just the text summary.
11. The next file 09\_Database\_Update.R doesn’t actually update the database. This was more of an adaptation from the original effort. The problem we have is that we can’t write directly from R to the SQL database because the RODBC drivers don’t operate using the additional memory available in a 64-bit system (my understanding of this is foggy at best and probably laughable by someone who actually understands such things). Basically every time you try to use the sqlSave function you get the R error that it cannot allocate a vector of size x. Previously I got around this by splitting the data frame into three parts because the offending columns were really the Summary and the Comments. This time however, even those weren’t writing out. So I saved the whole table to a .csv instead. Then I imported it into Access and then from SQL Server Management Studio was able to import it from Access. Weird. But it works.
12. Then after all of that my time ran out and we were not able to spend enough time reviewing the final results and it was determined that we would not be finishing this work at all and reverting back to the first draft and only make minor edits to respond to some of the minor comments. This wasn’t too much of a problem for the analysis since we hadn’t overwritten anything and we could just leave the Redo Draft alone. It was a little more complicated for Mike since he had been building the StationUseList and the geodatabase station feature class based on the assumption that a certain set of stations should be included. Instead he had to differentiate the new stations we had gone through the effort of locating from the ones we actually used in the first draft. Then as it is we may not even be submitting the first draft as a final attempt and abandon the 2012 entirely. Oh well! As of 8/21/2014.