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 those 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 that may have to be excluded (5/28/2014). The list of stations then to be located from the EPA include all the other organizations except for Environmental Monitoring and Assessment Program (& for now ‘National Aquatic Resource Survey Data’).
6. Next, dealing with spatial duplicates still has to be taken care of because there is overlap