Downloading & Using Data from the STORET Warehouse: An Exercise

- - - - - - An Exercise for the BEACH Program to QA IDs - - - - August 2012

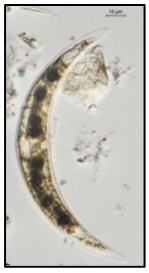
Parts 1 & 2 of this exercise walk you through the steps to generate and download a report to QA ORG_ID, PROJECT_ID (same as BEACH_ID), STATION_ID from the STORET Data Warehouse. The 3rd part describes how to import the data into Microsoft (MS) Excel, a software that can be used to QA your IDs. For questions, contact STORET Technical Support at 1-800-424-9067 toll free or storet@epa.gov; or for BEACH Program questions: kramer.bill @epa.gov, 202-566-0385.

The original STORET document upon which this is based can be found on: http://www.epa.gov/storet/Downloading_STORET_Data.pdf It has a 4th part that explains how to import data into MS Access to chart monitoring trends and a final section that shows how to download data by watershed using the Watershed Summary application. Please contact STORET Technical Support at 1-800-424-9067 toll free or storet@epa.gov for comments or questions on the original document.

Part 1: How to Query and Download the Data

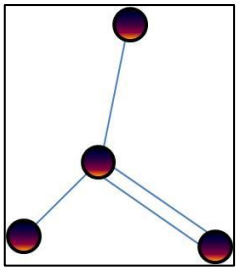
The STORET database can be used to access data on specific water resource chemical, physical and biological characteristics and parameters as well as methods used in assessments. All Data downloaded from the STORET warehouse references a data owning organization, or the organization responsible for collecting the data. STORET can be queried for monitoring location information, as well as the data collected at monitoring locations. STORET queries generate a file containing result data files, metadata files, and available reference documents associated with the data owning organizations. Downloaded data and document files are compressed in a (zipped) .tar.gz file format.

.



Biological Data

Habitat Data



Chemical Data

For this example, we will search for Louisiana Beach monitoring data from all stations with 2009 as the specified monitoring timeframe. The downloaded file will be a compressed (zipped) .zip file containing result data files, and metadata files; you will be primarily looking at the results file.

Step 1. Access the STORET Data Warehouse

a.) Go to the STORET main page: http://www.epa.gov/storet/

b.) Under the box titled Features, click the **Download Data** link

c.) Under ***The STORET Data Warehouse***, click the yellow button titled ***Browse or Download Modernized STORET Data.***

TIP: Legacy (pre-1999) data is also available (flat filed) by State/ County for easy downloading

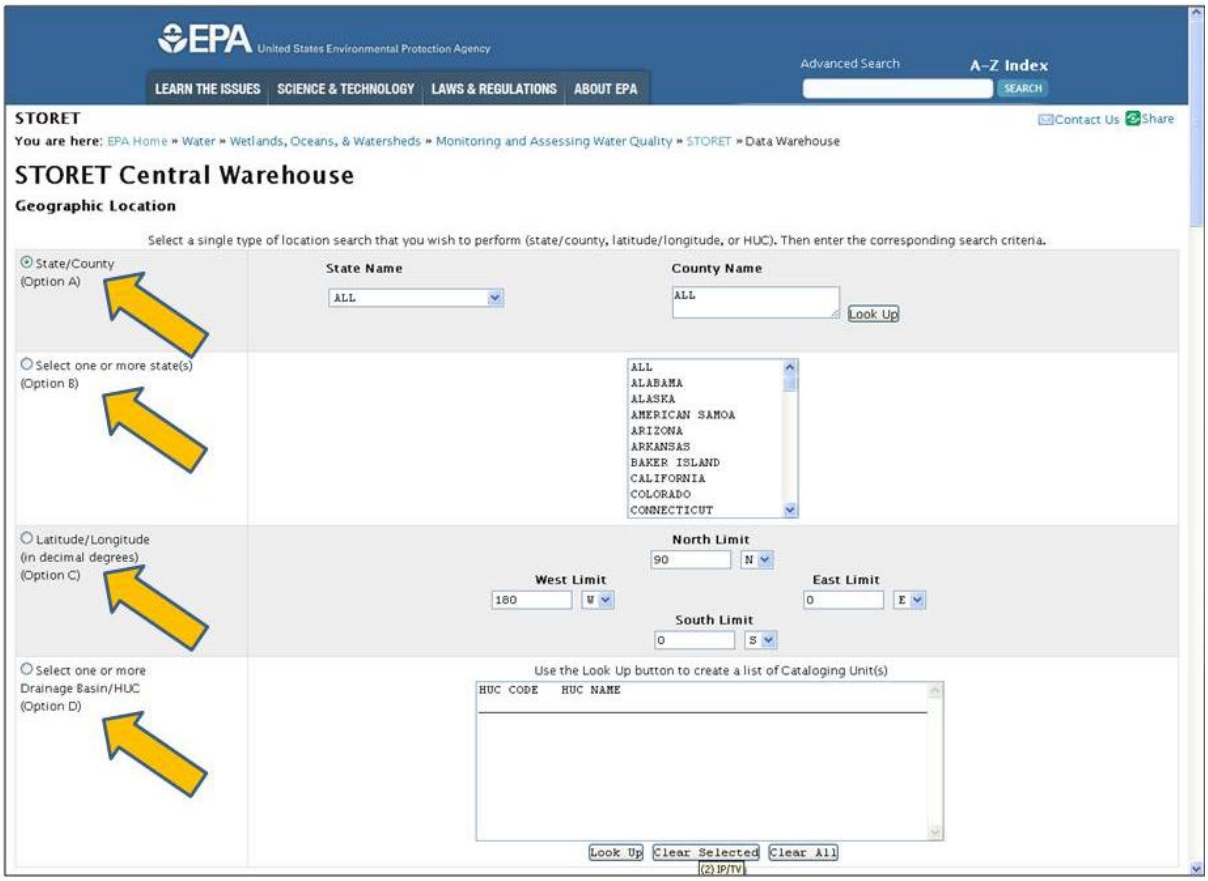
Step 2. Define Your Query

Queries can be defined by one or multiple parameters including *Geographic Location, Organizations, Stations, Projects, Station Types, Date, Activity Intent, Community Sampled, Characteristics and Activity Medium.*

d.) Under ***STORET/WQX Warehouse Reports-STORET Results*** click the link titled **Results Download**

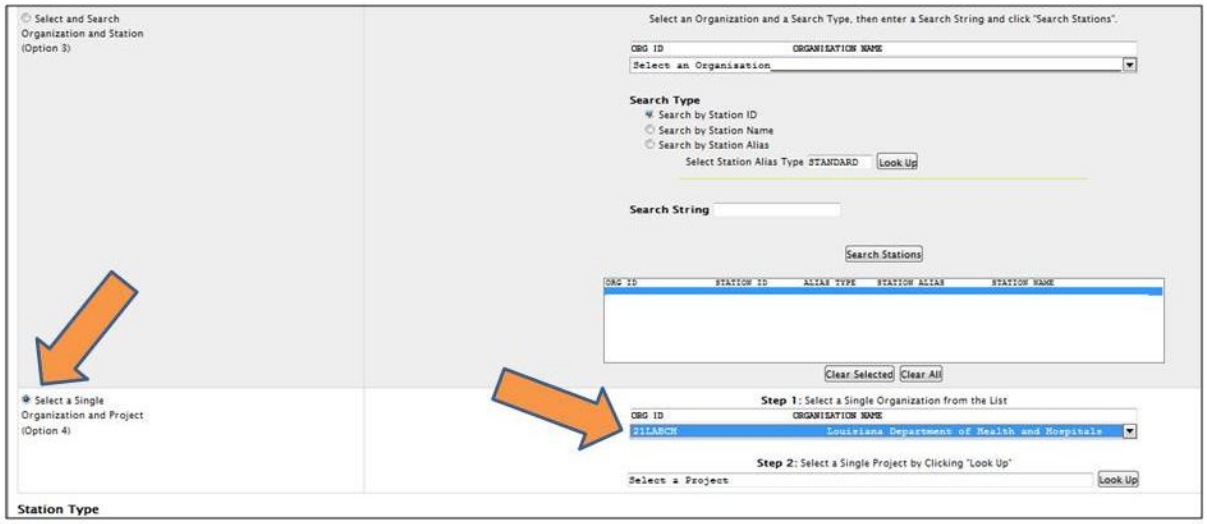
TIP: Dissolved oxygen is a physical/ chemical (regular) parameter (not bio or habitat)

e.) Geographic Location information can be further subdivided by State/County, Latitude Longitude Coordinates or drainage basin/HUC.



f.) Do not indicate a specific Geographic Location and move on to step g).

g.) Organization, Station and Project data can be further subdivided by Organization Type, Organization(s), Station(s), or Project(s). Select the Organization type **Select Option 4 (a Single Organization and Project)** and from the ORG ID drop down list enter the search criteria **21LABCH –Louisiana Department of Health and Hospitals.**



h.) Under **Station Type**, leave the defaults; this will capture all station types

TIP: Use the <Ctrl> key to select all BEACH-Program specific Station Types (3/4' s to bottom of the picklist)

i.) Under **Date**, **change the start date to JAN 1, 2009 and the end date to DEC 31, 2009.**

TIP: Leaving the default will capture all date ranges of STORET Data Warehouse (non-Legacy datasets)

j.) Under ***ACTIVITY MEDIUM*** , Select "***Water***"

k.) Under ***ACTIVITY INTENT*** and ***COMMUNITY SAMPLED***, leave the defaults: this will capture all sampling intents.

l.) Under ***Characteristic***, do nothing

TIP: The ***Characteristic Search*** box utilizes a *Beginning with* type search. For example, with regards to dissolved oxygen, typing in leading characteristics such as "disso" will return a list of matching parameters with characteristic names starting with "disso" like dissolved oxygen.

TIP: The percent sign "%" is a wildcard search prefix that searches for parts of a characteristic name if the full name is unknown (example: typing "%disso" or " %oxy" or "%DO" all work)

m.) Click the ***Continue*** button at the bottom of the screen

Step 3. Download Your Query Results

n.) Note the number of records found

TIP: MS Excel holds one million records per sheet, MS Access holds more but is limited to 3 gigabytes

TIP: If there are too many records, you may have to go back to narrow your query

TIP: You can adjust your query by narrowing the date fields or limiting characteristics

o.) Select the report types (***REGULAR, BIOLOGICAL and HABITAT***), "Regular" is the default

p.) In the ***Please select the appropriate user profile*** box Indicate the, specify your profile.

q.) Type your email address in the ***Please enter your email address*** box

TIP: Emails are used to notify you that your download is first processing and then completed

r.) Type a three character prefix like "XYZ" in ***Please specify three characters to prefix your report name*** box

TIP: This prefix will help to identify your download file later

s.) Scroll to the bottom of the screen to the **Select Data Elements for Report** checklist **– for the purpose of this example, of a state QA of the key IDs, click "Clear All", then select: Org ID, Beach ID/Project ID, Station ID, activity start date, Characteristic and Result Value as Text.**

TIP: These are the fields you will see in your report, you can select/de-select any you choose

t.) Under **Batch Processing**, click the **Immediate** button and note the URL and see if it has your 3 character prefix

TIP: ***Immediate*** and ***Overnight*** Reports follow the same directions but small (<300K records)

Immediate Reports are available in 1-15 minutes while ***Overnights*** (600K max) are next day

u.) Go to your e-mail account. You are waiting for 2 e-mails, including a Processing e-mail and a Completed e-mail.

i. When you receive the PROCESSING email, open and check that the URL matches the earlier URL

TIP: If you click the URL now you may go to an error page because the file is not ready yet

ii. When you receive the COMPLETED email, your file is ready to download; click on the URL

v.) Note the filename and click the **Save** button; save download file to your desktop or other directory; click **Close**

w.) DONE.

Part 2: Making Sense of Your Downloaded File

Now that you have your downloaded file, what is it and what do you do with it? This section will first explain how to remove the data in your downloaded file, answers common questions about the downloaded file, and lastly tells you how to identify the various files by their conventions. The result file will be renamed to be used in Parts 3 and 4. (Note: This exercise was written using WINZIP® 9.0. Some features may be different for other versions.)

Step 1. Retrieve your results textfile from the download

a.) Navigate to the directory where you saved the downloaded file from the STORET Data Warehouse

b.) Create a folder in this directory and name it **storet_data**

c.) Double-click the downloaded file to open it and click the **Yes** button when asked to decompress the file

TIP: Most compression engines like WINZIP® will be able to open the .zip file

d.) Extract all the files to your new folder named **storet_data**

TIP: Files with the Data_ prefix denote Regular, Biological, Habitat, or Metadata results

(Metadata is provided so users can explain data. Each one of the Regular, Biological, Habitat, or Metadata Results will have a separate file.

TIP: Metadata Results contain information to help you determine the quality of the data

TIP: Reference documents such as .pdf documents, and Files with the RefDoc_ prefix denote project-level reference documents associated with the organizations that own the data

e.) Rename your **Data_XYZ_....._RegResults.txt** file to **storet_data.txt**; this file contains your requested data.

f.) **DONE**.

QUESTIONS ABOUT THE FILES IN THE DOWNLOAD

What is a .ZIP file, anyway, and why isn't my download a textfile?

• As noted in Part 1, the downloaded .ZIP file is a compressed (zipped) file. This means that you will need compression software like WINZIP® to open the file. It was necessary to move to this compressed format for both ***Immediate*** and ***Overnight*** downloads as all downloads now contain multiple files, including your results textfile.

Why are there more files than just my query results in the downloaded file?

• In addition to the results data that you queried, you now automatically receive the metadata file and any (if any) project-level reference documents associated with the organizations that own the data. This is so that you can better determine the quality of the data you downloaded. The result file(s) contain the raw data that you queried; regular, biological, and habitat result queries are found in individual files. The metadata file includes information about the organizations that own the data including contacts, methods, labs, and other info. The reference document will also contain links associated with results and data owning organizations encompassing pictures, datalogger results, QAPPs, .pdfs or any project-level documents.

I can't make sense of my results file when I open it.

• All result (RegResult, BioResult, HabResult) files and metadata files are in a tab "" delimited format, the default format for Microsoft Word, Access or Excel. The delimiter format allows a database (Excel., Access) to organize a file (make it readable). Step by step instructions for importing data are given in part 3.

Is there any useful information in the metadata file?

• The metadata file contains the following summaries that can be used to contact the data owners, create a station list, describe the methods and procedures used, qualify the labs, correctly cite the data, and generally determine the quality of the data for yourself:

Organization summary Cooperating Organization Summary
Project Summary Sample Collection/Creation Procedure Summary
Sample Gear and Equipment Configuration Summary Sample Preservation and Handling Profile Summary
Analytical Procedure and Equipment Detail Summary Laboratory Summary

Lab Sample Preparation Procedure Summary Bibliographic Citation Summary

CONVENTIONS:

The files found in the download have four main components

1) Type of Document ___: Prefix denoting the document file is a data or reference document (Data_, RefDoc_)
2) Unique Identifier ___: 3 char ID given, followed by the date/time stamp (_XYZ_'year*mm*dd'_'24hr*mm*ss'_)
3) Type of Data ___: Suffix denoting the document contains results data, metadata, or reference data (_RegResults, _BioResults, HabResults, _Metadata, _Project_'PROJECTID'_"filename")
4) Type of File ___: Extension denoting the format of the document file (.txt, .pdf, .bmp, .gif, .jpg)

Examples:

Data_XYZ_20070322_205714_Metadata.txt

Data_XYZ_20070322_205714_RegResults.txt

Part 3: How to Import and Analyze the Data in Microsoft Excel

Now we're going to import the downloaded data into Microsoft (MS) Excel, perform some rudimentary analysis, and graph the data for one station in the dataset. MS Excel can only hold 1 million records per sheet (Note: This exercise was written using MS Excel 2007. Some features may be different for other versions).

Access a video tutorial of opening a delimited file on Microsoft Excel: http://www.epa.gov/STORET/tutorial/Analyze_delimited_file.swf.html.

Step 1. Import the data into Microsoft Excel

a.) Open Microsoft Excel

b.) From the main toolbar, click ***Data>From Text***

c.) In the ***Import Text File*** window

i. Navigate to the directory or folder where you have saved your downloaded STORET file

ii. In ***Files of Type*** textbox, select **"All Files"** or **"Text Files"** from the drop-list

iii. Select and double click your saved STORET File (renamed **storet_data.txt** in Part 1 of exercise)

d.) In ***Text Import Wizard***

i. The STORET Warehouse query application delivers data requests in a tab delimited, text file format. Since the Microsoft Excel Text Import Wizard has default settings set at tab delimited, for Step 1 click **Next** when prompted. **This also means that if you are submitting to WQX/STORET and you use a tab in any field, then that tab should be replaced with a ¶ (paragraph symbol) in your submission file.**

ii. For Step 2 Click the **Next** button and finally for Step 3 click the **Finish** button.

iii. For Step 4 the default is to place the data beginning in column A, row 1, if this is ok, click OK.

THIS HAS BEEN AN EXAMPLE TO FAMILIARIZE YOU WITH HOW TO CREATE AND DOWNLOAD A STORET REPORT AND PREPARE IT FOR ANALYSIS IN EXCEL. YOU ARE FREE TO SELECT THE SOFTWARE THAT BEST MEETS YOUR NEEDS TO QA THE ORG, PROJECT (BEACH), AND STATION IDS. Please review your IDs. All Beach Program data for your state should be present, correct, and associated with one BEACH_ORG_ID for your state; and with the national project "EPABEACH" and BEACH Program Station Types, so as to facilitate nation-wide data pulls.

All PROJECT_IDs (aka BEACH_IDs) should be present, correct, and associated with your BEACH_ORG_ID

All STATION_IDs should present, correct, unique in your state, and associated with the correct BEACH_ID(s) and with the ORG_IDs

Step 2. Sort the data

e.) In upper-left corner of the worksheet, click the grey box to select the whole worksheet

f.) From the main toolbar click ***Home>Format> Autofit Column Width***

g.) Browse the column names and rows of the data to familiarize yourself with the data

h.) From the main toolbar click ***Data>Sort***

i. Using the ***Sort by*** drop-list, select ***"Station ID",*** sort on "**Values**", and set the order as "**A to Z**".

ii. Click on "**Add Level**"

iii. Using the **Then by** drop-list, select "**Characteristic Name**", sort on "**Values**", and set the order as "**A to Z**"

iv. Click on "**Add Level**"

v. Using the ***Then by*** drop-list, select ***"Activity Start"*** , sort on "**Values**" and set the order as "**A to Z**"

vi. **Click OK**

ANALYZE THE DATA

i.) Locate **CNST1** in the station ID column and Characteristic Name **Enterococcus.** Select and highlight the numerical (no header) values under the **Result Value as Text** column (rows 2-34).

i.) From the main menu click Home, and from the icon toolbar click the ▼arrow to the right of the **Σ** button

i. From the drop-list select and click ***"Average".*** The average or mean value of the selected values will appear when u scroll to the bottom of the Result Vale as Text column. Repeat the above process to find the **"Max"** and **"Min"**

ii. Scroll to the bottom of the Result Value as Text column to find Average, Max, and Min calculations.

iii. Label **Average**, **Max**, and **Min** respectively. You can experiment with other analysis functions and even write your own equations

GRAPH THE DATA

j.) **For Station ID *CNST1* and Characteristic Name *Enterococcus* (rows 2-34), place the cursor in the select the *Result Value as Text* column (Column BX), and select the numerical values (not including the header). Then holding the Control key, select the Activity Start Dates (Column AE – not including the header) that correspond with Station ID *CNST1* and Characteristic Name *Enterococcus.***

k.) From the main toolbar select **Insert, within the charts menu, click on the icon for Line charts. Select** the line chart syle of your choice..

l.) **Double Click in the chart area.**

i. **Under *Chart Layouts*, choose the style that provides a chart title on top, the axis labels on the left side and bottom and a legend on the right side.**

ii. **Click on the chart title and change it to *"Enterococcus".***

iii. **Click on the left axis label and change it to "MPN"**

vi. **Click on the bottom axis label and change it to "Date"**

ANALYZE THE GRAPH

o.) Click and drag the graph to the bottom of the spreadsheet near the labeled **Average, Max**, and **Min** values

p.) Resize the Graph to be easier to read by clicking and dragging the small marks in any corner of the graph

q.) General questions regarding the graph

i. Is there a regular pattern to the data? Are there any breaks in the pattern?

ii. Between what values do most of the data points lie? Any outliers?

iii. What does this tell you about the Dissolved Oxygen at this Station?

r.) **DONE**.

- - - - - - - - - END OF BEACH EXAMPLE - - - - - - - - - -

Part 4: How to Import and Analyze the Data in Microsoft Access

Now you are going to import the downloaded data into Microsoft (MS) Access and perform some rudimentary analysis for every station in the dataset. MS Access can hold more records than MS Excel and is only limited to a file size of two gigabytes, more current versions of MS Access may have graphing features. STORET data can also be imported into statistical and GIS software packages, but is not covered in this exercise. (Note: This exercise was written using MS Access 2003. Some features may be different for other versions.)

Step 1. IMPORT THE DATA INTO ACCESS

a.) Open Microsoft Access and click on the icon for **Blank Database**

b.) On the right side of the screen, name the database "**storet_data.accdb**" and browse to the location that you want it on your hard drive by clicking on the folder icon. Click **OK**, then Click **Create**.

c.) Close the blank table that initially opens when creating a new database.

d.) From the main toolbar, click "**External Data**", then choose "**Text File**".

i. Browse to and double click on your downloaded STORET file (renamed **storet_data.txt** in Part 1 of exercise)

ii. Click **OK**

TIP: You can also import data from .xls or .xlsx spreadsheets if you choose "**Microsoft Excel**".

e.) From the Import Text Wizard

i. Click ***Delimited*** radio button then click ***Next*** button

ii. Click **Tab** radio button

TIP: STORET Data Warehouse Result downloads are all tab "□" delimited textfiles

iii. Check the **First Row Contains Field Names** checkbox then click the ***Next*** button

iv. Use the horizontal slide bar to peruse and click each column in turn, noting the **Data Type**: drop-list

v. Change **ALL** field values in the **Data Type**: box to **"Text**" **EXCEPT** any **Latitude** or **Longitude** fields, and the **Result Value as Number** field.

vi. Click the **Next** button then click the **No primary key** radio button then the **Next** button

TIP: This step is not strictly necessary but useful if you plan to add additional downloads to the same dataset (if a larger dataset was partitioned by date, for example)

vii. Click the ***Finish*** button then click the ***Close*** button

TIP: If you get any import errors, repeat the above steps and re-check step e.vi (Data Types)

ix. Click **storet_data** table and familiarize yourself with the data, this table contains your requested data

Step 2. Analyze The Data

f.) Under **Create**, click the **Query Design** button

g.) Choose the table that you imported by double clicking it and click the **Close** button.

h.) Double click on the following fields to put them in the query design view: "**Station Id**", "**Characteristic Name**", and "**Result Value as Number**". Put in the "**Result Value as Number**" field in 2 additional times. So that you have 5 fields in your query.

i. Under Criteria in the "Characteristic Name" field, type in "Enterococcus"

i.) Click the **Design** menu, select the **Summary** button

J.) Change the "**Group By**" in the "**Total**" line of the query under each "**Result Value as Number**" field to **Avg. Min.** and **Max,** respectively.

k.) Run your Query by clicking on the **Run** button with the red exclamation point on the **Design** menu.

l.) Adjust the column sizes by clicking and dragging the edges of the columns

m.) Familiarize yourself with the new information

i. Do most of the values fall within the same range across the stations?

ii. Are there any data gaps? Are there any outliers?

iii. What does this tell you about Dissolved Oxygen across the County?

n.) **DONE**.