

avenDownloading & Using Data from the STORET Warehouse: - - - - - An Exercise for the BEACH Program to QA IDs - - - - February 2008

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, as one software in which you could do a QA of the IDs, and how to QA your IDs. For questions, contact owsupport@rti.org, 800-844-0638 or 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 new Watershed Summary. 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

For this example, we will search for Michigan Beach monitoring data from all stations with 2007 as the specified monitoring timeframe. The downloaded file will be a compressed (zipped) .tar.gz file containing result data files, and metadata files; you will be primarily looking at the results file.

GO TO THE STORET DATA WAREHOUSE

- 1) Go to STORET main page: <http://www.epa.gov/storet/>
- 2) Under Features, click the **Download Data** link
- 3) Under Modernized STORET, click the **Browse or Download Modernized STORET Data** button
 - i. TIP: Legacy (pre-1998) data is also flat-filed by State/County for easy downloading

DEFINE YOUR QUERY

- 4) Under Query Options, under STORET Regular Results, click the **Regular Results by Project** link
 - i. TIP: More direct queries are possible given a pre-existing station list
- 5) Under **Project**, select the ORGANIZATION NAME or ORG ID – for this example 21MIBCH_WQX
 - i. TIP: Use Select a Project to select one or more beaches
- 6) Under **Station Type**, leave the defaults; this will capture all station types
 - i. TIP: use new BEACH-Program specific Station Types (at bottom of pick-list)
- 7) Under **Date**, for this example, change the start date to JAN 1, 2007
 - i. TIP: leaving the default will capture all date ranges of STORET Data Warehouse (non-Legacy datasets)
- 8) Under **Activity Medium**, leave as “Select All”

Because this example uses a Beach Organization (21XXBCH_WQX), all results will be relevant.

9) Under **Characteristic**, do nothing

Because this example uses a Beach Organization (21XXBCH_WQX), all results will be relevant.

10) Click the **Continue** button at the bottom of the screen

DOWNLOAD YOUR QUERY RESULTS

11) Note the number of records found

i. TIP: MS Excel holds 65K records per sheet, MS Access holds more but limited to 3 gigabytes [FOR THIS EXAMPLE, EXCEL IS FINE]

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

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

12) Type your email address in the **Please enter your email address** box

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

13) Type a three character prefix like “XYZ” in **Please specify three characters to prefix your report name** box

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

14) Scroll to the bottom of the screen to the **Select Data Elements for Report** checklist – for the purpose of this example, and for your state QA of the key IDs,

a. “Clear All”, and select: at least Org ID, Beach ID, Station ID, Station Name, Station Location info. Station lat/long. [this example will produce duplicate lines because multiple monitoring results were entered for each station]

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

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

i. 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 [FOR THIS EXAMPLE, IMMEDIATE WORKS FINE]

16) Go to your email account. You are waiting for a PROCESSING email and a COMPLETED email

a. When you receive the PROCESSING email, open and check that the URL matches the earlier URL (in case you are downloading one of multiple files)

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

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

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

18) DONE.

Part 2: Making Sense of Your Compressed Downloaded File

Now that you have your compressed downloaded file, what is it and what do you do with it? This section will first explain how to uncompress the data in your downloaded file, then 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.)

RETRIEVE YOUR RESULTS TEXTFILE FROM THE DOWNLOAD

- 1) Navigate to the directory where you saved the downloaded file from the STORET Data Warehouse
- 2) Create a folder in this directory and name it **storet_data**
- 3) Double-click the downloaded file to open it and click the **Yes** button when asked to decompress the file
 - i. TIP: Most compression engines like WINZIP® will be able to open the .tar.gz file
- 4) Extract all the files to your new folder named **storet_data**
 - i. TIP: Files with the Data_ prefix denote **Regular Results**, Biological, Habitat, or Metadata Results **[THIS EXERCISE FOCUSES ON REGULAR RESULTS]**
 - ii. TIP: Metadata Results contain information to help you determine the quality of the data
 - iii. TIP: Files with the RefDoc_ prefix denote Project-level Reference Documents associated with the organizations that own the data
- 5) Rename your **Data_XYZ_....._RegResults.txt** file to **storet_data.txt**; this file contains your requested data.
- 6) **DONE.**

----- YOU WILL OPEN THE FILES IN PART 3, - PAGE AFTER NEXT -----

QUESTIONS ABOUT THE FILES IN THE DOWNLOAD

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

- As noted in Part 1, the downloaded .tar.gz 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 documents can be pictures, datalogger results, QAPPS, or any project-level documents associated with the data owning organizations.

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

- All result (RegResult, BioResult, HabResult) files are in a tilde "~" delimited text format, which just means that we use the character "~" to separate the fields of the dataset. This is important when loading files into Excel, Access, or another database that can organize the file (make it readable) using the delimiter. How to do this is explained later.

Additionally, metadata files are in a tab "□" delimited format and can be easily opened in Microsoft Word or Excel.

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_'yearmdd'_'24hrmmss'_)
- 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
RefDoc_XYZ_20070322_205714_Project_COPSBDC1_cua0001.pdf
RefDoc_XYZ_20070322_205714_Project_SWMM_streamphoto.jpg

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 65,000 records per sheet. (Note: This exercise was written using MS Excel 2003. Some features may be different for other versions.)

IMPORT THE DATA INTO EXCEL

- 1) Open Microsoft Excel
- 2) From the main toolbar, click **Data>Import External Data>Import Data**
- 3) In **Select Data Source** window
 - a. Navigate to the directory or folder where you have saved your downloaded STORET file
 - b. In **Files of Type** textbox, select “**All Data Sources**” or “**Text File**” from the drop-list
 - c. Select and double click your saved STORET File (renamed **storet_data.txt** in Part 1 of exercise)
- 4) In **Text Import Wizard**
 - a. Click **Delimited** radio button then click **Next** button
 - b. Uncheck **Tab** checkbox then check **Other** checkbox and type the tilde character “~ “ in the **Other** field
 - i. TIP: STORET Data Warehouse result downloads are all tilde “~ “ delimited textfiles
 - c. Click **Next** button then click **Finish** button
 - d. Click **Existing Worksheet** radio button then click **OK** button

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 either the national project “EPABEACH”, or 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

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

SORT THE DATA

- 5) Select the whole worksheet by clicking the grey box in upper-left corner of the worksheet
 - a. From the main toolbar click **Format>Column>Autofit Selection**
 - b. Familiarize yourself with the data by browsing the column names and rows of the data
- 6) Select the whole worksheet by clicking the grey box in upper-left corner of the worksheet

7) From the main toolbar click **Data>Sort**

- a. Using the **Sort by** drop-list, select “**Station ID**” and leave the default **Ascending** radio button
- b. Using the **Then by** drop-list, select “**Activity Start**” (default **Ascending** radio button) and click **OK**

ANALYZE THE DATA

8) Select the numerical (no header) values under the **Result Value as Text** column for Station **156** (rows 2-160)

9) From the icon toolbar, click the ▼ arrow to the right of the Σ button

- a. From the drop-list select and click “**Average**”
- b. From the drop-list select and click “**Max**”
- c. From the drop-list select and click “**Min**”
- d. Scroll to the bottom of the **Result Value as Text** column to find **Average, Max, and Min** calculations
- e. Label **Average, Max, and Min** respectively
 - i. You can experiment with other analysis functions and even write your own equations

GRAPH THE DATA

10) Select the **Dissolved Oxygen** values under the **Result Value as Text** column for Station **156** (rows 2-160)

11) From the main toolbar choose **Insert > Chart** or click the **Chart Wizard** button.

- a. Under **Chart Type**, select **Line** and click **Next** [Places Dissolved Oxygen values along the Y-axis]

12) Select the **Series** tab and click button at the end of **Category (X) axis labels:** text field; note popup window

- a. From the worksheet, locate **Activity Start** column; select dates (no header) for Station **156**(rows 2-160)
- b. Click the button at the end of the **Chart Wizard** popup window [Places Dates along X-axis]

13) Click **Next** in the main window to go to **Step 3 of 4** of the Chart Wizard

- a. In the **Chart title:** field, type “**Dissolved Oxygen for Station 156**”
- b. In the **Category (X) axis:** field, type “**Dates**”
- c. In the **Value (Y) axis:** field, type “**mg/l**” and click **Next** button

- i. Determine the appropriate unit value by locating the **Units** column on your spreadsheet

14) Click **Finish** Button

ANALYZE THE GRAPH

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

16) Resize the Graph to be easier to read by clicking and dragging the small black box in any corner of the graph

17) General questions regarding the graph

- a. Is there a regular pattern to the data? Are there any breaks in the pattern?
- b. Between what values do most of the data points lie? Any outliers?
- c. What does this tell you about the Dissolved Oxygen at this Station?

18) **DONE.**

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.)

IMPORT THE DATA INTO ACCESS

- 1) Open Microsoft Access
- 2) From the main toolbar, click **File>New** then click the **Blank database** link on the right side of the window
- 3) In the **File New Database** window
 - a. Navigate to the directory where you want to save your database
 - b. In the **Filename:** textbox, rename the database to **storet_data.mdb**
- 4) From the main toolbar, click **File>Get External Data>Import**
 - a. Navigate to the directory or folder where you have saved your downloaded STORET file
 - b. In **Files of Type** box, select **“Text File”** from the drop-list
 - i. TIP: You can also import data from .xls spreadsheets if you choose **“Microsoft Excel”**
 - c. Select and double click your saved STORET File (renamed **storet_data.txt** in Part 1 of exercise)
 - d. From the **Security Warning** window, click the **Open** button
- 5) From the Import Text Wizard
 - a. Click **Delimited** radio button then click **Next** button
 - b. Click **Other** radio button and type the tilde character “~” in the **Other** textfield
 - i. TIP: STORET Data Warehouse Result downloads are all tilde “~” delimited textfiles
 - c. Check the **First Row Contains Field Names** checkbox then click the **Next** button
 - d. Click **In a New Table** radio button then click **Next** button
 - i. TIP: If adding a second file (ex. date partitioned), use **In an Existing Table:** radio button
 - e. Use the horizontal slide bar to peruse and click each column in turn, noting the **Data Type:** drop-list
 - f. Change **ALL** field values in the **Data Type:** box to **“Text” EXCEPT** any **Latitude** or **Longitude** fields
 - i. TIP: This avoids any difficulties MS Access has translating the Data Type
 - g. Click the **Next** button then click the **No primary key** radio button then the **Next** button
 - i. 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)
 - h. Click the **Finish** button then click the **OK** button

i. TIP: If you get any import errors, repeat the above steps and re-check step 5f (Data Types)

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

ANALYZE THE DATA

6) Under **Objects**, click **Queries**

7) Click **Create query by using wizard**

8) In the **Simple Query Wizard**

a. Under **Available Fields**: select **“Station Id”** and click the > button

b. Under **Available Fields**: select **“Result Value as Text”** and click the > button

c. Click the **Next** button

9) Click the **Summary** radio button then click the **Summary Options** button

10) Check the **Avg, Min, and Max** check boxes then click the **OK** button

11) Click the **Next** button and then the **Finish** button

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

13) Close the table (red **X**) and click the **Yes** button

14) Reopen the Query and familiarize yourself with the new information

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

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

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

15) **DONE.**

Downloading a Watershed Summary from the STORET Warehouse: An Exercise

This exercise walks you through the steps to generate and download a typical query for general categories of characteristics found in a given Watershed from the STORET Data Warehouse. Watershed Summary downloads can be treated identically to those from the STORET Warehouse Query page (import/analysis). Please contact STORET Technical Support at 1-800-424-9067 toll free or email storet@epa.gov for comments or questions.

GO TO THE STORET DATA WAREHOUSE AND WATERSHED SUMMARY

- 1) Go to STORET main page: <http://www.epa.gov/storet/>
- 2) Under Features, click **Download Data** link
- 3) Under Modernized STORET, click **Browse or Download Modernized STORET Data** button
- 4) Under **STORET Warehouse Watershed Summary**, click **Summary Results by Watershed** link

DETERMINE YOUR WATERSHED BY COUNTY [Skip if you know your Hydrologic Unit Code]

- 5) On the **STORET Warehouse Watershed Summary** page, read the description page
 - a. From the main toolbar click **Tools>Pop-up Blocker>Turn Off Pop-up Blocker** [unless already off]
 - i. TIP: This checks that the Pop-up blocker is off and allows the results Pop-up window to open
 - b. Click the **Surf your Watershed** link then click the **Locate Your Watershed** icon
- 6) Scroll to the bottom of the page and find the **Locate by Geographic Unit** box
 - a. Under **Locate by Geographic Unit** search box, select “**County Name**” from the drop-list
 - b. Under **Your Entry:**, type “Denver” in the textbox and click the **Submit** button
- 7) Under **List of Counties Produced by Search**, click **Denver, Colorado** link
 - a. Note the three 8-digit Hydrologic Unit Code (HUC) numbers that cross Denver County on the map
 - i. TIP: Write the three 8-digit HUCs down to be used to **DEFINE YOUR QUERY** later
 - b. Return to the STORET Warehouse Watershed Summary page by clicking the **Back** button four times

DEFINE YOUR QUERY

- 8) On the **STORET Warehouse Watershed Summary** page, click **Go to Watershed Summary** link
- 9) Under **Drainage Basin/HUC**, select “**10190002 – Upper South Platte**” from the drop-list
 - a. This is the information you wrote down from **step 7** above in **DETERMINE YOUR WATERSHED**
- 10) Under **Search By**, click the **Characteristic Type** radio button then click the **Query** button
- 11) Click the **Physical** link

- a. Browse the characteristics to see which characteristics (like Dissolved Oxygen) fall into this category
 - b. Browse the records list to see how many records of each characteristic exist in this particular HUC
 - c. Click **Close** button and repeat **step 11** above for any other category links that you are curious about
- 12) Click the **Select All** button; this gives you all data in this particular HUC [not just data in the county]
- i. TIP: You can narrow your query by checking only categories you care about
- 13) Click the **Get Results** button

DOWNLOAD YOUR QUERY RESULTS

- 14) Note the number of records found
- i. TIP: MS Excel holds 65K records per sheet, MS Access holds more but limited to 3 gigabytes
 - ii. TIP: If there are too many records, you may have to go back to narrow your query
 - iii. TIP: You can adjust your query by narrowing the Date fields or limiting characteristics
- 15) Type your email address in the **Please enter your email address** box
- i. TIP: Emails are used to notify you that your download is first processing and then completed
- 16) Type a three character prefix like “XYZ” in **Please specify three characters to prefix your report name** box
- i. TIP: This prefix will help to identify your download file later
- 17) Scroll to the bottom of the screen to the **Select Data Elements for Report** checklist
- a. Note the different Elements selected and leave the defaults checked
 - i. TIP: These are the fields you will see in your report, you can select/de-select any you choose
- 18) Under **Batch Processing**, click the **Immediate** button and note the URL and see if it has your 3 character prefix
- i. 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
- 19) Go to your email account. You are waiting for two separate emails, a PROCESSING and COMPLETED email
- a. When you receive the PROCESSING email, open and check that the URL matches the earlier URL
 - i. TIP: If you click the URL now you will go to an error page because the file is not ready yet
 - b. When you receive the COMPLETED email, your file is ready to download; click on the URL

- 20) Note the filename and click the **Save** button, then save the download file to your desktop or other directory
- 21) Re-**DEFINE YOUR QUERY** for the remaining HUCs (from **step 7** above) for all the data crossing your county
- 22) **DONE.** [Use what you learned in Parts 2, 3, & 4 above to import and analyze your data.]

C:\00KRAMER\ABEACHES\STORET\Downloadingstationlat-long.doc