Downloading & using data from the STORET Warehouse: an exercise

July 2010

This exercise addresses querying or searching for specific water resource data, and the respective methods used in collecting and analyzing data for a given state and county. The following exercises will guide you through steps to generate, download, access and organize a typical query for the STORET Data Warehouse. **Please contact STORET Technical Support at 1-800-424-9067 toll free or storet@epa.gov for comments or questions**

Part 1: **How to query and download the data -** step by step guide to generating and downloading a typical query for the STORET Data Warehouse

Part 2: **Making sense of your downloaded file –** description of the format of the downloaded file, and how to extract data in your file

Part 3: **How to import and analyze the data in Microsoft Excel**

Part 4: **How to import and analyze the data in Microsoft Access**

Part 5. **Downloading a watershed summary from the STORET warehouse: an exercise**

.

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. 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. Below is an example detailing a query for dissolved oxygen (DO) values from all monitoring sites in Denver County, Colorado, with no specified timeframe.

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 specific stations, projects or geographic locations

d.) Under ***STORET Central Warehouse- STORET Results***, click the link titled ***Results by Geographic Location***

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

e.) ***Geographic Location*** can be further subdivided by State/ County, latitude longitude coordinates or drainage basin/ HUC. Under ***Geographic Location*****,** select the ***State/County*** radio button

i. Under the ***State Name*** drop-list, select **"Colorado"**

ii. Under ***County Name*****,** click on the ***Look Up*** button and select ***"Denver"*** from the drop list

TIP: Leaving the State/County defaults gives you a national query (but is a very large dataset)

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

g.) Under **Date**, leave defaults; this will capture all date ranges of STORET Data Warehouse (non-Legacy datasets)

TIP: Use ***Date*** to refine a large query result into multiple smaller downloads

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

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

j.) Under ***Characteristic***, place the cursor in the ***Characteristic Search*** box

i. Type **"dissolved oxygen"** in the ***Characteristic Search*** box and click the ***Search*** button

TIP: The ***Characteristic Search*** box utilizes a *Beginning with* type search. For example, in regards to dissolved oxygen, typing in leading characters 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)

ii. Select **"dissolved oxygen (DO)"** from the ***Characteristic Name*** list and click the ***Select*** button

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

Step 3. Download your query results

l.) Note the number of records found

TIP: MS Excel holds 65K records per sheet, MS Access holds more but 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

m.) Select the report types (***REGULAR, BIOLOGICAL, and HABITAT***)

n.) 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

o.) 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

p.) 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

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

q.) 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

r.) 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

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

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

t.) 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 .tar.gz 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
TIP: Metadata Results contain information to help you determine the quality of the data
TIP: 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 .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 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
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.)

Step 1. Import the data into excel

a.) Open Microsoft Excel

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

c.) 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)

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

- a. Click ***Delimited*** radio button then click ***Next*** button
- b. Tab should be the only box checked in the delimiter field

TIP: STORET Data Warehouse result downloads are all tab "□" delimited text files. (Tabs are the symbols used to organize files).

- c. Click ***Next*** button then click ***Finish*** button
- d. Click ***Existing Worksheet*** radio button then click ***OK*** button

Step 2. Sort the data

e.) Select the whole worksheet by clicking the tan 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

f.) Select the whole worksheet by clicking the grey box in upper-left corner of the worksheet

g.) 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

h.) Highlight the numerical (no header) values under the ***Result Value as Text*** column for Station ***156*** (rows 2-172)

i.) From the icon toolbar, click the ▼ arrow to the right of the **Σ** button

- a. 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"**
- b. Label **Average**, **Max**, and **Min** respectively
  - i. You can experiment with other analysis functions and even write your own equations

GRAPH THE DATA

j.) Select the ***Dissolved Oxygen*** values under the ***Result Value as Text*** column for Station ***156*** (rows 2-172)

k.) 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]

l.) 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]

m.) 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

n.) Click *Finish* Button

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 black box 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**.

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 filesize 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

b.) From the main toolbar, click ***File>New*** then click the ***Blank database*** link on the right side of the window

c.) In the ***File New Database*** window

i. Navigate to the directory where you want to save your database

ii. In the **Filename:** textbox, rename the database to **storet_data.mdb**

iii. Click ***OK***

d.) From the main toolbar, click File>Get External Data>Import

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

ii. In ***Files of Type*** box, select **"Text File"** from the drop-list

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

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

iv. From the ***Security Warning*** window, click the ***Open*** button

e.) From the Import Text Wizard

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

ii. Tab should be the only box checked in the delimiter field

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

iii. Select ***Next***

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

iv. Click ***In a New Table*** radio button then click ***Next*** button

TIP: If adding a second file (ex. date partitioned), use ***in an Existing Table:*** radio button

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

vi. Change ***ALL*** field values in the **Data Type:** box to **"Text"** ***EXCEPT*** any ***Latitude*** or ***Longitude*** fields

TIP: This avoids any difficulties MS Access has translating the Data Type

vii. 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)

viii. Click the ***Finish*** button then click the ***OK*** button

TIP: If you get any import errors, repeat the above steps and re-check step 5f (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 ***Objects***, click ***Queries***

g.) Click ***Create query by using wizard***

h.) 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

i.) Click the ***Summary*** radio button then click the ***Summary Options*** button

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

k.) Click the ***Next*** button and then the ***Finish*** button

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

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

n.) Reopen the Query and 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?

o.) **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.**

Step 1. Access the STORET Data Warehouse and watershed summary

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

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

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

d.) If the HUC of your watershed is available, skip to step 3.

e.) The HUC can be determined with the zip code, or the geographic unit of the watershed. If the respective zip code of the watershed is available, in the box titled ***Find Your Watershed Enter your ZIP***, enter the zip code and select GO

f.) To determine the HUC of your watershed by geographic unit, under the box titled ***Features***, click the ***Watershed summary*** link

Step 2. Determine your watershed by geographic unit (i.e. County)

g.) On the **STORET Warehouse Watershed Summary** page, read the description page

i. From the main toolbar click **Tools>Pop-up Blocker>Turn off Pop-up Blocker** [unless already off]

TIP: This checks that the Pop-up blocker is off and allows the results Pop-up window to open

ii. Click the ***Surf your Watershed*** link, or if you know your HUC, scroll to the bottom of the page and select the link titled **"Watershed Summary".**

h.) Scroll down the **Surf Your Watershed** page to locate your watershed by geographic unit or state

i. Find the ***Pick your geographic unit*** box

ii. Under ***Pick your geographic unit*** search box, select ***"County Name"*** from the drop-list

iii. Under ***Enter your geographic information***: type "Denver" in the textbox and click the ***Submit*** button

i.) Under ***List of Counties produced by Search***, select the ***Denver, Colorado*** link

i. Note the three 8-digit Hydrologic Unit Code (HUC) numbers that cross Denver County on the map

TIP: Write the three 8-digit HUCs down to be used to define your query later

ii. Return to the STORET Warehouse Watershed Summary page by clicking the **Back** button four times

Step 3. Define your query

j.) On the ***STORET Warehouse Watershed Summary*** page, click ***Go to Watershed Summary*** link

k.) Under ***Drainage Basin/HUC***, select **"*10190002 – Upper South Platte*"** from the drop-list

a. This is the information you wrote down from **step 2** above in ***Determine your watershed by geographic unit***

l.) Under ***Search By***, click the ***Characteristic Type*** radio button then click the ***Query*** button

m.) 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 3m** above for any other category links that you are curious about

n.) Click the ***Select All*** button; this gives you all data in this particular HUC [not just data in the county]

TIP: You can narrow your query by checking only categories you care about

o.) Click the ***Get Results*** button

DOWNLOAD YOUR QUERY RESULTS

p.) Note the number of records found

TIP: MS Excel holds 65K records per sheet, MS Access holds more but 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

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

i. Note the different Elements selected and leave the defaults checked

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 email account. You are waiting for two separate emails, a PROCESSING and COMPLETED email

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 will 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, then save the download file to your desktop or other directory

w.) Re-define your query for the remaining HUCs (from **step 2h** above) for all the data crossing your county

x.) **DONE. [Use what you learned in Parts 2, 3, & 4 above to import and analyze your data**