NOAA Technical Memorandum NOS CS 44

INVENTORYING USGS OCEANOGRAPHIC GEOSPATIAL DATASETS FOR INCLUSION AT NOAA'S NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION

Silver Spring, Maryland
December 2020

noaa National Oceanic and Atmospheric Administration

U.S. DEPARTMENT OF COMMERCE
National Ocean Service
Coast Survey Development Laboratory

Office of Coast Survey
National Ocean Service
National Oceanic and Atmospheric Administration
U.S. Department of Commerce

The Office of Coast Survey (OCS) is the Nation's only official chartmaker. As the oldest United States scientific organization, dating from 1807, this office has a long history. Today it promotes safe navigation by managing the National Oceanic and Atmospheric Administration's (NOAA) nautical chart and oceanographic data collection and information programs.

There are four components of OCS:

> The Coast Survey Development Laboratory develops new and efficient techniques to accomplish Coast Survey missions and to produce new and improved products and services for the maritime community and other coastal users.
>
> The Marine Chart Division acquires marine navigational data to construct and maintain nautical charts, Coast Pilots, and related marine products for the United States.
>
> The Hydrographic Surveys Division directs programs for ship and shore-based hydrographic survey units and conducts general hydrographic survey operations.
>
> The Navigational Services Division is the focal point for Coast Survey customer service activities, concentrating predominately on charting issues, fast-response hydrographic surveys, and Coast Pilot updates.

NOAA Technical Memorandum NOS CS 44

INVENTORYING USGS OCEANOGRAPHIC GEOSPATIAL DATASETS FOR INCLUSION AT NOAA'S NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION

Emma Bonanno, Diego Burgos, Robert Verstraete, Meredith Westington
Earth Resources Technology and NOAA's Integrated Ocean and Coastal Mapping Program
Silver Spring, Maryland

December 2020

noaa National Oceanic and Atmospheric Administration

U. S. DEPARTMENT OF COMMERCE
Wilbur L. Louis Ross, Jr.
Secretary

National Oceanic and Atmospheric Administration
Dr. Neil J. Jacobs
Acting Under Secretary

National Ocean Service
Nicole LeBoeuf
Acting Assistant Administrator

Office of Coast Survey
Rear Admiral Shepard M. Smith
Director

Integrated Ocean and Coastal Mapping Program
Ashley Chappell
Program Coordinator

NOTICE

Mention of a commercial company or product does not constitute an endorsement by NOAA. Use for publicity or advertising purposes of information from this publication concerning proprietary products or the tests of such products is not authorized.

Table of Contents

List of Figures

List of Tables

EXECUTIVE SUMMARY

The Integrated Ocean and Coastal Mapping (IOCM) Program at NOAA seeks to “map once, use many times.” In order to fulfill this goal across many mapping sectors, IOCM advocates that everyone share their mapping data with officially recognized, central repositories such as NOAA’s National Centers for Environmental Information (NCEI), the national archive for geophysical ocean and coastal data. Bathymetry is one type of data that has broad application and is logistically difficult to collect, so every survey mile collected is valuable and potentially useful for multiple purposes.

From July to August 2020, NOAA’s IOCM program sponsored a geophysical data mining effort of U.S. Geological Survey (USGS) websites. The USGS Coastal and Marine Hazards and Resources Program (CMHRP) has high quality ocean mapping data of different types that are not fully mirrored at NCEI, including bathymetry. Archiving copies of the USGS CMHRP’s raw and processed bathymetric data at NCEI would increase public accessibility and allow the data to be incorporated into products like the U.S. Bathymetry Gap Analysis, which tracks how much of U.S. waters remain unmapped, and other data assemblies, e.g., Global Multi-Resolution Topography and the General Bathymetric Chart of the Oceans. This memorandum documents the methods used to mine data from USGS CMHRP websites and provides recommendations to others who may seek to extract future bathymetry data from these websites.

1. INTRODUCTION

The Integrated Ocean and Coastal Mapping (IOCM) program advocates the sharing of mapping data with officially recognized, central repositories, in particular NOAA's National Centers for Environmental Information (NCEI). Bathymetry is one type of data that has broad application and is logistically difficult to collect, so every survey mile collected is potentially valuable for a multitude of uses. Archiving copies of U.S. Geological Survey (USGS) raw and processed bathymetric data at NCEI would increase public accessibility and allow the data to be incorporated into products like the U.S. Bathymetry Gap Analysis,[1] which tracks how much of U.S. waters remain unmapped, and other data assemblies, e.g., Global Multi-Resolution Topography and the General Bathymetric Chart of the Oceans. This project supported IOCM's goals to "map once, use many times" in an effort fill U.S. bathymetry data gaps as shown in Figure 1.

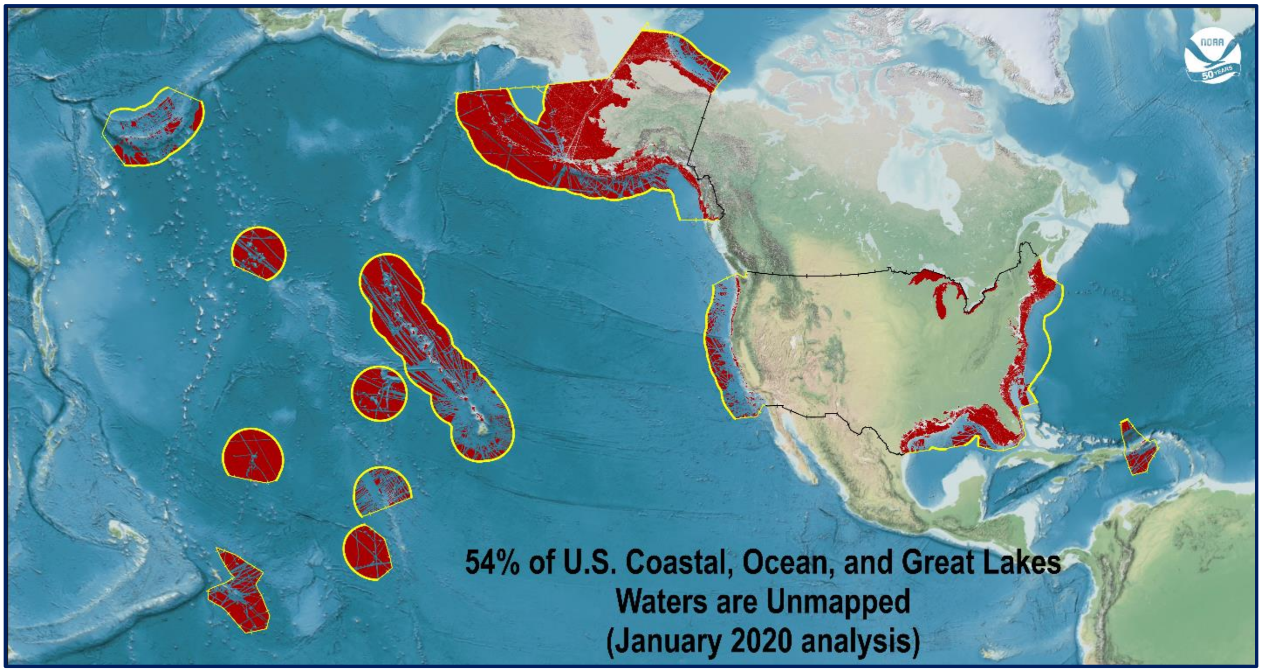

Figure 1. Distribution of U.S. waters that are considered unmapped per the January 2020 version of the U.S. Bathymetry Gap Analysis, shown in red. The analysis is based on openly accessible bathymetry at NCEI.

From July to August 2020, we conducted a data-mining project to identify USGS geophysical and oceanographic survey datasets that NCEI did not already possess in its repositories. We were primarily concerned with multibeam bathymetry, but we also sought single beam and interferometric bathymetry, sidescan and water column data, and even bottom samples and photographs. These datasets were the result of USGS projects and scientific research conducted primarily on coastal change and hazards in the Atlantic, Pacific, and Gulfs of Mexico and Alaska.

Gathering an inventory of useful bathymetric and other geospatial data from USGS data websites was a large task. USGS Coastal and Marine Hazards and Resources Program (CMHRP) is a publications-based organization that publishes its data in conjunction with a report on the research and conclusions drawn from the data. This means that surveys are often grouped together by the publication or associated dataset. The data are processed into products, such as digital elevation models, for the purposes of each publication. Additionally, the metadata often describe the publication as a whole rather than each survey themselves, which can make it difficult to determine if the desired raw data are useful and/or available.

NCEI is a data repository and seeks to archive raw survey level data to maintain for posterity and make available for further scientific study and use. USGS rarely releases raw survey data, which poses a problem when it comes to sending the data to NCEI. As we began our work, there was no systematic ability to transfer raw USGS data for ingest by NCEI, hence the importance of this project—finding and recording important data and figuring out ways to deliver it to NCEI repositories.

The methods section of this paper explores the two approaches we took to locate USGS' published datasets: manual keyword searching on USGS websites and a Python web-scraping tool to automate the process. While the approaches provide roughly the same search results, each can be particularly useful depending on the needs of the researcher. Additionally, we provide best practices for interpreting survey and publication metadata, organizing survey inventories, and utilizing keyword terminology that could be useful for future searches for bathymetric/geospatial data.

The results of this project yield a wealth of USGS bathymetric data that are not at NCEI and recommendations for increased data sharing between USGS and NOAA.

2. DEFINITIONS

The principle bathymetric data types sought in this project were multibeam, single beam, and interferometric echosounders:

Single beam sonar bathymetry uses a single echosounder to produce very accurate depth measurements. This equipment is most commonly used in shallow waters and smaller areas such as beaches or rivers.

Multibeam sonar bathymetry uses multiple echosounder beams simultaneously to provide extremely detailed and accurate measures of depth over a large area. Due to the relatively fixed angle of the swath, multibeam equipment is less efficient in shallow water, as it does not cover as large an area as in deep water.[2]

Interferometric sonar involves the use of phase-discriminating sensors, which measure the sum of returns on each side. They provide excellent imaging, but their depth detection is not as accurate as multibeam bathymetry, making them ideal for obstacle detection.[3]

3. METHODS

With the primary task to locate raw bathymetric data in USGS portals, the inventorying and mining of surveys involved multiple steps and approaches. This section explains each step taken during the data-mining phase and describes the two approaches we used to explore USGS data websites.

Each intern, working remotely due to COVID-19 restrictions, selected a USGS CMHRP data center from which to find data, i.e., Pacific Coastal and Marine Science Center (PCMSC), Woods Hole Coastal and Marine Science Center (WHCMSC), or St. Petersburg Coastal and Marine Science Center (SPCMSC). Figure 2 shows the geographic location of each USGS CMHRP science center.

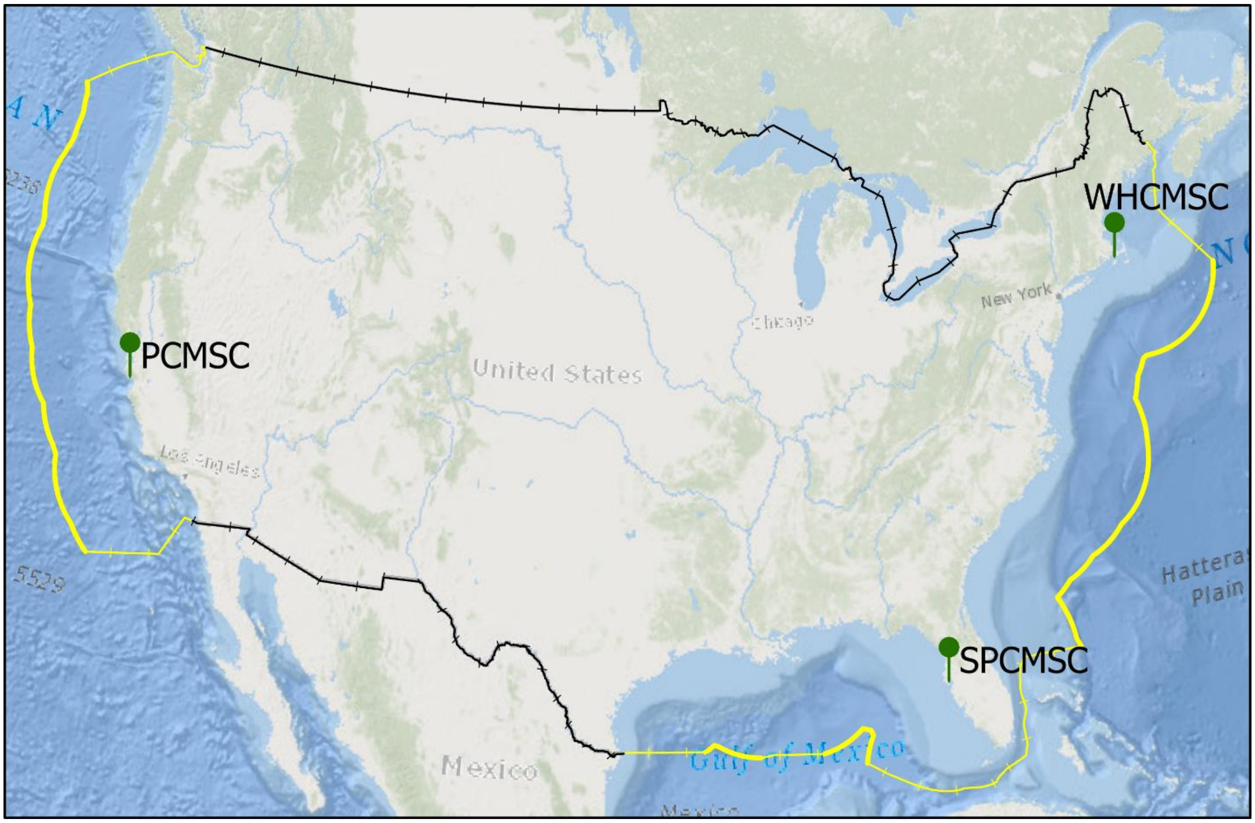

Figure 2. Geographic locations of each USGS coastal and marine science center. The extents of U.S. waters and land for the contiguous U.S. are depicted as yellow and black lines, respectively.

3.1. Step 1: Approaches to finding datasets

We employed two main approaches to inventorying published data from the USGS websites: manual searching and a web-scraping script. The two approaches complement each other and are particularly useful in unique ways. We also found that keywords and equipment types were very important when finding the desired data and have included our list of keywords and a few notes on equipment types. To focus our search, we sought data collected between 2000 and 2020.

3.1.1. Manual searching

Description: The manual search method essentially involves using keywords and other website filters to find public datasets on USGS sites. We started with the USGS Coastal & Marine Geology Data Catalog website,[4] moved on to using the USGS Publications Warehouse[5] and finally the ScienceBase Catalog.[6] After these steps, we discovered another USGS website with a list of metadata records organized by CMHRP science center,[7] where we were able to find metadata for every published dataset from each USGS science center. After going through the complete metadata list, we felt confident that we had found every published dataset for each USGS CMHRP data center. This method was applied to PCMSC and SPCMSC.

Pros: The user can be confident that all published datasets are found. The user can filter search results by keywords, dates, and geographic area. The metadata are easily accessible.

Cons: This approach can be time consuming and labor intensive as each search result has to be read through carefully. The user has to read lots of metadata. In addition, there are often multiple metadata files associated with the same survey because the collected raw data may have been processed into different products for each publication.

3.1.2. Web scraping

Description: The main site used for the Python web-scraping script was the USGS Coastal & Marine Geology Data Catalog website.[4] Running on a Jupyter Notebook, the plugins used were Pandas, Selenium Webdriver, and BeautifulSoup. The script took in a URL link, went through the metadata of each entry in the Data Catalog, and generated a row of metadata information for each entry. BeautifulSoup was used to extract certain entries of data such as the title, date, and relevant links through the metadata's HTML tags. The script then exported this data frame as a spreadsheet in Excel format. This method was applied to the WHCMSC.

Pros: The script was able to successfully take each entry in the USGS Coastal & Marine Geology Data Catalog and reformat to a spreadsheet. The most important features such as date of the survey, the metadata title, and relevant links made it easier to organize. Python can be used to quickly filter out duplicates or irrelevant data entries out of the data frame. The script was extremely useful in getting the titles, publication link, download links, and the metadata link for each entry. Key identifiers associated with many of these titles, such as a NOAA survey ID or USGS Field Activity Number (FAN), were used to categorize each entry. The USGS Coastal & Marine Geology Data Catalog keywords were useful to specify the search results for the spreadsheet. Having all the metadata titles made it very easy to compare against the other USGS website with a list of metadata records organized by science center. [7]

Cons: It was not possible to automate the laborious manual searching method using this code. Some values that the web scraper extracted were inconsistent. One of the most inconsistent values was the Issue Number that was contained in each metadata entry. These entries would vary with the USGS Issue Number, the DOI, the FAN, or another unrelated number. One would have to use the Field Activity search to pinpoint the FAN associated with each metadata entry. As a result, even if automating by script saved some time, one would still need to use manual methods in order to create a comprehensive inventory catalog.

See **Appendix A** to view the web-scraping script and instructions for using it.

3.1.3. Keywords and mixed use

We used the following keywords for our search: bathymetry, multibeam, single beam, sidescan, water column, bottom sample, and interferometric. It was particularly tricky to get proper results from the term “multibeam” because NCEI uses the term in relation to multibeam echosounders while USGS groups any swath bathymetry equipment in this category.

To decipher the appropriate bathymetry type, we were instructed to pay attention to the type of equipment referenced in the metadata. SWATHplus products often created some confusion. NCEI generally categorizes this equipment as interferometric, but the system also has sidescan capabilities. This dual use often led to these data being categorized as “Sonar: Sidescan" or even “Sonar: Multibeam" in USGS FAN Activity Details. For a general rule of thumb, Reson SeaBat 8111, 7111, and 7160 multibeam echosounders are some the most common multibeam bathymetry equipment found in PCMSC surveys.

3.2. Step 2: Interpreting metadata

USGS metadata typically followed the same layout for each dataset; however, our web scraper tool found that the metadata often lacked the desired information or had inconsistent XML tags. Due to these issues, manually reading the information from each metadata record was the most reliable method for developing the inventory.

The Abstract and the Processing Steps are particularly useful sections in USGS metadata to gain an understanding of the data. The metadata abstract is similar to the abstract of any other publication, briefly explaining the goals and results of the publication or project depending upon the science center. The processing steps detail the technical steps taken to process the survey data, often beginning with the methods of collection for the data and mentioning the equipment and vessel used to perform the survey. If the survey results are published as a raster DEM, this section often includes the raw data processing steps and GIS processing steps, e.g., geographic coordinate systems, projections, interpolations.

To find out whether NCEI already possessed the data, we checked the following NCEI data viewers/indexes: Bathymetric Data Viewer,[8] Trackline Geophysical Data Viewer,[9] Water Column Sonar Data,[10] and Index to Marine and Lacustrine Geological Samples.[11]

Searching by survey start and end dates and vessel name and comparing the resultant track lines was the most accurate way to determine if NCEI had the datasets.

3.3. Step 3: Inventory the data

Each intern created a spreadsheet inventory of the surveys associated with their data centers and recorded the key information shown in Table 1:

Table 1. Primary metadata elements captured for USGS data inventory spreadsheet.

Element Name	**Element Description**
Field Activity Number (FAN)	Unique survey number that USGS uses to recognize each survey. Two general formats: YYYY-###-FA (year and survey number, current format) and L-##-YY-LL (first letter of vessel/platform, survey number, year, and two letter abbreviation of the region surveyed; older format, transitioned in mid 2010's). A third type also gets used: YYLLL## (year, project acronym, and an incrementally increasing number representing the survey number; used to split activities that fall under the current format).
Digital Object Identifier (DOI)	Each publication is assigned a unique DOI in the format of 10.3133/#### (indicated a series publication) or 10.5066/#### (indicates a data release).
Survey year	Year in which the survey was performed. Format YYYY.
Ship name	Title of the ship or other watercraft used to conduct the survey. For nearshore waters, the vessel may be a jet-ski, kayak, or on foot.
Primary data type	Main data type described in the metadata. Multibeam, single beam, interferometric, sidescan, water column, sediment samples, photos, videos.
Other data type	Includes other collected data types as listed in the Field Activity Details page.
Metadata links	URL form. Useful when referencing inventoried surveys in the future. May be for the entire publication and not be survey specific.
Comments	Stores any notes about the survey that may be important.

As discussed in Section 3.1.3, equipment type was important to decipher dual use of keywords, but it was not necessary to include in the inventory columns. The equipment type was the most reliable means of confirming the primary data type of the survey and was often found in the Processing Steps of the metadata. However, one notable piece of equipment, SWATHplus, created confusion because it is capable of collecting both interferometric and sidescan data. The metadata specified the primary data type collected, while the USGS Field Activity Details page for each survey listed all data types collected for that particular survey.

3.4. Step 4: Download the data

Once the datasets were inventoried, we sought to download them to see what information was stored in the datasets. Our goal was to identify and download raw and processed datasets to be shared with NCEI repositories.

A download script, written in PyCharm using Python 2.7, was used to loop through the inventory and download each multibeam dataset we found. The script looped through an Excel spreadsheet containing the FAN numbers (without dashes) and the download links for each dataset. Each dataset was downloaded and unzipped to a folder labeled with the FAN.

See **Appendix B** for the download script and instructions for using it.

After the datasets were downloaded, we could inspect more closely to see if the data were raw, processed, or product files. Generally, if any raw multibeam datasets were found, the download size surpassed 50 gigabytes. Depending on the data type, the file formats and sizes for each dataset varied. It was important to read the metadata in order to know what each file type represented. Table 2 shows the file types associated with the one multibeam survey we identified through this project with downloadable access to raw data. For this record and others, there could be additional files available for download, such as maps or other images in PNG or JPEG format.

Table 2. File types for raw multibeam echosounder survey 2016-625-FA using 44kHz Reson SeaBat 7160.

File extension	File type
.7K	raw bathymetry data
.HSX	raw bathymetry data with ship motion included
.RAW	raw navigation data

3.5. Step 5: Submit raw data to NCEI

When submitting data to NCEI, there are standards for organizing the data into a specific directory structure. The multibeam bathymetry database at NCEI primarily maintains raw data files in the instrument's vendor-specific format. (e.g., .all, .s7k, .xse). However, other supplemental data (sound speed profiles, tides, vessel offsets, cruise reports, etc.) and/or processed versions or products of the multibeam data are also accepted. In all submissions, the data files and cruise/survey should be well-documented using metadata. For more information, we recommend visiting NCEI's Data Submission Guidelines page or contacting a data officer, as each submission will vary.[12]

For the one raw multibeam dataset that we downloaded, we discovered that the files were organized by Julian day and not organized per NCEI’s standards. To address this issue, the files from each day were shifted as follows: the HSX data went into the multibeam category; the 7K data were compressed and moved into the multibeam category; and the RAW data were compressed and moved into the ancillary files category.

3.6. Step 6: Visualize the data (optional)

We did not have the capability to view or process raw data; therefore, the GIS product downloads available on USGS websites were useful to visualize the nature and extent of the search results.

Due to the processed nature of the publicly available USGS datasets, loading them into ArcGIS was simple. Multibeam, interferometric, and sidescan sets were compiled into Digital Elevation Model rasters. Single beam and other data types were downloaded as point shapefiles. Some datasets were already projected, while some needed to have a geographic coordinate system and vertical datum applied in order to view on a map. See Figure 3 and Figure 4 for visual examples of multibeam and single beam datasets, respectively, downloaded from USGS.

Again, these datasets are not as useful for NCEI, which seeks primarily raw and processed data rather than products, like digital elevation models, but the GIS products can be ingested as ancillary files with the raw survey data.

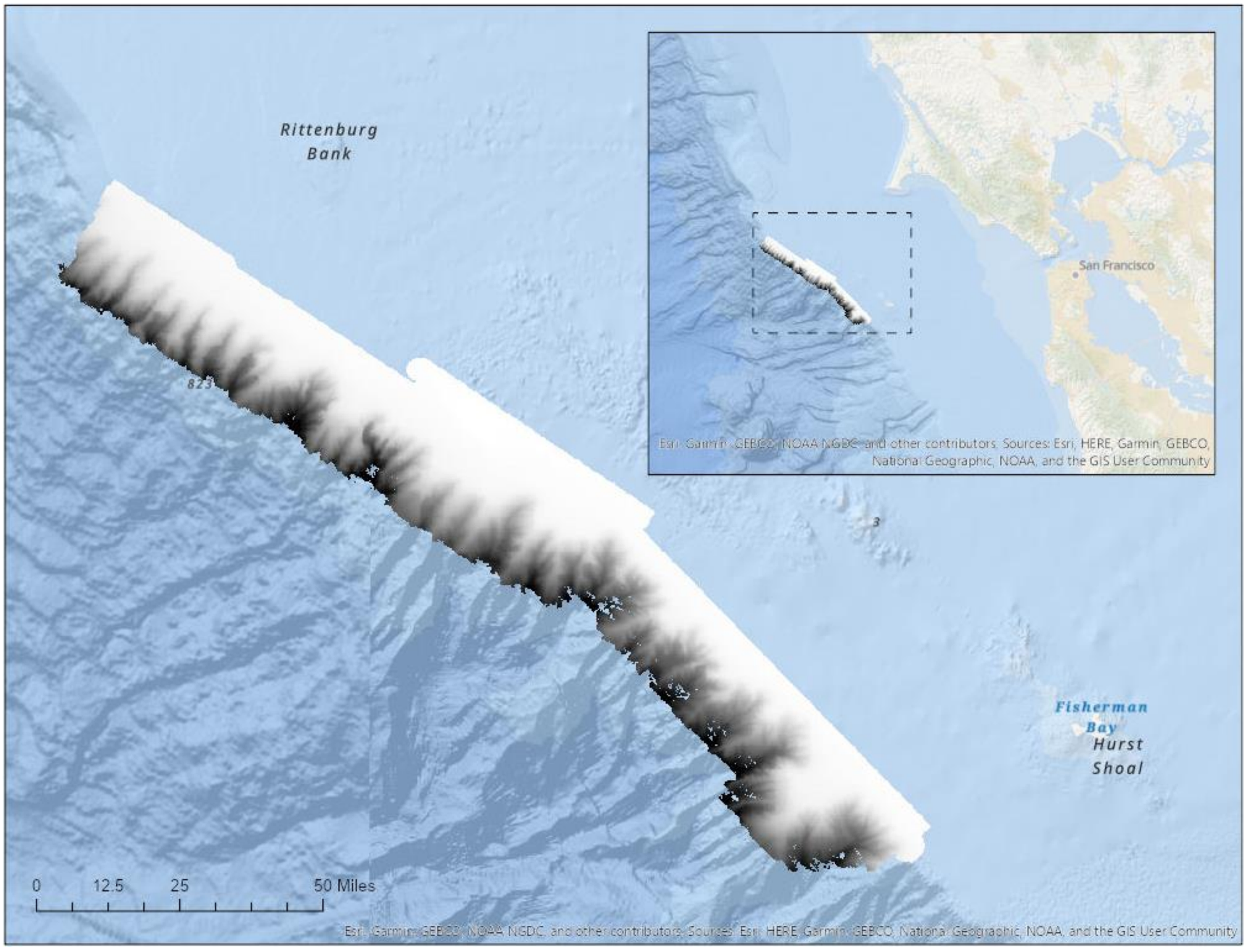
Rittenburg Bank
823
3
Fisherman Bay
Hurst Shoal
San Francisco
0
12.5
25
50 Miles
Esri, Garmin, GEBCO, NOAA NGDC, and other contributors, Sources: Esri, HERE, Garmin, GEBCO, National Geographic, NOAA, and the GIS User Community
Esri, Garmin, GEBCO, NOAA NGDC, and other contributors, Sources: Esri, HERE, Garmin, GEBCO, National Geographic, NOAA, and the GIS User Community

Figure 3. Example of a multibeam dataset, specifically a digital elevation model (DEM) downloaded directly from USGS. Survey F-01-11-NC.

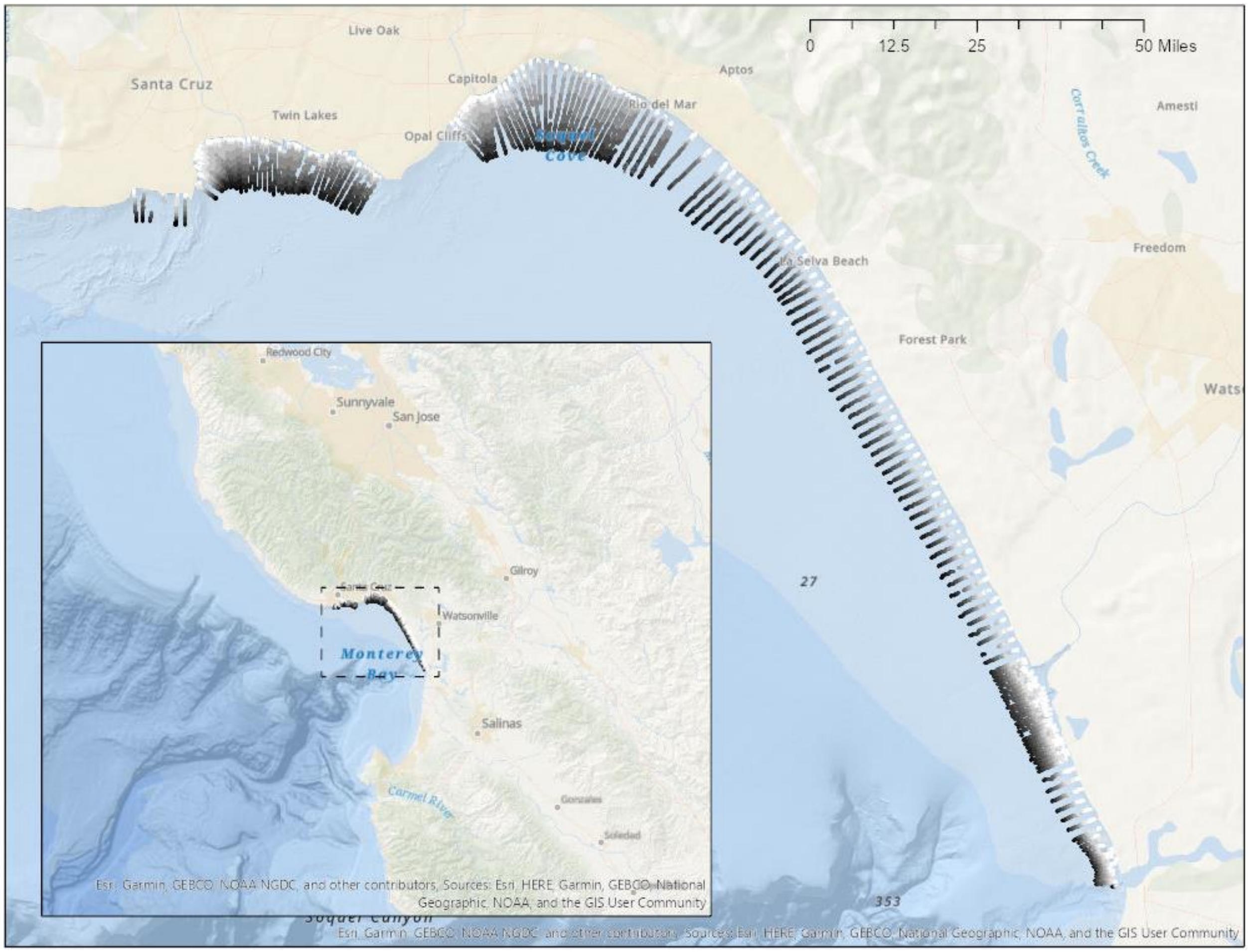
0
12.5
25
50 Miles
Live Oak
Santa Cruz
Capitola
Aptos
Twin Lakes
Opal Cliffs
Rio del Mar
Soquel Cove
Amesti
Corralitos Creek
La Selva Beach
Freedom
Forest Park
Wats
27
353
Soquel Canyon
Redwood City
Sunnyvale
San Jose
Gilroy
Santa Cruz
Watsonville
Monterey Bay
Salinas
Carmel River
Gonzales
Soledad
Esri, Garmin, GEBCO, NOAA NGDC, and other contributors, Sources: Esri, HERE, Garmin, GEBCO, National Geographic, NOAA, and the GIS User Community
Esri, Garmin, GEBCO, NOAA NGDC, and other contributors, Sources: Esri, HERE, Garmin, GEBCO, National Geographic, NOAA, and the GIS User Community

Figure 4. Example of a single beam data set downloaded directly from USGS showing point data in lines perpendicular to the shore. Survey 2016-627-FA.

4. RESULTS

We discovered 169 unique surveys from PCMSC, 161 from SPCMSC, and 101 from WHCMSC for a grand total of 431 USGS surveys of interest to NCEI. Table 3 breaks down the total number of datasets found at each science center by datatype. The ancillary datasets column refers to sediment samples, conductivity-temperature-depth (CTD) measurements, and bottom photos and videos.

Through this project, one raw multibeam survey from PCMSC was found and can be directly ingested by NCEI repositories. In addition, the IOCM program and NCEI now have documentation of every published USGS bathymetric survey between 2000 and 2020 that is not already in the NCEI databases. These 431 surveys will be sought for NCEI download and ingest in the coming months, which will help to fill in gaps in the U.S. Bathymetric Gap Analysis, among other uses.

Table 3. Number of unique USGS Field Activities inventoried for NCEI.

Science Center	Multibeam datasets	Single beam datasets	Interferometric datasets	Sidescan datasets	Sub-bottom profiles	Ancillary datasets
PCMSC	10	40	47	9	18	45
SPCMSC	21	77	33	0*	26	4
WHCMSC	35	2	41	8	15	0
Total	**66**	**119**	**121**	**17**	**59**	**49**

* This data type was collected as ancillary data to the principal data types that were desired, i.e., multibeam, single beam, and interferometric bathymetry.

5. CONCLUSION

Both techniques for data mining are effective and yield the same results. The main distinction lies where the researcher personally feels comfortable spending the bulk of their time. Regardless of approach, when selecting useful datasets from USGS, the user will have to read considerable amounts of metadata.

We recommend beginning the search with specific data goals. Identify your desired data type, equipment, and geographic area prior to diving into the search. The method you choose depends on your comfort level.

As a first step, try manually searching the USGS websites mentioned above to get a feel for the kinds of results you will get. You may already find what you need that way.

We recommend using the scraper tool if…

- you want multiple surveys that cover a larger area or an entire data type;
- you are comfortable with XML and Python— the scraper tool is quite user friendly;
- you enjoy coding; or
- you are only interested in publication-level information.

We recommend using the manual searching approach if…

- you are comfortable with repetitive tasks;
- you are searching for survey level information— manual searching will allow you to personally decide whether each survey is useful for you or not.

Following from this inventory, NOAA's IOCM team plans to develop a data submission agreement with USGS and provide assistance in the organization, documentation, and transfer of the existing data discovered because of this project via hard drives. Also, it was discovered during the drafting of this technical memorandum that USGS has unpublished datasets that were not picked up during this project to inventory published data. The team will seek more information about the amount and type of unpublished data.

In the longer term, the recommended path forward is to expand upon the data submission agreement between USGS and NOAA whereby a copy of USGS' raw data would transfer to NOAA within 90 days of the conclusion of each USGS survey. These data could then be shared in a timelier manner through NCEI's central repository to benefit numerous stakeholders and support the IOCM goal to "map once, use many times."

ACKNOWLEDGMENTS

Thank you to Evan Dailey from the USGS Pacific Coastal and Marine Science Center for providing key information about the survey data as well as guidance on the general data publishing workflow.

Thank you to VeeAnn Cross and Matthew Arsenault from the USGS Woods Hole Coastal and Marine Science Center for providing key information on metadata information and XML tags that helped develop the web scraper.

Thank you very much to Arnell Forde from the USGS St. Petersburg Coastal and Marine Science Center for providing vital information about center bathymetric activities and keyword distinctions to properly identify surveys.

We also extend a gracious thank you to our project advisors and editors of this technical memo: Ashley Chappell and James J. Miller of the Office of Coast Survey; and Christiane Reiser and Jennifer Jencks of NCEI.

REFERENCES

1. NOAA Office of Coast Survey. 2020. United States Bathymetry Gap Analysis. Available: https://noaa.maps.arcgis.com/home/item.html?id=4d7d925fc96d47d9ace970dd5040df0a

2. NOAA National Ocean Service. Hydrographic Surveying. Available: https://oceanservice.noaa.gov/navigation/hydro/

3. R2 Sonic. Five Main Differences between Multibeam and Interferometric Side Scan Sonars. Available: https://www.r2sonic.com/interferometric-side-scan-sonar-vs-mbes/

4. United States Geological Survey. USGS Coastal and Marine Geology Data Catalog. 2020. Available: https://data.usgs.gov/cmgp/#q=*%3A*

5. United States Geological Survey. USGS Publications Warehouse. Available: https://pubs.er.usgs.gov/

6. United States Geological Survey. ScienceBase Catalog. 2020. Available: https://www.sciencebase.gov/catalog/

7. United States Geological Survey. List of Metadata Records by Science Center. Available: https://cmgds.marine.usgs.gov/catalog/record-list.php

8. NOAA National Centers for Environmental Information. Bathymetric Data Viewer. Available: https://maps.ngdc.noaa.gov/viewers/bathymetry/

9. NOAA National Centers for Environmental Information. Trackline Geophysical Data. Available: https://maps.ngdc.noaa.gov/viewers/geophysics/

10. NOAA National Centers for Environmental Information. Water Column Sonar Data. Available: https://www.ngdc.noaa.gov/maps/water_column_sonar/index.html

11. Curators of Marine and Lacustrine Geological Samples Consortium. Index to Marine and Lacustrine Geological Samples (IMLGS). NOAA National Centers for Environmental Information. doi:10.7289/V5H41PB8. Available: https://www.ngdc.noaa.gov/geosamples/

12. NOAA National Centers for Environmental Information. Data Submission Guidelines. Available: https://www.ngdc.noaa.gov/iho/SubmittingMarineGeophysicalData.pdf

APPENDIX A: PYTHON SCRIPT FOR AUTOMATING THE USGS DATA INVENTORY

The following script was used to create the inventory of data holdings at WHCMSC. Webdriver is used to access the Javascript portion of the USGS Science Data Catalog where the metadata links are located. BeautifulSoup is then used to extract HTML tags of relevant data inside the metadata page.

The function USGS requires a Pandas DataFrame, a URL from the USGS Science Data Catalog, and the number of results found in the Data Catalog search. The function USGS_scraper is dependent on the URL being a link to a USGS Science Data Catalog search. The size is determined by rounding the number of results to the nearest ten while adding 10 more to that. This is done to account for the final 10 results in the results page. The function gets information based on its HTML tag inside a metadata page. If someone wants to use this code for another website, the HTML tags need to be changed to ones that are present in said website.

```
#import packages
from google.colab import files
import requests
import pandas as pd
from bs4 import BeautifulSoup
from selenium import webdriver
options = webdriver.ChromeOptions()
options.add_argument('--headless')
options.add_argument('--no-sandbox')
options.add_argument('--disable-dev-shm-usage')



# open it, go to a website, and get results
wd = webdriver.Chrome('chromedriver',options=options)

#function to remove HTML tags
import re
TAG_RE = re.compile(r'<[^>]+>')

def remove_tags(text):
  return TAG_RE.sub('', text)



#function for web scraper
def USGS_scraper(url_loop, df, size):
  wd_loop = webdriver.Chrome('chromedriver',options=options)
  for i in range(0,size,10):
    temp = 'start='+str(i)

    if i != 0:
       url_loop = url_loop.replace(original,temp)

    original = 'start='+str(i)
    wd_loop.get(url_loop)
    soup_range = BeautifulSoup(wd_loop.page_source)

    for  link  in  soup_range.find_all('a',  class_  =  'resultTitle',  href  =
True):
      link_ = link.get('href')
      link_ = link_.replace("full.html#","")

      #then goes into the individual xml link to extract
      page_loop = requests.get(link_)
      soup_loop = BeautifulSoup(page_loop.content)
      title = soup_loop.find('title')
      ID = soup_loop.find('issue')
      year_published = soup_loop.find('pubdate')
      geoform = soup_loop.find('formspec')
      startdate = soup_loop.find('begdate')
      enddate = soup_loop.find('enddate')
      metadata = link_

      if soup_loop.citeinfo is not None:

         # Try to get the CSRF token
         links = soup_loop.citeinfo.find_all('onlink')
      else:
         # Token not found. Replace 'pass' with additional logic.
         links = 'Not Found'

       entry=pd.DataFrame({'Issue':[ID],'Title':[title],'Date
Published':[year_published],'Start Date':[startdate],'End Date':[enddate],
'filetype':[geoform],'Links':[links], 'Metadata':[metadata]})
       df = df.append(entry, ignore_index = True)

   df['Issue'] = [remove_tags(str(text)) for text in df['Issue']]
   df['Title'] = [remove_tags(str(text)) for text in df['Title']]
   df['Date Published'] = [remove_tags(str(text)) for text in df['Date
Published']]
   df['Start Date'] = [remove_tags(str(text)) for text in df['Start Date']]
   df['End Date'] = [remove_tags(str(text)) for text in df['End Date']]
   df['Links'] = [remove_tags(str(text)) for text in df['Links']]
   df['filetype'] = [remove_tags(str(text)) for text in df['filetype']]
   df['Links'] = [remove_tags(str(text)) for text in df['Links']]
   df['Metadata'] = [remove_tags(str(text)) for text in df['Metadata']]

   df = df.drop_duplicates(subset = ['Title'])
   df = df.reset_index(drop = True)

 return df
```

```
# Block of code to generate a dataframe of multibeam data from the Woods
Hole Region

df_bathy_multi = pd.DataFrame(columns = ['Issue','Title','Date Published',
'Start Date', 'End Date', 'filetype','Links'])

url_bathy_multi =
'https://data.usgs.gov/datacatalog/#fq=keywords%3A(%22multibeam%20bathymetr
y%22)&fq=datasource%3A(%22Woods%20Hole%20Coastal%20and%20Marine%20Science%2
0Center%22)&q=*%3A*&start=0'

df_bathy_multi = USGS_scraper(url_bathy_multi, df_bathy_multi, 180)

df_bathy_multi
```

APPENDIX B: PYTHON SCRIPT FOR DOWNLOADING DATA FROM USGS SITES

The following script was used to download the one raw multibeam survey found at USGS. It does not work for large datasets that need to be downloaded from the cloud and require multistep verification for download. If the script encounters a large dataset, it may take some time to download it. The script requires an Excel spreadsheet with a column of FAN numbers without dashes (“FAN”) and download links (“data_download”). It also includes a column of additional details to identify FAN with multiple datasets (“name_add”). For example, “5m” for 5 meter resolution creates a dataset titled “S2210MB_5m”, and “chenega” for an Alaskan survey with two locations “2014622FA_chenega”.

```
# import packages
import pandas as pd
import requests
import os
from zipfile import ZipFile
import numpy as np

# import excel file and convert to dataframe
xls = pd.ExcelFile('multi_download_links.xlsx') # import excel file storing
download links
df = pd.read_excel(xls, 'downloadable multibeam data')  # select sheet to
convert to dataframe
df = df.dropna(subset=['data_download']) # drops rows with empty values
df = df.reset_index() # creates new index for df
df = df.drop(columns='index') # drops old index

print(df)

# checking data type of cell from spreadsheet
print(str(type(df.name_add[0]))=="<type 'unicode'>")
print(str(type(df.name_add[0])))

# create function that downloads datasets from URLs in spreadsheet
def download(url, FAN, detail, dest_folder):  # url = data download url, FAN
= FAN number, dest_folder = desired folder name

    filename = FAN + '.zip' # designate download filename
    path = 'multi_downloads'  # designate path to desired file location

    # create file path for download
    if str(type(detail))=="<type 'unicode'>": # check the data type
        file_path = os.path.join(path, dest_folder, dest_folder + '_' +
detail)
```

```
    else:
        file_path = os.path.join(path, dest_folder)
    if not os.path.exists(file_path):  # check if folder already exists
        os.makedirs(file_path)  # create folder if it does not already exist

    r = requests.get(url, allow_redirects=True)  # accesses the URL

    if r.ok:  # checks if the request is valid
        print('saving to' + os.path.abspath(file_path))
        with open(file_path + '\\' + filename, 'wb') as f:
            f.write(r.content)  # opens the download URL and saves it to
correct file path
    else:
        print('Download failed: status code {}\n{}'.format(r.status_code,
r.text)) # print status code if the URL does not work

    # extracting the zip files
    # opening the zip file in READ mode
    with ZipFile(file_path + '\\' + filename, 'r') as zip:
        # printing all the contents of the zip file
        zip.printdir()

        # extracting all the files
        print('Extracting all the files now...')
        zip.extractall(file_path)
        print('Done!')

    # delete leftover zipfile
    if os.path.exists(file_path + '\\' + filename):

        os.remove(file_path + '\\' + filename)

    else:
        print(filename + ' does not exist')

# run the download function for a single FAN
download(url=df.data_download[0], FAN=df.FAN[0], detail=df.name_add[0],
dest_folder=df.FAN[0])

# create loop that iterates through the links spreadsheets utilizing the
function above
index = range(0,len(df)) # create array of length of dataframe
for i in index: # loop through code
    download(url=df.data_download[i], FAN=df.FAN[i], detail=df.name_add[i],
dest_folder=df.FAN[i])
```