Data Processing

Jethan d'Hotman
South African Environmental Observation Network
js.dhotman@saeon.nrf.ac.za

Question:

Have you ever been to an aquarium / museum and seen a display without an informational graphic and didn’t know what you were looking at?

- The animal / object is in front of you, all the data is there, so why don’t you know? what's missing?

Metadata

Data (information) about your data

- Gives context to your data
- Provides information about the quality of your data
- Helps with troubleshooting issues
- Free
    - Its information you already have

Metadata cont.
Example:

Project:	SAEON	**Variables:**	1
Sub Project:	Algoa Bay Long Term Research	**Deploy Depth (m):**	30
Time Series:	WCE_30m_024	**Bottom Depth (m):**	30
Series Code:	ABLTMRP_WCEAS_UTR_30m_1/4_024	**Bin No:**	1
Storage location	C:\Data Store\Algoa Bay Sentinel Site (ABSS)\ Continuous Monitoring Platform (CMP)\Staging\ UTRs\20210311_dHotman	**Total Bins:**	4
Location:	Woody Cape	**Start Date:**	2020-09-15
Site:	East	**Start Time:**	14:00:00
GPS Name:	UTR - WOODY	**End Date:**	2021-03-10
Mooring Type:	UTR Array	**End Time:**	14:00:00
Mooring No:	3	**Instrument:**	Onset Hobo Pro v2 Water Temperature Logger
Latitude:	-33.75	**Serial No:**	20174977
Longitude:	26.22	**Comment:**	
Sample Interval (hr):	1		
No of samples:	4225		

Back to Data Processing

What is data processing?

Wikipedia

Data processing is, generally, "the collection and manipulation of items of data to produce meaningful information."[1] In this sense it can be considered a subset of *information processing*, "the change (processing) of information in any manner detectable by an observer."[note 1]

Britannica

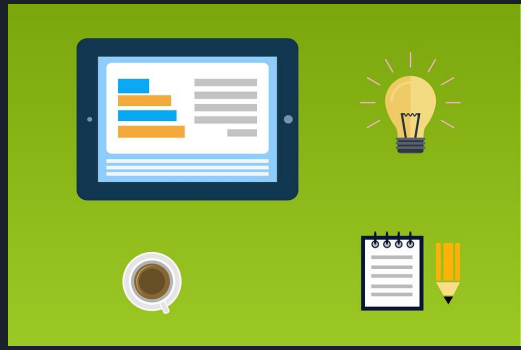
data processing, manipulation of data by a computer. It includes the conversion of raw data to machine-readable form, flow of data through the CPU and memory to output devices, and formatting or transformation of output.

Nature

Data processing is a set of methods that are used to input, retrieve, verify, store, organize, analyse or interpret a set of data.

Why process your data?

- **Validation**
  - To ensure your data is correct
- **Sorting**
  - Arrange the data in some sort of sequence
- **Classification**
  - Separate the data into various categories
- **Summarization**
  - Provide statistical values to reduce detail
- **Aggregation**
  - Combine multiple data sets
- **Analysis**
  - Interpretation of the results
- **Reporting**
  - List and present the detail of the results



Types

Manual

Pros

- No complicated setup
- No specialised skills

Cons

- Consistently time consuming
- Prone to mistakes
- Limited to smaller datasets
- Limited to basic processing

Automatic

Pros

- Quick and easy (once setup)
- Can perform more complicated processing
- Can process large datasets

Cons

- Long to set up
- Need specialised skills
- Mistakes can affect multiple datasets

Data processing basic steps

1. Conversion into a user friendly format (e.g. csv, netcdf etc)
2. Add metadata
3. Remove data collected outside the deployment period
    a. Top and tail
4. Quality assurance / control
    a. Add quality flags
5. Remove outliers / impossible values
    a. More advanced

Keep a record of any processing you do

Data conversion

December CTD004.hex - Notepad

File Edit Format View Help

```
* Sea-Bird SBE19plus  Data File:
* FileName = C:\Documents and Settings\FIELD\Desktop\December CTD004.hex
* Software version 2.2.0 SB_Terminal.dll
* Temperature SN = 6802
* Conductivity SN = 6802
* System UpLoad Time = Dec 06 2013 12:48:43
* <ApplicationData>
* <SeatermAF>
* <SoftwareVersion>2.1.0</SoftwareVersion>
* <BuildDate>25-Jul-2012 22:05:08 GMT</BuildDate>
* </SeatermAF>
* </ApplicationData>
* cast   4 05 Dec 2013 05:50:00 samples 738 to 21509, avg = 1, stop = mag switch
*END*
067CC109C81908154B4E2E968C00008E9747B32F95FB0C
067CC209C81708154B4E2F969100008EA447BE3C25FB08
067CC509C81808154C4E2E969300008EAC47F14486FB0B
067CC309C81908154E4E2E9695000C8EA747D3444AFB0B
067CC409C81808154E4E2E969B00088EAC47A842A7FB07
067CC709C81708154D4E2E969A00008EAA478B3E9BFB07
067CD209C81808154E4E2E969600008EAB475C32C9FB03
067CE009C81708154F4E2E969700008EAC4749238AFB12
067CE509C8170815514E2D969F00008EAB47291841FB0D
067CE509C8170815514E2D969E00008EB4470A12FBFB04
067CE909C81708154E4E2D969900008EA9474D1C58FB04
067CF109C8180815504E2C969F00008EA5464227AAFB0F
067CFE09C8170815524E2C96A700008EA646883325FB0A
067D0009C8170815524E2C969900018EA4469029BAFB09
067D0009C8170815524E2C96A200008EA247081E2AFB09
067D3C09C81F0815654E2C969900018E1173161852FB0A
```

```
# </Sensors>
# datcnv_date = Oct 22 2021 14:28:27, 7.26.7.129 [datcnv_vars = 11]
# datcnv_in = C:\Users\JethanD\Downloads\Depth\December CTD004.hex C:\Users\JethanD\Downloads\Depth\Previous years data no pH.xmlcon
# datcnv_skipover = 0
# datcnv_ox_hysteresis_correction = yes
# datcnv_ox_tau_correction = yes
# file_type = ascii
*END*
  16.740098    14.7252     -0.049     0.0146    11.2717     -0.049      9.3935    980.6668          0 339.243056          1  0.000e+00
  16.677730    14.7251     -0.049     0.0146    11.2743     -0.050     11.8472    980.6058          0 339.243058          2  0.000e+00
  16.708914    14.7249     -0.047     0.0146    11.2761     -0.048     13.4837    980.6516          0 339.243061          3  0.000e+00
  16.740098    14.7250     -0.044     0.0146    11.2778     -0.044     13.4379    980.6516          0 339.243064          4  0.000e+00
  16.708914    14.7250     -0.044     0.0146    11.2816     -0.044     13.1182    980.5905          0 339.243067          5  0.000e+00
  16.677730    14.7248     -0.045     0.0146    11.2822     -0.046     12.3278    980.5905          0 339.243070          6  0.000e+00
  16.708914    14.7241     -0.044     0.0146    11.2805     -0.044     10.0191    980.5295          0 339.243073          7  0.000e+00
  16.677730    14.7232     -0.042     0.0146    11.2804     -0.042      7.0413    980.7584          0 339.243076          8  0.000e+00
  16.677731    14.7228     -0.038     0.0146    11.2835     -0.038      4.8372    980.6821          0 339.243079          9  0.000e+00
  16.677731    14.7228     -0.038     0.0146    11.2824     -0.038      3.8072    980.5447          0 339.243082         10  0.000e+00
  16.677731    14.7226     -0.043     0.0146    11.2794     -0.044      5.6360    980.5447          0 339.243084         11  0.000e+00
  16.708914    14.7221     -0.040     0.0146    11.2814     -0.040      7.8470    980.7126          0 339.243087         12  0.000e+00
  16.677731    14.7212     -0.036     0.0146    11.2855     -0.036     10.0893    980.6363          0 339.243090         13  0.000e+00
  16.677731    14.7211     -0.036     0.0146    11.2798     -0.036      8.2498    980.6210          0 339.243093         14  0.000e+00
```

Ln 1, Col 1 | 100% | Windows (CRLF) | UTF-8

Add metadata

	Date Time, GMT+02:00	Temp, °C (LGR S/N: 10106788, SEN S/N: 10106788)
1	2012-11-05 12:00	25.331
2	2012-11-05 13:00	26.134
3	2012-11-05 14:00	25.671
4	2012-11-05 15:00	27.014
5	2012-11-05 16:00	28.866
6	2012-11-05 17:00	29.690
7	2012-11-05 18:00	27.284
8	2012-11-05 19:00	24.412
9	2012-11-05 20:00	23.689
10	2012-11-05 21:00	22.705
11	2012-11-05 22:00	22.058
12	2012-11-05 23:00	21.581
13	2012-11-06 00:00	21.151
14	2012-11-06 01:00	20.793
15	2012-11-06 02:00	20.484
16	2012-11-06 03:00	20.222

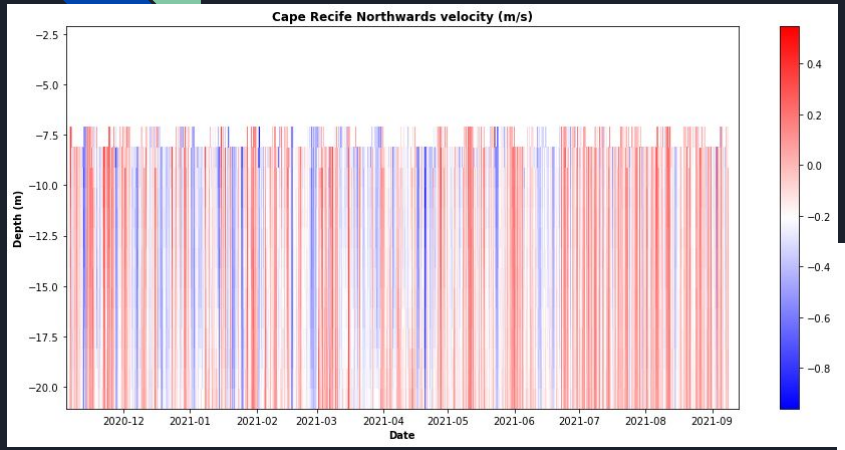
# Project:	SAIAB
# Sub Project:	ATAP
# DeployCode:	CP001_001
# Station:	CP001
# Longitude:	nan
# Latitide:	nan
# Deployment time (GMT+2):	2012-11-10 06:54
# Recovery time (GMT+2):	2013-06-22 13:24
# Depth (m):	
# Acoustic Release SN:	120142
# Sample interval (min):	60
# Mumber of samples:	5383
# Instrument:	Onset Hobo Pro v2 Water Temperature Logger
# Serial No:	10106788
# Comment:	
Date Time, GMT+02:00	Temp
2012-11-10 07:00	20.793
2012-11-10 08:00	22.106
2012-11-10 09:00	11.2
2012-11-10 10:00	10.956
2012-11-10 11:00	11.053

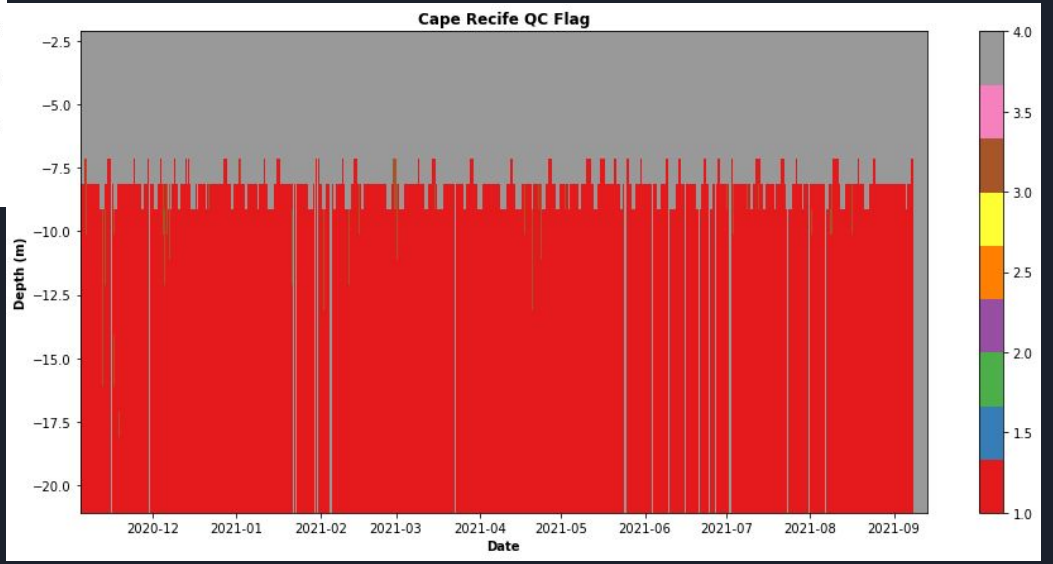
Top and tail

# Project:	SAIAB	
# Sub Project:	ATAP	
# DeployCode:	CP001_001	
# Station:	CP001	
# Longitude:	nan	
# Latitide:	nan	
# Deployment time (GMT+2):	2012-11-10 06:54	
# Recovery time (GMT+2):	2013-06-22 13:24	
# Depth (m):		
# Acoustic Release SN:	120142	
# Sample interval (min):	60	
# Mumber of samples:	5383	
# Instrument:	Onset Hobo Pro v2 Water Temperature Logger	
# Serial No:	10106788	
# Comment:		
Plot Title: Cape Point		
#	Date Time, GMT+02:00	Temp, °C (LGR S/N: 10106788, SEN S/N: 10106788)
1	2012-11-05 12:00	25.331
2	2012-11-05 13:00	26.134
3	2012-11-05 14:00	25.671
4	2012-11-05 15:00	27.014
5	2012-11-05 16:00	28.866
6	2012-11-05 17:00	29.690
7	2012-11-05 18:00	27.284

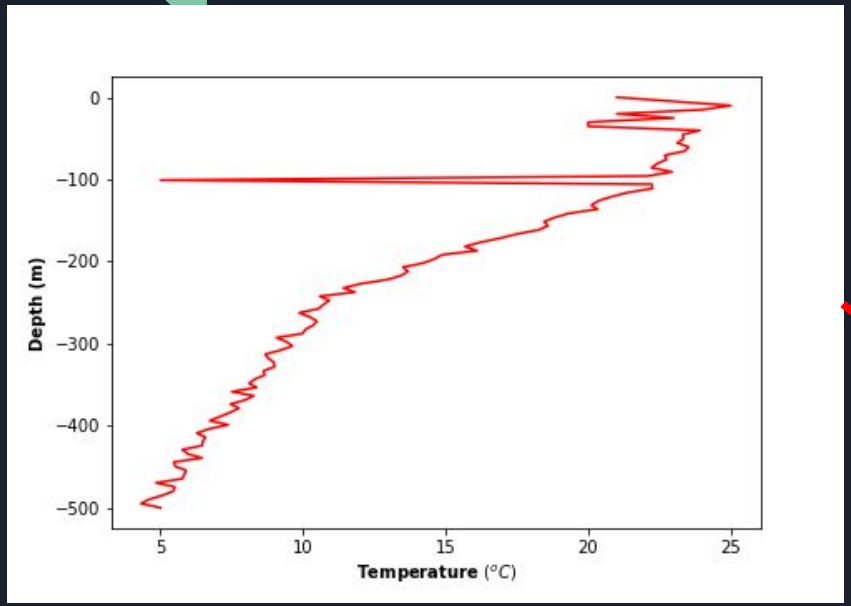
# Project:	SAIAB	
# Sub Project:	ATAP	
# DeployCode:	CP001_001	
# Station:	CP001	
# Longitude:	nan	
# Latitide:	nan	
# Deployment time (GMT+2):	2012-11-10 06:54	
# Recovery time (GMT+2):	2013-06-22 13:24	
# Depth (m):		
# Acoustic Release SN:	120142	
# Sample interval (min):	60	
# Mumber of samples:	5383	
# Instrument:	Onset Hobo Pro v2 Water Temperature Logger	
# Serial No:	10106788	
# Comment:		
Date Time, GMT+02:00	Temp	
2012-11-10 07:00	20.793	
2012-11-10 08:00	22.106	
2012-11-10 09:00	11.2	
2012-11-10 10:00	10.956	
2012-11-10 11:00	11.053	

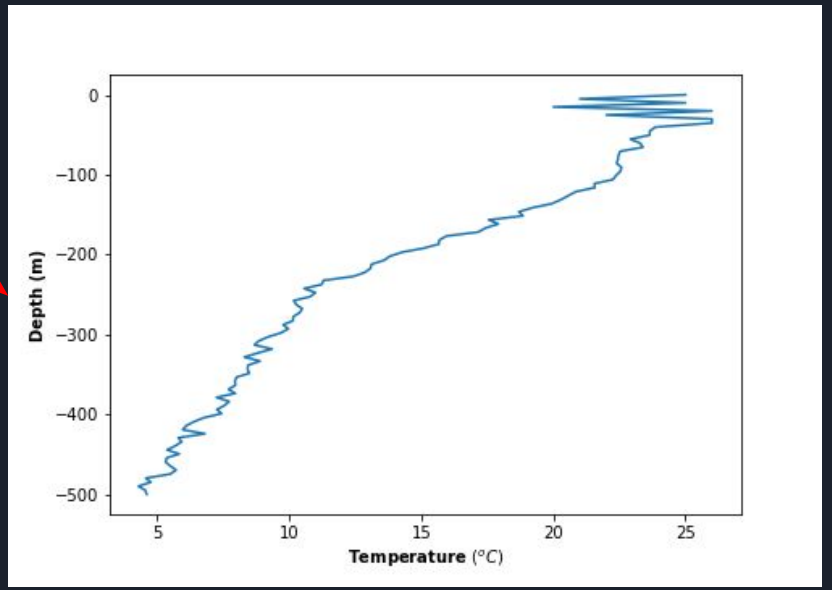
Quality assurance / control





Removing outliers

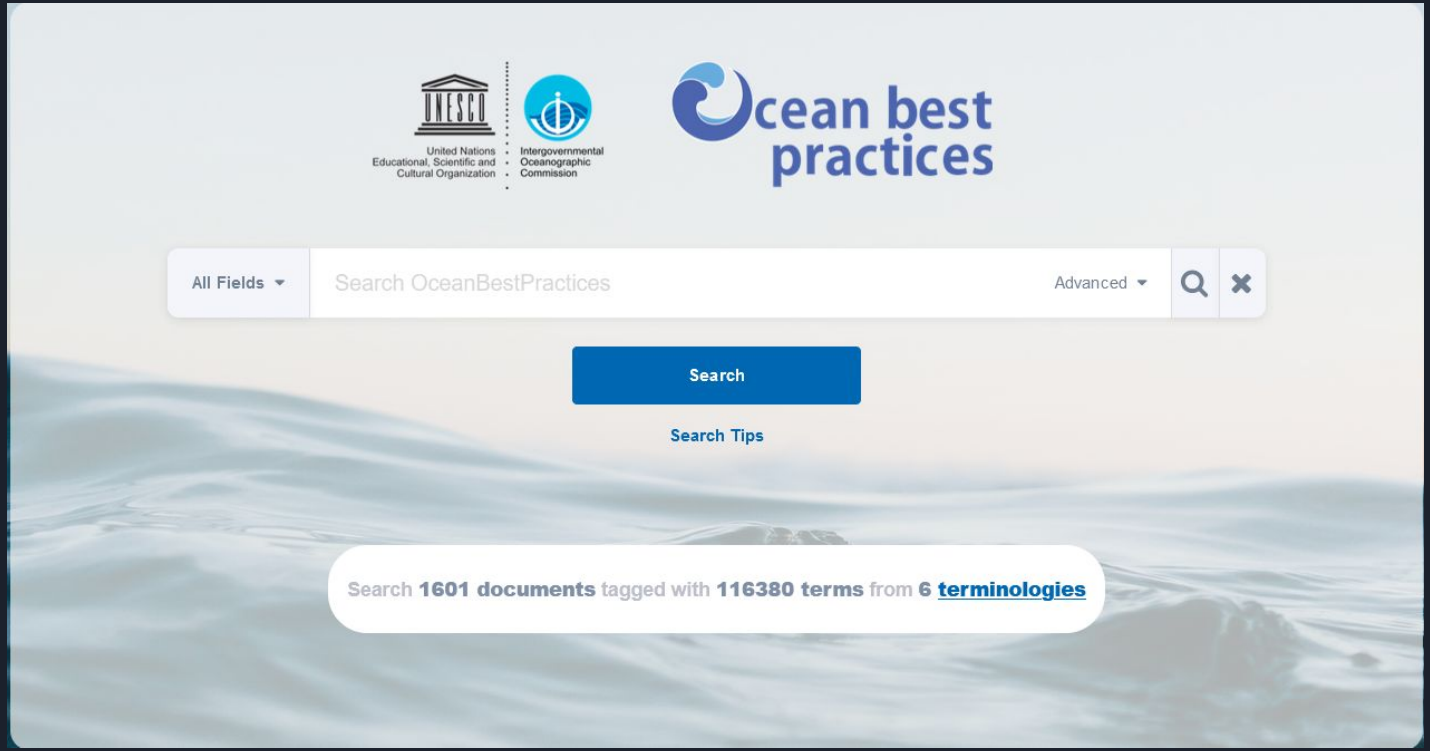




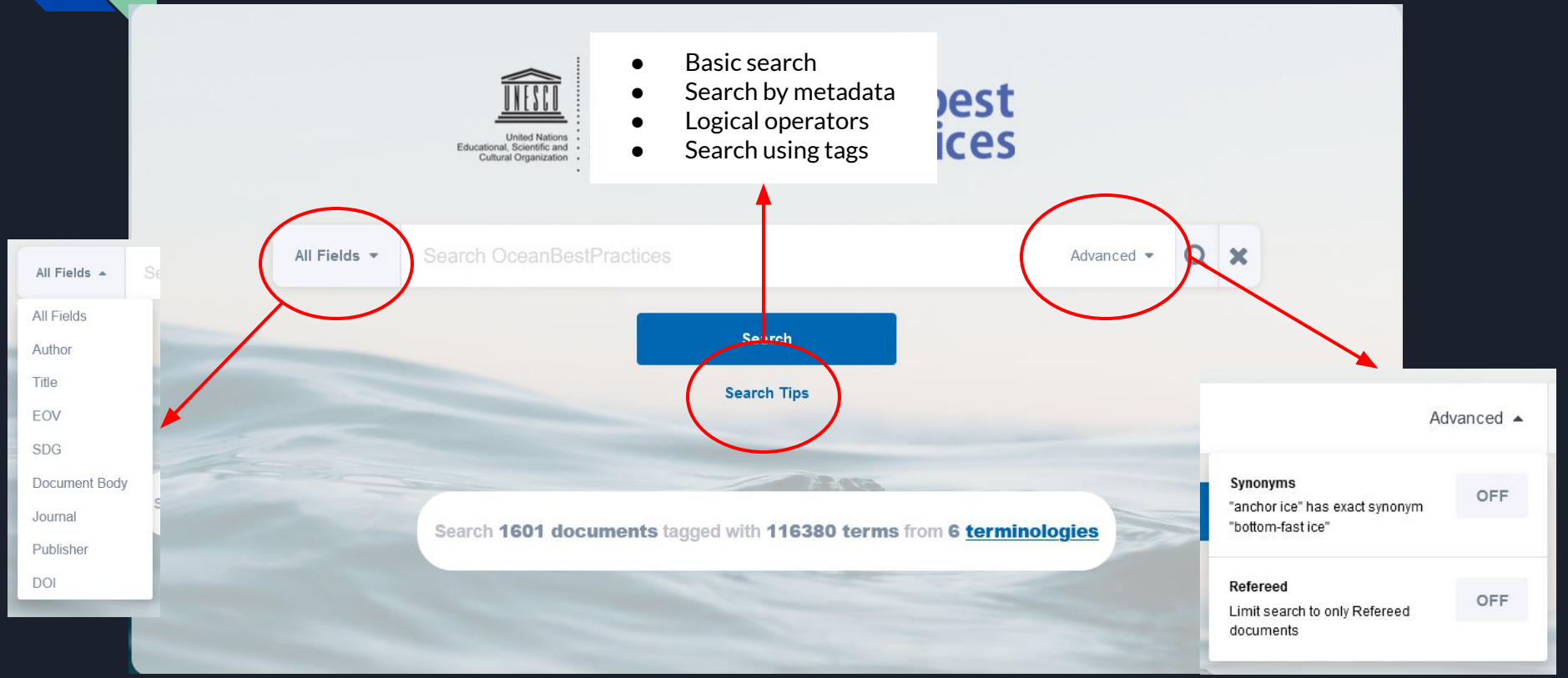
Examples / guides

1. Instrument user guides
    a. SeaBird data processing user manual
2. Program documentation / guides
    a. Argo data cookbook
    b. Go-Ship Hydro manual
3. Scientific and technical reports
    a. Thomson, R.E. and Emery, W.J. (2014). Data Analysis Methods in Physical Oceanography. San Diego, Ca, Usa: Elsevier Science.
4. Programs and toolboxes
    a. QARTOD
    b. IMOS
5. Ocean Best Practices Repository
    a. https://search.oceanbestpractices.org/
    b. Links to more than 1000 data processing documents / guides

Using the OBPS repository



Using the OBPS repository cont.



Summary

1. Metadata Metadata Metadata
2. Data processing serves many functions
    a. Including validation and analysis
3. There are many data processing guidelines and tools to help you get started
4. 5 main/common processing steps
    a. Data convection
    b. Add metadata
    c. Top and tail
    d. Quality control
    e. Remove outliers
5. **Keep a record of any data processing you do**

Thank you !!!

Project:		Sub Project:	
Time Series:		Series Code:	
Site:		GPS Name:	
Mooring Type:		Mooring No:	
Latitude:		Longitude:	
Deploy Depth (m):		Bottom Depth (m):	
Bin No:		Total Bins:	
Start Date:		Start Time:	
End Date:		End Time:	
Interval (hr):		No of Samples:	
Instrument:		Serial No:	
No of Variables:		Variables	
Last calibrated:		Raw / processed:	
Supporting documents:		Contact info	
Storage Location / DOI:		Comment:	