

Taming an Ocean of Data at AOOOS

End to end data lifecycle management

Background - Axiom

- Cyberinfrastructure technology development:
 - environmental, biological and geoscientific data
- AK Headquarters, OR and RI satellite offices
- Support federal, university and NGO groups
- Mission driven



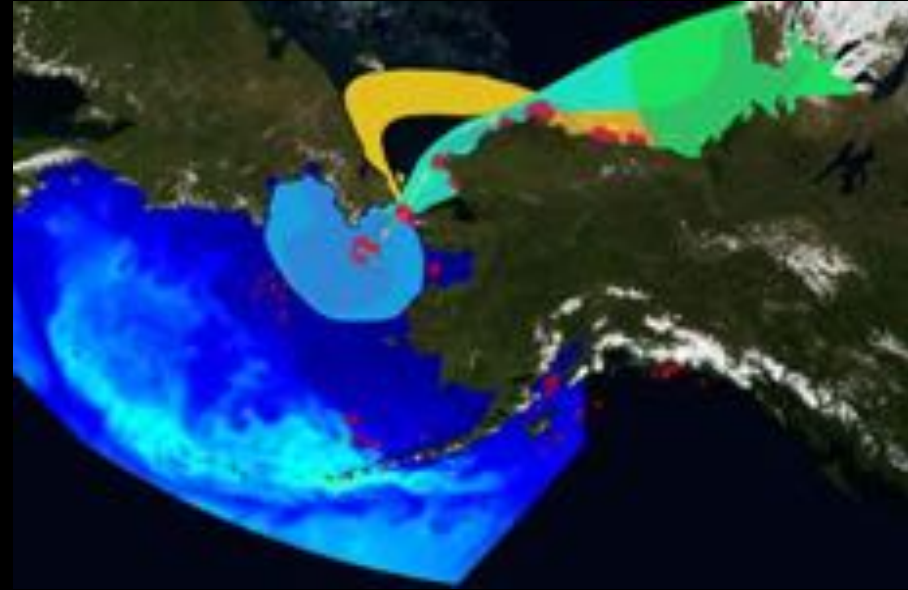
Background - Axiom

- Shared cyberinfrastructure approach
- Community developed software, standards and protocols
- Scalable compute and storage infrastructure (HPC)
 - 5 petabytes storage; 3,000 processing cores



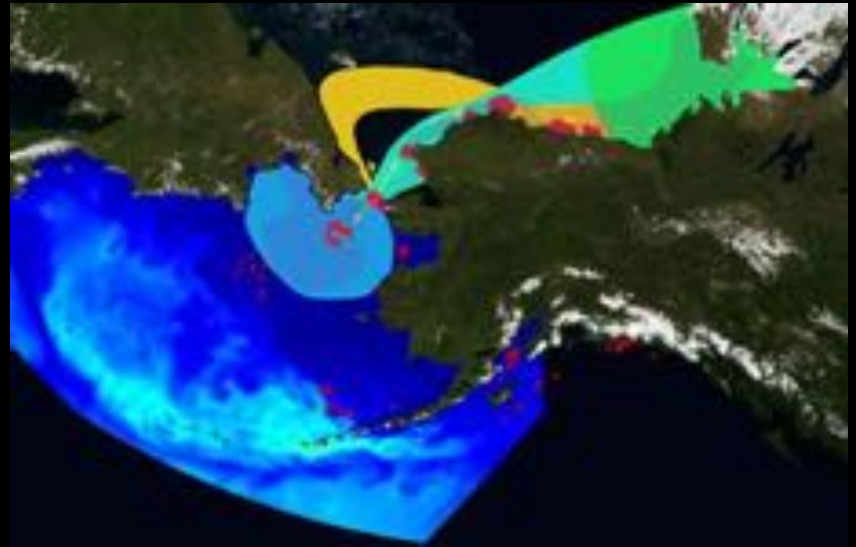
Background - Alaska Ocean Observing System

- Officially established in 2005
- Regional member of the Integrated Ocean Observing System (IOOS)
- Network of critical ocean and coastal observations, data and information products
- “Eye on Alaska’s coasts and oceans”



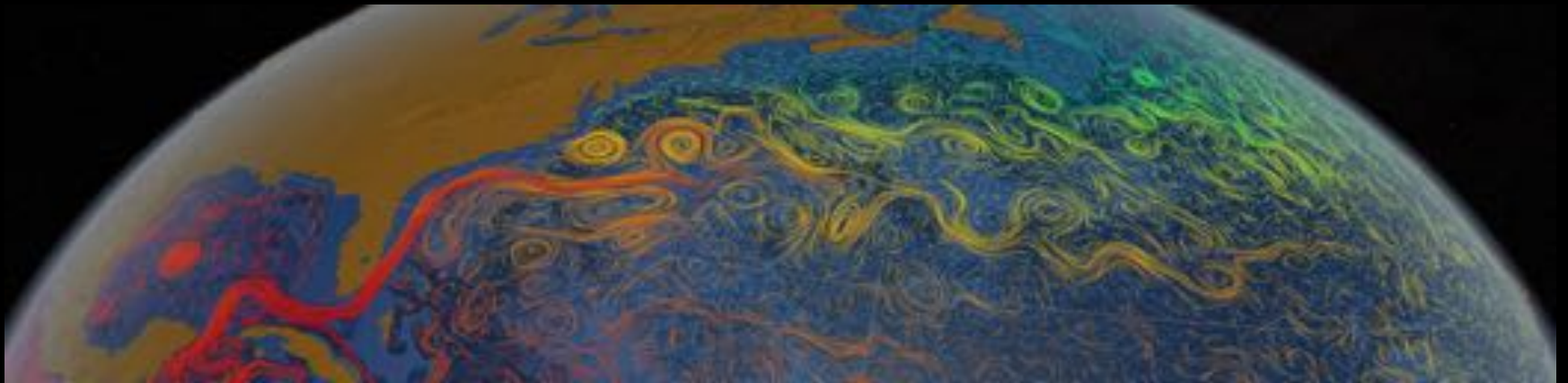
Goals - Alaska Ocean Observing System

- Increase access to existing coastal and ocean data
- Package information and data in useful ways to meet the needs of stakeholders
- Increase observing and forecasting capacity in all regions of the state, with a priority on the Arctic and Gulf of Alaska



Challenges - Alaska Ocean Observing System

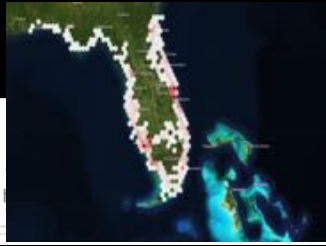
- Lots of existing data and research efforts in the region
 - Varying data/metadata quality
 - Limited sharing beyond specific research effort
 - Often not ready for public consumption
 - Complex data formats
 - Lack of context/metadata
 - Hard to find (not discoverable)
 - Isolated (not interoperable/comparable/synthesizable)



Data Types

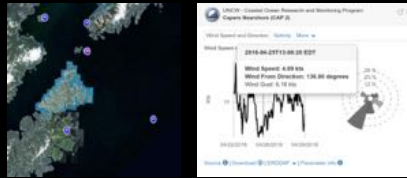
Biodiversity

count, richness, diversity indices



Platforms

moorings, shore stations



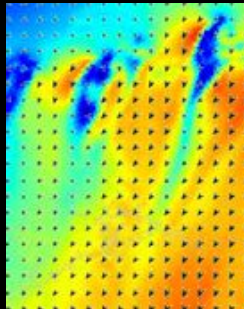
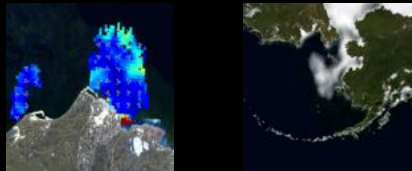
Products

skill assessment, shoreline change, etc.



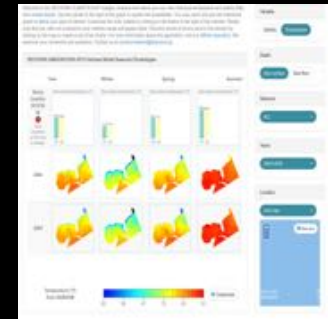
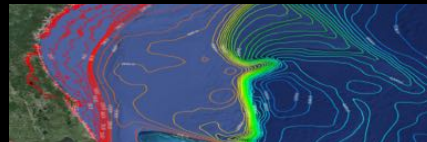
Grids

models, satellite, radar

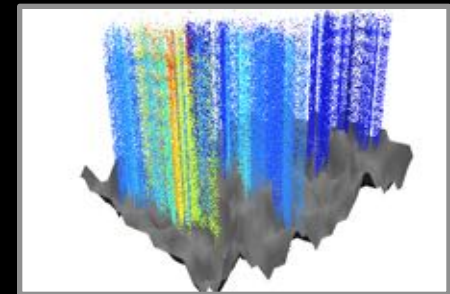
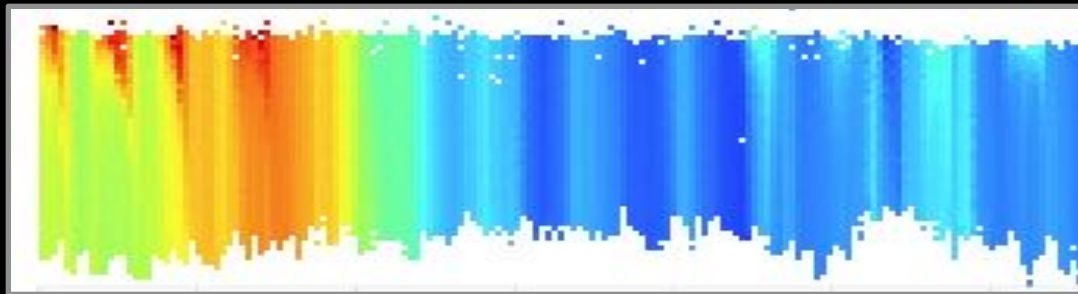


GIS

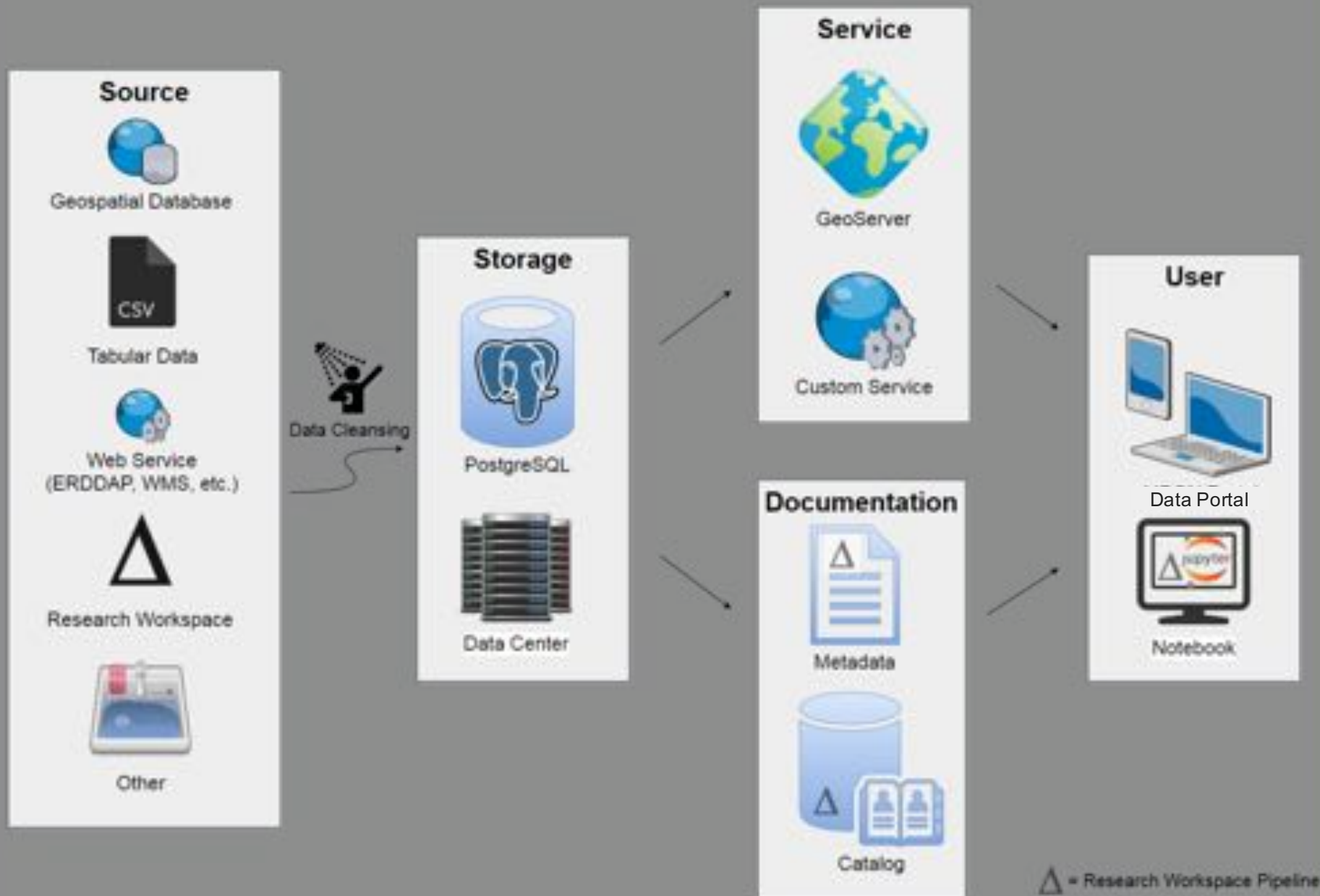
Habitat types, bathymetry, fishing zones, etc.



Gliders



Data Pipeline



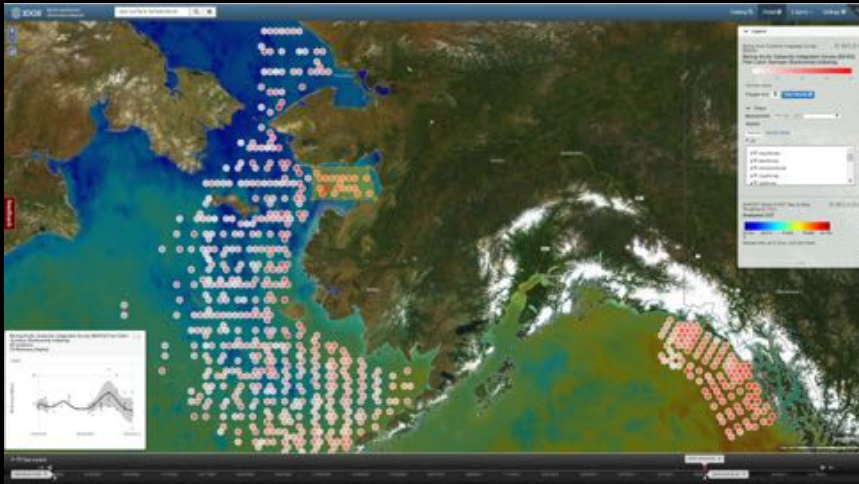
Data Cleaning/Upgrades

- Data
 - Structural/syntax problems
 - Quality Control/Quality Assurance (QA/QC)
 - Clean up or flag invalid/suspect data
 - Monitor data stream sources for outages
- Metadata
 - Make sure observable properties are clearly defined
 - Make sure units are clearly defined
 - Make sure spatial and temporal axes are defined
 - Use community standards when possible
 - CF conventions/standard names
 - ACDD metadata attribute conventions
 - **Dataset should be self describing!!**
 - **Even to people outside of the domain**

Data Portal - Public Data Exploration and Access

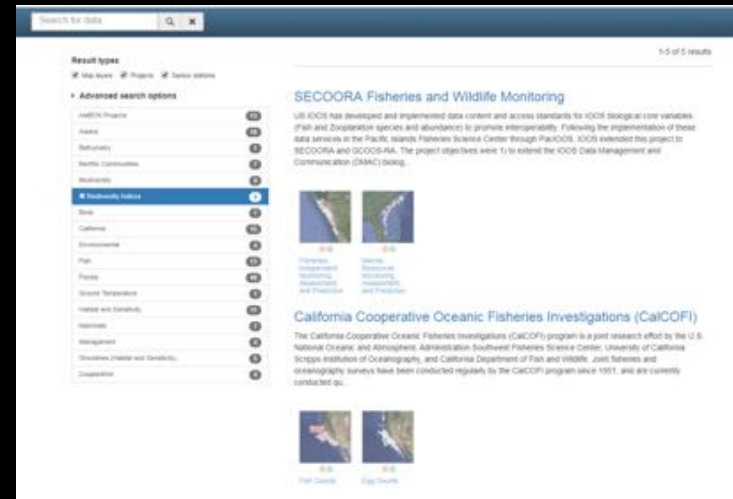
Map

Integrate & visualize data from many sources



Catalog

Search, metadata, & data download



Data Views

Rapidly assimilate & compare different data streams



Downloads Using Interoperability Services

Egg Counts

Metadata URL: <https://openviewwg.noaa.gov/metadata/>

This layer includes fish egg counts and standardized counts for eggs captured in CaCO₃ otoplancton nets (primarily vertical [Calvet or Parrelve], oblique [oblique tow], and surface tow [delta net]). Surface tows are normally standardized to count per 1,000 m³ strained. Oblique tows are normally standardized to count per 10 m² of the surface sampled. Egg densities include only tows where one or more eggs were captured for the species selected by the user, i.e. no "zero" tows. The "Egg Count" data set includes all tow species, i.e. both positive and negative tows.

Filter options:

Alpha: takes the average of event values within the selected area
Gamma: groups all events within the selected area and treats them as a single sample
Beta: Gamma*Alpha

Richness: Count of distinct species
% Dominance (Singer-Parker): Numerical importance of the most abundant species
Shannon-Wiener Diversity: This index quantifies the uncertainty associated with species prediction
Pielou's Evenness: Species evenness quantifies how close in count each species is within a sampling event

California Cooperative Oceanic Fisheries Investigations (CCOPI)
Egg

Download

ncWMS
Shapefile
CSV
THREDDS
netCDF
OPenDAP
ERDDAP

Richness (gamma)

Bering Arctic Subarctic Integrated Survey (BASIS) Fish Catch Surveys Biodiversity Indexing
Privacy selection:

Calculated data

Binned years

Raw data

CSV <https://data.arctic.cohgs.gov/ifs?service=WFS&version=1.0.0>

Shape file <https://data.arctic.cohgs.gov/ifs?service=WFS&version=1.0.0>

JSON <https://data.arctic.cohgs.gov/ifs?service=WFS&version=1.0.0>

Download

ERDDAP



ERDDAP

Easier access to scientific data

Brought to you by NOAA NMFS SWFSC EAD

ERDDAP > tabledap > Make A Graph

Dataset Title: **Granite Crk**

Institution: SnoTel (Dataset ID: gov_usda_nrcs_wcc_snotel_963)

Range: longitude = -145.395 to -145.395°E, latitude = 63.945 to 63.945°N, depth = 0.0 to 0.9396m, time = 2018-05-05T13:00:00Z to 2018-05-22T08:00:00Z

Information: [Summary](#) | [License](#) | [FGDC](#) | [ISO 19115](#) | [Metadata](#) | [Background](#) | [Data Access Form](#)

Graph Type: **linesAndMarkers**

X Axis: **time**

Y Axis: **air_temperature**

Color: **longitude**

Constraints

time	▼
▼	
▼	
▼	
▼	

Optional
Constraint #1

time	▼	2018-05-15T00:00:00Z
▼		
▼		
▼		
▼		

Optional
Constraint #2

time	▼	2018-05-22T08:00:00Z
▼		
▼		
▼		
▼		

Server-side Functions

distinct

▼	▼	▼
▼	▼	▼

Graph Settings

Marker Type: **Filled Square** Size: **5**

Color:

Color Bar: **▼** Continuity: **▼** Scale: **▼**

Minimum: Maximum: N Sections: **▼**

Y Axis Minimum: Maximum: Ascending: **ascending**

Redraw the Graph (Please be patient. It may take a while to get the data.)

Optional

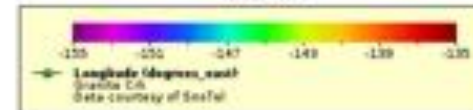
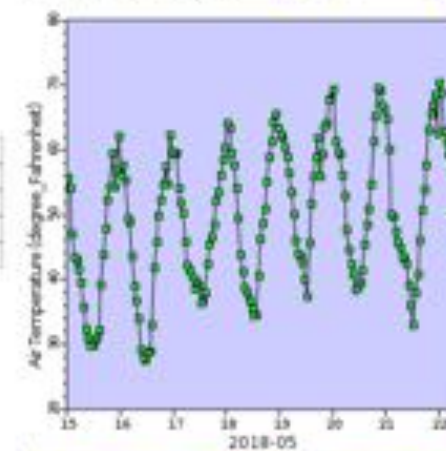
Then set the File Type: **htmlTable** (File Type information)

and [Download the Data or an Image](#)

or view the URL: http://erddap.xos.org/erddap/tabledap/gov_usda_nrcs_wcc_snotel_963.htmlTable?time=1

([Documentation](#) / [Bypass this form](#))

Time range: **7** day(s)



That's great for structured, predictable data. What about research programs, synthesis projects, citizen science, etc.?

RESEARCH WORKSPACE

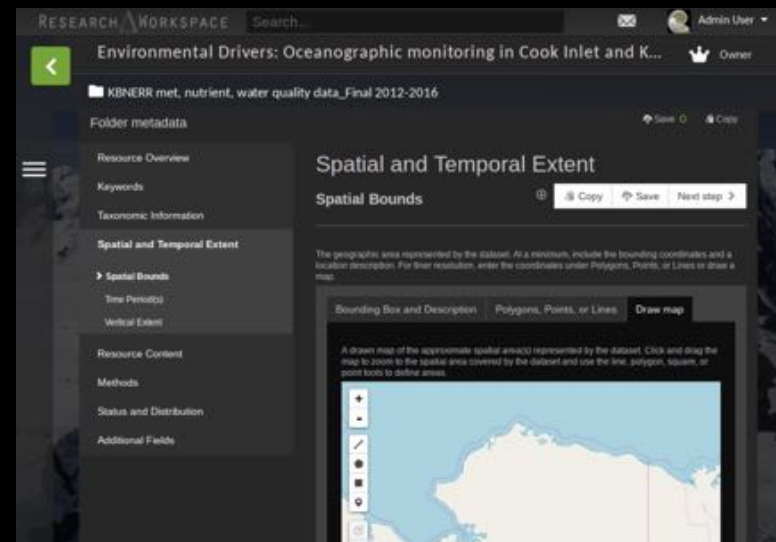
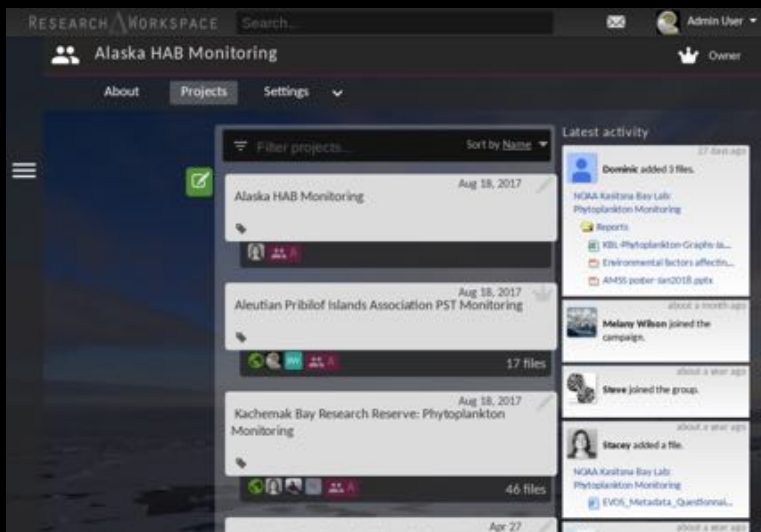
~web-based platform for collaboratively managing science projects through the entire data lifecycle~

Share
Analyze
Preserve



RESEARCH WORKSPACE

- Organize into projects, research campaigns, and organizations
- Manage sharing through advanced security permissions
- Coordinate data exchange across networks, groups, programs
- ISO 19115/19110 standards metadata editor
- Execute server side Python and R numeric workflows (Jupyter) on uploaded data AND any data in Axiom CI stack
- Archive pathway to DataONE & Datacite DOI minting



Supporting the entire data lifecycle

DATA COLLECTION & QUALITY CONTROL

Scientists or Ingestion

STORAGE

Research Workspace

DESCRIPTION

Metadata Editor

ARCHIVE & PRESERVATION

*Repository submission
pathway*

ACCESS & DISCOVERY

*Data portals & search
catalogs*

REUSE & TRANSFORMATION

*Jupyter Notebook & data
analyses*

Research Workspace - Metadata

- We help build large collections of diverse datasets.
- Unorganized, undocumented data collections benefit no one.
- Metadata tells the story (who, what, when, where, why, and how) of the data.
- Metadata makes data discoverable.

Research Workspace - Metadata

- The best metadata is written by the people closest to the data.
- Many researchers aren't familiar with authoring it.
- Metadata can be difficult to write well (existing standards are complex and confusing).
- Researchers are also very busy people.

Resource Overview

Basic Overview

Contacts

Category and Form

Keywords

Taxonomic Information

Spatial and Temporal Extent

Resource Content

Methods

Status and Distribution

Additional Fields

Resource Overview

Basic Overview



Copy

Save

Next step >

This section provides an overview about the dataset and any associated file(s).

Resource Title

A descriptive title for the data that includes the subject matter, where data was collected, and when it was collected.

Assessing abundance of beluga whales in Bristol Bay using genetic mark-recapture methods, 2002-2011

Abstract

A summary of the key aspects of the dataset that includes when, where, why, and how it was collected, as well as a brief description of its variables and file formats.

This project estimated the abundance of beluga whales within the Bristol Bay stock using genetic mark-recapture methods and combined genetic data with aerial survey data to develop an unbiased correction factor for use in future aerial surveys. The project was started in 2004 by the Alaska Beluga Whale Committee, which funded sample collection from 2004 until 2012 and genotyping from 2004 until 2011, and continued through funding from the North Pacific Research Board (NPRB 1516) from 2015 through 2017. The data for this project were generated using genetic markers from skin biopsies of beluga whales Bristol Bay from 2002 to 2011 using mark-recapture methods.

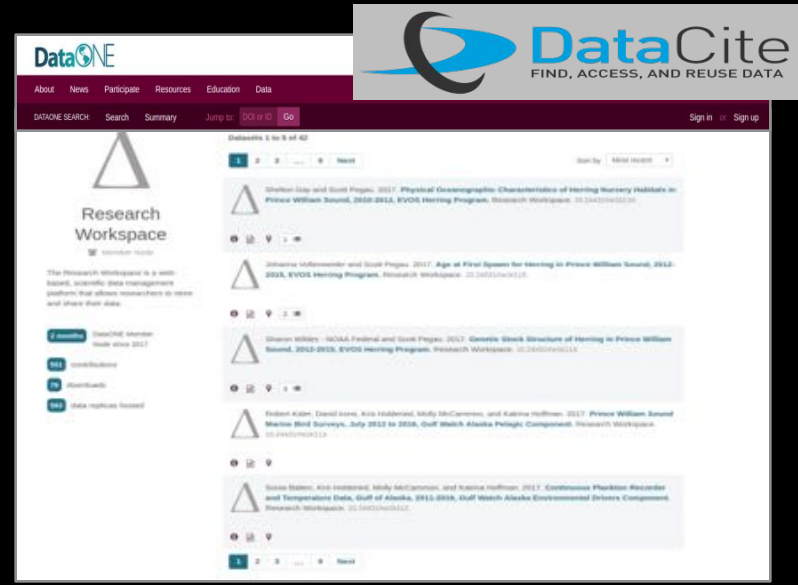
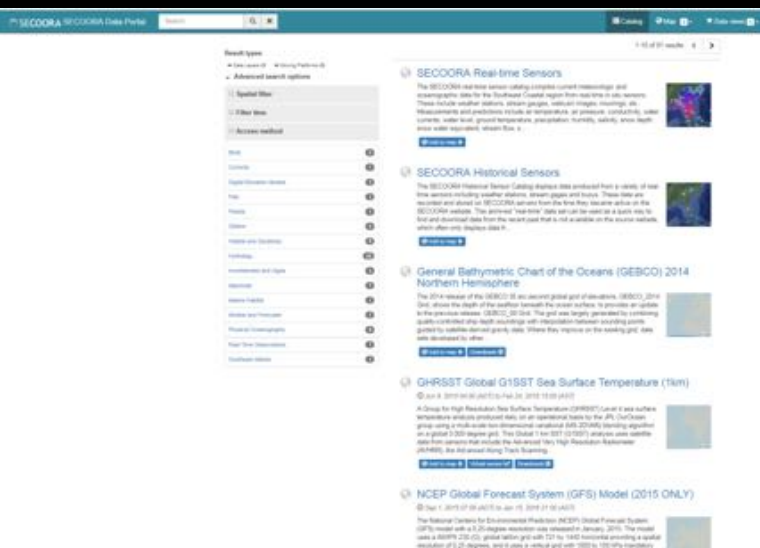
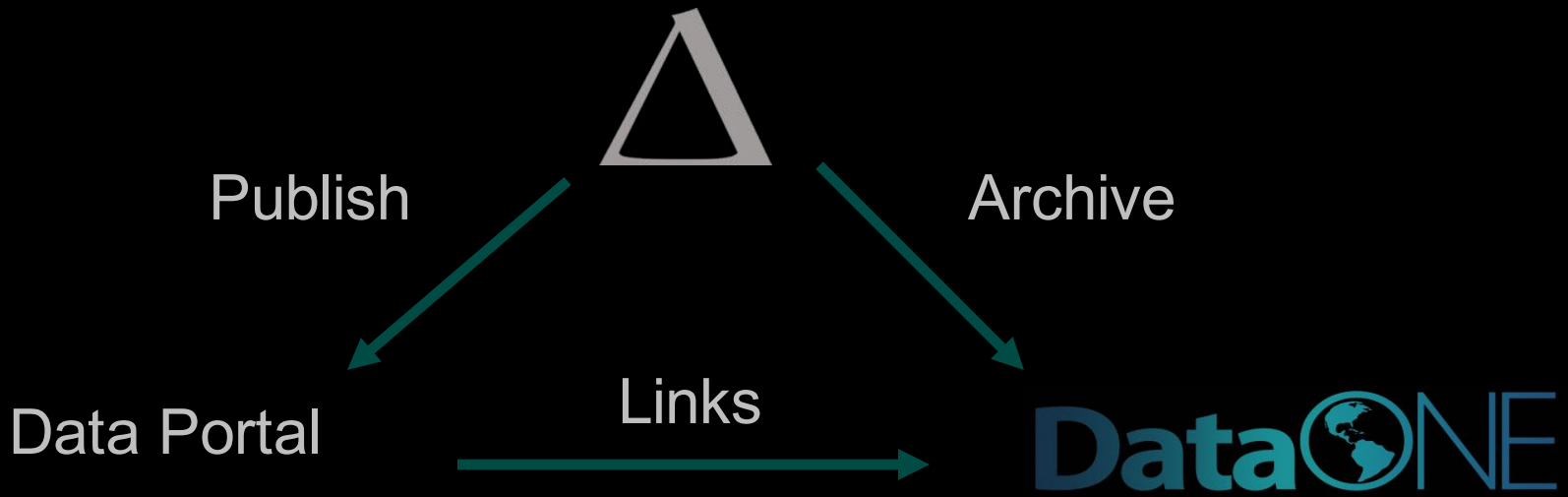
Data from this project consists of 2 .csv data files archived here (NPRB_1516_Bristol_Bay_beluga_whale_abundance_data_sample_list.csv and NPRB_1516_Bristol_Bay_beluga_whale_abundance_data_matching_file.csv).

Purpose

The intention of the dataset and why it was collected or developed, as well as a statement about the dataset's relevance to any larger project or effort.

The Bristol Bay beluga whale stock is genetically distinct from other stocks and tagging studies show it is restricted to Bristol Bay year-round. Quantifying the abundance of belugas in the Bristol Bay stock is important for their management and is critical information for upcoming stock status reviews. This is the first estimate of abundance of belugas in Bristol Bay with appropriate confidence limits.

Research Workspace - Publish & Archive



DataONE

Mission:

“Enable new science and knowledge creation through universal access to data about life on earth and the environment that sustains it.”

How:

Cyberinfrastructure + Community

DataONE - Networked Repositories

Member Nodes



Coordinating Nodes



DataONE - Member Nodes

Member Nodes

- Defined policies
- Persistent IDs
- Immutable content
- Standardized metadata
- Resource maps (bag-it, etc)
- Implement DataONE API

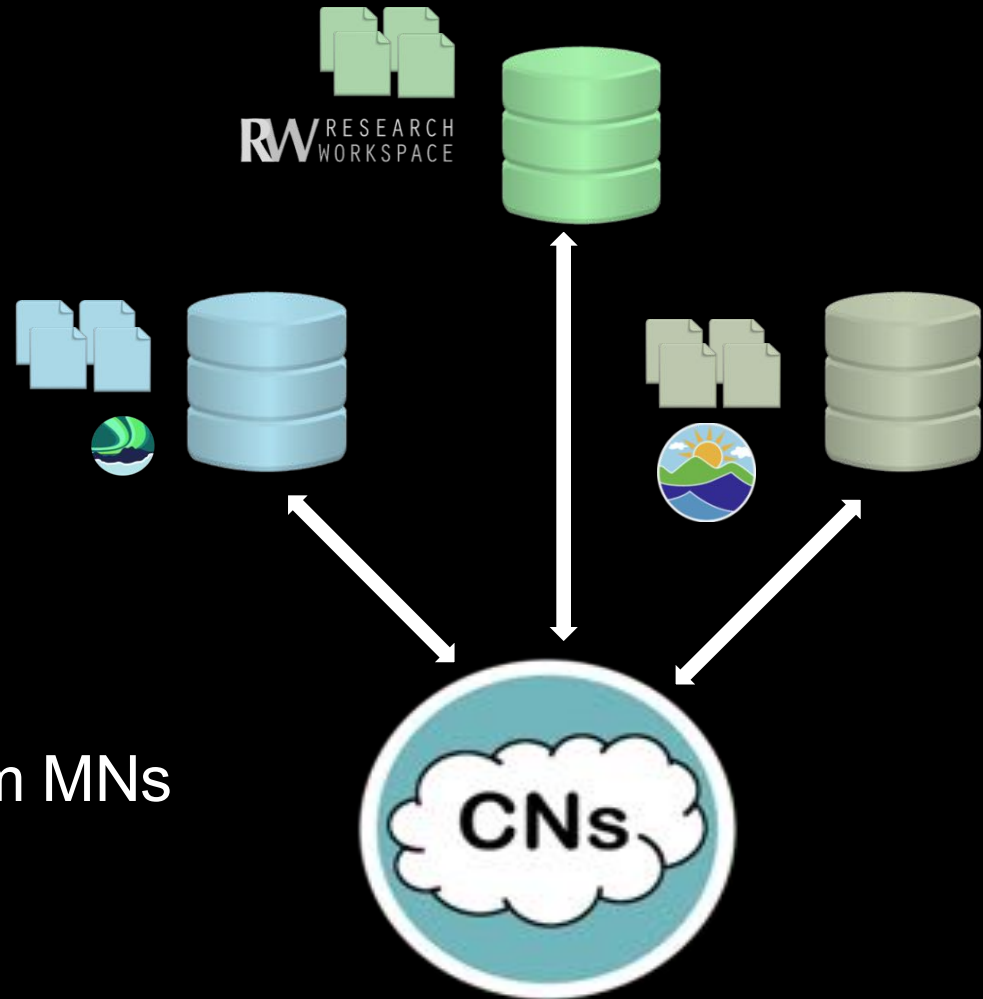


DataONE - Coordinating Nodes

Member Nodes
(Preservation Repo
+ DataONE API)

Coordinating Nodes

- Sync metadata from MNs

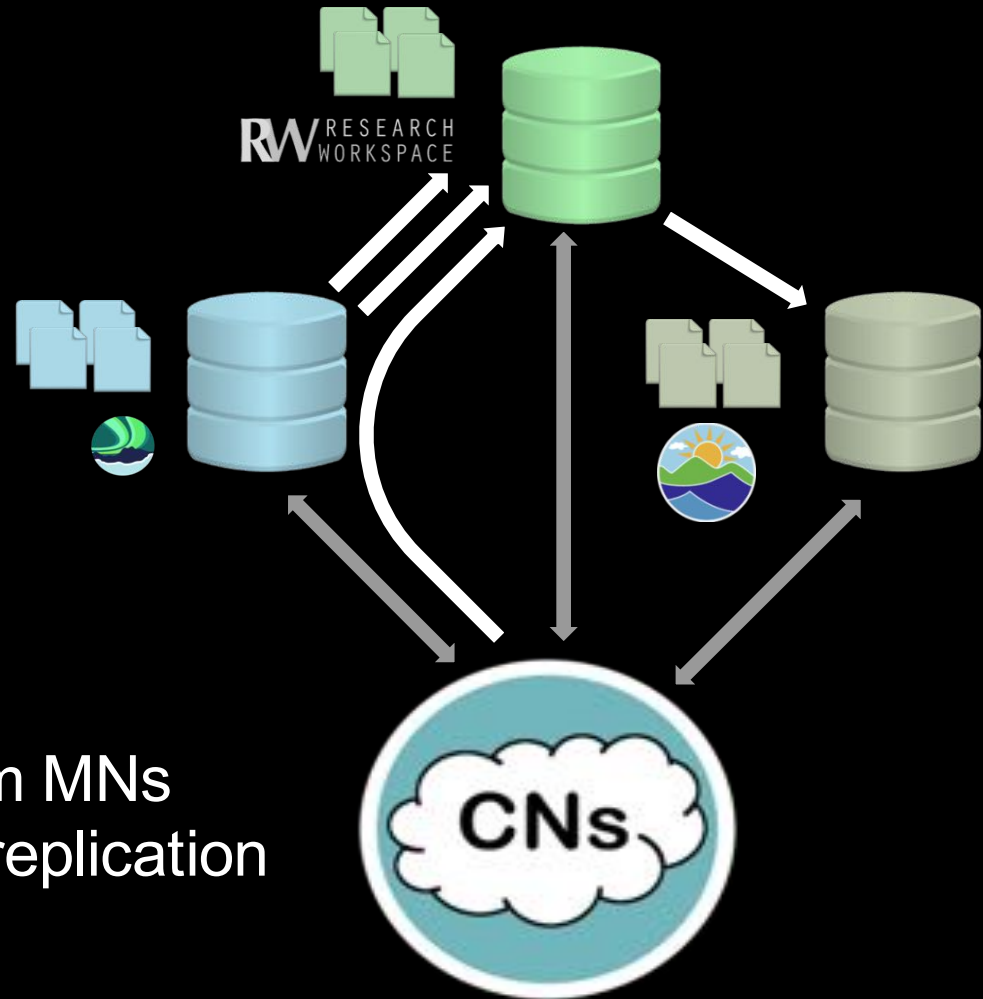


DataONE - Coordinating Nodes

Member Nodes
(Preservation Repo
+ DataONE API)

Coordinating Nodes

- Sync metadata from MNs
- Control MN to MN replication



DataONE - Federated Search

The image shows a screenshot of the DataONE Federated Search interface. The main search area is highlighted with a red box. The search results are displayed in a list format, with filters on the left and a data table on the right.

Search ⓘ

Search phrase

My Search

Creator: Jones

Filter by:

- Data attribute
- Data files
- Member Node
- Creator
- Year
- Identifier
- Taxon
- Location

Name

Year

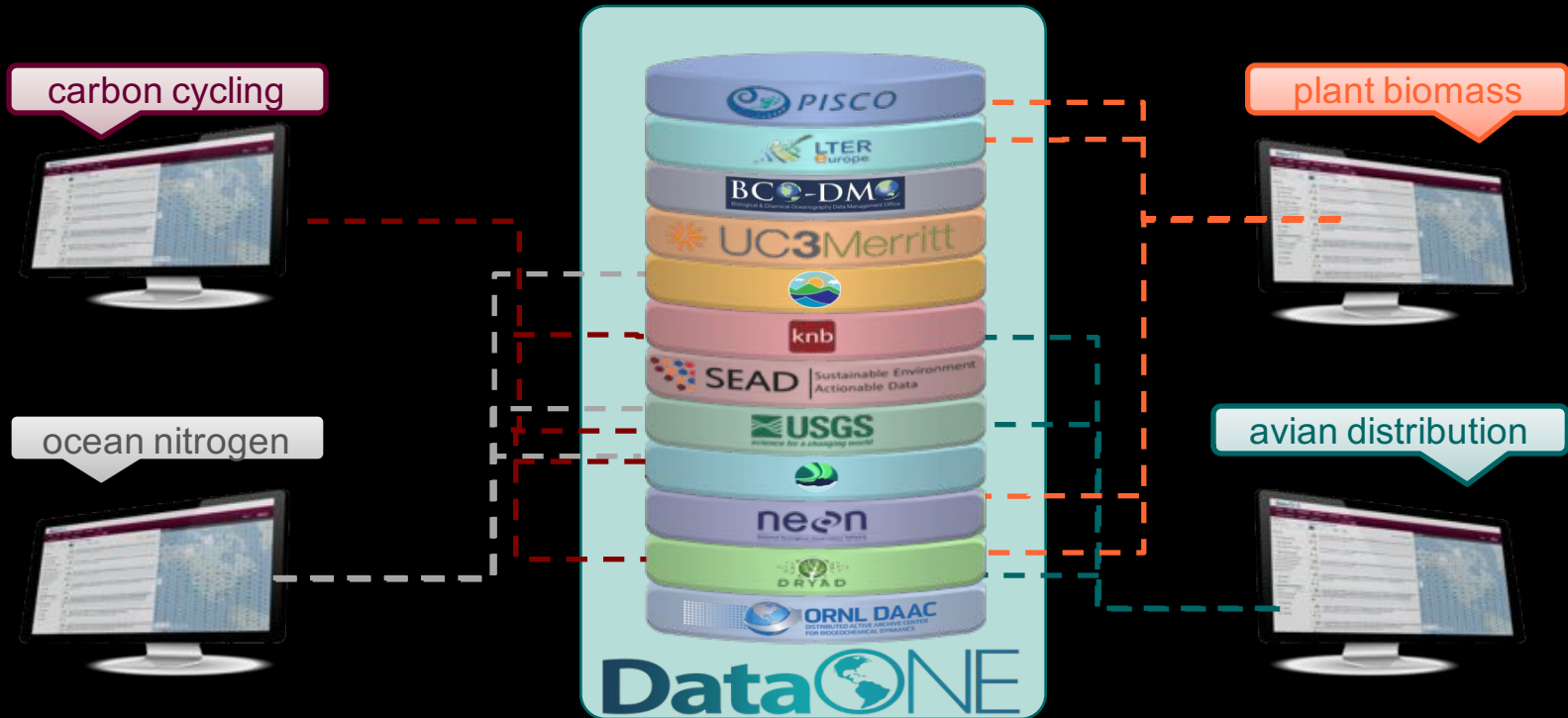
Identifier

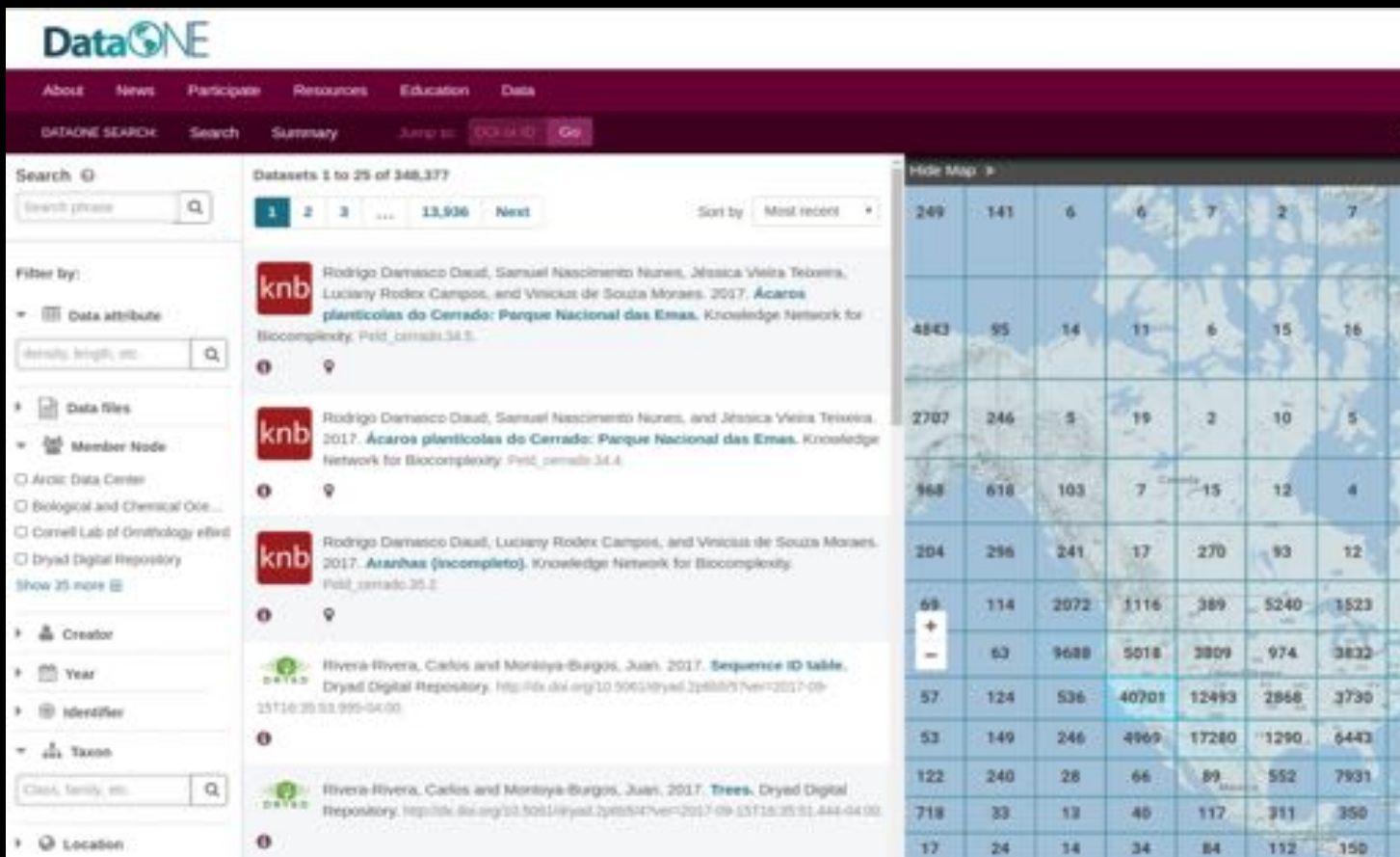
Taxon

Location

	0335	08	14	6	8	9	8	11	21	56
2185	191	6	12	2	10	6	10	56	33	
597	212	76	6	5	10	5	6	7	67	
125	164	99	7	263	91	6	10	29	134	
44	68	1632	389	301	4923	5434	76	54	152	
80	36	9130	4510	3607	396	2875	7798	254	155	
39	75	427	39579	4723	1844	2585	4914	425	135	
34	125	187	3916	6280	784	6167	760	173	87	
64	165	19	52	56	298	2221	642	75	63	
281	19	11	29	66	87	116	271	1230	24	
9	19	12	17	43	66	198	57	85	64	
17	34	17	33	33	33	114	81	429	39	
10	67	21	9	15	14	56	21	265	122	

DataONE - Discoverability



Dataset
ReviewFlag for
ArchiveReserve
DOIUpdate
MetadataPackage
DatasetPush to
RW MNCN Index
and Sync

The screenshot displays the DataONE website interface. The top navigation bar includes links for About, News, Participate, Resources, Education, and Data. Below the navigation bar is a search bar with the text "DATAONE SEARCH" and a "Go" button. The main content area shows search results for datasets, with a list of 1 to 25 of 348,377 datasets. The results are sorted by "Most recent". The first three results are from the Knowledge Network for Biocomplexity (knb) and describe datasets related to "Ácaros plânticos do Cerrado: Parque Nacional das Emas". The fourth result is from the Dryad Digital Repository and describes a "Sequence ID table". The fifth result is also from the Dryad Digital Repository and describes "Trees".

Search results for datasets:

- 1. Rodrigo Damascio Daudi, Samuel Nascimento Nunes, Jéssica Vieira Teixeira, Luciany Rodex Campos, and Vinicius de Souza Moraes. 2017. **Ácaros plânticos do Cerrado: Parque Nacional das Emas**. Knowledge Network for Biocomplexity. [Field_cerrado.34.3](#).
- 2. Rodrigo Damascio Daudi, Samuel Nascimento Nunes, and Jéssica Vieira Teixeira. 2017. **Ácaros plânticos do Cerrado: Parque Nacional das Emas**. Knowledge Network for Biocomplexity. [Field_cerrado.34.4](#).
- 3. Rodrigo Damascio Daudi, Luciany Rodex Campos, and Vinicius de Souza Moraes. 2017. **Aranhas (incomplete)**. Knowledge Network for Biocomplexity. [Field_cerrado.35.1](#).
- 4. Rivera-Rivera, Carlos and Montoya-Burgos, Juan. 2017. **Sequence ID table**. Dryad Digital Repository. <https://doi.org/10.5061/dryad.2p88877ver=2017-09-25T16:25:51.444-04:00>
- 5. Rivera-Rivera, Carlos and Montoya-Burgos, Juan. 2017. **Trees**. Dryad Digital Repository. <https://doi.org/10.5061/dryad.2p88877ver=2017-09-25T16:25:51.444-04:00>

On the right side of the screenshot, there is a "Hide Map" button and a grid of data points overlaid on a map. The grid consists of 7 columns and 7 rows of cells, each containing a numerical value. The values in the grid are:

249	141	6	6	7	2	7
4843	95	14	11	6	15	16
2707	246	5	19	2	10	5
968	618	103	7	15	12	4
204	296	241	17	270	93	12
68	114	2072	1116	389	5240	1523
+	63	9688	5018	3809	974	3832
57	124	536	40701	12493	2868	3730
53	149	246	4965	17280	1290	6443
122	240	28	66	89	552	7931
718	33	13	40	117	311	350
17	24	14	34	84	112	150

- Locate, identify and cite research data



The screenshot displays the Research Workspace interface for a project titled "Environmental Drivers: Oceanographic monitoring in Cook Inlet and Kachemak Bay". The interface includes a search bar, navigation tabs for "Files", "Archives", and "Settings", and a sidebar menu. The main content area shows a list of data archives with their respective descriptions and DOIs. Three red arrows point to the DOI links for each archive entry.

RESEARCH  WORKSPACE Search...

Environmental Drivers: Oceanographic monitoring in Cook Inlet and Kachemak Bay

Files Archives Settings

Archives

Holderied, K. (2017). Oceanographic Monitoring in Cook Inlet and Kachemak Bay, CTD Data, 2012-2016, Gulf Watch Alaska Environmental Drivers Component (Version 1) [Data set]. Axiom Data Science. <https://doi.org/10.24431/rw1k1d>

Annual CTD data files, 2012-2016 10.24431/rw1k1d

Doroff, A. (2017). Oceanographic Monitoring in Cook Inlet and Kachemak Bay, Water Quality, Meteorological, and Nutrient Data collected by the National Estuarine Research Reserve System's System-wide Monitoring Program (NERRS SWMP), 2012-2016, Gulf Watch Alaska Environmental Drivers Component (Version 1) [Data set]. Axiom Data Science. <https://doi.org/10.24431/rw1k1c>

KBNERR met, nutrient, water quality data_Final 2012-2016 10.24431/rw1k1c

Doroff, A. (2017). Oceanographic Monitoring in Cook Inlet and Kachemak Bay, Zooplankton Data, 2012-2015, Gulf Watch Alaska Environmental Drivers Component (Version 1) [Data set]. Axiom Data Science. <https://doi.org/10.24431/rw1k12>

Zooplankton_Final, 2012-2016 10.24431/rw1k12

Research Workspace - Analysis/Synthesis

- Challenges
 - Data Availability
 - Compute Resources
 - Barriers to Entry
- Solution: Jupyter Notebooks
 - Local availability of large datasets
 - TB+ model/satellite data
 - Real time sensor system
 - No need to download data to analyze
 - Powerful compute resources
 - High bandwidth/throughput
 - Powerful CPUs / large RAM
 - Hardware optimized for numerical computation
 - No software management burden!



- Create and share documents that contain code, equations, and visualizations
- Reproducible numerical simulations and statistical modeling
- Access uploaded data stored in the Workspace or data portal



Richness

the number of distinct species found in a sample

$$S = \sum (p_i > 0)$$

% Dominance (Berger-Parker)

the ratio between the number of individuals belonging to the most abundant species and the total number of individuals in the sample

$$\text{Dominance} = \max(p_i)$$

Shannon-Wiener Diversity

index quantifies the uncertainty associated with species prediction

$$H' = - \sum_{i=1}^S p_i \ln(p_i)$$

Pielou's Evenness

species evenness quantifies how close in count each species is within a sampling event

$$J' = \frac{H'}{\ln(S)}$$

```
In [17]: # create dominance and shannon-weaver diversity indices
p_i_stats = aggregated_df.groupby(['location_id', dateField])[['p_i']].agg(
    'sw_diversity': lamb

diversity = aggregated_df.groupby(['location_id', dateField]).agg({
    'species_tsn': np.count_nonzero,
    'lat_station': np.mean, # why are decimals truncated?
    'lon_station': np.mean, # why are decimals truncated?
}).rename(columns={'species_tsn': 'richness'})

diversity = diversity.merge(p_i_stats, left_index=True, right_index=True)

# add Pielou's Evenness Index
diversity['evenness'] = diversity['sw_diversity']/np.log(diversity['richness'])

diversity = diversity.reset_index(level=[dateField, 'location_id'])
diversity
```

Richness

the number of distinct species found in a sample

$$S = \sum (p_i > 0)$$

% Dominance (Berger-Parker)

the ratio between the number of individuals belonging to the most abundant species and the total number of individuals in the sample

$$\text{Dominance} = \max(p_i)$$

Shannon-Wiener Diversity

index quantifies the uncertainty associated with species prediction

$$H' = - \sum_{i=1}^S p_i \ln(p_i)$$

Pielou's Evenness

species evenness quantifies how close in count each species is within a sampling event

$$J' = \frac{H'}{\ln(S)}$$

```
In [17]: # create dominance and shannon-weaver diversity indices
p_i_stats = aggregated_df.groupby(['location_id', dateField])['p_i'].agg(
    'sw_diversity': lamb

diversity = aggregated_df.groupby(['location_id', dateField]).agg({
    'species_tsn': np.count_nonzero,
    'lat_station': np.mean, # why are decimals truncated?
    'lon_station': np.mean, # why are decimals truncated?
}).rename(columns={'species_tsn': 'richness'})

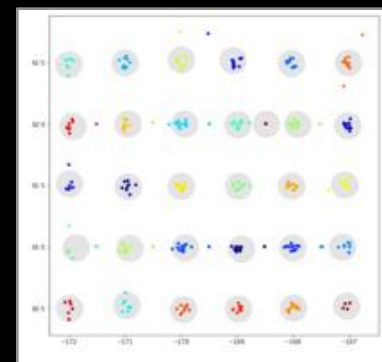
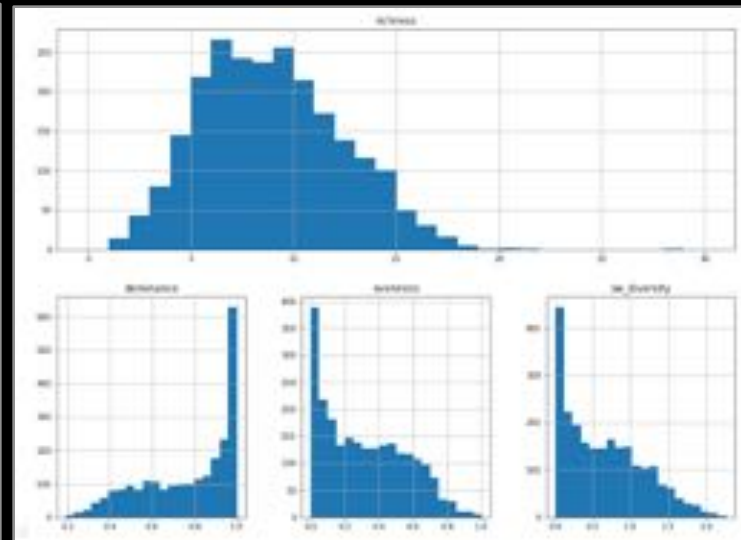
diversity = diversity.merge(p_i_stats, left_index=True, right_index=True)

# add Pielou's Evenness Index
diversity['evenness'] = diversity['sw_diversity']/np.log(diversity['richness'])

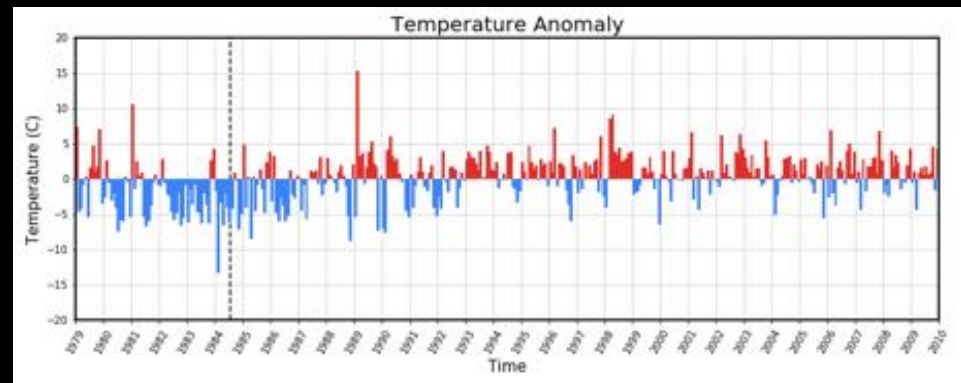
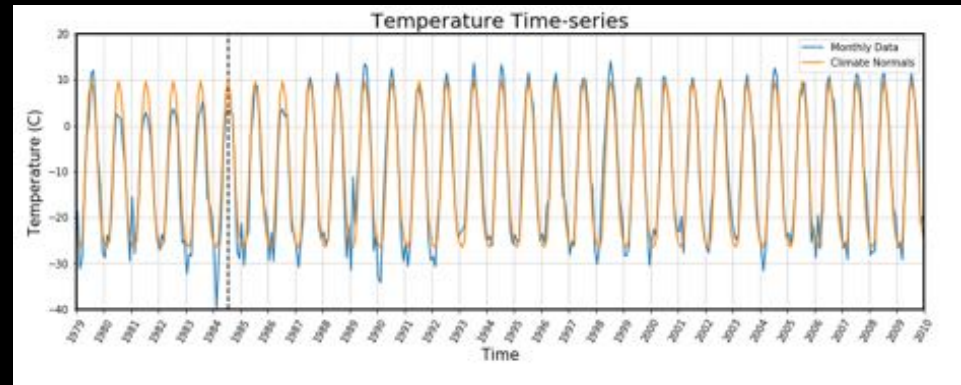
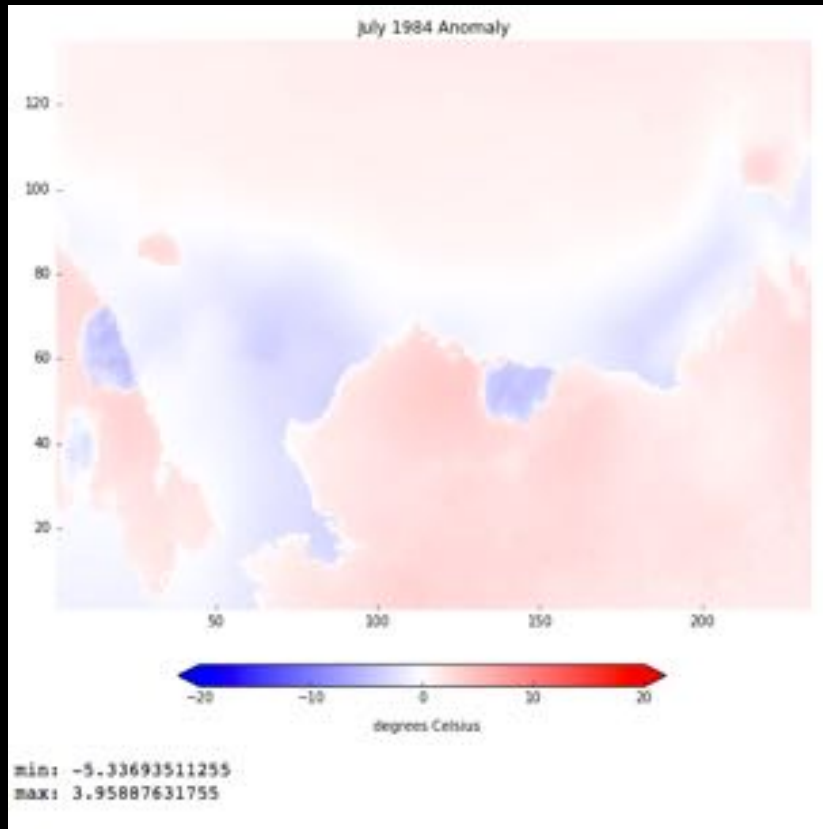
diversity = diversity.reset_index(level=[dateField, 'location_id'])
diversity
```

```
Out[17]:
```

	location_id	start_date	lat_station	richness	lon_station	sw_diversity	dominance	evenness
0	1	2002-09-19 18:30:00	51.295745	4	-178.344080	1.036659	0.587719	0.74
1	4	2002-09-12 20:08:00	51.718330	4	-179.724165	0.352634	0.920635	0.25
2	8	2002-09-20 01:07:45	51.837745	3	-177.017800	0.880010	0.741005	0.45



Research Workspace - Time-series Anomalies



- Calculate climate normals on a 31-year long, multi-terabyte dataset
- Then plot temperature anomalies over a region



Thank you! Questions?