MBON Data Portal (v2.0)

Stacey Buckelew, Axiom Data Science
Really “Big Data” – Data volume is increasing exponentially

By the year 2020, the digital universe will reach 44 zettabytes – that’s a 10-fold increase from 2013.

Global scientific output doubles every nine years

07 May 2014, 16:46 BST | Posted by Richard Van Noorden | Category: Policy, Publishing

It’s a common complaint among academics: today’s researchers are publishing too much, too fast. But how fast is the mass of scientific output actually growing?
DATA MANAGEMENT CHANGES

• Data explosion is reforming data management and governance
• Leveraging existing data systems in different ways - architecture, storage, databases
• Emphasis on open data as crucial to reliable science
• Integration, visualization, and dissemination are tools, not by-products
Ocean Data is Big Data

- Shared cyberinfrastructure approach
  - Leverage applications, systems & hardware across partners

- Community developed software, standards and protocols
  - Standardize systems and interfaces across partners

- Scalable compute and storage infrastructure (HPC)
  - 5 petabytes storage; 3,000 processing cores
Customizable data environment allowing scientists, managers, and the public others to discover, interact, and access biologically-relevant data from different sources.
REUSE & TRANSFORMATION
Jupyter Notebook & data analyses
REUSE & TRANSFORMATION
Jupyter Notebook & data analyses
DATA ACCESS & DISCOVERY
Data portals & search catalogs

REUSE & TRANSFORMATION
Jupyter Notebook & data analyses
Map
Integrate & visualize data from many sources

Catalog
Search, metadata, & data download

Data Views
Rapidly assimilate & compare different data streams
MBON Data Pipeline

Source
- Geospatial Database
- CSV
- Tabular Data
- Web Service (ERDDAP, WMS, etc.)
- Research Workspace
- Other

Service
- GeoServer
- Custom Service

Storage
- PostgreSQL

User
- MBON Portal
- Notebook

Documentation
- Metadata
- Catalog

Data Cleansing

△ = Research Workspace Pipeline
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.

Research Workspace

Axiom Data Science
Custom Data Views (‘dashboards’)

BASIS (Bering Arctic Subarctic Integrated Survey): Fish Species Richness and NSIDC Sea Ice Concentration, Bering Sea (2002 - 2015)

Saved charts

1. Richness (gamma)
   Bering Arctic Subarctic Integrated Survey (BASIS) Fish Catch Surveys; Biodiversity Indexing
   Polygon selection

2. Sea Ice Concentration
   NSIDC Sea Ice Concentration
   Polygon selection
Custom Data Views (‘dashboards’)

Manatee Mortality and Sea Surface Temperature (SST), Florida (1974-2013)

Saved charts

Comparison chart

Sea Surface Temperature

Water Temperature (°F)

Manatee Mortality (count)

Axiom DATA SCIENCE
MBON Data Catalog

SECOORA Fisheries and Wildlife Monitoring

US IOOS has developed and implemented data content and access standards for IOOS biological core variables (Fish and Zooplankton species and abundance) to promote interoperability. Following the implementation of these data services in the Pacific Islands Fisheries Science Center through PacIOOS, IOOS extended this project to SECOORA and GCOOS-RA. The project objectives were 1) to extend the IOOS Data Management and Communication (DMC) biolog.

California Cooperative Oceanic Fisheries Investigations (CalCOFI)

The California Cooperative Oceanic Fisheries Investigations (CalCOFI) program is a joint research effort by the U.S. National Oceanic and Atmospheric Administration Southwest Fisheries Science Center, University of California Scripps Institution of Oceanography, and California Department of Fish and Wildlife. Joint fisheries and oceanography surveys have been conducted regularly by the CalCOFI program since 1951, and are currently conducted qu...

MBON Portal (http://mbon.ioos.us)
Data Catalog

Content Pages
- Summary
- Data availability
- License / usage
- Variable / layer preview
- Contact information
- Original-source link
Downloads Using Interoperability

-.ncWMS
-Shapefile
-CSV
-THREDDS
-netCDF
-OPenDAP
-ERDDAP
Downloads Using Interoperability

### Richness (gamma): Binned years

<table>
<thead>
<tr>
<th>Start date</th>
<th>End date</th>
<th>Standard deviation</th>
<th>Mean</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
<th>25th Percentile</th>
<th>75th Percentile</th>
<th>Records (count)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002-01-01T00:00:00Z</td>
<td>2002-01-01T00:00:00Z</td>
<td>2.1722</td>
<td>7.16</td>
<td>7</td>
<td>13</td>
<td>0</td>
<td>9</td>
<td>9</td>
<td>80</td>
</tr>
<tr>
<td>2003-01-01T00:00:00Z</td>
<td>2004-01-01T00:00:00Z</td>
<td>1.9140</td>
<td>8.07</td>
<td>4</td>
<td>12</td>
<td>7</td>
<td>16</td>
<td>0</td>
<td>72</td>
</tr>
<tr>
<td>2004-01-01T00:00:00Z</td>
<td>2004-12-31T00:00:00Z</td>
<td>2.1998</td>
<td>7.33</td>
<td>2</td>
<td>11</td>
<td>6</td>
<td>9</td>
<td>9</td>
<td>59</td>
</tr>
<tr>
<td>2005-01-01T00:00:00Z</td>
<td>2006-01-01T00:00:00Z</td>
<td>1.8770</td>
<td>6.83</td>
<td>8</td>
<td>12</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>59</td>
</tr>
<tr>
<td>2006-01-01T00:00:00Z</td>
<td>2007-01-01T00:00:00Z</td>
<td>2.1352</td>
<td>8.21</td>
<td>8</td>
<td>12</td>
<td>5</td>
<td>8</td>
<td>12</td>
<td>71</td>
</tr>
<tr>
<td>2007-01-01T00:00:00Z</td>
<td>2008-01-01T00:00:00Z</td>
<td>2.1691</td>
<td>8.03</td>
<td>1</td>
<td>12</td>
<td>5.75</td>
<td>8</td>
<td>8</td>
<td>72</td>
</tr>
<tr>
<td>2008-01-01T00:00:00Z</td>
<td>2009-01-01T00:00:00Z</td>
<td>2.2287</td>
<td>5.29</td>
<td>2</td>
<td>11</td>
<td>3.25</td>
<td>6.75</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>2009-01-01T00:00:00Z</td>
<td>2010-01-01T00:00:00Z</td>
<td>2.3254</td>
<td>4.75</td>
<td>1</td>
<td>10</td>
<td>3</td>
<td>6</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>2010-01-01T00:00:00Z</td>
<td>2010-01-01T00:00:00Z</td>
<td>2.5559</td>
<td>4.76</td>
<td>4.5</td>
<td>12</td>
<td>3</td>
<td>6</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>2011-01-01T00:00:00Z</td>
<td>2012-01-01T00:00:00Z</td>
<td>1.9375</td>
<td>4.39</td>
<td>1</td>
<td>9</td>
<td>3</td>
<td>6</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>2012-01-01T00:00:00Z</td>
<td>2012-12-31T00:00:00Z</td>
<td>2.4812</td>
<td>6.44</td>
<td>6</td>
<td>2</td>
<td>14</td>
<td>5</td>
<td>8</td>
<td>77</td>
</tr>
<tr>
<td>2013-01-01T00:00:00Z</td>
<td>2015-01-01T00:00:00Z</td>
<td>3.1619</td>
<td>8.64</td>
<td>8</td>
<td>3</td>
<td>16</td>
<td>6</td>
<td>11</td>
<td>81</td>
</tr>
</tbody>
</table>

Showing 1 to 13 of 13 entries
Data Lifecycle

- **Data Creation & Quality Control**: Scientists or Ingestion
- **Data Storage**: Workspace
- **Data Description**: Metadata Editor
- **Data Access & Discovery**: Data portals & search catalogs
- **Reuse & Transformation**: Jupyter Notebook & data analyses
- **Archive & Preservation**: Repository submission pathway

[Diagram showing the lifecycle of data with various steps and tools related to data creation, storage, discovery, and reuse.]
~web-based platform for collaboratively managing science projects through the entire data lifecycle~

Share
Analyze
Preserve
Organize into projects, research campaigns and organizations

Coordinate data exchange across networks, groups, programs

ISO 19110/19115-2 standards metadata editor

Execute server side R and Python numeric workflows (Jupyter) on uploaded data AND any data in Axiom CI stack

Archive pathway to DataONE & Datacite DOI minting
Research Workspace: Metadata

- Integrated ISO 19110/19115-2 standards metadata editor
- Author metadata alongside data
- Implement labor-saving tools

Folder metadata

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

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 the format.

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 in 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, 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.
Locate, identify and cite research data

Archives


3. KBNERR met, nutrient, water quality data_Final 2012-2016


5. Zooplankton_Final, 2012-2016
Create and share documents that contain code, equations, and visualizations

Reproducible numerical simulations and statistical modeling

Access uploaded data stored in the Workspace
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

\[ Dominance = \max(p_i) \]

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]:
<table>
<thead>
<tr>
<th>location_id</th>
<th>start_date</th>
<th>lat_station</th>
<th>richn...4</th>
<th>lon_station</th>
<th>sw_diversity</th>
<th>dominance</th>
<th>evenness</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2002-09-10</td>
<td>51.295745</td>
<td>4</td>
<td>-179.344060</td>
<td>1.036659</td>
<td>0.587719</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>1830000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2002-09-11</td>
<td>51.718330</td>
<td>4</td>
<td>-179.724165</td>
<td>0.350034</td>
<td>0.920635</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>2039000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2002-09-12</td>
<td>51.327575</td>
<td>5</td>
<td>-179.24533</td>
<td>0.833616</td>
<td>0.781035</td>
<td>0.46</td>
</tr>
</tbody>
</table>
Live Demo

http://mbon.ioos.us/