GeoResJ 6 (2015) 195–201

Contents lists available at ScienceDirect

GeoResJ

journal homepage: www.elsevier.com/locate/GRJ

Bringing dark data into the light: A case study of the recovery of Northwestern Atlantic zooplankton data collected in the 1970s and 1980s

Peter H. Wiebe *, M. Dickson Allison

Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA

ARTICLE INFO

Article history:
Received 10 October 2014
Revised 4 March 2015
Accepted 6 March 2015
Available online 6 April 2015

Keywords:
Data rescue
Zooplankton biomass
Zooplankton species abundance
Dark data
North Atlantic Gulf Stream Rings

ABSTRACT

Data generated as a result of publicly funded research in the USA and other countries are now required to be available in public data repositories. However, many scientific data over the past 50+ years were collected at a time when the technology for curation, storage, and dissemination were primitive or non-existent and consequently many of these datasets are *not* available publicly. These so-called "dark data" sets are essential to the understanding of how the ocean has changed chemically and biologically in response to the documented shifts in temperature and salinity (aka climate change). An effort is underway to bring into the light, dark data about zooplankton collected in the 1970s and 1980s as part of the cold-core and warm-core rings multidisciplinary programs and other related projects. Zooplankton biomass and euphausiid species abundance from 306 tows and related environmental data including many depth specific tows taken on 34 research cruises in the Northwest Atlantic are online and accessible from the Biological and Chemical Oceanography Data Management Office (BCO-DMO).

1. Introduction

Recent changes in National Science Foundation (NSF) and other agency data policies (NSF11060 [1]; Office of Science & Technology Policy (OSTP) memo 2013 [2]) mandating timely and open access to data and information generated in the course of US funded research have resulted in a relatively rapid change in the culture of data sharing. Technological advances, policy changes, and increased awareness of the need for and benefits of well-curated data make it much more likely that recently generated research results will be made publicly available and in a timely manner (Wallis et al. [3]; Hook et al. [4]). However, many scientific data were generated at a time when the technology for curation, storage, and dissemination were primitive or non-existent, and data sharing was not viewed as essential. In addition, many of the datasets were collected and stored by individuals as small projects that make up the "long tail" of the science enterprise (Heidorn [5]). These smaller projects, in contrast to large projects that involve many investigators, form the bulk of the projects funded by agencies such as NSF. Data from these projects, large and small, have in the past been poorly curated and thus less visible to other scientists, largely not publicly available online, and hence named "Dark Data" (Heidorn [5]). But as Sinha et al. [6] emphasize, without access to the types of historical observations or legacy data that make up the "dark data" in the "long tail" of science, emerging scientific challenges will not be addressable. "...making these data available on demand must be one of the highest priorities for any enterprise seeking to develop a cyberinfrastructure capable of promoting new ways to examine the earth system through time" (Sinha et al. [6]). One international project designed to rescue historical oceanographic data was the IOC/IODE GODAR project, which focused mainly on physical data (Conkright et al. [7]; Caldwell [8]). More recently, the paucity of marine ecosystem data available to conduct cutting edge research and the critical need for the rescue of past data were also highlighted in a recent EarthCube End-User Domain Workshop Report "Articulating Cyberinfrastructure Needs of the Ocean Ecosystem Dynamics Community" (Kinkade et al. [9]) and by Banse [10].

There are significant dark datasets currently unavailable from multidisciplinary programs funded in the 1970s and 1980s such as those that studied the Northwest Atlantic cold-core and warm-core rings (The Ring Group [11]; Joyce and Wiebe [12]).

The Cold-Core Rings (CCR) studies took place between 1972 and 1977, and the Warm-Core Rings (WCR) Program occurred in 1981 and 1982. Large oceanic eddies or rings form when Gulf Stream

* Corresponding author. Tel.: +1 508 289 2313; fax: +1 508 457 2169.
E-mail address: pwiebe@whoi.edu (P.H. Wiebe).

http://dx.doi.org/10.1016/j.grj.2015.03.001
2214-2428/

Table 1
Metadata being sought in the zooplankton data rescue effort. Modified from Anon [38], Annex 3).

Metadata type	Metadata sub-category descriptions
Cruise metadata	• Name of the ship • Investigator-designated Cruise Identifier • Associated Project • Associated Institute • Principal investigator(s) for cruise • Other responsible investigators, and their variable(s) • Cruise or data report
Station metadata	• Station latitude and longitude • Station Month, Day, Year • Station Time (designated as "local", "GMT/UTC", "ship", etc) • Investigator-designated Station Identifier • Meteorological Observations (atmospheric conditions, sea state) • Station Sounding (bottom depth) • Information about any other supplementary/complementary data collected at the same time (same station)
Sampling gear metadata	• Describe the sampling gear used, providing a literature reference if available • If using a "standard" net (e.g., a NORPAC net) was used, be sure to note any modifications to this net • What net mesh size was used (usually in microns) • What was the net opening shape (square or circular) and the opening mouth area or diameter • Was a flowmeter used? When and how was it calibrated?
Net tow metadata	• Towing Method (horizontal, vertical, oblique) • Towing depth-range (a range of starting and ending depths for each net or bottle), or the wire angle and wire out during the tow • Towing Duration (minutes or hours) • Towing Distance (in meters) • Average Towing Speed (knots or meters per second)
Sample processing metadata	• What volume of water was filtered to yield this sample • How were samples preserved, and in what (e.g., 5% buffered formalin) • How were samples processed (summarize the counting, weight, or volume method)? • Was the sampled split (via Folsom splitter or other method)? What was the size of the final aliquot? • Were large plankton removed prior to making biomass measurements? Was a size or volume criteria used in deciding what to remove and what could remain? • Investigator-designated tow, net, or sample identifier
Sample metadata	• Provide the units for each measurement (e.g., #/liter, #/m^3, #/m^2, mg/m^3, mg/haul, ...) • If taxonomic codes, symbols, or abbreviations are used in the data, provide a translation table to help reduce possible misunderstandings of the taxa (e.g., "CfcV" = "Calanus finmarchius copepodite V", ...) • Is an estimate of final uncertainty of the data known?

waters first meander, then separate, forming a ring of Gulf Stream water around a core of cold Slope Water or a core of warm Sargasso Sea water. The CCRs move south or southwest from their point of origin into the Sargasso Sea and are initially 150–300 km in diameter and 2500–3500 m deep. They can persist as identifiable features for up to 2 years. WCRs move to the west/southwest in the Slope Water north of the Gulf Stream. They are 100–200 km in diameter, extend to at least 1500 m deep, and exist for a shorter period of time (usually less than a year) before gradually breaking up and rejoining the Gulf Stream. Both of these kinds of rings form about 5 to 8 times a year.

Rings are particularly interesting to the biologist because species living north and south of the Gulf Stream are distinctly different (Wiebe et al. [13]; Wiebe et al. [14]). Arctic boreal and temperate species from the Slope Water or tropical–subtropical species from the Sargasso Sea are isolated during ring formation within their particular ring structure. Thus, a community of animals from one area is expatriated in the territory of another community of animals. As a ring decays, the water gradually takes on the physical and chemical characteristics of the surrounding non-ring water. Species outside the ring invade the ring habitat while those expatriated go to local extinction (Wiebe and Flierl [15]). This phenomenon provides for a large-scale natural ecological experiment that was the focus of the rings studies.

Data collected during the 1970s in the CCR program were managed by each individual PI separately. For processing and plotting, the data were put onto punch cards and processed by main frame computers such as the Honeywell Sigma 7. Collaborators would meet face to face to discuss the scientific results and share data in the form of written data reports. In the 1980s, the WCR program had a program service office and began to provide some data management services. Most investigators were using microcomputers (manufactured by Commodore, Apple, IBM, and others) and some data were stored on floppy disks. Collaborations between the investigators were conducted at week-long workshops (Wiebe [16]). Some, but not all of the investigators' data and information were stored on a Digital Equipment Corporation minicomputer (VAX 11/780), but when that computer was phased out ~1995, the data were stored on 9-track tapes and they subsequently disappeared. Some of the WCR zooplankton data were summarized in a technical report (Barber and Wiebe [17]). The CTD physical data from many of the cruises were submitted to NODC, but locating the data from these programs is quite difficult without an in-depth knowledge of the program's deployments, etc.

The objective of this paper is to describe the efforts to recover the zooplankton biomass and euphausiid species counts and related environmental data from 34 cruises to the Northwest Atlantic Ocean that were locked in notebooks and old digital file formats, and deposit them into a modern publically available data repository (e.g. the Biological and Chemical Oceanography Data Management Office – BCO-DMO).

1.1. BCO-DMO repository

The Biological and Chemical Oceanography Data Management Office (BCO-DMO) was created and funded by the National Science Foundation (NSF) in 2006 to serve investigators funded by the NSF Biological and Chemical Oceanography Sections to support the scientific research community through improved access to marine biogeochemical and ecological data and information (Anonymous, 2013 [18]). BCO-DMO provides research scientists and others with the systems necessary to work with data from heterogeneous sources with increased efficacy. The BCO-DMO data management system is composed of a metadata database, the distributed client–server JGOFS/GLOBEC data system (Flierl et al. [19]; Glover [20]; Wiebe et al. [21]), and a Web browser with text-based and map-based user interfaces accessing the information and data available from the repository. The metadata database is implemented using the Drupal content management system. These metadata provide the means to discover, access, and reuse data managed by BCO-DMO. The JGOFS/GLOBEC data system provides the means to manage and retrieve the actual data, and any standard Web browser can access the metadata and data. BCO-DMO is a repository for managing data on short- and medium-term time scales; data are routinely submitted to the appropriate national archive.

BCO-DMO uses established controlled vocabularies and ontologies that enable data interoperability, advanced search and discovery (Leadbetter et al. [22]), and the linking of existing data