GeoResJ 6 (2015) 17–20

Contents lists available at ScienceDirect

GeoResJ

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

Rescued from the deep: Publishing scientific ocean drilling long tail data

Jamus Collier [b,*], Stefanie Schumacher [a], Cornelia Behrens [b], Amelie Driemel [a], Michael Diepenbroek [b], Hannes Grobe [a], Taewoon Kim [c], Uwe Schindler [b], Rainer Sieger [a], Hans-Joachim Wallrabe-Adams [b]

[a] Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research (AWI), Germany
[b] University of Bremen Center for Marine Environmental Sciences (MARUM), Germany
[c] Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Japan

A R T I C L E   I N F O

Article history:
Received 10 October 2014
Revised 25 January 2015
Accepted 27 January 2015
Available online 14 March 2015

Keywords:
Data rescue
Scientific ocean drilling
Long tail data
IODP/ODP/DSDP
PANGAEA
Open Access data publishing

A B S T R A C T

Scientific ocean drilling began in 1968 and ever since has been generating huge amounts of data, including that from shipboard analysis of cores, in situ borehole measurements, long-term subseafloor hydrogeological observatories, and post-expedition research done on core samples and data at laboratories around the world (Smith et al., 2010). Much of the data collected aboard the drilling vessels are captured in a number of program databases (e.g., Janus), and eventually archived in a long-term data repository. However, data resulting from researchers' analyses on core samples in the post-cruise period are generally confined to journal articles and scholarly literature or, particularly for raw or processed data sets, to the hard drives of those researchers. Thus, knowledge of and access to long tail research data that constitutes a significant portion of the overall output of scientific ocean drilling is at risk of remaining lost to the multidisciplinary Earth sciences community.

In order to address the issue of long tail data from scientific ocean drilling, the Integrated Ocean Drilling Program Management International (IODP-MI) partnered with the data publisher PANGAEA hosted by the Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research (AWI) and the University of Bremen Center for Marine Environmental Sciences (MARUM) to conduct a post-cruise data rescue project. This collaboration began in 2007 and continued until 2013. This report summarizes the goals, methods, results, and lessons learned from the IODP Post-Cruise Data Capture Project.

1. Introduction

Scientific ocean drilling represents one of the longest running and most successful international scientific collaborations in the Earth sciences [1]. It can be said to have begun in 1961 with Project Mohole, a mission documented for LIFE Magazine by onboard observer, author Steinbeck [2]. Although those pioneering researchers succeeded in drilling more than 183 m into the Earth's crust below approximately 3500 m of water, Project Mohole did not achieve many of its ambitious goals and was eventually discontinued [3]. Nonetheless, Project Mohole contributed significantly to technological advances that enabled future scientific ocean drilling research, including the development of a dynamic positioning system that was the forerunner of the modern dynamic positioning systems used throughout scientific and industrial ocean drilling today.

Scientific ocean drilling did not end with the termination of Project Mohole. In 1966, the U.S. National Science Foundation (NSF) established the Deep Sea Drilling Project (DSDP) to build upon the promise of Project Mohole science and technology. DSDP met with nearly immediate success, both in terms of scientific and technological achievement and also in instituting reporting formats for shipboard data [4]. In 1975, DSDP expanded to include international partners, making it a pioneer in international scientific collaboration in the Earth sciences. The DSDP continued operation until 1983, conducted 96 expeditions, when it was transitioned to the international Ocean Drilling Program (ODP).

The ODP began drilling in 1985 and continued to advance deep sea drilling science and technology. Between 1985 and 2003, ODP conducted 110 expeditions, drilled more than 1797 holes and recovered more than 222,400 m of core material, and produced extensive volumes of downhole logging data and shipboard analytical data [5]. The ODP established new policies for samples, data, and publications, and attempted to collect each expedition's scientific literature into comprehensive volumes. In 2003, the ODP was transformed into the fully international Integrated Ocean Drilling Program (IODP). The IODP operated from 2003 until 2013, conducted 48 expeditions, drilled more than 500 holes, and recovered

* Corresponding author.

http://dx.doi.org/10.1016/j.grj.2015.01.003
2214-2428/© 2015 The Authors. Published by Elsevier Ltd.

more than 70,000 m of core. In 2013, it was de-integrated to form the International Ocean Discovery Program which will operate until 2018 or beyond.

Considering the nearly 50 years of international scientific collaboration in scientific ocean drilling, amounting to hundreds of expeditions, hundreds of thousands of meters of recovered core material, millions of samples provided to researchers, more than 30,000 DSDP/ODP/IODP-related publications, and a truly enormous amount of shipboard and post-cruise data, it is no surprise that scientific ocean drilling has, and will continue to create, a long tail of scientific data. An unknown, but surely large quantity of this data exists exclusively in printed form as tables and charts in peer-reviewed and the scholarly literature, not effectively identified or indexed as data, thus presenting a massive problem of undiscoverable dark data [6]. For scientists working in the field of marine geology, it is nearly impossible to get an overview about the availability of research data from scientific ocean drilling. Therefore, a project to rescue these data from obsolescence was deemed to be worthwhile given recent advances in the scientific community's understanding of the value of long tail data in terms of economic efficiency, scientific opportunities, and policy relevance [7–9].

From the inception of the DSDP, sample and data management was a fundamental component of scientific planning, shipboard operations, and the post-cruise moratorium period [10]. Policies were implemented to encourage researchers to inform the program operators of their peer-reviewed publications related to expedition samples and data. While not all shipboard data was digitized or systematically archived (e.g., hand-drawn paper barrel sheets), vast quantities of shipboard data from core analysis, downhole logging, and other measurements were collected and eventually made freely available for distribution in electronic formats by IODP partners at Texas A&M University and Columbia University Lamont-Doherty Earth Observatory.

As the program transitioned into an international program operating three drilling vessels, new data management systems were developed to keep pace with advances in information technology and the needs of the different drilling vessels and their operators. Since 2003, each of the three IODP drillship operators, known as implementing organizations, maintains its own data management systems for shipboard data, as well as for core and sample material. Currently, there are no integrated data management systems in place for submission, storage, and archiving of post-cruise analytical data related to program samples or data.

The majority of publications (with related primary data) are produced after a 1 year moratorium and many are published years after an expedition. Those post-cruise data are distributed in various journals related to marine geosciences from major publishers and in smaller journals. Insufficient data management strategies, practices and systems in place for data from post-expedition experiments and analyses, including a lack of metadata and information on discoverability of such data, hinders the re-use of what has been arguably some of the most scientifically valuable data generated by the scientific ocean drilling programs [11]. In order to address these issues of dark long tail data from scientific ocean drilling, IODP-MI and PANGAEA developed a framework comprised of workflows, components, and interfaces to discover, curate, and archive legacy scientific ocean drilling data into an integrated portal for discovery of and access to shipboard and post-expedition data.

2. Implementation

The approach of the IODP Post-Cruise Data Capture Project was organized into iterative tasks to research and identify legacy data from the scientific literature that are not included in the data systems of the implementing organizations or their related repository archives, capturing and, where necessary, digitizing those data, generating ISO 19115 standard metadata [12] for each data supplement, performing QA/QC processes on captured data to ensure accuracy in capture and digitization, and publishing these data and metadata with appropriate citations in the Open Access data archive PANGAEA [13] and through the IODP Scientific Earth Drilling Information Service (SEDIS) [14] data portal. The PANGAEA Data Publisher for Earth and Environmental Sciences provided the technical infrastructure, as well as the data curators and workflows, for identifying, extracting, digitizing, and archiving the captured data sets. The decision was made to begin systematically searching geosciences journals starting with those published in 1969 and proceeding chronologically due to the fact that data from later publications were more likely to be available in digitized formats online. This process was started with highly ranked journals, continued through successive stages of lower tier peer-reviewed journals, and then to data presented in DSDP Initial Reports Part II and ODP Scientific Results. Some older articles required manually searching through printed journals.

The process of researching and identifying legacy data in post-cruise publications included manual searches of journals from relevant disciplines (e.g., marine geology, sedimentology, palaeontology, geochemistry, petrology, geophysics) discovered via major reference databases (i.e. AGI Georef, ScienceDirect, etc.). Citations for DSDP/ODP/IODP-related publications were collected and the articles were manually searched for relevant data sets that were not captured or accessible in any IODP data system. The primary method for determining whether a data set from a published paper existed in the PANGAEA archive was by comparing the source publication's references and metadata with the IODP publications databases and metadata.

Once a publication was confirmed to contain DSDP/ODP/IODP-related data not contained in a SEDIS-accessible data provider, the workflow prescribed a method for extracting, and where necessary digitizing, the data from tables, charts or other formats in the publication. Extracting and digitizing of the data was performed such that numerical data or data tables were transformed into an ASCII format. Graphs with specified numerical data in the publication were examined and, where possible, the data was transposed to numerical form in an ASCII format. ISO19915-compliant metadata was generated for all extracted data, with citations for the extracted data supplement, references to the source publication, and a digital object identifier (DOI) assigned to each extracted data set. After a QA/QC process to confirm accuracy of the extracted material, the data and metadata were published in the PANGAEA systems and harvested by the SEDIS system where they made discoverable and accessible for re-use.

3. Outcome of post-cruise data rescue efforts

The post-cruise data capture project was conducted in two phases, starting in 2007 and finishing in 2013. The project relied on a collaborative effort between subject matter experts from the natural sciences (e.g., geological and environmental sciences) and informatics and computer sciences. The project team calculated that the amount of effort per data set published within this project was between three to four person-hours per publication/data supplement loaded to PANGAEA.

Between 2007 and 2013, the PANGAEA data curators examined more than 6000 journal articles and reports in a search for DSDP/ODP/IODP-related data. Approximately half of all DSDP/ODP/IODP-related articles reviewed were found to have extractable data sets in tables, appendices, and supplementary materials. Where an article contained more than one extractable data set, each extractable data set was digitized and archived individually