BCO-DMO
Biological and Chemical Oceanography Data Management Office

Automating Data Submission to a National Archive

AGU Poster IN21A-1324
AGU Fall Meeting
December 14, 2010

Tobias T. Work[1], Cynthia L. Chandler[2], Robert C. Groman[1], Molly D. Allison[1], Stephen R. Gegg[2]
1~Biology Department, 2~Marine Chemistry and Geochemistry Department, Woods Hole Oceanographic Institution, Woods Hole MA 02543

Abstract

http://www.agu.org/cgi-bin/wais?ss=IN21A-1324

In late 2006, the U.S. National Science Foundation (NSF) funded the Biological and Chemical Oceanographic Data Management Office (BCO-DMO) at Woods Hole Oceanographic Institution (WHOI) to work closely with investigators to manage oceanographic data generated from their research projects. One of the final data management tasks is to ensure that the data are permanently archived at the U.S. National Oceanographic Data Center (NODC) or other appropriate national archiving facility. In the past, BCO-DMO submitted data to NODC as an email with attachments including a PDF file (a manually completed metadata record) and one or more data files. This method is no longer feasible given the rate at which data sets are contributed to BCO-DMO. Working with collaborators at NODC, a more streamlined and automated workflow was developed to keep up with the increased volume of data that must be archived at NODC. We will describe our new workflow; a semi-automated approach for contributing data to NODC that includes a Federal Geographic Data Committee (FGDC) compliant Extensible Markup Language (XML) metadata file accompanied by comma-delimited data files. The FGDC XML file is populated from information stored in a MySQL database. A crosswalk described by an eXtensible Stylesheet Language Transformation (XSLT) is used to transform the XML formatted MySQL result set to a FGDC compliant XML metadata file. To ensure data integrity, the MD5 algorithm is used to generate a checksum and manifest of the files submitted to NODC for permanent archive. The revised system supports preparation of detailed, standards-compliant metadata that facilitate data sharing and enable accurate reuse of multidisciplinary information. The approach is generic enough to be adapted for use by other data management groups.

Motivation

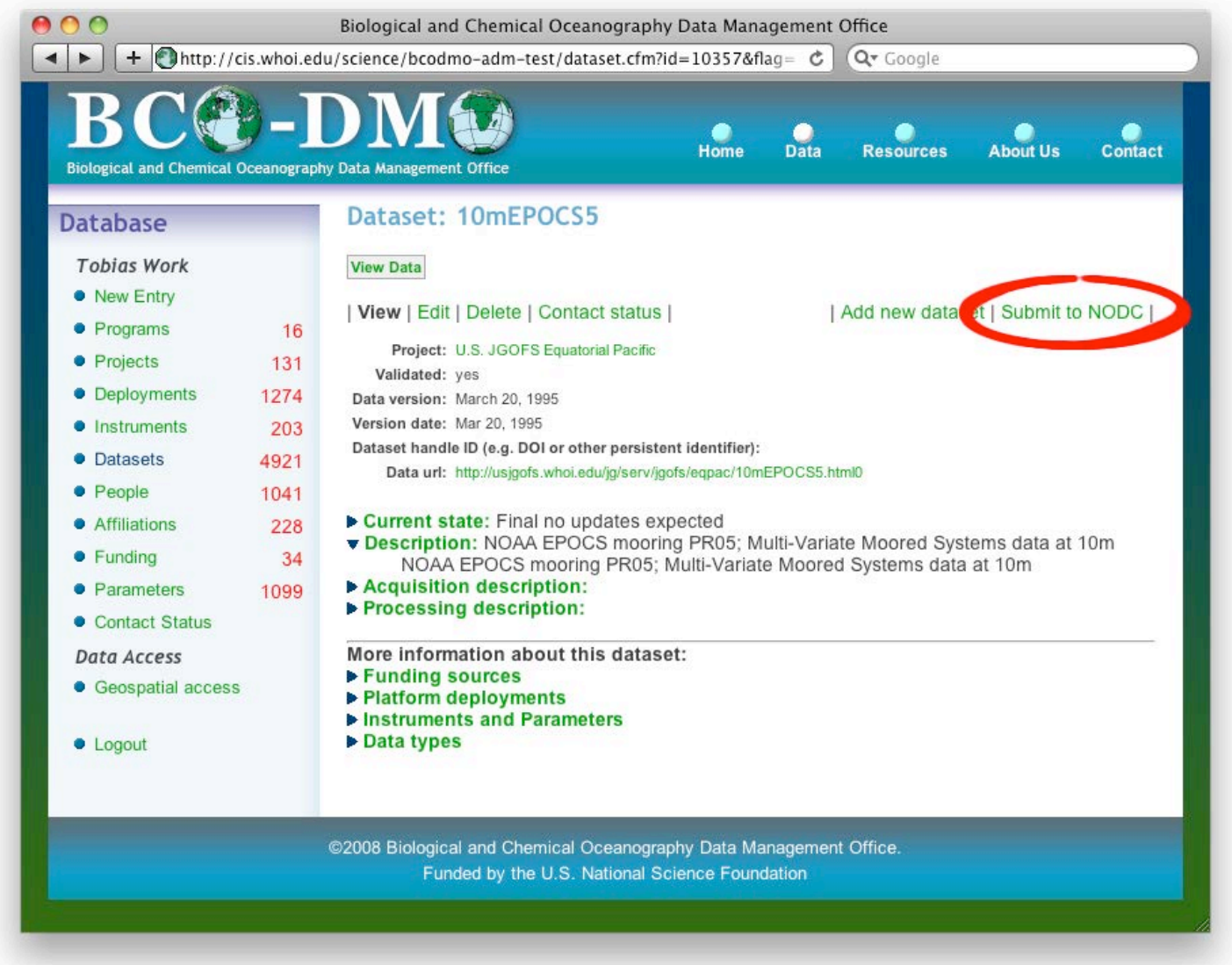
The motivation for a more automated data submission system is straightforward: the old way (filling out a PDF by hand and sending it in an email) is simply not fast enough given the volume of data that BCO-DMO is required to manage and therefore submit to NODC or other appropriate archiving facility. The new way involves much less effort by the data manager and yields an improved product that includes an FGDC-compliant metadata record in an XML-structured file that simplifies data ingest by the archivist.

Acknowledgments

This work has been funded by the National Science Foundation. We are grateful for the contributions of Julie Allen for her applications programming work, Jacqueline Mize (National Oceanic and Atmospheric Administration National Coastal Data Development Center) for her metadata training course, and Steven Rutz and Jonathan Blythe (National Oceanographic Data Center) for working with BCO-DMO to develop this streamlined workflow.

Archive Process

1 Select & Customize


Biological and Chemical Oceanography Data Management Office
http://cis.whoi.edu/science/bcodmo-adm-test/dataset.cfm?id=10357&flag=

Database

Tobias Work

- New Entry
- Programs 16
- Projects 131
- Deployments 1274
- Instruments 203
- Datasets 4921
- People 1041
- Affiliations 228
- Funding 34
- Parameters 1099
- Contact Status

Data Access

- Geospatial access
- Logout

Dataset: 10mEPOCS5

View Data

| View | Edit | Delete | Contact status | | Add new dataset | Submit to NODC |

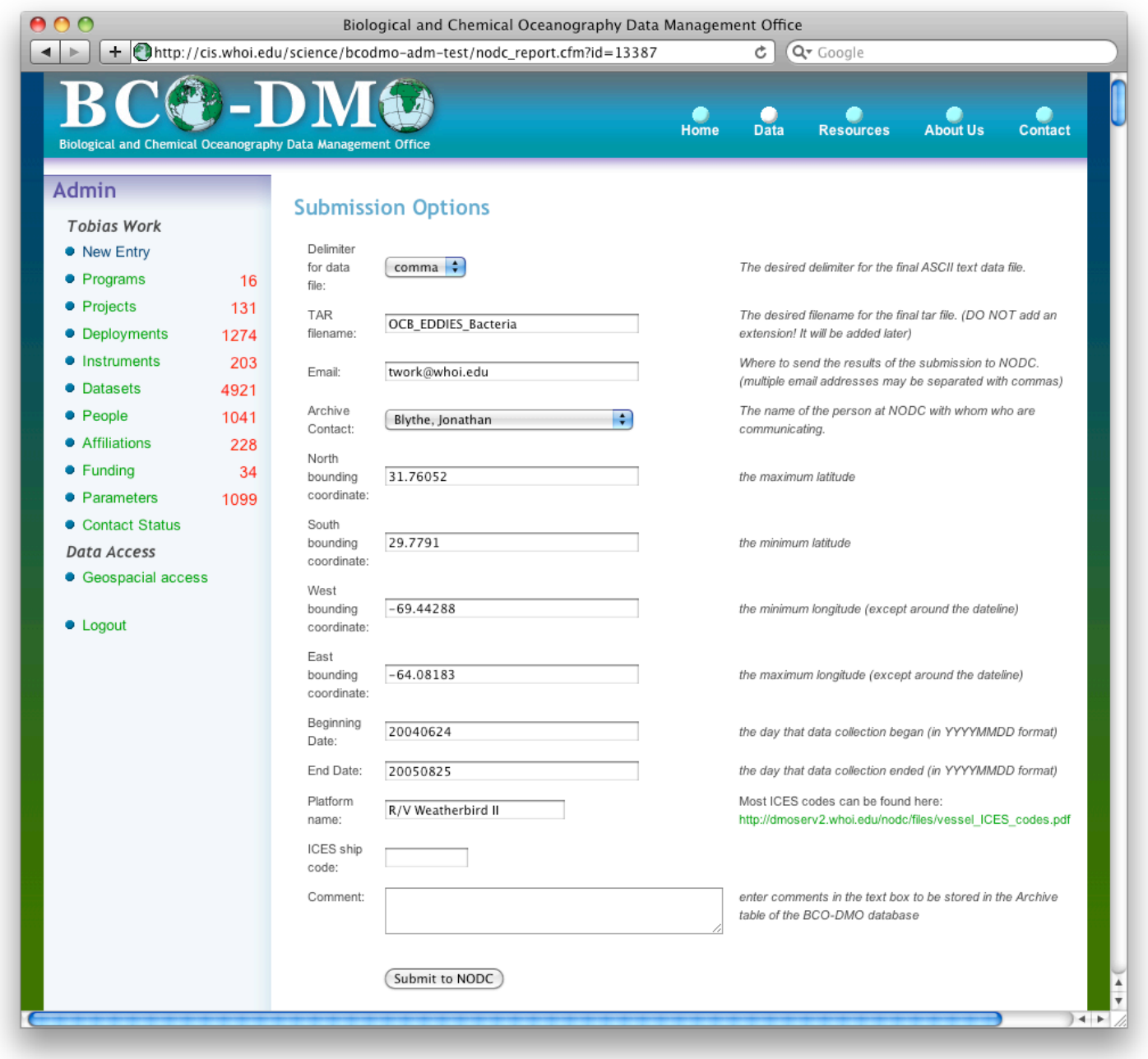
Project: U.S. JGOFS Equatorial Pacific
Validated: yes
Data version: March 20, 1995
Version date: Mar 20, 1995
Dataset handle ID (e.g. DOI or other persistent identifier):
Data url: http://usjgofs.whoi.edu/jg/serv/jgofs/eqpac/10mEPOCS5.html0

▶ Current state: Final no updates expected
▼ Description: NOAA EPOCS mooring PR05; Multi-Variate Moored Systems data at 10m
NOAA EPOCS mooring PR05; Multi-Variate Moored Systems data at 10m
▶ Acquisition description:
▶ Processing description:

More information about this dataset:
▶ Funding sources
▶ Platform deployments
▶ Instruments and Parameters
▶ Data types

©2008 Biological and Chemical Oceanography Data Management Office.
Funded by the U.S. National Science Foundation

Using the BCO-DMO metadata database web interface, the data manager identifies the dataset that is to be archived at NODC or other appropriate national archiving facility.


Biological and Chemical Oceanography Data Management Office
http://cis.whoi.edu/science/bcodmo-adm-test/nodc_report.cfm?id=13387

Admin

Tobias Work

- New Entry
- Programs 16
- Projects 131
- Deployments 1274
- Instruments 203
- Datasets 4921
- People 1041
- Affiliations 228
- Funding 34
- Parameters 1099
- Contact Status

Data Access

- Geospatial access
- Logout

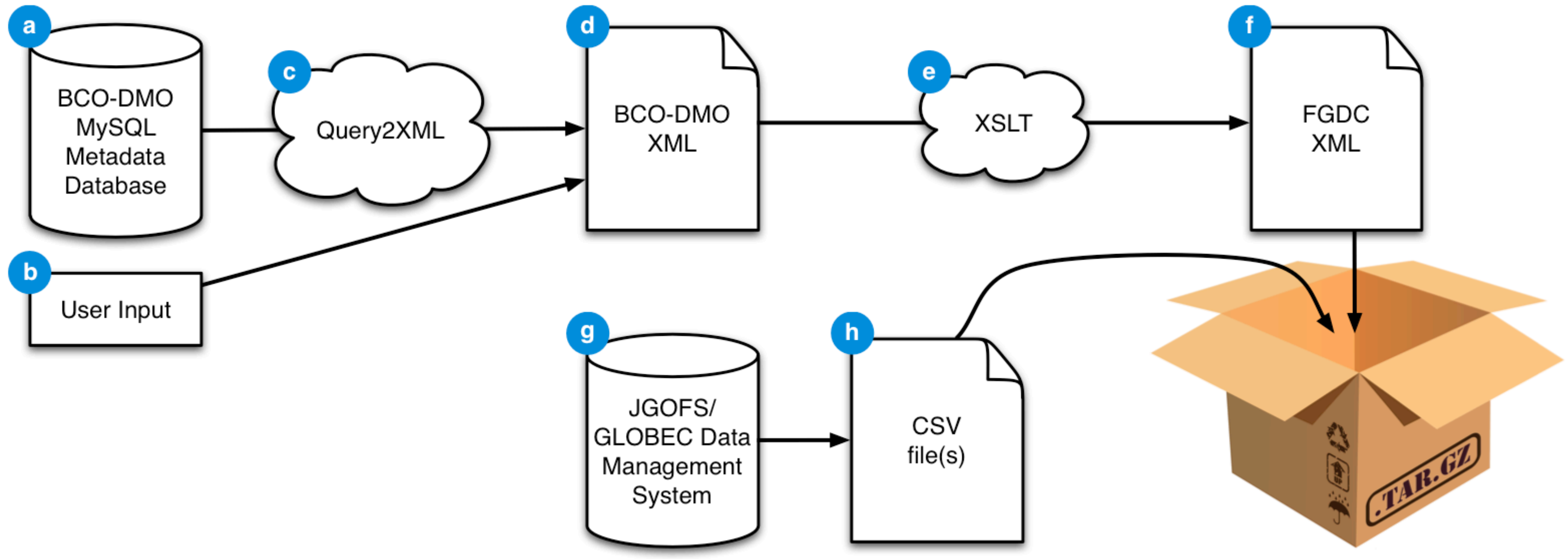
Submission Options

Delimiter for data file:	comma	*The desired delimiter for the final ASCII text data file.*
TAR filename:	OCB_EDDIES_Bacteria	*The desired filename for the final tar file. (DO NOT add an extension! It will be added later)*
Email:	twork@whoi.edu	*Where to send the results of the submission to NODC. (multiple email addresses may be separated with commas)*
Archive Contact:	Blythe, Jonathan	*The name of the person at NODC with whom who are communicating.*
North bounding coordinate:	31.76052	*the maximum latitude*
South bounding coordinate:	29.7791	*the minimum latitude*
West bounding coordinate:	-69.44288	*the minimum longitude (except around the dateline)*
East bounding coordinate:	-64.08183	*the maximum longitude (except around the dateline)*
Beginning Date:	20040624	*the day that data collection began (in YYYYMMDD format)*
End Date:	20050825	*the day that data collection ended (in YYYYMMDD format)*
Platform name:	R/V Weatherbird II	Most ICES codes can be found here: http://dmoserv2.whoi.edu/nodc/files/vessel_ICES_codes.pdf
ICES ship code:		
Comment:		*enter comments in the text box to be stored in the Archive table of the BCO-DMO database*

Submit to NODC

Options pertaining to the submission package may be specified at this stage of the archive process. For instance, the data manager may choose either tab or comma separated for the format of the data file that is to be archived at the data center. Other options include bounding coordinates of the dataset, start and end dates of data collection, and a comment to accompany the record of the archive process in the BCO-DMO metadata database. The majority of the information in this form is pre-filled by scripts that are automatically executed when the page is loaded.

2 Package



Packaging is the most complex step of the archive process. The methodology in this stage is loosely based on the methodology detailed in a paper titled *A Solution to Metadata: Using XML Transformations to Automate Metadata*[1].

a he BCO-DMO MySQL metadata database provides the majority of the information that populates the FGDC XML metadata file (see f

b formation unavailable from the MySQL metadata database but required for a complete FGDC XML metadata record comes from the user input gathered in step . 1

c nce MySQL is used for the metadata database, a method of converting SQL results into XML was needed. In order to accomplish this, a PHP library, Query2XML[2], was used. This library is highly configurable and saved us from having to “reinvent the wheel”.

d he XML file generated using the Query2XML PHP library in c ovides most of the information required for a complete FGDC XML metadata record. However, not every piece of information required by the FGDC standard is in the metadata database. The remaining information is supplied by the user and added to the BCO-DMO XML document. This is simple given PHP's ability to read, write, and manipulate XML.

e n XSLT crosswalk was written to transform the XML we produced in d to a FGDC XML file. Note that suitably constructed XSLT's can be used at this stage to generate other standards compliant XML files. It should also be noted that to apply this transformation, we are using a PHP library, XML_XSLT2Processor[3], in conjunction with the Java based SAXON XSLT and XQuery 2.0 Processor[4].

f is step produces a standards compliant FGDC XML file, ready to be packaged and sent to NODC. This file is validated against the FGDC XML schema to ensure accuracy and conformity to the standard.

g CO-DMO not only manages scientific metadata, but scientific data as well. Scientific data are managed by the JGOFS/GLOBEC Data Management System[5].

h sing the JGOFS/GLOBEC Data Management System API, the data files for the package are formatted as either tab or comma separated files.

[1] Mize, Jacqueline. "A Solution to Metadata: Using XML Transformations to Automate Metadata." *Oceans 2009 Conference MTS/IEEE Biloxi*. 2009. PDF.

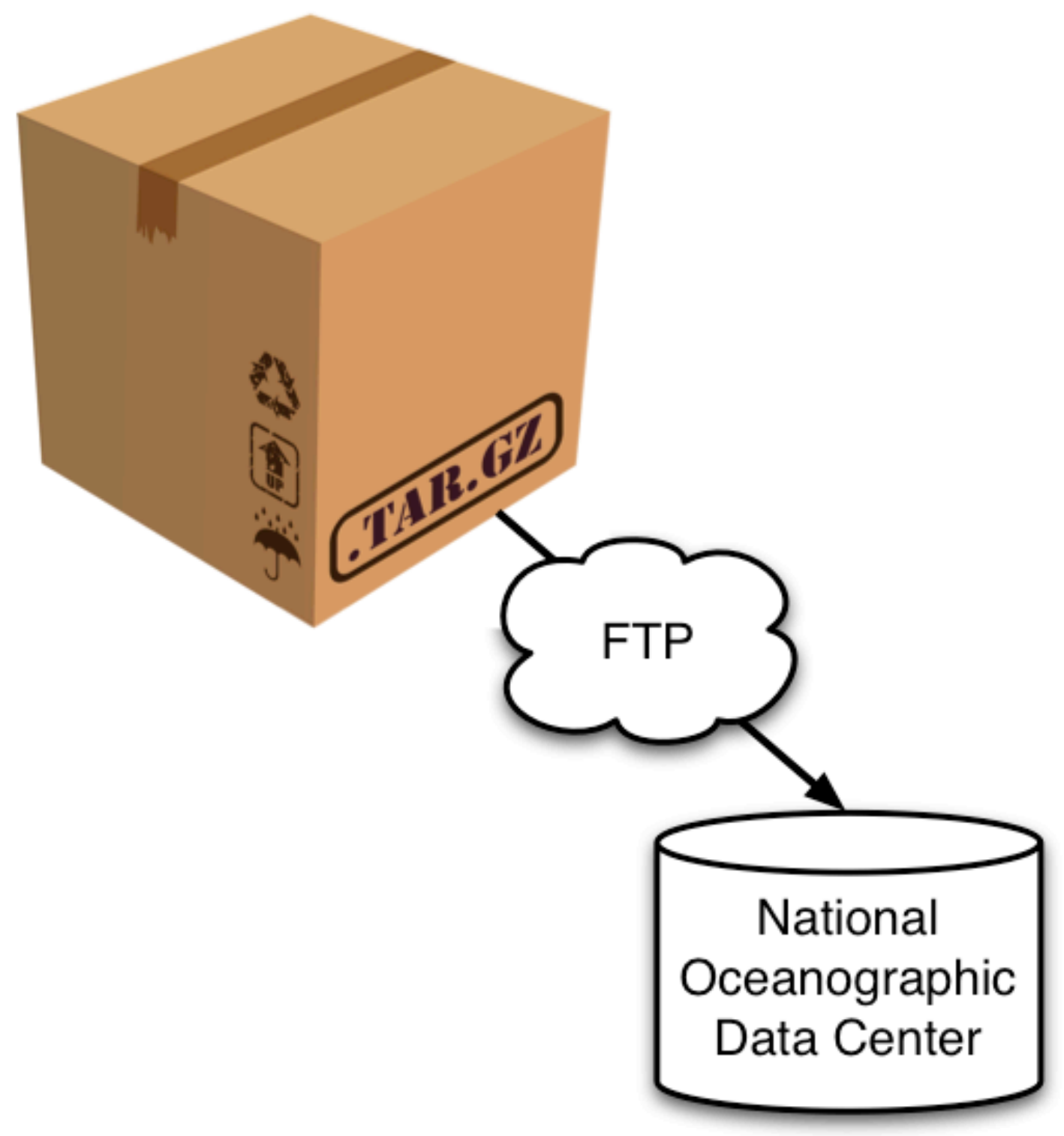
[2] See http://query2xml.sourceforge.net/docs/html/XML_Query2XML/_Query2XML.php.html for more information and downloads.

[3] See http://sourceforge.net/projects/xslt2processor/ for more information and downloads.

[4] See http://saxon.sourceforge.net/#F9.2HE for more information and downloads.

[5] See http://globec.whoi.edu/globec-dir/doc/datasys/jgsys.html for more information and downloads.

3 Send



In the final stage, the gzipped tar file containing the the metadata, data, and other relevant files (not pictured) could be transferred to NODC via File Transfer Protocol (FTP). Since the process is still being developed, the packages are inspected by hand before they are sent to the archive to ensure they are complete. Although FTP is supported by the software, currently the data manager sends the package as an email attachment. Every package archived is also accompanied by a text file containing the MD5 hash of the gzipped tar file in order to ensure data integrity.

In the end, this improved workflow has saved a significant amount of time for both BCO-DMO and NODC. The amount of time required by a BCO-DMO data manager to archive a dataset at NODC has been reduced from hours or even days to minutes. In addition, this approach easily can be adapted to submit data to other archive facilities.