An Introspective Approach to Enabling Discovery, Understanding and Appropriate Use of NOAA Hydrographic Data for New and Emerging Clients

Jeremy McHugh Dan Neumann John Tucker James Hiebert

April 4, 2008

Abstract

Hydrographic survey data is acquired at great expense and can be extremely valuable for uses beyond nautical charting. The gateway to the widest and best use of that valuable data is hinged on the quality of and accessibility to its metadata.

The standard metadata document for the National Oceanic and Atmospheric Administration's (NOAA's) hydrographic surveys is and has been a narrative style report to accompany each survey. NOAA is developing an efficient way to create a new generation of XML-based reports to attain consistently higher quality and increased utility of the reports and to more easily meet our internal and external metadata requirements.

NOAA has analyzed content requirements for the reports and has formalized our internally required metadata into encoded XML schemas. Schemas which describe our metadata to external entities are in development. The ultimate goal is to have our internal schemas included in a superset of metadata that will be compliant with International Standards Organization's 19139 Technical Specification and the Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata.

This article focuses on the development of XML schemas that will hold NOAA's internally required metadata and the development of the software interface used to populate metadata instance documents.

Introduction

Hydrographic survey data is acquired at great expense and should be thought of as a precious resource for at least two reasons. First, it represents a very large capital investment that deserves prudent care and responsible use. Costs for data acquisition are on the order of tens of thousands of U.S. dollars per square nautical mile of coverage. Second, it often functions as a base that is both related to and interpreted in the context of a host of other types of marine and ecosystems science data. The traditional and still primary use of NOAA hydrographic data is for nautical charting. However, there are many new clients, including diverse groups of marine scientists, who are finding hydrographic data and associated metadata (ie, documentation) increasingly useful and often necessary. Together, the large capital investment and the increasingly widespread utility of the hydrographic data for a growing client base establish a strong case for effective and responsible data management.

Lastly, Metadata is a major component of a modern holistic management strategy for scientific data (Marine Metadata Interoperability, 2007). It is critical that we provide open public access to clear, meaningful, and useful metadata that will allow potential clients to discover, understand and appropriately use our data.

Status Quo and Motivation for Improvement

For the past two centuries, NOAA and it's predecessor agencies have produced loosely structured narrative reports describing the conditions under which the various data were acquired and processed for each hydrographic survey. These reports were fine in their day because the hydrographic surveys were not distributed outside the highly specialized community who (1) already understood implicitly what a hydrographic survey is and (2) were proficient at using that survey data as source material to compile final products. In the past, only the final products were distributed to the public. But, in recent years, there has been an explosion of final products that can be created from the processed hydrographic data and the Internet has enabled direct access by anyone worldwide to those products as well as to the underlying data. In order for an anonymous user to judge the fitness for his or her use (whatever that may be) of any particular hydrographic survey, extensive metadata describing the technology of data acquisition and the checks for data quality must be provided and tightly linked with the base information. In the U.S., several significant government reports have (1) explained the need for such metadata to support integrated and efficiently coordinated ocean and coastal mapping activities among numerous entities and (2) charged NOAA to take the lead in these activities (Mayer et al, 2004; U.S. Commission on Ocean Policy, 2004; Bush Administration, 2004).

Therefore, NOAA has recognized a need to modernize our hydrographic survey report to enable us to:

1. produce the reports more efficiently
2. reduce inconsistency and eliminate redundancy both within and between related reports by having a single report serve as a consolidated source for all survey documentation
3. generate a consistent reporting product for all NOAA hydrographic surveys
4. meet our external metadata requirements more easily and efficiently
5. eventually populate a single repository or resource from which all U.S. hydrographic survey data can be searched and retrieved (the planning for such a repository is underway)

Development

Rationale and design principles

From the outset, we chose to develop reports encoded in the eXtensible Markup Language (XML). Why XML? At a fundamental level, this decision was made because XML offers simplicity by using text that is both readable and understandable by humans and it offers flexibility in that it is well-suited for representing hierarchical data structures (Bray et al, 1997 & 1998; Aiken & Allen, 2004). Within the U.S. federal government, XML is considered a key technological component of the Federal Enterprise Architecture and is being used by numerous agencies, departments and branches in a variety of creative ways (Sall, 2004). One particularly well-suited and interesting example is the use of XML to draft highly structured legislative documents in the U.S. House of Representatives (Carmel, 2002).

Additionally, it was recognized that the Internet is the most familiar way for the rest of the world to view and interact with our metadata and XML is the de facto standard way to share structured information on the Internet (Organization for the Advancement of Structured Information Standards (OASIS), 2007). Numerous metadata standards have been promulgated by various organizations around the world, but the world's metadata producers seem to be converging on the use of international standards with XML implementations (Federal Geographic Data Committee, 2007; International Standards Organization, 2007).

We formerly espoused the relational database model for encapsulating virtually all hydrographic-related data in NOAA's Office of Coast Survey. However, XML offers numerous improvements. It is designed to handle semi-structured information and does so better by providing direct links between metadata and data, rather than the relational 'join' concept which exists only in processing scripts, not the data structures themselves. It also eliminates the need for artificial placeholders for defined fields when the information isn't available; in XML, if it isn't there, it isn't there. Our legacy data is littered with miscellaneous, machine-dependent, undocumented, unavailable, and invalid placeholders. Processing that data is one of our biggest challenges. XML makes versioning, the process of accommodating the inevitable changes to data and metadata and formats, to be easier, less disruptive, and more coherent. The main disadvantage discovered so far is that XML is more voluminous, but with today's prices for memory and speeds of data transfer this proves to be easily overcome and only reinforces our abilities to document in more detail.

To that end, we embarked on an effort to develop a semi-automated XML-based hydrographic survey report that writes itself to the greatest extent possible. The report will begin during the project planning phase and be passed through our data pipeline all the way through the archival of data and metadata. Along the way, appropriate offices will add their parts and employ others. The semi-automation will be achieved by harvesting information from existing data streams used by data acquisition and processing software packages.

There are three overriding principles that govern this development process. First, to eliminate redundancy and other inefficiencies, we only want to enter a given piece of information once. This requires thoughtful design up front. If the same information exists in separate parts of the XML encoding, we want links and references between them, rather than redundant copies. Second, wherever possible, we want to harvest information that is automatically generated during routine data acquisition and processing procedures. This will reduce the amount of manual input -thus reducing blunders- and streamline the production of the reports. Finally, we want to enter or harvest our information in as close to real-time as possible. The best and most useful metadata is acquired concurrently with the data that it supports.

The next three subsections will describe the (1) development of the rules that govern the formalization of the survey reports (ie, the XML schemas), (2) the development of the software tool that will be used to generate XML documents, and (3) how we plan to implement XML survey reports.

XML schema development

XML schemas can be thought of as a formalization of the metadata. They specify the structure, content, and semantics of the XML survey reports which we will be producing. Our schemas are written in the RELAX NG schema language (OASIS, 2002; RELAX NG, 2003).

Before developing the schemas, it was necessary to analyze our content requirements. We asked ourselves the following questions. Exactly what pieces of information do we want to be in an ideal hydrographic survey report? Is there anything we have traditionally documented that is no longer necessary? Is there some new piece of documentation that we should be harvesting? The end result was a comprehensive list of ‘ingredients’ required to make an ideal report.

We then broke our ‘ingredients’ (content requirements) down into eight separate XML schemas (Figure 1). The breakdown of the XML content requirements was governed by two design limitations related to the hierarchical structure inherent to XML. First, XML is governed by single-inheritance relationships and is not conducive to representing multiple-inheritance relationships where a child element inherits features from multiple parents. This problem can be overcome by dividing our XML code into multiple schemas that can be cross-referenced. Second, we wanted to be able to insert information into XML only once and avoid duplication of effort. Reuse of the XML code is made possible by the division and cross-referencing of the XML encoded metadata. Figure 1 shows the individual XML schemas that will comprise an XML-based survey report and the information that is included.

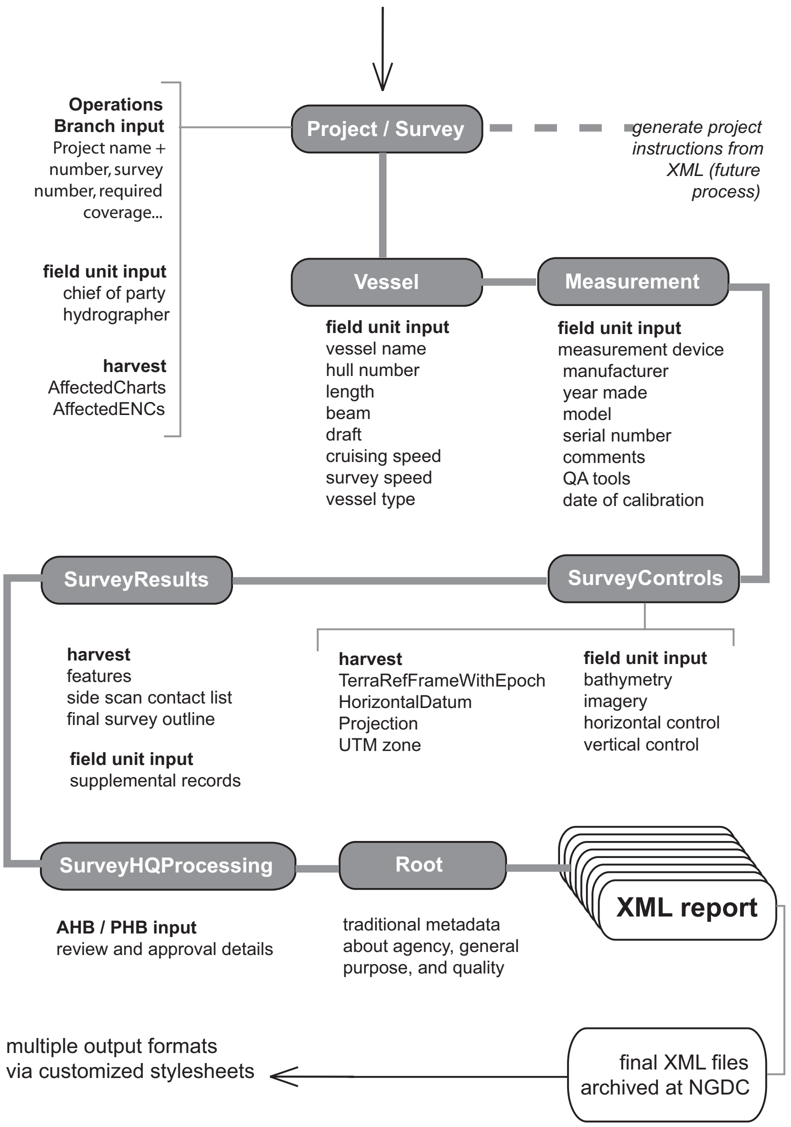


Figure 1: Flow diagram representing the development of an XML-based survey report. Gray shapes represent XML schemas.

An example of the logical model of the 'Vessel' schema is shown in Figure 2 to help convey how the the heirarchical nature of XML applies to the documentation of a hydrographic survey vessel.

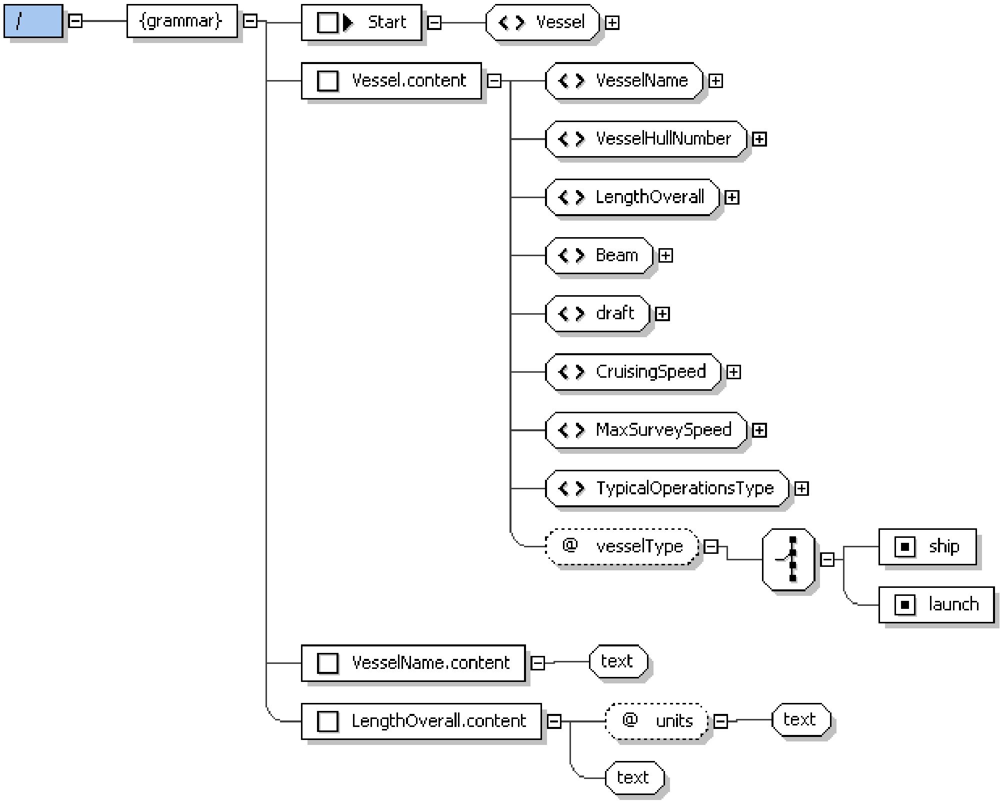


Figure 2: Logical Model of the ‘Vessel’ XML schema.

Input wizard development

In order for NOAA personnel to create an XML-based hydrographic survey report, it is required that there be some convenient interface with which to populate the set of XML instance documents. The requirements of such an interface are as follows:

1. It must effectively inform the user as to what metadata is necessary. This can also be thought of as asking the right questions of the data. If there is something specific that we want to know about the data, we want to explicitly ask that question via the interface rather than passively expecting the answer to be covered in a narrative report.

2. It must enforce the constraints which we have formalized. For example, if a certain element is required, the user must be properly warned if he or she violates that requirement.

3. It must accommodate dynamic metadata requirements and be as extensible as the underlying XML upon which it is built.

4. It should provide mechanisms to automatically harvest metadata that already exist as part of the NOAA data processing pipeline. It should minimize the amount of manual user input and eliminate redundant manual input.

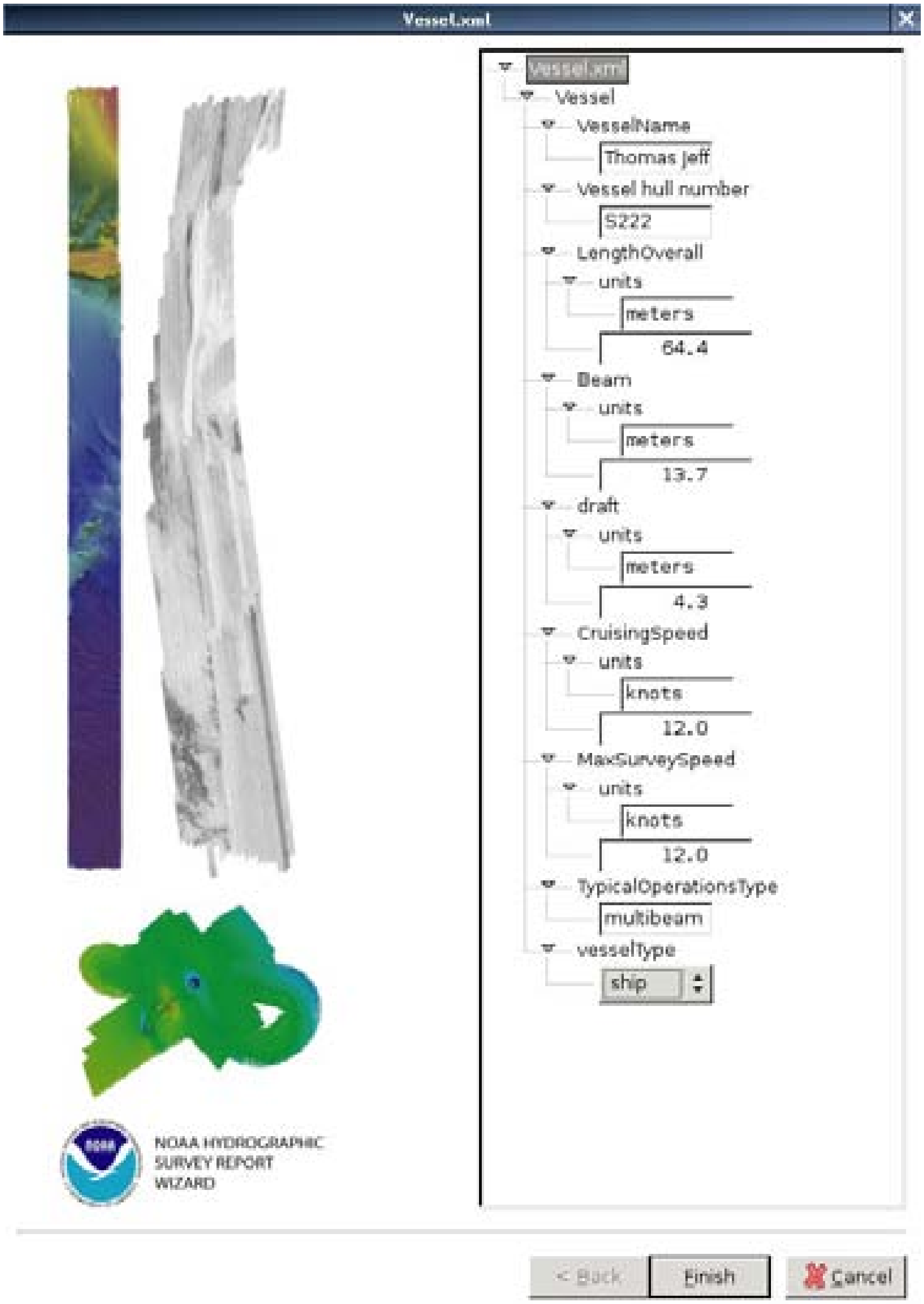
5. It may need to include access and/or version control to manage out-of-flow changes to the metadata. For example, it is possible for project instructions to change while a survey is underway. Cases such as this need to be properly managed.

To meet requirements number one through three, we have begun to design and develop a metadata input wizard. Its premise is simple: at run-time it reads one of the formalized schemas and then automatically generates a set of input dialogs. An example of this is shown in figures 3 and 4. The schema that is shown in figure 3 is read by the input wizard and then converted into a series of input dialogs as shown in figure 4. This system ensures that requirements one through three are met, assuming that the formalized schemas remain up to date and contain self-describing elements which adequately inform the user as to their semantic meaning.[1]

```
<element name="Vessel">
  <element name="VesselName">
    <text/>
  </element>
  <element name="VesselHullNumber">
    <text/>
  </element>
  <element name="LengthOverall">
    <attribute name="units" >
      <data type="string"/>
    </attribute>
    <data type="integer"/>
  </element>
  ...
</element>
```

Figure 3: This excerpt from the code listing ‘Vessel.rng’ shows the codified formalization of the metadata required to describe a survey vessel.

[1] While not entirely relevant to the discussion, the wizard is implemented in the python programming language with window management code in wxpython. The wizard can run on a variety of platforms including Linux and Windows.



Vessel.xml
Vessel
VesselName
Thomas Jeff
Vessel hull number
S222
LengthOverall
units
meters
64.4
Beam
units
meters
13.7
draft
units
meters
4.3
CruisingSpeed
units
knots
12.0
MaxSurveySpeed
units
knots
12.0
TypicalOperationsType
multibeam
vesselType
ship

Figure 4: Screenshot of input wizard for the 'Vessel' schema.

Implementation

XML will allow us flexibility in the display of information through the use of stylesheets tailored to functional groups within our data pipeline. A variety of stylesheets will be created with which the resulting metadata may be displayed. For example, one stylesheet could display the metadata in a manner similar to the traditional narrative report format, while another could just filter out the specific features to create a report of dangers to navigation. We could also output the report as an HTML web document while another could write a pdf file for printing purposes. Stylesheets with targeted design will allow users access to the particular information they want to see in a more efficient way than searching through a digital narrative document. An example of an output format from a Vessel instance document is shown in figure 5.

vessel name	**NOAA Ship THOMAS JEFFERSON**	hull number	**S222**
length	**63.4 m**	cruising speed	**12 knots**
beam	**13.7 m**	maximum survey speed	**12 knots**
draft	**4.3 m**	typical operations	**multibeam and side scan sonar**

Figure 5: Example of an output format from a Vessel instance document.

An XML-based survey report will help us meet our metadata requirements more efficiently. We intend to have our internally required XML metadata contained within a superset of XML metadata envelopes compliant with external metadata standards. That superset will meet the requirements of the International Standards Organization's (ISO's) 19115 metadata standard and associated XML implementation (ISO technical specification 19139) as well as the Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata. The work to develop that superset of XML envelopes for external data is underway.

Conclusions

The use of XML in a documentation system for highly technical work, such as that involved in hydrographic surveying, offers rich opportunities for gains in quality and efficiency of production of metadata.

Future Work

We will continue developing a fully functional wizard interface to be used by office- and ship-based personnel to create hydgrographic survey reports. Once tested, that interface will be used to generate XML documents compliant with internal and external (FGDC and ISO) metadata requirements. Those documents will be put into a metadata repository for U.S. hydrographic survey data. Planning is underway for such a repository and for the superset of XML schemas into which our internally required metadata will fit. In the near future, the public will have open access to metadata for NOAA hydrographic surveys via the geodata.gov (Geospatial One-Stop) portal. This portal is designed to promote the sharing of geospatial data in the U.S. (Geodata.gov, 2008).

References

Aiken, P & Allen, D 2004, ‘XML for Data Management’, Elsevier/Morgan Kaufmann, Amsterdam/Boston, viewed 10 March 2008, <http://www.netlibrary.com/>.

Bray, T & Paoli, J & Sperberg-McQueen, CM 1997, ‘Extensible Markup Language’, World Wide Web Journal, volume 2, number 4, pages 29-66.

Bray, T & Paoli, J & Sperberg-McQueen, CM 1998, Extensible Markup Language (XML) 1.0 - W3C Recommendation 10-Feb-98, Cambridge, Massachusetts, viewed 10 March 2008, <http://www.w3.org/TR/1998/REC-xml-19980210.pdf>.

Bush Administration 2004, ‘U.S. Ocean Action Plan; The Bush Administration’s Response to the U.S. Commission on Ocean Policy’, Washington, DC, viewed on 4 April 2008 <http://ocean.ceq.gov/actionplan.pdf> .

Carmel, J 2002, ‘Drafting Legislation Using XML at the U.S. House of Representatives’, XML 2002 Conference, viewed 11 March 2008, <http://www.idealliance.org/papers/dx_xml03/papers/05-01-04/05-01-04.pdf>.

Federal Geographic Data Committee 2007, Content Standard for Digital Geospatial Metadata, Reston, Virginia, viewed 10 March 2008, <http://www.fgdc.gov/metadata/geospatial-metadata-standards>.

Geodata.gov 2008, viewed on 4 April 2008, <http://geodata.gov/>

International Standards Organization 2007, Technical Comittee 211 Geographic information/Geomatics, Honefoss, Norway, viewed 10 March 2008, <http://www.isotc211.org/>.

Marine Metadata Interoperability 2007, ‘Improving Data Sharing to Improve Science’, viewed 11 March 2008, <http://marinemetadata.org/for-scientists>.

Mayer, L & Barbor, K et al 2004, ‘A Geospatial Framework for the Coastal Zone: National Needs for Coastal Mapping and Charting’, National Adademies Press, Washington, DC, viewed on 14 March 2008 <http://www.nap.edu/openbook.php?isbn=0309091764>

Organization for the Advancement of Structured Information Standards 2007, viewed 10 March 2008 <http://www.oasis-open.org/home/index.php>.

Organization for the Advancement of Structured Information Standards 2002, RELAX NG Technical Committee home page, viewed on 11 March 2008, <http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=relax-ng>.

RELAX NG 2003, RELAX NG Home Page, viewed 11 March 2008, <http://www.relaxng.org/>.

Sall, K 2004, 'How the U.S. Federal Government is Using XML', XML Conference and Exposition 2004, Philadelphia, Pennsylvania, viewed 11 March 2008 <http://lists.oasis-open.org/archives/ihc/200409/pdf00000.pdf>.

U.S. Commission on Ocean Policy 2004, 'An Ocean Blueprint for the 21st Century; Final Report', Washington, DC, viewed on 4 April 2008 <http://www.oceancommission.gov>.

Author Biographies

Jeremy McHugh

Jeremy has worked with NOAA's Office of Coast Survey out of Silver Spring, Maryland for the past five years as a physical scientist planning hydrographic surveys on the Atlantic and Gulf coasts. Before that, he earned a M.S. degree in Structural Geology from the University of Nevada, Reno, Mackay School of Mines.

NOAA's National Ocean Service, Office of Coast Survey
1315 East-West Highway, Station 6726
Silver Spring, MD 20910
t. 301 713 2700 x117
f. 301 713 4533
jeremy.mchugh@noaa.gov

Dan Neumann

Dan has worked with NOAA's Office of Coast Survey since 1974. He is a senior information technology specialist in The Data Acquisition and Control Branch of the Hydrographic Surveys Division. He has a Masters Degree in Geography from the University of California, Riverside.

John Tucker

John, one of three original designers of the S-57 standard for digital nautical data now in use worldwide, would prefer to stay at home, make beautiful things out of wood, pet his cats, take pictures of his beautiful wife, and run circles around the fat cats on Wall Street.

James Hiebert

James has a background in Computer Science having earned his M.S. from the University of Oregon in CS with foci in peer-to-peer networking, Internet routing and ecological modeling and simulation. He has worked for NOAA's Office of Coast Survey since 2007.