Using a Cruise Report to Generate XML Metadata

Briana M. Sullivan
The Center for Coastal and Ocean Mapping/Joint Hydrographic Center
University of New Hampshire
Durham, USA
briana@ccom.unh.edu

Abstract— Since 2005 metadata generation at the Center for Coastal and Ocean Mapping/Joint Hydrographic Center has slowly evolved from a painful and tedious process of copying and pasting, to generate hundreds of files, to using an automated system that generates 90% of the needed metadata from the data collected on cruises. However there remained one piece missing to the automated system- the wordy part of the metadata that deals with information such as the attribute accuracy report, abstract and the process description. This information cannot be mined from the raw survey data. This paper illustrates how to generate a template from a Microsoft Word based cruise report that can be used in conjunction with another template (generated from the raw data collected on a cruise) to create XML metadata ready for submission to the NOAA/National Geophysical Data Center.

***Index Terms*—Informatics, metadata.**

I. Introduction

Many research ships now routinely collect multibeam bathymetry and acoustic backscatter data. For this data to be submitted to the NOAA/National Geophysical Data Center (NGDC), it needs to be accompanied by XML (eXtensible Markup Language) metadata, the data about the data. Each survey line of data that is collected needs its own metadata file, and with hundreds of lines collected on each cruise, generating metadata files can become quite tedious. Automating the metadata generation would save time and reduce errors. It is trivial to extract concrete data (i.e., the geographical bounding box and start and end times) from the raw data files themselves. However, contained within the metadata are also paragraphs of explanations, (i.e. the purpose of the mission and an overview of what was accomplished on the mission). These parts of the metadata are usually found in accompanying cruise reports. Mining the necessary paragraph information directly from the cruise reports eliminates the need for redundant typing and reduces errors introduced with cutting and pasting.

Most of the chief scientists at the Center for Coastal and Ocean Mapping/Joint Hydrographic Center use Microsoft Word to write their cruise reports. To make it less intrusive to them, it was evident that leveraging Microsoft Word as a tool for the data mining was the best route to take. After organizing the report, the parts that contain information needed in the metadata template (the Area Level template) would need to be "tagged".

```
#AREA INFORMATION (same across all metadata files for this
area)

area: Necker Ridge
cruiseID: KM1121
ship: RV Kilo Moana
PI_or_Chief_Scientist: Dr James V. Gardner
instrument: Kongsberg Maritime EM122 multibeam echosounder,
Knudsen 3260 chirp subbottom profiler
Map_Projection: unprojected
False_Easting: 0
False_Northing: 0
Latitude_Resolution: 0.00005
Longitude_Resolution: 0.00005

Title: Raw <instrument> <dataType> data for the <area>,
southwest of Hawaiian Ridge.

Abstract: <instrument> <dataType> data were collected in
<dataFormat> format.  This metadata file is for a specific
survey line of this cruise (<Line_Number>).
Purpose: The National Oceanic and Atmospheric Administration
(NOAA), on behalf of the United States Government, has a
requirement to carry out surveys of specific regions on the
continental margin (slope and rise) in the Pacific Ocean in
order to obtain <dataFormat>. These data are required to
support any potential claim for extended jurisdiction by the
```

Fig. 1. Example of the Area Level Template.

In order for the entire metadata generation system to work, two plain text templates are needed: the Area Level template (generated from the cruise report) and the Cruise Level template (generated from a script that processes the raw data files). These templates have evolved, since 2005, to contain exactly what NGDC needs for FGDC (Federal Geographic Data Committee) compliant data submission. The remainder of this paper describes each of the templates in detail and the Python script used to combine them together to get the resulting XML metadata.

II. The Area Level Template

It was noticed over the years that certain cruise datum remains constant each time an area is revisited (i.e. the name of the area, the map projection, certain keywords describing the features of the area, etc.). The same applies for cruise level data that is attached to each line item in a survey. Because of this, a template was created for both the area level of data and the cruise level of data. This reduces redundancy and promotes reuse of the templates. The Area Level Template can be reused each time work is done in that area. Likewise, the Cruise Level Template can be reused for each line of data that is collected during that cruise.

The Area Level Template (shown in Fig. 1) is created by using Microsoft Word in conjunction with an eXtensible Stylesheet Language Transformation (XSLT) and an XML Schema Definition (XSD) file (explained in separate sections below).

```
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform" version="1.0">
  <xsl:template match="/">
    #AREA INFORMATION (same across all metadata files for this area)
    Title: <xsl:value-of select="//title"/>
    PI_or_Chief_Scientist: <xsl:value-of select="//pi"/>
    area: <xsl:value-of select="//areaname"/>
    Purpose: <xsl:value-of select="//purpose"/>
    Abstract: <xsl:value-of select="//overview"/>
    instrument_mb: <xsl:value-of select="//mbesinstrument"/>
    Vertical_Positional_Accuracy_Report:<xsl:value-of select="//verticalposaccuracy"/>
    Horizontal_Positional_Accuracy_Report:<xsl:value-of select="//horizontalposaccuracy"/>
    Attribute_Accuracy_Report:
    instrument_seismic: <xsl:value-of select="//seismic"/>
    instrument_gravity: <xsl:value-of select="//gravity"/>
    Process_Description: <xsl:value-of select="//procdesc"/>
    Supplemental_Information: md5_checksum value: &lt;md5_checksum&gt;.  The cruise report
    Map_Projection:
    False_Easting: 0
    False_Northing: 0
    Latitude_Resolution: 0.00005
    Longitude_Resolution: 0.00005
    Place_Keyword_Thesaurus:NASA/GCMD Location Keywords
    Place_Keyword:Ocean -> Pacific Ocean -> Central Pacific Ocean
    Place_Keyword_Thesaurus:GEBCO Gazetteer of Undersea Feature Names
    Place_Keyword:
    Theme_Keyword_Thesaurus:Extended Continental Shelf Project Glossary of Terms
    Theme_Keyword:Foot of the slope
    Theme_Keyword_Thesaurus:None
    Theme_Keyword:Offshore Ridge
  </xsl:template>
</xsl:stylesheet>
```

Fig. 2. Example of the XSLT file.

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:simpleType  name="isodate">
    <xs:restriction base="xs:date">
      <xs:pattern value="YYYY-MM-DD"/>
    </xs:restriction>
  </xs:simpleType>
  <xs:simpleType  name="yeardate">
    <xs:restriction base="xs:date">
      <xs:pattern value="YYYY"/>
    </xs:restriction>
  </xs:simpleType>
  <xs:element name="cruisereport">
    <xs:annotation>
      <xs:documentation>Defines the cruise report as a collection of elements.<
    </xs:annotation>
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="cover" />
        <xs:element name="toc" type="xs:string"></xs:element>
        <xs:element ref="intro" />
        <xs:element ref="area" />
        <xs:element ref="systems" />
        <xs:element ref="data" />
        <xs:element ref="log" />
        <xs:element ref="references" />
        <xs:element ref="appendicies" />
      </xs:sequence>
    </xs:complexType>
```

Fig. 3. Example of the XSD file.

Figure 1 gives an example of how both the Area Level and Cruise Level Templates use name/value pairs, separated by a colon. The Python script uses the colon to identify where a label ends and the value for that label begins.

A. The Transformation File - XSLT

XSLT is a style sheet language for XML documents and describes how the data in an XML file will be transformed and rendered. Knowing the format for the desired end result is the first step to creating the transformation file. In this case, it is a very simple plain-text file, copying the layout of the Area Level Template. An example of the XSLT file is shown in Fig. 2.

B. Attaching the XSD (Schema)

An XML Style Definition (XSD - an XML schema Fig. 3) defines the layout of the cruise report with objects. Each object is a specific item that will be tagged - such as the ship's name, the principal investigator, etc. The XSD also defines the relationship between the objects, the data type for each element and can restrict the type of data allowed.

Once an XSD has been created, it needs to be attached to the Microsoft Word version of the cruise report.

This can be done once the "Developer" tab is visible. To make the "Developer" tab visible, click on the "Office Button", click on "Word Options" then in the "Popular" tab check the box "show Developer tab".

To attach the XSD to the cruise report, select the "Developer" tab and then click on "Structure", doing so expands a side bar (on the right) for the XML structure (Fig. 4). Click the link ("Templates and Add-Ins") in the XML Structure window, then click the "Add Schema" button and find the XSD file that will be used. Fill in the "Schema Settings" with any Uniform Resource Identifier (URI, this is just a tag name, and should be something descriptive) and an alias to shorten the URI, if desired. The schema selected will then be displayed in the list of available XML schemas; select it to make sure it is used. After OK is clicked, the list of elements within that schema is then seen in the XML Structure window (Fig. 5).

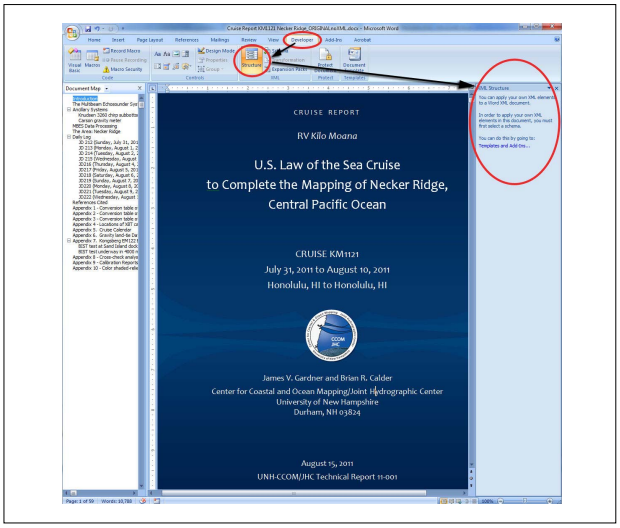

Fig. 4. Showing the Developer Tab in Microsoft Word.

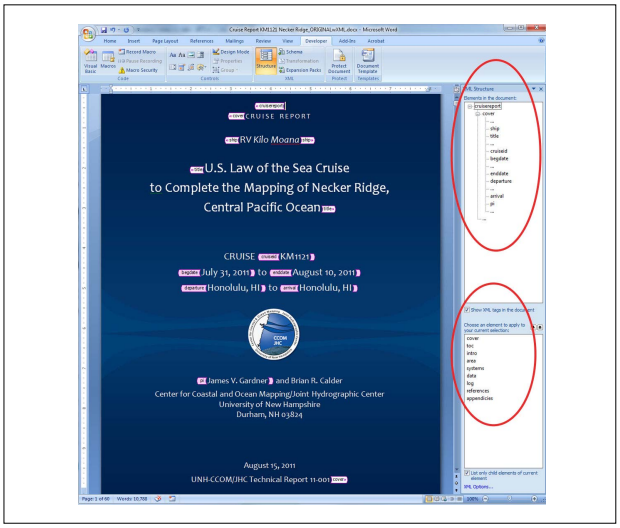

Fig. 5. Marking up the report.

C. Using the Schema to Mark Up the Cruise Report

Once the schema is attached, the lower window in the "XML Structure" window will be populated with the elements available for use. The upper window in the "XML Structure" window will be empty until a selection in the document is made and an element is applied to it from the hierarchical list of elements. Because of its hierarchical nature, tagging should follow the order of the layout and should start with the outermost tags first. So, for this example, start by selecting the entire cruise report (ctrl-A), then click on the desired element (i.e., "cruisereport") and apply it to the whole document. Doing this will wrap the selected element in the (<cruisereport>) tag. To view the tags as they are placed, go to the "XML Structure" window and check the box that says "Show XML tags in the document". As elements are applied their sub elements will be visible (i.e., when the "cover" tag is selected nine new tags are revealed) which can then be applied to other elements within the report (see Fig. 5).

Clicking on the elements in the XML Structure window highlights text in the report that is associated with that tag. Using the elements as a guide will ensure that everything needed for generating the metadata is tagged. There is also an option (when "Schema" on the Developer tab is clicked) in the "Templates and Add-ins" dialog box on the "XML Schema" tab to "validate document against attached schemas" which will help to ensure that all elements that should be used are used and in the correct order.

Since the XSD is built to emulate the structure of the report, the main structure tags should match headings and subheadings, making it intuitive to mark up and easier to find the data that needs tags. This ensures that the report is written in a consistent format and also makes it possible for another person to do the marking up.

D. Generating the Area Level Template

After the needed elements of the report have been properly marked up, it can then be transformed into the Area Level Template. To make the transformation, select "Save As" and choose "Word 2003 XML Document (*.xml)" as the "Save As" type. Word 2003 needs to be the selected type of XML document because they are uncompressed XML files. The 2007 format bundles images and other files together into a zip file. The "Apply Transform" box must then be checked and the "Transform" button to find the XSLT file must be clicked. The XSLT file must be applied to produce the plain text Area Level Template.

III. The Cruise Level Template

The Cruise Level Template (Fig. 6) is very similar to the Area Level Template; however, it contains two different sections separated by the first '---'marker. Each line of the cruise is separated by subsequent '---' markers. This portion of the template can be generated with any scripting language that processes the raw data files and parses out the desired data. The top portion is information that changes for each cruise and only three of the eight items need to be filled in by hand (the process

Fig. 6. The Cruise Level Template.

date, publication date and cell size); the others can be generated from the cruise report.

IV. Combining the Templates with the Python Script

The Python script is used to combine the Area and the Cruise Level Templates together in the FGDC format that NGDC has outlined on their website (http://www.fgdc.gov/metadata/csdgm/index_html). With a simple command, "%python areaFile.txt cruiseFile.txt", the script will take the area file and the top portion of the Cruise Level File and hold it in memory, then loop through the bottom portion of the Cruise Level File at each new '---' marker and apply the saved area and cruise data to it, generating one XML metadata file for each line of data collected during the cruise. In a matter of seconds, hundreds of files are created with the appropriate data, only changing what is specific to that line of data.

V. Conclusion

Generating metadata for each line collected in a bathymetric survey can be time consuming, error prone and tedious. This project used Microsoft Word to test the theory that metadata generation could begin with the cruise report needed at the end of each mission. Using the developer tools within Microsoft Word it was shown (with data from the Center for Coastal and Ocean Mapping/Joint Hydrographic Centers Law of the Sea cruise to Necker Ridge in 2011 [1]) that it is possible to extract metadata information from the cruise report. With the power of XML and XSLT the data extracted from the cruise report can be transformed into any format needed. The transformed data can then be combined with data extracted from each survey line to generate the final metadata file.

This system of metadata generation has made the metadata more consistent with the cruise report, helped to organize the cruise report, and ultimately has generated the metadata directly from the sources (the cruise report and the survey line data) - helping to reduce redundancy and errors.

ACKNOWLEDGMENTS

I thank Jim Gardner and Paul Johnson for their help in testing this theory and understanding the technical aspects of the cruise report, and Maureen Claussen for comments that greatly improved this manuscript.

REFERENCES

[1] J. V. Gardner, "U.S. Law of the Sea Cruise to Complete the Mapping of Necker Ridge, Central Pacific Ocean," UNH CCOM/JHC, Durham, 2011.