



NOAA Technical Memorandum NODC-NCDDC-1

Automated Metadata Generation Using Extensible Markup Language (XML) Techniques

Jacqueline Mize
Eric Roby
Kathy Martinolich
Lenny Collazo
David E. Sallis

January 2009

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
National Environmental Satellite, Data, and Information Service
National Oceanographic Data Center ■ National Coastal Data Development Center

Notice

Mention of a commercial company or product does not constitute an endorsement by the National Coastal Data Development Center. Use of information from this publication concerning proprietary products or the test of such products for publicity or advertising purposes is not authorized.

This publication should be cited as follows:

Mize J., E. Roby, K. Martinolich, L. Collazo, and D. E. Sallis (2009). Automated Metadata Generation Using Extensible Markup Language (XML) Techniques. NOAA Technical Memorandum NODC-NCDDC-1, National Coastal Data Development Center, Stennis Space Center, MS 39529.

Automated Metadata Generation Using Extensible Markup Language (XML) Techniques

Jacqueline Mize¹
Eric Roby²
Kathy Martinolich¹
Lenny Collazo³
David E. Sallis³

¹*Radiance Technologies, Inc.
350 Wynn Drive, Huntsville, AL 35805*

²*NOAA National Coastal Data Development Center,
Building 1100, Room 101, Stennis Space Center, MS 39529
www.ncddc.noaa.gov*

³*General Dynamics Information Technology
294 Thames Avenue, Bay St. Louis, MS 39520*



NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
National Environmental Satellite, Data, and Information Service
National Oceanographic Data Center ■ National Coastal Data Development Center

January 2009

Abstract

In 1990, the Office of Management and Budget (OMB) Circular A-16 called for Federal Agencies to create and maintain metadata, in accordance to the Federal Geographic Data Committee (FGDC) standards, for any spatial data that is collected, produced, acquired, maintained, distributed, used, or preserved (Office of Management and Budget 2002). President Bill Clinton signed Executive Order 12906 in 1994 to strengthen OMB Circular A-16, which was subsequently revised (Clinton 1994).

Metadata, the standardized documentation of data, comes in a variety of standards apart from the FGDC standard. Some standards predate the FGDC Content Standard for Digital Geospatial Metadata (CSDGM), and others were created to meet the specific needs of particular audiences. Many discipline-specific user communities, especially from the private and academic sectors, developed their own metadata standards—Directory Interchange Format (DIF), Ecological Metadata Language (EML), and International Organization for Standardization (ISO), to name a few. Metadata creation often is time consuming because many metadata standards are complex and difficult to implement. This variety of available standards has created some interoperability and compatibility issues. Many conventional metadata creation and validation methods in use today do not readily address interoperability issues.

Using Extensible Markup Language (XML) techniques to automate metadata creation provides a way to overcome numerous obstacles to producing and maintaining relevant metadata. Programmatic metadata generation provides many other benefits, such as reduced effort, enhanced accuracy, and improved efficiency.

Overall, automation of metadata using XML technologies proved successful and provided many benefits. Over 17,000 FGDC CSDGM compliant metadata records were produced relatively quickly requiring little resource. Conventional methods of creating this metadata, using current metadata editing tools and template techniques, would have taken much longer and would have required additional resources. Thus, the automation has far better results in terms of resources and time while increasing productivity.

Contents

ACRONYMS	4
INTRODUCTION	1
METHODS AND MATERIALS.....	3
TEST CASES.....	8
Florida Fish and Wildlife Conservation Commission	8
Louisiana Department of Natural Resources	8
RESULTS	8
CONCLUSIONS.....	9
REFERENCES	11

Figures

Figure 1. Metadata Automation Process	2
Figure 2. Simple concat function	5
Figure 3. Translation of source schema to target schema.....	5
Figure 4. User-defined functions show conversion of latitude and longitude	6
Figure 5. Constants added to mapping.....	6
Figure 6. XSLT generated from MapForce	7

Acronyms

ADO	ActiveX Data Objects
CSDGM	Content Standard for Digital Geospatial Metadata
DIF	Directory Interchange Format
EML	Ecological Metadata Language
FGDC	Federal Geographic Data Committee
FWRI	Fish and Wildlife Research Institute
GAME	Geospatial Assessment of Marine Ecosystems
GOS	Geospatial One-Stop
HTML	HyperText Markup Language
ISO	International Organization for Standardization
LA DNR	Louisiana Department of Natural Resources
MERMAid	Metadata Enterprise Research Management Aid
NCDDC	National Coastal Data Development Center
NODC	National Oceanographic Data Center
ODBC	Open Database Connectivity
OMB	Office of Management and Budget
PHINS	Priority Habitat Information System
SONRIS	Strategic Online Natural Resources Information System
SQL	Structured Query Language
XML	Extensible Markup Language
XSLT	Extensible Stylesheet Language Transformations
XSD	XML Schema Definition

Automated Metadata Generation Using Extensible Markup Language (XML) Techniques

INTRODUCTION

Any organization involved in data collection must carry out their due diligence with regard to addressing data stewardship. This ultimately involves the generation of metadata to serve as the official record for the data. There are many obstacles to creating and maintaining quality metadata. Producing and validating records against a particular standard can be both challenging and time consuming. More often than not, manually created records contain omissions and errors caused by poor record management tools and quality control measures. Some metadata standards—such as the Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM) (Federal Geographic Data Committee 1998)—are quite complex, necessitating the use of dwindling resources to train personnel in their proper use.

To protect the initial investment made, an organization must commit resources in maintaining the official record for their data. Falling short of this, an organization runs the risk of degrading future potential data sharing and discovery.

Some core issues that must be considered include interoperability between systems and user communities, and compatibility among different metadata standards. Organizations may need to distribute metadata in a variety of formats and standards to a diverse array of systems. For instance, an organization may have meticulously documented all of their data using the Ecological Metadata Language (EML) standard. Publishing to a clearinghouse, such as the Geospatial One-Stop (GOS), may be required. GOS requires that metadata be submitted using the FGDC or ISO standards (Geospatial One-Stop 2006). To meet this challenge, the organization must now address the requirement without posing a further burden on organization resources.

These solutions often lead to complex processes that become unsupportable. One such solution, crosswalks between standards, may address interoperability issues but usually results in the manual mapping of elements of one metadata standard to equivalent elements of another standard (Chan and Zeng 2006), usually within a spreadsheet application. Nonmappable elements are often left out, leading to a loss of information, or elements are mapped to nonequivalent elements and substandard metadata records are generated.

The majority of elements required for various metadata standards often already exist within databases or are digitally documented from other sources, such as data dictionaries or standard operating procedures. FGDC metadata is a compilation of information about a data set in a particular format. If this metadata information is already digitally stored, it might be pulled from these sources to populate the metadata record, avoiding duplication issues. Metadata creation in these cases would require transferring the information provided from the databases and other sources to the metadata standard elements. Format

conversions from the source to the target format may be necessary. It was theorized that a programmatic process would be the ideal method of creating metadata.

Metadata automation, the programmatic process of creating and updating metadata, is the key to providing accurate metadata while addressing the various challenges that an organization faces in balancing data stewardship needs with fiscal realities.

Automated metadata can be generated by using XML techniques. A representative document of what a source contains, i.e., a data model, can be mapped to a representative document of the desired output, or target. These representative documents, called *schemas*, can be created from XML, as subsequently described in detail. This mapping between the source and the target defines a *transform*. The transform is then applied to the source XML to create the desired output. Figure 1 is an overview of the process that the National Coastal Data Development Center (NCDDC) developed using XML techniques to automate metadata creation.

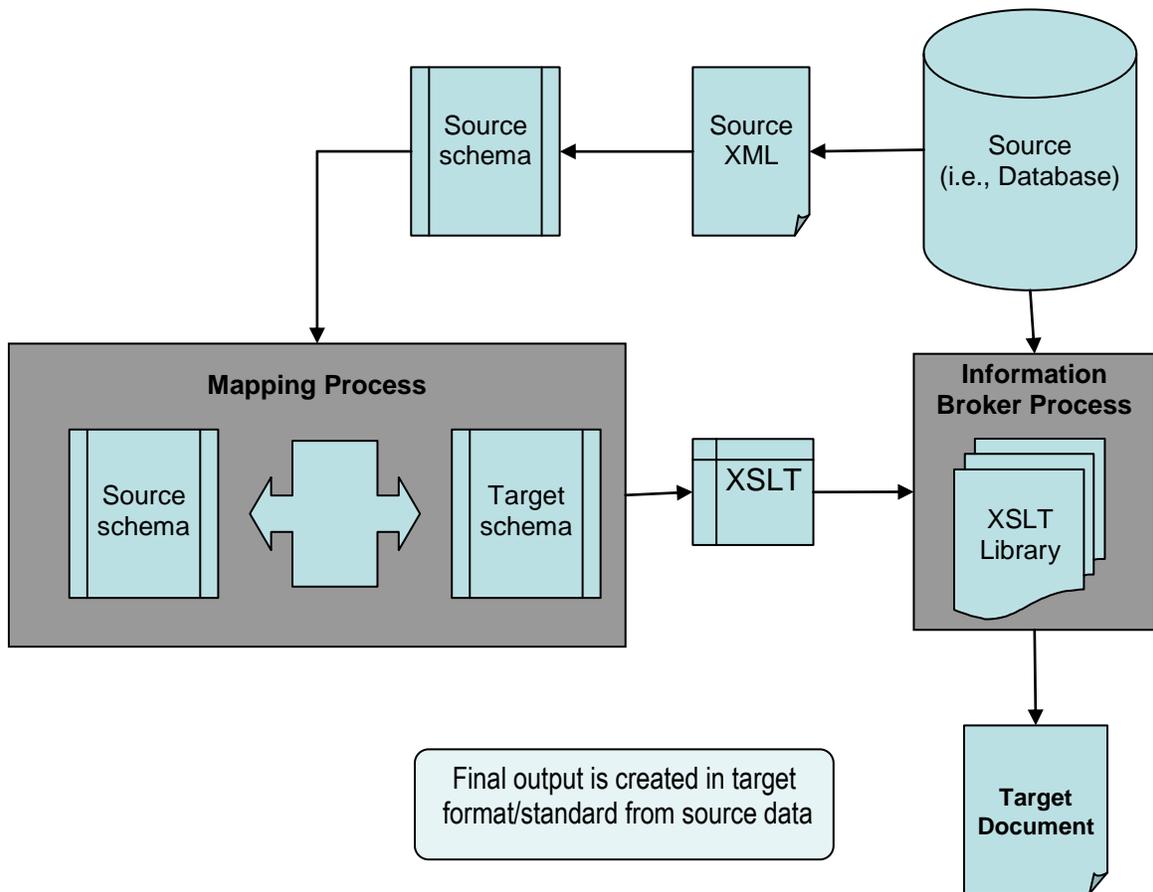


Figure 1. Metadata Automation Process.

METHODS AND MATERIALS

The use of XML technologies facilitates the exchange of complex structured data between applications. It protects the investment by being flexible and easily adapted to technological changes, as well as being interoperable across the spectrum (Tittel et al. 2002). It provides a rich set of standards that collectively support the automation of metadata, resulting in discrete process-steps that can be easily replicated and hosted on any platform. Additionally, XML tools are widely available that make the XML technology accessible to a broad range of implementers.

Microsoft Access[®] is used to develop the XML representation of the database. First, the data must either be imported into Access or a link created from Access and exported to the data source. Next, a query must be created that is representative of all the data attributes. For example, if the data has collection stations and observation values in one table and coordinates for the stations in another table, the two tables must be joined to generate an output that includes all the properties for the data. The EXPORT function from the FILE menu in Access can then be used to create the XML and the XML schema definition (XSD) files for the data. These files are mapped to the metadata attributes.

Schemas of the data models are used to map information from other metadata standards either from a metadata-like database or from databases. Schemas of the XMLs are needed to define the structure, content, and semantics of the XML documents. The schema represents the data's model and defines the objects, attributes, and relationships while defining the rules for the structure and content of the XML document (Altova 2003a). The proliferation of metadata schemas has provided a wide range to choose from as different communities attempt to meet the specific needs of users (Chan and Zeng 2006). Schemas can also be created from an existing XML document if there is a need for customization.

The development of schemas can be accomplished using an XML editor, such as XMLSpy[®] (Altova 2008a). XMLSpy supports many features, including the ability for a user to create a schema based upon connections to external relational databases. XMLSpy supports several of the most popular relational databases including Microsoft Access, Microsoft Structured Query Language (SQL) Server, Oracle[®], ActiveX[®] Data Objects (ADO) compatible, and some Open Database Connectivity (ODBC) databases. If no connection is possible, sample XML documents can be loaded into the editor and a schema can be generated (Altova 2003a).

The database schema, known as the source schema, can be mapped to the desired target schema. Mapping schemas can be accomplished with the use of a visual programming tool, such as XMLSpy's companion software, Altova MapForce[®] (Altova 2008b). MapForce is an essential integration tool for XML and database development that requires little or no programming knowledge and skill; however, a working knowledge of schemas and metadata standards is recommended.

Once the source schema and the target schema are selected, the mapping process can begin. MapForce contains multiple libraries with individual functions. Depending on the desired programming language output, supporting functions appear as boxes that can be simply selected and dragged into place within the mapping.

Some elements may have 1:1 relationships, but many do not. Simple functions may be required to produce the desired outcome. Concat is an example of a simple function that combines two or more elements from the source schema and places the result in a single element in the target document as shown in Figure 2 (Altova 2008b). The reverse process may also be needed. Single elements from the source can be divided through various string and logic functions and mapped to one or more target elements.

Selective elements from the source schema may need to be interpreted and translated to conform to specified standards of the target schema. As Figure 3 depicts, the user may select a field from the source schema, instruct how to interpret the input, and add the resulting interpretation(s) to the target schema.

Some mappings may require additional process steps not supported by the default library of functions. MapForce supports the creation of User-defined functions to address this need. Once defined, these new functions are available in the same manner as the default functions. To address the complexity of large process chains, user-defined functions can be developed which encapsulate several steps, reducing the visual clutter in the interface and improving readability of mapping details.

Complex mathematical functions and recurring translations are prime candidates that can benefit from creating user-defined functions. Figure 4 shows a commonly occurring conversion of latitude and longitude from degrees, minutes, and seconds to decimal degrees. Instead of creating a function with multiple steps each time the conversion needs to take place, these steps were added to the user function library. This provides a simple drag-and-drop function analogous to the MapForce default libraries for the selected output language.

Occasionally, the source schema does not contain all of the data needed to comply with the target schema. Constants may be required to create boilerplates, which are practical and convenient for static elements. Boilerplates are also useful for providing missing information for elements that do not map (see Fig. 5).

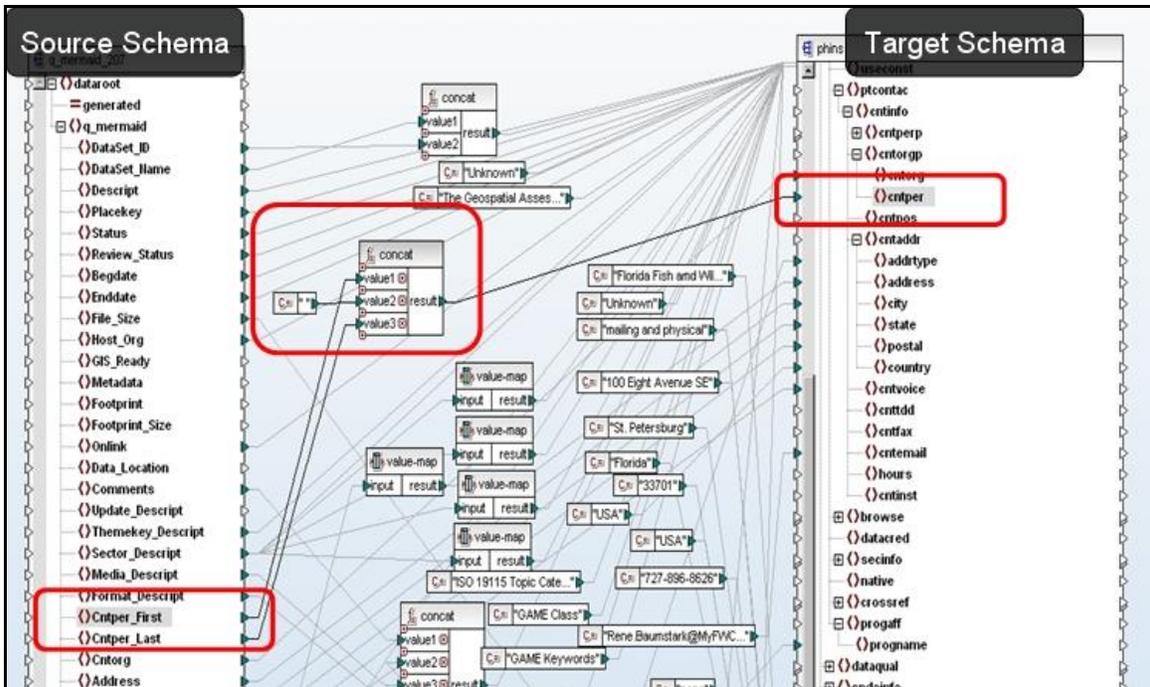


Figure 2. Simple concat function (from Altova 2008b).

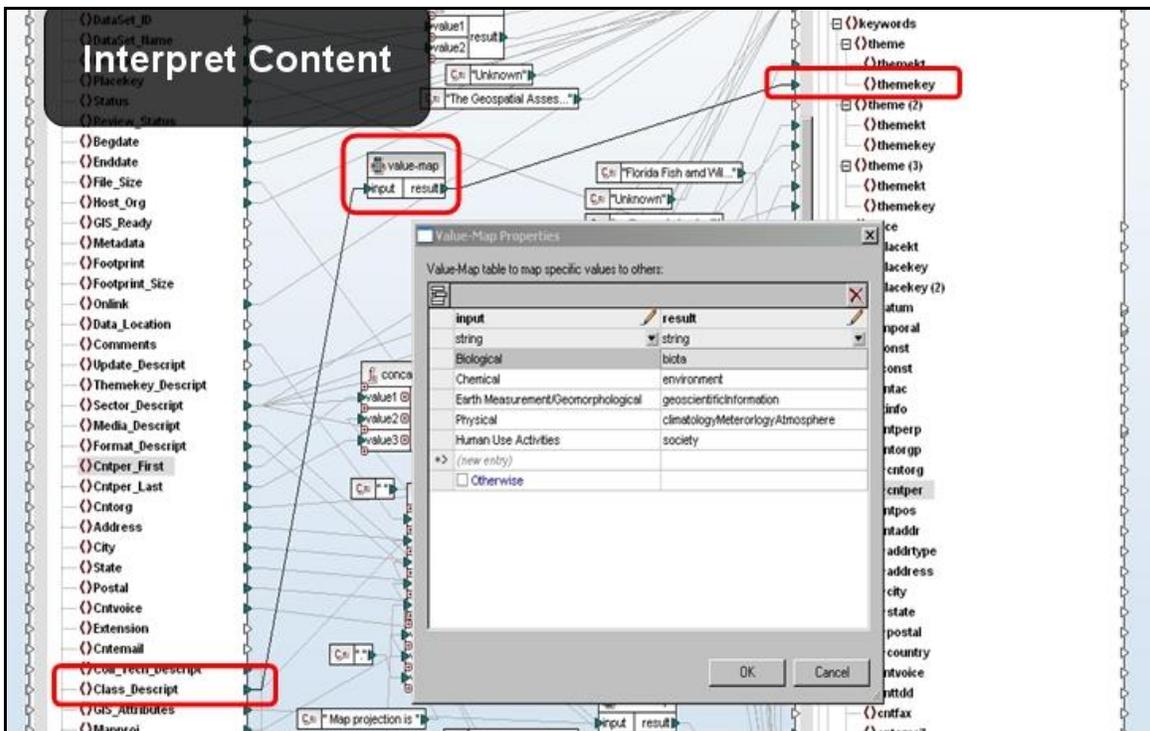


Figure 3. Translation of source schema to target schema (from Altova 2008b).

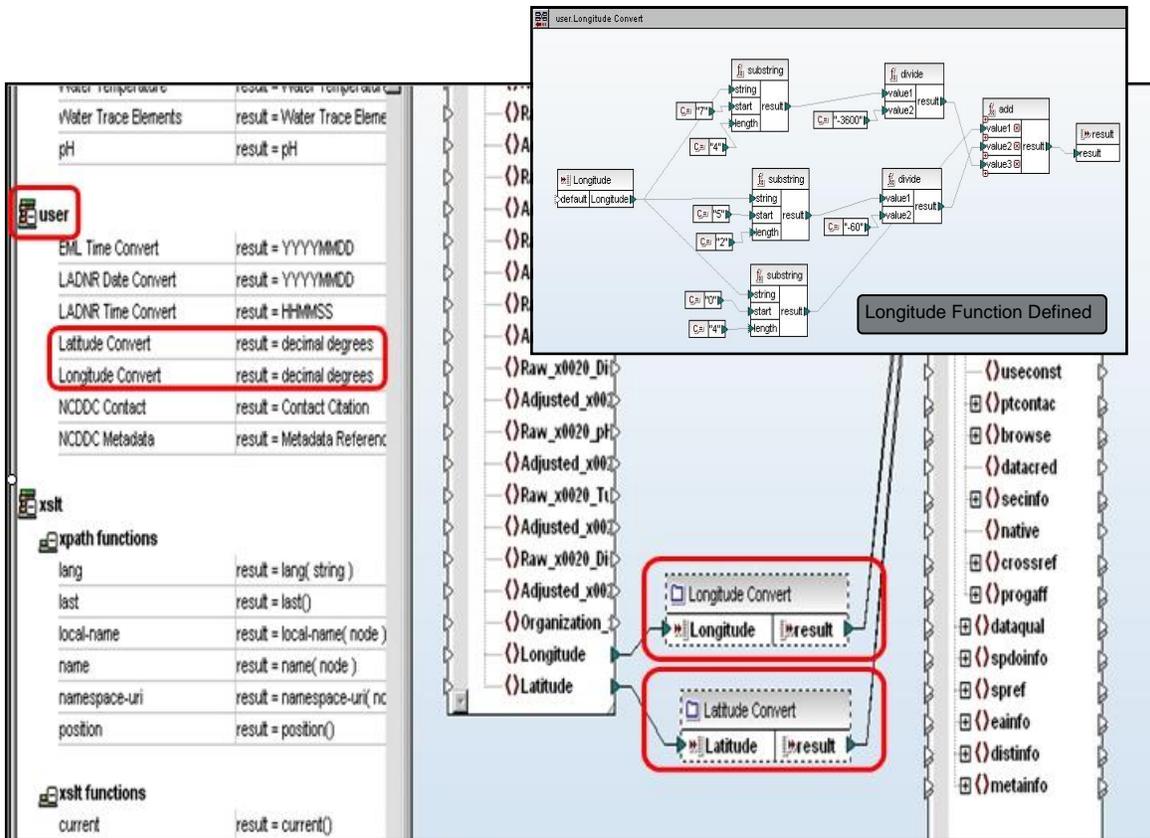


Figure 4. User-defined functions show conversion of latitude and longitude (from Altova 2008b).

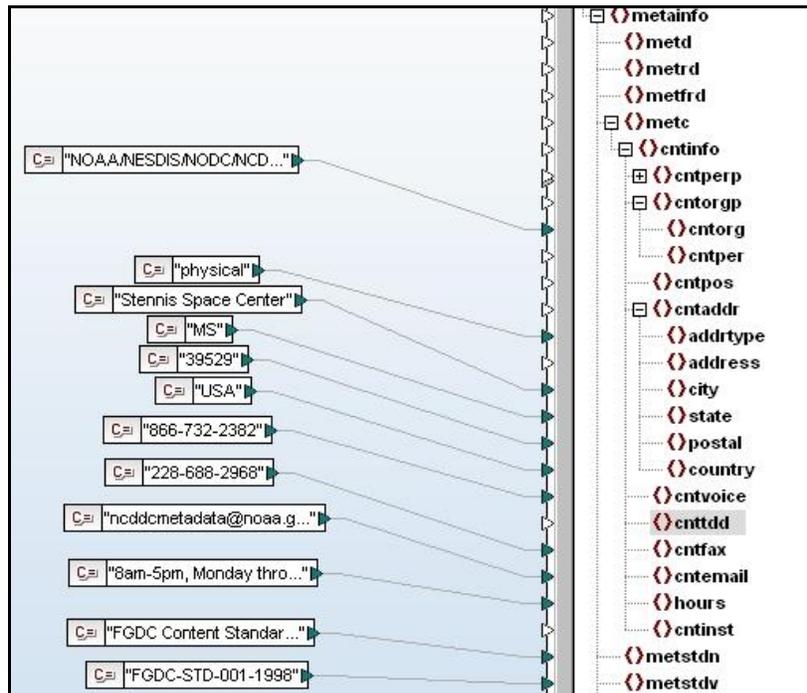


Figure 5. Constants added to mapping (from Altova 2008b).

```

1  <?xml version="1.0" encoding="UTF-8"?>
2  <!--
3  This file was generated by Altova MapForce 2008sp1
4
5  YOU SHOULD NOT MODIFY THIS FILE, BECAUSE IT WILL BE
6  OVERWRITTEN WHEN YOU RE-RUN CODE GENERATION.
7
8  Refer to the Altova MapForce Documentation for further details.
9  http://www.altova.com/mapforce
10 -->
11 <xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform" xmlns:xsi="
    http://www.w3.org/2001/XMLSchema-instance" xmlns:xs="http://www.w3.org/2001/XMLSchema"
    exclude-result-prefixes="xs xsi xsl">
12   <xsl:output method="xml" encoding="UTF-8" indent="yes"/>
13   <xsl:template match="/dataroot">
14     <metadata>
15       <xsl:attribute name="xsi:noNamespaceSchemaLocation">
16         <xsl:value-of select="'G:/Metadata/GAME_schema/JacI/Working/GAME/phins.xsd'"/>
17       </xsl:attribute>
18       <idinfo>
19         <citation>
20           <citeinfo>
21             <origin>
22               <xsl:value-of select="'Louisiana Department of Natural Resources (LDNR)'/>
23             </origin>
24             <pubdate>
25               <xsl:value-of select="'Unknown'"/>
26             </pubdate>
27             <xsl:for-each select="q_marsh_veg_2007">
28               <xsl:for-each select="Station_x0020_ID">
29                 <xsl:variable name="Vvar146_result" select="concat('Herbaceous Marsh Vegetation Data
    from Station ', )"/>
30                 <xsl:variable name="Vvar147_result" select="concat('$Vvar146_result, ' from the Louisiana
    Department of Natural Resources (LA DNR)')"/>
31               <title>
32                 <xsl:value-of select="$Vvar147_result"/>

```

Figure 6. XSLT generated from MapForce (from Altova 2008b).

Throughout the mapping process, MapForce checks the validation of the mappings against the assigned target schema. If errors exist (Altova 2003b), the generation, of an Extensible Stylesheet Language Transformation (XSLT), will abort. Once the mapping is complete and valid, the visual programming tool builds the transform (XSLT), which can support a variety of languages, as seen in Figure 6. MapForce supports creation of an XSLT in XSLT, XSLT2, XQuery, Java, C#, and C++ (Altova 2008b). The generated XSLT can be used either “as is” or further edited in an XML editor. Output from XSLTs can be XML, HTML, or plain-text documents as well.

Once the XSL transform is complete, it can be added to NCDDC’s Information Broker Service transform library. The Information Broker uses a third-party XSLT engine to perform XSL transformations. Clients of the Information Broker construct calls to its transform(s) method by providing the type of the incoming content (e.g., “eml”), the type of the resulting content (e.g., “fgdc”) and the XML content to be transformed. The result of this service call is the transformed content.

TEST CASES

The mapping process was used to fully automate the creation of valid FGDC-compliant metadata in two cases.

1. The Florida Fish and Wildlife Conservation Commission (FWC)

The FWC's Fish and Wildlife Research Institute (FWRI) had a large volume of records stored in their Geospatial Assessment of Marine Ecosystems (GAME) database. GAME records needed to be published to the Priority Habitat Information System (PHINS), a metadata clearinghouse that requires FGDC CSDGM compliant metadata records (Florida Fish and Wildlife Conservation Commission 2008).

An XML representation of the GAME database was used to create a schema of the database. The GAME schema and the FGDC CSDGM schema were loaded into MapForce. Sixty-six elements from the GAME database schema were mapped to the FGDC CSDGM schema. The remaining elements were static and boilerplates were created for constant elements, such as Metadata Standard Name and version.

An XSLT was generated once the target mapping passed validation and the resulting XSLT was added to the transform library within the information broker Service. The information broker programmatically processed the data from the database using the GAME XSLT. After several months, new data and updated contact information were added to the GAME database. The updated contents of the database was re-processed and submitted to the Information Broker transform service resulting in new FGDC records.

2. Louisiana Department of Natural Resources (LA DNR)

LA DNR's (2008) Strategic Online Natural Resources Information System (SONRIS) represented a large volume of coastal data that required FGDC CSDGM compliant metadata.

LA DNR's coastal data involved seven different databases, one for each data collection type. XML representations of each database were used to create schemas in XMLSpy. Seven XSLT's were built, one for each of the data types. Dynamic elements were mapped from the source to the target. Most of the information for the metadata records was thoroughly documented in various other documents. Stationary links to these documents were added to boilerplate elements as required. The resulting XSLTs were added to the transform library within the Information Broker Service.

RESULTS

FWRI GAME records resulted in the automated generation of 3,493 FGDC CSDGM metadata records. The resulting records contained much more information than the minimally required Identification Information and Metadata Reference Information

sections. The records also included Spatial Data Organization and Distribution Information sections.

The resulting metadata was subjected to various quality assurance/quality control (QA/QC) techniques. All 3,493 of the resulting metadata records were validated against the FGDC CSDGM schema using the NCDDC Metadata Enterprise Research Management Aid (MERMAid) tool. MERMAid rigorously checks for validation errors (National Coastal Data Development Center 2005). All 3,493 GAME records passed validation. Record content was visually inspected for erroneous errors.

These records were then made publically available via various metadata clearinghouses. As the database is edited, the automation process is performed periodically, and the updated records are broadcast to affected clearinghouse nodes, updating existing record inventories. The metadata process, on average, took approximately 0.8 seconds of real time per record. Time spent mapping the schemas to create the XSLT took one person familiar in metadata standards about one week to complete. The process has resulted in complete and current metadata that is publically searchable.

LA DNR records resulted in the automated generation of 13,565 FGDC CSDGM metadata records. The resulting LA DNR records contained much more information than the minimally required Identification Information and Metadata Reference Information section. The records also included Data Quality Information, Entity and Attribute Information, and Distribution Information.

The resulting metadata records were subjected to QA/QC techniques. Random sampling was conducted on four percent of the resulting LA DNR records to check for validation against the FGDC CSDGM schema, also using MERMAid. All randomly sampled records passed validation. Record content was visually inspected by several LA DNR and NCDDC staff for the randomly selected records.

Time spent mapping the schemas and creating the XSLTs took one metadata specialist about two weeks. The metadata generation process for LA DNR records, on average, took approximately 0.9 second of real time per record. Only a few exceedingly large files within one of the LA DNR databases took much longer to process than the rest. These rare cases are considered outliers and are not included in the average processing time of the majority of LA DNR files.

CONCLUSIONS

Automation of metadata using XML technologies proved to be successful and provided many benefits. Over 17,000 FGDC CSDGM compliant metadata records were produced quickly, dramatically reducing record management overhead for the organization.

Programmatic generation of metadata allowed for greater consistency among the records. The amount of errors and omissions can be limited by the automation process.

All records processed using the same transform will be processed in the same consistent manner.

Record maintenance was effectively reduced to maintaining the accuracy of the resulting XSLTs. Updates can be applied in one location and applied to all records. As changes occur at databases, such as corrections, additions, or deletions of data, the entire record inventory can be reprocessed programmatically. Automated metadata can be scheduled to rerun periodically as the database is updated so that the metadata remains current and accurately reflects any changes in the data. Any changes that affect the existing accuracy of the transform(s) can be updated within the transform(s), and the entire record inventory can be reprocessed programmatically. The Florida GAME trial case has confirmed that current metadata can be produced and maintained in this manner with minimal resource.

Interoperability between systems, interoperability between user communities, and compatibility among different metadata standards can be made easier through automation of metadata. Automation makes the transition to other metadata standards a manageable process. Multiple transforms can be applied to a source to create output in a variety of formats and standards. Granularity of metadata records, either collection level or individual records, can be addressed programmatically, depending on how the database is queried. The current process generates a metadata record for each row in the query output. If the number of records is too large, then the EXPORT tool in MS Access cannot handle the process, and the output has to be divided into multiple sets by changing the conditions in the query. For example, if the data includes multiple years, the query can be adjusted to create separate files for each year.

Direct connections to the databases and subsetting large files could improve efficiency and reduce processing time. No direct connections to the databases were established for the test cases because of the precautionary measures taken for testing purposes. The process of creating the XML and XSD files for the data is simple, but can be time consuming because the EXPORT tool is extremely slow when dealing with a large number of records.

Adding conditions to the query for limiting the number of records to generate the metadata conversion map could possibly decrease the processing time. Exceedingly large records may also increase processing time.

The greatest potential for error occurs during the mapping process. The users' level of familiarity with the source data and the target schema greatly affect the accuracy. The success of this process hinges ultimately on the accuracy of the mapping effort. At this stage of the process, it is vital to utilize QA/QC techniques and review to achieve quality products.

REFERENCES

- Altova (2003a). *XmlSpy 2004, Enterprise Edition: User & Reference Manual*. United States of America: Altova GmbH & Altova, Inc., Beverley, MA.
- Altova (2003b). *MapForce 2004. User & Reference Manual*. United States of America: Altova GmbH & Altova, Inc., Beverley, MA.
- Altova (2008a). MapForce [Computer software]. Professional Edition, version 2008 sp1. Altova, Inc., Beverley, MA.
- Altova (2008b). XMLSpy [Computer software]. Professional Edition, version 2008 sp1. Altova, Inc., Beverley, MA.
- Chan, Lois Mai, and Marcia Lei Zeng (2006). Metadata Interoperability and Standardization—A Study of Methodology, Part I. Achieving Interoperability at the Schema Level. *D-Lib Magazine* 12 (6), Accessed 19 Dec 2008. <<http://www.dlib.org/dlib/june06/chan/06chan.html>>.
- Clinton, Bill (1994). United States. Executive Order 12906 of April 11, 1994. Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure . Federal Register, GPO, 1994. *Federal Register* 59 (71), <<http://www.archives.gov/federal-register/executive-orders/pdf/12906.pdf>>. Accessed 22 Dec 1998.
- Federal Geographic Data Committee (FGDC). FGDC-STD-001-1998. Content standard for digital geospatial metadata (revised June 1998). Federal Geographic Data Committee. Washington, D.C.
- Florida Fish and Wildlife Conservation Commission, Fish and Wildlife Research Institute (2008). *Florida GAME*. http://research.myfwc.com/features/category_sub.asp?id=6899.
- Geospatial One-Stop (2006). Creating and Publishing Metadata in Support of the Geospatial One-Stop and the NSDI. *Geospatial One-Stop* 31 July 2006. <<http://www.geodata.gov/gos/metadata/CreatePublishMetadata.pdf>>. Accessed 6 Jan 2009.
- Louisiana Department of Natural Resources (2008). SONRIS. <http://sonris-www.dnr.state.la.us/www_root/sonris_portal_1.htm>. Accessed 22 Dec 2008.
- MERMAid. (2005). [Computer software]. version 1.2. NOAA National Coastal Data Development Center, Stennis Space Center, Mississippi.
- National Aeronautics and Space Administration (2008). Directory Interchange Format Writer's Guide (2008). Global Change Master Directory. <<http://gcmd.nasa.gov/>>. Accessed 21 Jan 2009.

NOAA Coastal Services Center (2008). CSDGM XML Schema Document Representation. <<http://www.fgdc.gov/metadata/fgdc-std-001-1998.xsd>>. Accessed 22 Dec 2008.

Office of Management and Budget (2002). Coordination of Geographic Information and Related Spatial Data Activities. OMB Circular No. A-16, Revised 19 Aug 2002. <http://www.whitehouse.gov/omb/circulars_a016_a016_rev/#1>. Accessed 22 Jan 200.

Tidwell, Doug (2008). *XSLT: Mastering XML Transformations* (2nd ed.). Cambridge, MA: O'Reilly.

Tittel, Ed, Natanya Pitts, and Frank Boumphrey (2002). *XML for Dummies* (3rd ed.). New York: Hungry Minds, Inc.