ELSEVIER

Computers & Geosciences 32 (2006) 1646–1653

COMPUTERS & GEOSCIENCES

www.elsevier.com/locate/cageo

Using XML to improve the productivity and robustness in application development in geosciences☆

Ulisses T. Mello*, Liqing Xu

IBM T. J. Watson Research Center, 1011 Kitchawan Road, Yorktown Heights, NY 10598, USA

Received 4 November 2005; received in revised form 12 March 2006; accepted 12 March 2006

Abstract

In this paper, we describe an approach to apply Extensible Markup Language (XML) technologies to improve the robustness of geological and geophysical applications as well as to increase the efficacy in the application development process. Geological and geophysical applications are often data centric, I/O intensive and their development is incremental. Therefore, significant amount of development resources is devoted to the design and reengineering of the container data structures that store data. This process is time consuming, mechanical and error prone. Normally, ad hoc parsers are necessary for reading inputs, as well as numerous filters, or adapters to transform the data for integration with other legacy applications. Most of this can be avoided by using XML-related technologies. XML has a type system schema that can be used to define input parameters and constraints. The XML parser can validate the input data using the constraints defined in the schema. Exporting results in XML format allows the use of Extensible Stylesheet Language Transformations (XSLT) to transform XML output to any other format necessary for integration with legacy applications. Additionally, XML-data binding code can be automatically generated in specified languages such C++ and Java. We used this approach to develop applications for seismic ray-tracing and basin modeling with great success, and the major benefits of this approach were the significant gains in productivity during the developement and application robustness.

Keywords: XML; C++ binding; Basin modeling; Ray-tracing; Application development

1. Introduction

Geological and geophysical applications are often data centric, I/O intensive and their development is incremental. Therefore, a significant amount of development resources is devoted to the design and reengineering of data structures that store data, called containers. This process is time consuming, mechanical and error prone. In this paper, we describe an approach to apply Extensible Markup Language (XML) technologies to improve the robustness of geological and geophysical applications as well as to increase the productivity in the application development process. XML is becoming the de facto format for information exchange among applications over the internet (see www.xml.org). In addition, industries are defining open standards for data exchange using XML (e.g.,

☆Code available from server at http://www.iamg.org/CGEditor/index.htm.

*Corresponding author. Tel.: +1 914 945 1268; fax: +1 914 945 3434.

E-mail addresses: ulisses@watson.ibm.com (U.T. Mello), liqingxu@us.ibm.com (L. Xu).

0098-3004/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2006.03.003

WITSML for well site information transfer at www.witsml.org). XML is powerful because it provides self-describing tags and extensibility for defining documents. Many standards have been defined around XML such as Extensible Stylesheet Language (XSL), Extensible Stylesheet Language Transformations (XSLT), XML Query Language (XQuery), XML Path Language (XPath), and the Extensible HyperText Markup Language (XHTML).

The structure, content, and semantics of XML document is defined by an XML document type definition, or a schema language, which can be used to parse and validate XML documents according to schema rules. XML Schema languages such as W3C XML Schema, Document Type Definition (DTD), Schematron and Relax-NG can be used for this purpose (see http://www.oasis-open.org/cover/schemas.html for a more detailed survey of schemas). In this paper, we used the W3C XML Schema, which is becoming the most widely used and, for the sake of simplification, we refer to it in the rest of the text as XML schema.

There are many advantages of using XML over other document formats. It is simple and its regular and the consistent notation make it easier to build a program to process XML data. It is an open standard and many tools have been developed to parse, and manage these documents. It is extensible and anyone can create new tags that can be shared. It is self-describing, and its tags are meaningful. It has built-in support for international applications including texts in all the world's alphabets. It provides methods to declare and enforce document structures, contents and semantics. It enables platform, storage, and application independent data interchange, and finally, its plain text structured nature is a benefit to developers.

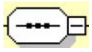
In geosciences, specific XML documents specifications have been defined to promote standards for parties to exchange documents. POSC (www.posc.org), *the Petrotechnical Open Standards Consortium*, is an organization to define open specifications for petrotechnical documents to address Exploration & Production information challenges and opportunities. POSC has since published XML document specifications for well site information data (WITSML), and production data (PRODML). The Open Geospatial Consortium, Inc. (OGC) (http://www.opengeospatial.org) is leading the development of standards for geospatial and location based services. The Commission for the Management and Application of Geoscience Information (CGI) (https://www.seegrid.csiro.au/twiki/bin/view/CGIModel/WebHome) is working on the design of **GeoSciML**. It is clear that more specifications are coming sooner than later.

Next we describe an approach to apply XML technologies to improve the robustness of geological and geophysical applications as well as to increase the efficacy in the application development process. In Section 2, we briefly introduce some XML basic information and resources. In Section 3, we discuss the data binding and how it can be automated for programming languages. In Section 4, we present an application case and in Section 5 we present the conclusions.

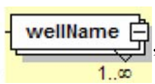
2. XML basics

XML documents are often heterogeneous and hierarchical in structure and have intrinsic order. For a good introduction to XML see Harold and Means (2004). The data model of an XML document can be visualized as a tree. A node on the tree can be an element, attribute, text, document, comment, processing instruction or a namespace. Fig. 1 shows a fragment of XML schema definition of a well log document from POSC using XML Spy notation. The notation of this graphic representation of a schema is similar to UML notation (Carlson, 2001). In this figure, each box icon represents either a complex element such as 

in open view mode or in collapsed view mode such as 

and the icon  indicates that the element contains a sequence of child elements. The icon that appears in the top left of the wellId element 

indicates that this element contains text data. An icon with notation 

, which appears at the botton of the wellName element 

, indicates that this element must appear at least once and can repeat as many times as necessary. The icon  indicates that only one of the possible elements in the list (GeopoliticalLocation, OffshoreLocation, SurveyLocation, and LegalDescription) can be used. In Listing 1, we show a fragment of this XML document with some details of well information such as operator name, well name, well location etc. In Listing 2, we show a fragment of the XML