Grammars and Parsers for Validating Binary File Formats

WILLIAM UNDERWOOD

Abstract: Automated tools are required for identifying and validating the formats of the huge number of files ingested into digital data and record archives. Data description languages for describing file formats have emerged to address these challenges. Projects such as JHOVE (JSTOR/Harvard Object Validation Environment) are emerging to create collections of programs for validation of file formats. This presentation reports progress in addressing the research question: Is it possible to extend the context-free grammars used to specify the syntax of programming languages to the specification of binary file formats and to use these grammars with parsers for validating the file formats of binary files?

This presentation describes the extension of context-free grammars from textual languages to binary files. Context-free binary file grammars are being used to specify such binary file formats as AVI, WAVE, WMA, WMV, BIFF (Microsoft Excel XLS), JPEG and TIFF. Examples of grammars that specify binary file formats are presented and it is explained how a parser generator can generate a program for validating a binary file format from the grammatical specification of the format.

This approach differs from that of data description languages, which are used to describe file formats, in that it specifies the format as a formal grammar. None of the descriptions of file formats created using data description languages are formal grammars. This approach differs from other projects that seek to develop tools for validating file formats in that it seeks the capability to generate a validation program directly from a grammatical specification of a binary file format through the use of a parser generator.

Acknowledgement: This research is sponsored by the Applied Research Division of the National Archives and Records Administration under Army Research Office Cooperative Agreement W911NF-10-2-0030.

About the author:

Dr. William Underwood is a Principal Research Scientist with the Information Technology and Decision Support Division of the Georgia Tech Research Institute in Atlanta. He earned his PhD in Computer Science from the University of Maryland. His current research interests are in developing formal, theoretical foundations for records management and archival science, experimental investigations of alternative digital preservation strategies, and the application of natural language processing, machine learning and knowledge-based reasoning technologies to the support of automated archival description, Freedom of Information Act (FOIA) review, and search and retrieval of records in digital archives. Dr. Underwood's research, sponsored by the National Archives and Records Administration (NARA), has involved prototyping of an archival repository and archival

processing system for accession, arrangement, preservation, review, description and retrieval of electronic records. His current research, sponsored by the Applied Research Division of NARA, seeks to develop tools for validation of and metadata extraction from binary file formats, reliable identification of file formats, automatic description of the content of record series, and decision support for review of Presidential records for Presidential Record Act restrictions and FOIA exemptions.