Investigations of Data Representation

From 2006-2011, this work was supported as part of “Innovative Systems and Software: Applications to NARA Research Problems”, a Cooperative Agreement between the US National Archives and Records Administration (NARA) and the National Center for Supercomputing Applications (NCSA) at the University of Illinois, Urbana-Champaign.

NCSA’s goals for the “Investigation of Data Representation” task focused in three areas: developing an open-source parser for DFDL with an emphasis on enhancing performance and scalability, integration of DFDL/Defuddle into the SHAMAN preservation architecture, and exploration of practical applications of DFDL and the development of a library of DFDL descriptions.

Intellectual Merit

Preservation can be thought of as communication with the future. The records we preserve today need to be accessible and displayable by future technology. Beyond maintaining the accessibility of the raw bits of the digital data, preservation requires maintaining an ability to interpret the data as meaningful structures, relationships, and visual representations.

This work contributed to the development of a preservation system that would dramatically lower the per-file-format effort required for preservation. In particular, we are contributing to the development a format description language (the Data Format Description Language) and format-independent parser (Daffodil) to support interpretation of arbitrary binary or ASCII formatted files in terms of well-defined logical models.

The explicit, declarative, descriptive model we are developing through this project significantly reduces the amount of machine and operating system dependent software that must be maintained to preserve access to file content and minimizes the effort needed to support new formats.

Broader Impacts

While preserving access to file content is a primary motivation for the development of DFDL and Daffodil, they are useful across the curation and preservation process and more broadly in e-Science in general.

The Technology

The Data Format Description Language (DFDL) Standard

Our team has participated in the development of the Data Format Description Language (DFDL), a new standard specification from the Open Grid Forum, released in January, 2011 http://www.ogf.org/dfdl/.

The DFDL is a language to describe existing data formats, both binary and text, in a manner that makes the data accessible through generic mechanisms. The DFDL specification is based on the XML Schema (http://www.w3.org/XML/Schema.html), which is used to define the structure and semantics of XML documents and to annotate schemas for the benefit of human readers and applications. The input is a sequence of bytes and the output is an XML Information Model.

For more information about DFDL, see Related Links below

Parser Development

In previous work, Talbott and others at Pacific Northwest Labs developed the Defuddle parser, which implemented an early version of the DFDL specification [6]. Subsequently, this project updated and extended the Defuddle parser [1-3].

At the time of the release of Version 1 of the DFDL Specification, we reviewed the Defuddle parser, and determined that it needed to be completed revised [4].

The Daffodil parser is a completely new implementation, based on Version 1 of the DFDL, as well as lessons learned from Defuddle [5]. The Daffodil parser is only partly implemented. It is released “as is” in August 2011.

Semantic Extensions

While the XML Schema language is well suited for describing the layout of data (the "syntax"), interoperability and robust archiving require semantic mark up as well. This project will extend the DFDL model to support mapping to semantic web languages (the Resource Description Framework (RDF) and the Web Ontology Language (OWL)). We are exploring a two-step mechanism based on the use of the Gleaning Resource Descriptions from Dialects of Languages (GRDDL) specification http://www.w3.org/TR/grddl/ to associate XML to RDF mapping instructions, written, for example, in XSLT, with the DFDL description file [2,3].

Team Members and Alumni

- Robert E. McGrath, NCSA
- Joe Futrelle, Woods Hole Oceanographic Institute
- Jim Myers, RPI
- Alejandro Rodriguez, Amazon
- Jason Kastner, NCSA

Acknowledgements

- U.S. National Achives and Records Administration (NARA)
  This work was supported through National Science Foundation Cooperative Agreement NSF OCI 05-25308 and Cooperative Support Agreements NSF OCI 04-38712 and NSF OCI 05-04064 by the National Archives and Records Administration.
- The EU Sustaining Heritage Access through Multivalent Archiving (SHAMAN) project, http://shaman-ip.eu/shaman/.

Citations

1. McGrath, R.E., J. Kastner, A. Rodriguez, and J. Myers. ``Defuddle: a Tool for Format Translation and Metadata Extraction (Poster)''. *Microsoft E-science Workshop* (2009).
2. McGrath, R.E., J. Kastner, A. Rodriguez, and J. Myers. ``Experiments in Data Format Interoperation Using Defuddle'', National Center for Supercomputing Applications, June, 2009, http://www.archives.gov/applied-research/ncsa/11-experiments-in-data-format-interoperation-using-defuddle.pdf
3. McGrath, R.E., J. Kastner, A. Rodriguez, and J. Myers. ``Towards a Semantic Preservation System'', National Center for Supercomputing Applications, June, 2009, http://arxiv.org/abs/0910.3152.
4. Rodriguez, A. and R. E. McGrath, ``Some Notes of comparison between DFDL and Defuddle''. National Center for Supercomputing Applications, October, 2010.
5. Rodriguez, Alejandro and Robert E. McGrath, ``Daffodil: A New DFDL Parser''. National Center for Supercomputing Applications, October, 2010
6. Talbott, T. D., K. L. Schuchardt, E, G. Stephan, and J, D. Myers, ``Mapping Physical Formats to Logical Models to Extract Data and metadata: The Defuddle Parsing Engine'', *International Provenance and Annotation Workshop*. 2006, Springer: Heidelberg. p. 73-81.

Related Links

- OGF Standards: Data Format Description Language (DFDL), http://www.ogf.org/dfdl/.
- The Open Grid Forum Data Format Description Language WG (DFDL-WG), http://redmine.ogf.org/projects/dfdl-wg.
- wikipedia, "Data Format Description Language". 2011, http://en.wikipedia.org/wiki/Data_Format_Description_Language.
- Defuddle (Old) Examples and code: http://sourceforge.net/projects/defuddle.