IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITB12855

This paper appears in the book, *Emerging Trends and Challenges in Information Technology Management, Volume 1 and Volume 2* edited by Mehdi Khosrow-Pour © 2006, Idea Group Inc.

A Digital Preservation Ingest Parsing Service for Complex Data Objects

Donald F. Flynn, Pacific Northwest National Laboratory, PO Box 999 K7-28, Richland, WA 99352,
T: 509-375-2570, F: 509-375-2443, Don.Flynn@pnl.gov

INTRODUCTION

Large scientific research efforts whether they are in academia, government or the private sector are usually comprised of diverse and geographically separated teams that collect data using various methods. The information that comprises a scientific study is often created in large quantities and vastly different formats and structures. Results may be produced from different software applications using a range of protocols, databases, and operating systems, creating a "Complex Data Object".

The computing industry, and more precisely, the field of knowledge management is faced with the significant challenge of how to best administer these digital assets in a fashion that provides efficient authoring and collaboration capabilities in the distributed team environment while still maintaining the historical scientific record. While ongoing efforts are addressing the problem of wide-area data management, less attention has been given to accurately capturing the scientific record of interest. In December 2004, at a Department of Energy (DOE) National Collaboratory Workshop, the "fossil record of science" was identified as a critical issue [1].

The purpose of this research is to prove that the complex data object (CDO) can be transformed into a Digital Preservation Object (DPO) in order to meet the requirements for data provenance and long-term archival. The second step of this research is to then prove that the inverse is also true, that a DPO can be extracted and transformed back to the same state as the original CDO for future inquiry.

ELECTRONIC LABORATORY NOTEBOOK

A CDO is comprised of some unique and distinguishable characteristics. While it can appear as a simple taxonomy containing pointers to flat data files, this is a trivial case and not where our interest lies. Instead, the real world is much richer in context, and we must base our work on a richer CDO that offers unique and interesting challenges.

In 1997, the Department of Energy launched an initiative called, "DOE2000" with the goal being "to bring innovation to, and accelerate the development of, communications systems, computational capabilities, and collaboration strategies that current and emerging technology make possible" [2]. As part of this initiative, the Pacific Northwest National Laboratory, along with Lawrence Livermore and Oak Ridge National Laboratories, developed the Electronic Laboratory Notebook (ELN) in an effort to provide an electronic alternative to the standard paper-based notebook.

Currently the ELN is a publicly available open source shareware product with over 1200 downloads to date. It is being used in academia and government laboratories across the country in such research areas as high energy physics, chemistry, biology, and molecular sciences. The ELN client application is a java based program that runs on any system with Java 1.4+ plugins enabled. It runs with the new Scientific Annotation Middleware (SAM) server or older Perl based server (both also publicly available) [3].

The ELN is based on an ontology that meets the definition of a rich CDO. It is structured in a top-down hierarchy where a notebook contains multiple chapters, and chapters may contain multiple pages. Each page has attributes that may consist of text documents, binary documents, relational databases, object oriented databases, static text (notes), images, or data from other software applications [4].

DIGITAL PRESERVATION OBJECT

Transforming a CDO to a DPO requires a standard format for storing not only the data from the CDO, but the preservation metadata as well. In order to meet the requirements set forth by the CDO, the digital preservation standard format needs to contain a robust ontology capable of handling all the possible attributes of the CDO. Likewise, the digital preservation standard must be exactly that, a standard, which all digital preservation systems can ingest and interpret correctly.

In order to responsibly preserve information for the long term, the digital asset itself needs to be documented with enough detail to provide for future retrieval and assembly. To accomplish this, three broad categories of data must be defined; descriptive metadata describing with the asset content pertains to, structural metadata describing how an asset is constructed and technical preservation metadata describing the data provenance.

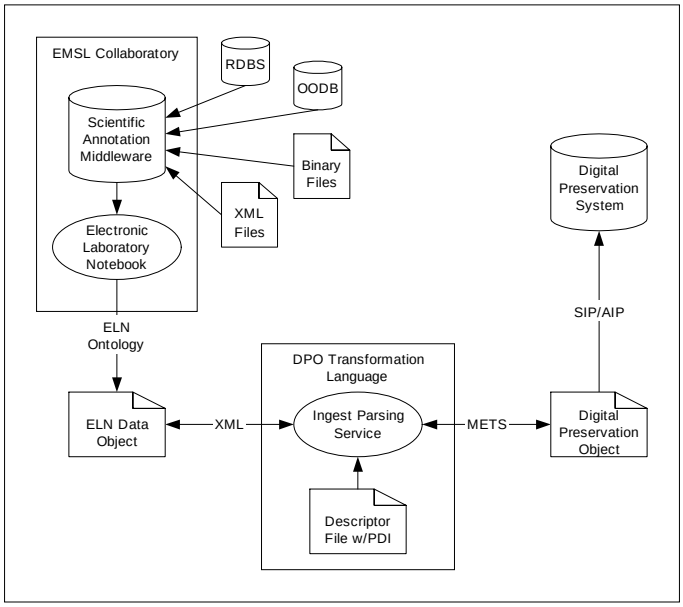
The Metadata Encoding & Transmission Standard (METS) is a Digital Library Foundation initiative and Library of Congress standard that provides an XML document format for encoding and exchanging digital asset metadata. A METS document is comprised of the following seven major sections; METS Header, Descriptive Metadata, Administrative Metadata, File Section, Structural Map, Structural Links, Behavior [5].

The METS format is flexible enough to support the requirements of the CDO and facilitate the management of those objects. Further, since it uses the XML schema language of the World Wide Web Consortium, it easily supports adaptations and provides the flexibility needed for future evolution. A good practical example of an implementation is the National Digital Newspaper Program (NDNP) which has extended METS with custom schema representations as an alternative solution for digitally preserving and archiving historical newspapers [6].

Complementing METS is the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), which provides an application independent interoperability framework based on metadata harvesting [7]. The Open Archives Initiative is funded by the Digital Library Foundation, the Coalition for Networked Information, and the National Science Foundation. Their mission has been to develop and promote interoperability standards primarily for the dissemination of content. It is important to obtain a basic knowledge of OAI-PMH and how it is designed in order to understand why the METS ontology is built the way it is.

OAI-PMH calls for three distinct types of "information packages". An information package is the complete METS document and supporting files. The Submission Information Package (SIP) is created by the information producer and is ingested into the digital preservation system either by the system itself or some external means. The Archival Information Package (AIP) is stored in the digital preservation system as an extension of the SIP, where the AIP includes preservation metadata for the digital asset. Lastly, the Dissemination Information Package (DIP) is a formalized method for retrieving and presenting an end user with an archived digital asset.

Figure 1.



RELATED WORK

In 2000, the Massachusetts Institute of Technology (MIT) and Hewlett-Packard joined forces in an attempt to attack the problem of maintaining and sharing digital content over the long-term. The result of this combined effort was the development of "DSpace", an open-source application that accepts digital materials, makes them available over the web, and stores them in a data management system to help preserve them for years to come [8].

The DSpace platform also supports the OAI-PMH version 2.0 as a data provider. This is accomplished by using various implementations of the Online Computer Library Center (OCLC) OAICat Framework. In addition, DSpace supports the use of hierarchical sets, where records can be contained in zero or more sets. DSpace exposes collections as sets, and since the organization of collections is likely to change over time, it is a less stable basis for selective harvesting. DSpace also has an incomplete, experimental METS export tool, planned for dissemination information packages.

DSpace is related to this research effort since it is the only practical implementation of a digital preservation system that uses the OAIS framework and accepts METS documents. While it establishes a baseline for this research, it is not yet fully functional and has not addressed the issues of the CDO.

INGEST PARSING SERVICE

The goal of this research is to prove that a "language" can be constructed to transform a CDO into a DPO. Inversely, the research will strive to also prove that the same "language" can transform the DPO back to the original CDO. This will be a contribution to the field that does not exist to date and serve as a basis for further research and implementations.

As part of this research effort, an XML web service has been written to act as an "Ingest Parsing Service". This service helps test the theory of the transformation concept and acts as a practical implementation. The service takes as input two files: an XML representation of the CDO and a descriptor file with Preservation Description Information (PDI) which defines the mapping between the CDO and the DPO.

The DOE2000 ELN serves as the source of the CDO since it is open source software and contains many attributes of a good ELN system. The ingest parsing service contains the logic necessary to transform the CDO into a DPO based on the METS ontology. METS defines what is necessary for both a SIP and an AIP, which prepares the DPO for inclusion in any digital preservation system that supports the standard OAI-PMH. Figure 1 provides an overview of the architecture.

The DPO transformation language is based on the principles of abstract algebra, and applying this, allows us to show that elements from a set *A* (such as the CDO) can be mapped successfully to elements in a set *B* (such as the DPO) and vice-versa. Here, a "set" is a well-defined collection of elements, such that it is possible to determine, for any given element *x*, whether or not *x* belongs to the set [9].

Using algebraic laws and proof by induction, we are able to create a set of operators that allow us to construct transformation "expressions", that is, string representations of transformations. For example, the most trivial case is the one-to-one relationship, where a single element from set *A* maps directly to set *B*. This can be represented with the string, "*A1->B1*". Of course the intent is (and our research confirms) that we can represent much more complicated transformations as well, however, the proofs of such representations exceeds the scope of this paper.

CONCLUSION

This research is a first step in encapsulating the complex data object and transforming it to an object capable of withstanding the test of time. As years pass, it is critical that the scientific record is maintained, not only for legal compliance issues, but for historical reference. Data provenance issues such as authenticity and accuracy are of utmost importance.

The result of this research effort will prove that a CDO can be transformed successfully and reliably to a DPO using a middleware component such as the Ingest Parsing Service. This research will also show that the DPO can be inversely extracted from an archived state back to the original CDO. This research provides a solution to an extremely complicated problem. This ongoing effort is attractive due to the distributed architecture, use of industry standards, and robustness based on proof using mathematical foundations. Further exploration of the contributions made by this research should open up even more exciting avenues of research for digital preservation.

REFERENCES

[1] "Opportunities for Distributed Science: A DOE National Collaboratories Program Workshop", sponsored by the National Center for Atmospheric Research (NCAR), Boulder, CO, December 1-3, 2004.

[2] DOE2000 Program Plan, March 1997. <http://www.mcs.anl.gov/DOE2000/plan/cn9608.htm>

[3] Talbott, Tara, Peterson, Michael, Schwidder, Jens, Myers, James D. May 2005. Adapting the Electronic Laboratory Notebook for the Semantic Era, Proceedings of the 2005 International Symposium on Collaborative Technologies and Systems (CTS 2005), May 15-20, 2005, St. Louis, MO

[4] Myers, James D. 2003. Collaborative Electronic Notebooks as Electronic Records: Design Issues for the Secure Electronic Laboratory Notebook (ELN). Proceedings of the 2003 International Symposium On Collaborative Technologies and Systems (CTS'03).

[5] The Library of Congress. METS: An Overview & Tutorial. May 2005. <http://www.loc.gov/standards/mets/metsoverview.v2.html>.

[6] Murray, Ray L. June 2005. Toward a Metadata Standard for Digitized Historical Newspapers. Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries: 330-331.

[7] The Open Archives Initiative Protocol for Metadata Harvesting, June 2002. <http://www.openarchives.org/OAI/openarchivesprotocol.html>

[8] Smith, Mackenzie. July 2005. How Can We Preserve Digital Files and Save our Collective Memory? IEEE Spectrum: 22-27.

[9] Bloch, Norman J. 1987. Abstract Algebra with Applications. Prentice-Hall, Inc.