MADRAS: A Metadata and Archival Description Registration and Analysis System for the Analysis of the Recordkeeping Capabilities of Metadata Sets

Lori Lindberg, Monique Leahey-Sugimoto, Nadav Rouche, and Holly Wang; University of California, Los Angeles; USA

Abstract

This paper discusses progress in the development by the Description Group of the InterPARES 2 (International research on Permanent Authentic Records in Electronic Systems) Project of MADRAS – Metadata and Archival Description Registration and Analysis System, a metadata schema registry and analysis system for the identification, registration, and analysis of existing and prospective metadata schemas, sets, and application profiles relevant to electronic recordkeeping and digital preservation. It updates a paper presented at last year's IS&T Archiving Conference in Washington, DC. This paper will update our progress in the development of the system, describe some of the findings to date, describe our relationship with the developers of ISO 23081 and their influence on MADRAS, and outline some challenges encountered in the development process. InterPARES is an international multi-disciplinary research collaboration emanating out of the archival community that has been working since 1999 to devise new models, methods and automated tools for ensuring the creation and preservation of reliable and authentic electronic records. The second phase of this project, InterPARES 2, which is due to be completed in 2006, integrates the disciplinary perspectives and concerns of the scientific and digital arts communities, as well as those of e-government, and is focusing in particular on the preservation of records generated by emergent interactive, experiential and dynamic systems and processes.

Introduction

The intellectual background and theoretical/analytical impetus behind MADRAS - the Metadata and Archival Description Registration and Analysis System, a research product and tool of the Description Cross-Domain of the InterPARES 2 Project (http://www.interpares.org), was outlined at last year's IS&T conference by Professors Anne Gilliland of the University of California at Los Angeles and Sue McKemmish of Monash University in Australia.[1] Their paper and presentation set up the context of the work and its place within InterPARES 2, summarized the then-current state of the prototype version of the system, and outlined the development of the metadata schema analysis process, all of which has served as the basis for the development of MADRAS and its present iteration. Interested persons should refer to that paper for this background information since our limited timeframe prohibits our repeating that information here. This paper will update our progress in the development of the system, describe some of the findings to date, describe our relationship with the developers of ISO 23081 and their influence on MADRAS, and outline some challenges encountered in the development process.

The Prototype System and the Development of the Web-enabled Beta Interface

The purposes of the prototype system and the research have remained constant and are still in effect for the beta version currently in development. These purposes are to:

- Describe relevant metadata schemas, etc. and their versions and features in a standardized and unambiguous way
- Capture information on relevant crosswalks and application profiles
- Assess how well registered schemas address recordkeeping and preservation requirements as expressed in particular analysis instruments
- Assist users in identifying metadata tools that may meet their specific needs
- Assist developers of existing and prospective metadata schemas and sets in assessing how well they address recordkeeping and preservation requirements.

The prototype system was developed within a Microsoft Access database environment. This environment allowed the testing of the controlled structure of metadata about metadata schemas developed within the research as an XML DTD, as well as serving as a test bed for the analysis process under development. Translating this prototype into a dynamic web-enabled interface has been a fairly smooth process with regard to the testing of the DTD. We are currently registering schemas and crosswalks and opening the interface to other researchers within the InterPARES project to gain their feedback and impressions of the system from the perspective of users relatively unfamiliar with the interface. Feedback from initial testing has raised a few issues that have been quickly addressed. Several of these issues are outlined later in this paper.

The Registry and Analysis Interfaces

The current beta environment for MADRAS is implemented with PHP, a server-side scripting language that provides web developers tools for building dynamic websites. The back-end web server is Apache 1.3 and the database server is MySQL 3.22. Both servers are hosted on a machine running the Unix operating system. PHP, Apache and MySQL are all open-source technology and are used by many database-driven web applications. The Education Technology Unit (ETU) from the Graduate School of Education and Information Studies at UCLA is hosting MADRAS and provides server-side support. Although we are required by ETU to implement MADRAS with the above technology, we also agree that it is a good and flexible option for our project.

Currently, MADRAS is designed to support the research goals of InterPARES 2. Based on research requirements, MADRAS is implemented in two major parts: a registry interface based on the XML DTD, within which schemas are registered and described, and an analysis interface where the analysis of schemas for their recordkeeping capabilities takes place. At the registry interface, InterPARES researchers log into the system and register metadata schemas. Each researcher who registers a particular schema "owns" that schema and has the privilege to edit, delete and duplicate the schema registration record within MADRAS. While all InterPARES researchers can browse registered schemas, only the researchers that own them can modify the schema registration records. The analysis interface is conceptually designed to lead a user through a series of specially constructed questions that methodically assess a schema for its recordkeeping capabilities. The questions are appropriately weighted as to importance and relevance to recordkeeping and preservation requirements as expressed in the analysis instruments upon which the questions are based. Results from the analysis questions are tallied and reports are produced that identify recordkeeping and preservation strengths and weaknesses within a schema. These reports can be further supplemented by suggestions for modification or addenda to bring the schema closer to meeting recordkeeping and preservation requirements.

We do not anticipate much more change to take place with the registry interface. We have just opened the system to all researchers within the project and anticipate additional feedback through the spring and summer to help us streamline the registration process and make it more transparent, particularly for those persons relatively unfamiliar with records and record keeping. At this point, registration of a schema takes a fairly large amount of time, with some complex and larger schemas taking in excess of an hour, and sometimes longer, to fully register. We are looking at options for a minimal set of data needed for a basic submission and the essential registration information needed to provide a researcher the means to analyze a schema. While our long-term goal is to build a registration and analysis interface that anyone, including the public, can use, at this time only InterPARES researchers can do registration and only Description Cross-Domain researchers do the analysis of schemas.

The MADRAS analysis interface is still under development. Researchers are still revising the analysis questions. To allow for more flexibility within this development environment, two analysis interfaces have been constructed; one a dynamic interface that can be changed at will by researchers with specific privileges, and a static interface that changes only when aspects of the dynamic interface are finalized. We designed a dynamic analysis interface within MADRAS for researchers to test the research findings that could be supported by a system with the highest flexibility, while leaving a static interface for continued population of the schema database and allow for schema analysis to take place within a particular "release" of the system. In the dynamic interface, researchers can modify analysis questions at any time, and the interface used to display the questions can be arranged and rearranged according to any agreed-upon requirements. This would understandably be confusing to users if changes were frequent, and conclusions would be difficult to draw from any particular release because of overlapping changes and lack of control. The skeleton of the static analysis interface has been constructed and it will be easier to move ahead and implement it once we have agreement on the analysis questions and how to display the questions to the users, constituting the first "release version" of the analysis.

We have about 30 tables in our current system. Prior to the present implementation, the prototype contained 28 registered schemas. Since the beta version of MADRAS has been operational we have registered 20 more schemas and 10 crosswalks. We anticipate completing the registration of all schemas on our present list of schemas to be registered and adding more as the year continues. In addition, we will look at registering application profiles as particular implementations of registered schemas. As a research project, most of our tables are relatively small, holding a limited amount of data. The current size of the MADRAS site is 20 megabytes (without appended documents) with around 100 PHP files. More files will be generated in conjunction with the development of the analysis interface. We expect that MADRAS will grow into a mid-sized application after processing more feedback from InterPARES researchers and adding more data and infrastructure. MADRAS is allowed 50,000 queries per hour from the database server, and MySQL 3.22 has a 4-gigabyte limit on table size We are not concerned that the size of MADRAS will challenge our present computing environment.

MADRAS and ISO 23081

Our liaison with the technical committee responsible for the development of ISO 23081 Parts 1, 2 and 3 is presenting particular challenges for a project such as MADRAS which has been developed and constructed by researchers of varying knowledge levels regarding records and recordkeeping and from disparate recordkeeping philosophies. Challenges include how to accommodate the various audiences and communities that may utilize MADRAS and providing a transparency of the analysis process to accommodate those without a recordkeeping background who are concerned about these issues but relatively unfamiliar with recordkeeping theory, processes and terminology. Another, more significant, challenge is how to construct and present questions that address the complexity of the metadata model behind ISO 23081 and the conceptual entities incorporated within the standard in a user-friendly manner. As the metadata counterpart to ISO 15489, the international records management standard, ISO 23081 is in itself quite detailed and complex, with multiple types of metadata accruing at various layers and at different times within a recordkeeping system. With ISO 23081 incorporating the significant findings about the authenticity of records developed within the InterPARES project as well as the conceptual recordkeeping model behind the Australian Recordkeeping Metadata Standard, itself the basis for ISO 15489, the assessment tool developed for MADRAS is planned to be the Part 3 assessment tool for ISO 23081. This assessment tool must accommodate both of the major models of records management currently in use in the archives and records management communities, the life cycle model as reflected in the InterPARES research and the continuum model developed in Australia. InterPARES and the ISO technical committee have established a formal relationship and our plan is to have all issues concerning the construction of the analysis questions as well as their coverage of the standard resolved before the summer of 2006 in order to provide the MADRAS system developers adequate time to implement the analysis system, test it, and ready the interface for an alpha release in September of 2006.

Another particular challenge is how to demonstrate the relevance of incorporating recordkeeping and preservation principles and practices within metadata schemas to those communities who have never considered it. Developers of imaging metadata standards, for example, may not realize that utilization of standards and aggregating images in a database that are described using those standards creates records. These records are simultaneously digital assets as well as records in the truest sense, being the products of activity and set aside for keeping. Many developers of imaging standards do not consider that there are layers of metadata essential to the long-term survival of these assets, not only for their preservation but their management, both at the point of creation as well as through time and across systems. The application of recordkeeping processes and preservation considerations within metadata will enable their metadata schemas to be more robust and serve multiple purposes at the same time. Thus, submitting publications such as this to conferences that focus on areas outside the records management and archival communities allows us to raise awareness of the importance of this work to them, fostering a better understanding and encouraging future collaborations.

Conclusion

The development of a complex system such as MADRAS involves a great many people from various backgrounds who have worked together to produce a product that has the potential to be a powerful tool for metadata schema developers and users. Our hope is that the ISO will provide the long-term sustainable support for the system and continue to develop it after the end of InterPARES 2 in December 2006. For further information on MADRAS and to see the system in its current iteration, please visit http://www.gseis.ucla.edu/us-interpares/madras/.

References

[1] Anne Gilliland-Swetland and Sue McKemmish, *A Metadata Schema Registry for the Registration and Analysis of Recordkeeping and Preservation Metadata.* A paper presented at Archiving 2005 – IS&T's 2005 Archiving Conference, Washington, DC.

Author Biography

Monique Leahey-Sugimoto and Holly Wang are current graduate students at UCLA, pursuing the Masters of Library and Information Science through the Department of Information Studies. Both are Graduate Student Researchers with the InterPARES 2 project. Nadav Rouche is a December 2005 graduate of the same program and a former Graduate Student Researcher with the InterPARES 2 project.

Lori Lindberg is a full-time Lecturer in the archives specialization at the School of Library and Information Science at San Jose State University and a Doctoral Student in Information Studies at UCLA. As part of her doctoral research, Lori is a Graduate Student Researcher with the InterPARES 2 project.