

**Federal CIO Council XML Working Group
Meeting Minutes, July 18, 2001
American Institute of Architects, Board Room**

Please send all comments on/corrections to these minutes to [Laura Green](#).

Working group co-chair Owen Ambur convened the meeting at 9:00 a.m. at the American Institute of Architects. Attendees introduced themselves.

Presentation: "Electronic Records Archive (ERA)"

Ken Thibodeau ([NARA](#)) briefed the WG on the National Archives and Records Administration's [Electronic Records Archive](#). This presentation is available at the XML.gov website in [PowerPoint](#) and [HTML](#) formats.

Mr. Thibodeau began his presentation by informing the WG that NARA is not committed to XML but is exploring the use of it. The Electronic Records Archive (ERA) is part of NARA's attempt to move into the future. NARA foresees a time when all government documents are created on computers and preserved electronically. So, if NARA wants to be able to continue as a comprehensive government archive for the 21st century, it has to be able to deal with any electronic document in any electronic format that agencies may create. Right now, ERA is a research and exploratory development program. The ERA team recognizes that there are so many challenges facing an archive such as this that it would be foolish to try to build a system that would do everything they want it to right now.

The goal of the ERA project is to build a successor to Archives II, to build the archives of the future to meet the standard set by Archives II.

There are three parts to the ERA Vision: to overcome technological obsolescence in a way that preserves demonstrably authentic records; to build a dynamic solution that incorporates the expectation of continuing change in information technology and in the records it produces; and to find ways to take advantage of continuing progress in information technology in order to maintain and improve both performance and customer service

NARA faces a critical challenge with the ERA project in that very few people have experience in doing digital preservation of this magnitude. ERA's capabilities and methods are limited. There are no COTS solutions to draw upon.

The ERA team wants to find a way to use mainstream technology. It hopes to be able to build on existing infrastructure and not go off into obscure technologies. It hopes to have as wide an application as possible in the digital preservation/retention arena. Currently, the project is collaborating with 13 national archives from around the world. Additionally, ERA is supporting a project that is attempting to scale down the software being developed for ERA so it can be applied to state or university archives. The ERA team wants to develop NARA-specific systems only when it confronts NARA-specific issues.

Mr. Thibodeau then showed the WG a [slide](#) illustrating the ERA Infrastructure Concept. ERA sees itself having repositories that spread across the internet and have different flavors. ERA needs to deliver tools for NARA employees and public users across the nation. This slide describes a complex environment, one that needs core technologies. The infrastructure will be scalable and therefore will require very high-speed access. Other requirements include security, mediation among systems (both on the inbound and outbound sides), and infrastructure independence. In such a system, it must be possible to change any or all technologies without affecting the information it is preserving.

Some of this infrastructure will be provided by NARA, some will be provided by the National Information Infrastructure under NARA's control, and some will be provided by the National Information Infrastructure outside of NARA's control.

Mr. Thibodeau then discussed NARA partnerships in the ERA project. The name of the game for NARA is collaboration. The presentation [slide](#) entitled "NARA Partnerships" lists major NARA partners.

Mr. Thibodeau then showed the WG a [slide](#) illustrating the ERA functional model, which is intended to be an Open Archival Information System (OAIS) implementation. The OAIS is very general; it assumes that the user is trying to manage some information that is to be preserved for a period of time. This information comes from outside the OAIS. The OAIS keeps it until and outside community of consumers asks for it.

The InterPARES project is the largest archival research project the world has ever seen. The researchers involved have been developing a model of the high level process of electronic records preservation. The "InterPARES: Preserve Electronic Records Model" [slide](#) shows the first level of decomposition.

Mr. Thibodeau then showed the WG a [slide](#) ("Information Management Architecture for Persistent Object Preservation") presenting another way of looking at the process, this time focusing on a Knowledge-Based POP. NARA does not preserve documents; it preserves records. If one wants to manage collections over time, one must work at three different levels: data, information, and knowledge. Information is a meaningful combination of data. Knowledge is a meaningful combination of information. On the slide, the three columns represent the three basic OAIS functions (ingest services, management, access services), while the blue areas represent software mediators. If one wishes to preserve e-records over time, one must separate the physical storage of data from the management of data. The storage resource broker is the middleware.

Mr. Thibodeau then showed the WG another illustration of the concept model ("ERA Concept Model"). This view describes a virtual workspace. Once again, there are three columns of functionality. A key component of this view is the desire to build 80% of the functionality that NARA uses all of the time into the workspaces while still allowing room to use special tools. For example, if one has a record that is subject to a specific law as to how it should be archived, the workspace should allow for this unique method. NARA wants to be in a position where if it does not need a particular tool, it can simply stop using it. Another key component is the desire to have repository records be in containers as neutral as those that hold paper containers. To do this, NARA must take some metadata out of the equation. This data will be taken out and placed in storage. NARA wants to be able to go from one format to another while preserving the basic structure of the records, taking them from one technology and delivering them into a very different one.

There are two basic processes in the ERA: bringing the records into the archive and providing access to them once they are there. For the first process, the system needs to figure out what archiving process applies to the incoming records. For the second, the system must figure out the opposite.

NARA's basic problem is deciding whether or not it will accept a submission. It has to know what an agency should have transferred, what the agency says was sent, and what really was sent. The best way to figure this out is to construct a model for what records should have come in.

It helps to have records be as self-defining as possible. This is where XML comes in. NARA would like for records, files, series, and record systems to all be self-defining.

Mr. Thibodeau then showed the WG a [slide](#) ("E-mail: Group-wise View") of a self-defining e-mail. The following slide displayed the e-mail in a text editor view, wherein the message ceases to be self-defining. Mr. Thibodeau showed the WG [slides](#) of the MIME-aware and tagged MIME-aware views. The last view uses XML.

NARA wants to construct models that capture the essential structure of the information. A DTD would be a good model of document structure.

Mr. Thibodeau then showed the WG a [slide](#) ("eXtensible Business Reporting Language Example") that illustrated an XBRL DTD, which he considered a nice, simple DTD that deals with a great amount of data.

One can therefore get an abstract model that can be expressed in any number of ways.

With XML, NARA can create a model of what a record from an agency should be that is as precise as it needs to be. Mr. Thibodeau then provided an example using ERA's method of determine the dates of records.

Another problem ERA faces is deciding how tight it wants these models to be. The user does not want to have to run two searches because the software changed at one point.

The next series of [slides](#) demonstrates how EWA would transform and aggregate documents.

Mr. Thibodeau then presented another example, where ERA receives a request for correspondence. The next few [slides](#) illustrate the flow of events in the system.

Mr. Thibodeau then discussed ERA's Persistent Object Preservation (POP) project. POP is aimed at creating independence of technological infrastructure. It will embed changes in a comprehensive information management architecture designed for preservation. It will be inherently extensible. It will facilitate use of future advanced technologies without requiring change in what is preserved.

If NARA uses XML, it will not have to tell agencies how to construct their records; the agencies will be able to tell NARA how they were constructed.

When asked how NARA implements relationships, Mr. Thibodeau replied that each agency establishes the relationships between documents. There may be some exceptions in which NARA creates topic maps, but the original agency usually will make the links, if there are any to be made.

Bill Suffrage remarked that many agencies have long term non-permanent records. He asked if NARA has looked into make this solution available to other agencies.

Mr. Thibodeau replied that it is. Initially, NARA concentrated on the preservation problem. But it will increasingly be looking at the front end of the lifecycle. It is in NARA's best interest to find solutions that individual agencies can also use.

Bill Morgan (GSA) remarked that NARA will have to use some automated mechanism to move archives over time.

Mr. Thibodeau commented that NARA will have to automate as much as it can.

Mr. Morgan asked how NARA will remove archives in an automative way.

Mr. Thibodeau responded that currently NARA is looking closely at the DoD 1515.2 standard. One of the challenges NARA faces deals with the archiving of e-mail: which e-mail messages should be preserved? Something as trivial as a cancellation for an afternoon's meeting will only take up space.

Bette Fugitt (FIRM) asked Mr. Thibodeau for his opinion of TIFF versus PDF.

Mr. Thibodeau replied that the determination of what is the best technology for an agency to use in keeping its records is up to the agency itself. He would not like to get NARA into the business of telling an agency what it should use.

Ms. Fugitt asked if TIFF format would adequately preserve and meet the requirements of accessibility from NARA's perspective.

Mr. Thibodeau responded that NARA did a collaborative study several years ago with the DoD. The bottom line of the study was that while there are many technologies out there, none is perfect.

Ms. Fugitt asked if this meant that agencies were going to go back to hardcopy transmission.

Mr. Thibodeau replied NARA may be accepting documents in TIFF format.

Mr. Suffrage remarked that one could wrap a TIFF image in a PDF wrapper.

Mr. Thibodeau replied that one thing that worries NARA is that a submitter could wrap anything in PDF and send it to NARA. NARA would have no idea that there might be something in the wrapping.

Matthew Kern (PCi) asked for the number of records NARA has now.

Mr. Thibodeau replied that there is no way to count each record, but NARA does use less than a terabyte's worth of physical storage.

Mr. Ambur asked what message the CIOs should take away from NARA's work.

Mr. Thibodeau replied that he would have to get back to Mr. Ambur on that.

More information on this project is available at the [ERA website](#).

The WG then broke for ten minutes.

Presentation: "XML Topic Maps"

Sam Hunting (eTopicality) briefed the group on XML Topic Maps. His presentation is available at <http://xml.gov/presentations/etopicality/home.html>.

There are several different standards bodies that can define what a topic map is and, fortunately, they all seem to agree. ISO published [ISO 13250](#), the foundation document of the topic map effort. OASIS is also working with topic maps. [Topicmaps.org](#) is an independent consortium of parties interested in developing the applicability of the Topic Maps Paradigm to the World Wide Web by leveraging the XML family of specifications as required. Topicmaps.org will probably be folded into OASIS within the year. [Topicmaps.net](#) is an informative site hosted by ISO editors. It is the home of the "bleeding edge" of the topic map movement.

There are three things to know about a topic map. First, a topic is a machine representation of a subject (a subject is something one wishes to talk about). Second, an association is a connection between topics. Third, a scope is what makes an association useful to a user.

There are other takes on what a topic map is. One sees a topic map as an "information overlay." There are many information resources out in the world. The topic map exists on a plane above that data pointing down to it without affecting it in any way. It is a knowledge network, a distributive representation of memory. There are heaps of data all over the web. To cope with it all, one can overlay a topic map on one heap. The topic map can be laid over any heap and is not specific to any particular heap.

While DTDs are as frozen as they can be, the topic map data model is not set in stone. The topic map data model is necessary because the DTD semantics are meaningless without context. One assumes that tag names are intuitive. If one wishes to link together disparate parts of the web, one cannot rely upon mere intuition. Topic maps can provide the semantics missing from DTDs.

One should think of the topic map data model as the web equivalent of a relational model. A topic map is a network with which one connects disparate sources of data.

In the topic map world, people are mapping resources that exist all over the internet. There may be multiple copies of the same resource. In this instance, one would not want to create fifty nodes for something with fifty repetitions. One can deal with this by using a topic naming constraint merge. Two topics with the same name and the same scope are considered the same topic. By inputting these topics as XML documents and merging them into one document, one can eliminate the repetition.

There are several topic-map-related efforts underway at this time. One is the RDF effort. Another is the WebDAV effort. The RDF effort is focused on statements, making it much less flexible than a topic map. One can drive topic maps from RDF, but one cannot derive RDF from topic maps. WebDAV affects data in a different manner than do topic maps.

Mr. Hunting stressed the fact that topic maps are based on the idea of a complex taxonomy. Topic maps are capable of

expressing such a taxonomy. They can even serve as tools for modeling business processes.

John Milligan (CSC) asked how a topic map would be used for re-purposing.

Mr. Hunting replied that one could have multiple overlays over the same data.

Mr. Milligan inquired as to how far along the Patent Office is in its topic map effort.

Mr. Hunting replied that he has not yet implemented topic maps at the Patent Office, primarily due to the complexity of the Office's business processes.

Presentation: "One Perspective on the Value of XML Topic Maps."

Ron Daniel, Jr. ([Interwoven](#)) briefed the WG on the value of XML topic maps. This presentation is available in [HTML](#) and [PowerPoint](#) formats.

Mr. Daniel began by stating that his company, Interwoven, is not a "topic map company." Its view on topic maps is that they can provide real value to people who run websites as well as their customers.

The value of topic maps can be summed up in one word: categorization. Categorization is an incredibly important activity. People categorize things all the time. Topic maps allow for multiple categorizations to be in effect at the same time.

Categorization is very important for Interwoven customers. Current categorization on websites is fairly rudimentary. The most useful websites will end up following the Five Laws of Library Science:

- Books (information, websites, records systems) are for use
- Every book its reader
- Every reader his or her book
- Save the reader's time
- The library is a growing organism

Topic maps will become more important as a consequence of this website growth.

Mr. Daniel showed the WG a [slide](#) depicting a possible topic map view. However, this view is a node and arc diagram, which is hard for the reader to follow.

Mr. Daniel then showed the WG a [slide](#) depicting another possible topic map view taken from a DVD website. There is categorization at this website, but it is sometimes hard to follow.

Another example, this one using hyperlinks, was displayed.

Several lessons can be gleaned from those examples. Topic maps will help visitors navigate and discover information on websites. User interfaces will only expose small portions of the topics at any one time. Someone has to create and maintain the categories and categorize the data.

Mr. Daniel showed the WG a [slide](#) illustrating another style of map. It is a more boring interface, but this makes it more suited to people desiring speed and accuracy. People who create topic maps and categorize things will need to use node-and-arc style displays. Mr. Daniel also showed the WG a topic map interface based on a map display. However, the only information that can be gleaned from a display like this is roughly how many documents are in one clump. There is no indication of the size or quality of the data.

Mr. Daniel then showed the WG a [slide](#) displaying the basic maintenance required to keep topic maps up to date.

Mr. Daniel concluded with the following points. Topic maps can deliver real value. They are not real "maps" and most users should not see a complex interface. Use of standard XTM formats and tools will bring down costs, which in turn

will be affected by the relative populations of site visitors, catalogers, maintainers, and analysts. No tool or format will make it easy to create a good classification scheme.

Mr. Daniel commented that Interwoven is also familiar with dealing with alphabetized lists. Different people have different views of hierarchies. His remarks are by no means to be taken as a blanket condemnation of graphical interfaces, just a call to examine how much value such an interface really has.

Mr. Hunting remarked that he is pleased to see the design aspect being discussed. He disagreed with the statement that the value is not the map. A map is a set of categories, not an actual picture.

Mr. Daniel replied that there is a lot of value in what the syntax of topic maps is trying to do. With regards to display, designers need to think carefully about what they are trying to do. The "right" display will vary from organization to organization.

Eliot Christian (USGS) commented that users do not know ahead of time what an arc label means. He asked where the taxonomy of all arcs would reside in a maps world.

Mr. Hunting replied that users will come to an agreement as to what the arcs will mean.

Mr. Christian asked how one would understand arcs without a hierarchy listing.

Mr. Hunting replied that topic maps can help the user understand.

Mr. Daniel pointed out that this is the debate between topic maps and RDF. Whatever one can do with RDF, one can pretty much do with topic maps. With RDF, one can provide archetypes but eventually one will get away from this world of symbols and find oneself wondering what they all mean.

Mr. Christian pointed out that this view is different from the one that the semantic web will be self-documented.

Mr. Morgan asked if information is lost upon convergence.

Mr. Daniel replied that the items still exist as distinct entities; they have just been categorized.

Mr. Morgan asked if information is lost upon merging records.

Mr. Daniel replied that the information still exists. Topic maps exist as a mean of expressing that information. What one does with the information is not a topic map issue.

Mr. Hunting added that it is a design feature of the topic map specification that one must respect the intent of the author of a map. One can always preserve the map from which the topic was merged so one can always get back to where he was before the merge.

Mr. Daniel remarked that when one makes a topic map, one is trying to display the subjects and their essential relations to one another. Whenever this is done, some information is thrown away. Classification schemes embody a point of view.

Ms. Fugitt told the WG about a presentation she attended where the [Washington state library](#) discussed its XML work. The library has developed a set of metadata that is XML based. The webmaster of the library website has a keyword section and a hierarchy tree. She suggested that WG members consider visiting the library to look at the webmaster's instructions.

Mr. Ambur remarked that some vendors would like to have us believe that machines can classify information better than people can.

Mr. Daniel replied that the word "better" is always used in relation to some standard. Interwoven sells software that can classify information. However, it does not exceed human performance in terms of accuracy. It is much faster than

humans at classifying. Machines are faster and more consistent, but they cannot exceed the level of performance of a human.

Mr. Ambur pointed out that every website is essentially a rudimentary topic map. He asked how one could improve those classifications and their processes.

Mr. Daniel replied that he could not presume to tell agencies how to improve.

Ms. Fugitt remarked that Washington state, Dow Chemical, and GE have used FAQ's, hit counts, and call answer centers to find potential areas of improvement on their sites.

Mr. Christian stated that there is a false notion that there exists some "best" taxonomy or browser tree. Taxonomies may not work in all languages, and they may be proprietary.

Mr. Daniel replied that some commercial tools are starting to emerge that will help with on-the-fly relations between differing taxonomies.

Final Announcements

Ms. Fugitt announced that the next FIRM Forum ("Meeting the Challenges of Electronic Information") will be held August 14 at the Agriculture Department. For more information, please visit the forum [website](#) or e-mail [Ms. Fugitt](#).

Mr. Ambur announced that the agenda for next month's meeting is still up in the air. He hopes to have someone give a presentation on e-books.

Mr. Royal reported a conversation with Bruce Bargmeyer, who has been working on different registry/repositories specific to the ISO 11179 standard, and who would like to try to normalize the relationship between ISO, ebXML, and UDDI registry/repositories. NIST will soon launch an XML repository at XML.gov. It is likely that the WG will receive a proposal from Mr. Bargmeyer. Mr. Royal recommended that the WG seriously consider sponsoring Mr. Bargmeyer's work.

Mr. Ambur remarked that that might make a good focus for the August meeting.

Mr. Royal added that Lisa Carnahan will be leading the ebXML registry/repository effort. She is also the chair of an OASIS committee gathering some of ebXML's work.

Next Meeting: August 15.

Last Name	First Name	Organization
Abel	Elizabeth	OFHEO
Ambur	Owen	Interior-FWS
Braaten	Ward	NSF
Campbell	Richard	FDIC
Collins	Jamey	GAO
Daniel	D. Ron	Interwoven
Fugitt	Bette	FIRM
House	Robert	State
Hunting	Sam	Topicality
Jordan	Rick	Peregrine Systems
Kantor	Bohdan	LOC

Kern	Matt	Pci
Kwari	Nicholas	ED
Levenson	Steve	US Courts
Milligan	John	CSC
Niemann	Brand	EPA
Poot	Lex	DTS
Revis	Nancy	ARL
Royal	Marion	GSA
Sallaway	Susan	OCC
Smith	Rick	MPG
Stockwell	Mel	IONA Technologies
Thibodeau	Ken	NARA
Threatte	Jackie	ARL
Todd	Mike	OSD
Vasko	John	FIRM
Warrington	Earl	GSA
Yee	Theresa	LMI
Zimmerman	Ann	Pure Edge Solutions