Archival Interoperability: The Intersection of Stability and Change
Kenneth Thibodeau

National Institute of Standards and Technology
Interoperability Week, Plenary Session, 24 April 2007

The mission of the National Archives and Records Administration is to serve American democracy by safeguarding and preserving the records of our Government, ensuring that the people can discover, use, and learn from this documentary heritage. The records of the U.S. Government are defined in law to include laws, policies, regulations, contracts, grants, personnel files, and other such documents that readily come to mind but also any information assets in any form that any agency creates or receives in doing its business that ought to be kept for some time "because of the informational value of data in them."[1] Thus, the scope of NARA's responsibility extends over any type of information any federal agency produces or uses in executing its mission. That includes audio/video materials, scientific data, geographic information systems, design and engineering data, virtual reality models, and many more.

NARA has two roles in helping the nation to realize the value in these information assets – records – held by the federal government. The first is to guide and assist agencies in creating and keeping records in the conduct of their business. The second is to preserve in the National Archives those records that have significant value, even after the original agencies' needs for them have been exhausted. In the first role, there is a dual emphasis on improving the conduct of government business and in identifying and ensuring timely transfer of records selected for preservation. In the second role, the emphasis is on providing access to preserved records to anyone with an interest and a right of access. In principle, NARA preserves such records permanently.

These two roles confront NARA with three major interoperability challenges. The first is on the supply side: the need for interoperability with systems in all other agencies from which we ingest electronic records. The second and third challenges surface on the demand or dissemination side. The second challenge is interoperability among records. The third interoperability challenge concerns discovery and delivery of preserved electronic records.

Supply side challenge

In essence, the first challenge is one of data transfer which is occasionally synchronized, and hopefully though not always contemporaneous. Up to now NARA has received transfers of electronic records mostly on physical media. We are moving towards online transfers but the carriers are not the main problems in transfers. The big problems are in the range of technologies from which the transfers originate and in the nature of the data transfers. It is practically a given than any technology – including any storage media – that exists is used somewhere in the federal government and that we probably need to effect a transfer from it. Compounding the problem, the U.S. Government does not always refresh technology in a timely

[1] 44 U.S.C. 3301

manner. It is not very frequent, but also not unusual for us to receive transfers on obsolete media and in obsolete formats. For example, several times in the last decade we received transfers on punch cards.

DiGenova materials

We will probably never eliminate this type of problem – e.g., presidential or congressional commissions – but we are working to reduce it, mainly through negotiated procedures, especially by getting agencies to transfer electronic records to us as soon as possible.

The challenge stemming form the nature of the data transfers is that we do not transfer pieces or packets of data that systems are designed to export. Rather we need large batches of data, often the entire user-created content of a system, or in the case of office automation of several systems, or a snapshot of an entire database. For example, we don't get individual email messages, or even all of a user's messages, but most often all of the email of all employees in an organization, often spanning several years. We are approaching this challenge in a variety of ways. NARA has endorsed the DoD standard for records management applications, DoD 5015.2-STD, for governmentwide use. The DoD Standard assumes electronic records will be managed in an application dedicated to records management, with its own digital repository. In collaboration with 19 other agencies and the Object Management Group, NARA has lead the development of requirements for an alternative approach, called Records Management Services (RMS). RMS will provide required records management functionality through services that appropriately control records that remain in the systems used to conduct business. RMS will be included in an upcoming release of the Federal Transition Framework from OMB. RMS is also being developed as a voluntary standard by the Object Management Group. NARA has issued regulations specifying formats and mechanisms for transfer of permanently valuable federal electronic records to the National Archives. In addition, Lockheed Martin is developing as part of the ERA system a Packaging Tool that agencies can use to extract electronic records from their own system and package them for transfer to the National Archives.

Demand Side Interoperability Challenges

Preservation: Reagan Moore, Director of Data and Knowledge Systems at the San Diego Supercomputer Center, defines preservation as communication with the future. I would refine that to specify that preservation is a method for communicating with the future such that the message that is received at any arbitrary time and place in the future is the same as the message sent in all essential respects. That is a non-trivial process because, in any context that depends on digital technology and involves a significant lapse of time, we must assume that the message which will be received will not be identical to the message sent for the simple reason that it will be received using technology that differs substantially from that on which it was sent. Moreover, over long times, it is probable that there will be differences in the way information is represented, necessitating some translation or interpretation of the original message. Thus, digital preservation entails the delivery of specific, stateful information objects to unknown future recipients across generations of changing information technology, and also evolving semantics, syntax and pragmatics.

Archival preservation – the preservation of records – is even more complex. A record is any piece of information that is made or in some way acquired in the course of some practical activity and that is kept by an actor on a presumption that it will be useful in subsequent activity or at lease w ill serve as a stable memory of either the action or the state of affairs in which it occurred, or state of affairs which it produced. In contrast to publications which are intended to be broadcast, if not to the general public at least to a designated community, records are instruments or byproducts of specific activities. They are created in a universe of discourse where there is a high level of shared knowledge, which is either the implicit "common knowledge" of participants in the action, or explicit in other, related records.

Interoperability Among Records

Records preserved in an archives – by which we mean not an inactive store but an institution or system specifically designed to preserve and provide access to records – are necessarily outside of the original universe of discourse in which the records were created. The archives, then, has the responsibility to provide the receiver of a preserved record with supplementary information crucial to correct interpretation of the record. The archives should provide descriptive information about the domain in which the record was created: what was the creating organization; what were its functions; who were the principal actors. Secondly, the archives needs to enable the receiver to discover and navigate to related records. Traditionally, "original order:" physical arrangement of paper documents into file folders and series of file folders was the predominant method for both identifying and accessing related records. In the world of hard copy, physical arrangement is practically the only effective method for instantiating relationships among records in a persistent fashion. Digital IT provides much richer ways for expressing relationships among records and navigating across them. One of the goals we have set for the Electronic Records Archives system we are developing is to capture such relationships and enable future receivers of preserved records to navigate to related records on the basis of these relationships.

In more abstract terms, the aggregate of the records of a single actor constitutes a partially ordered set, as does any archivally proper subset, where archivally proper refers to relations established by the record creator. The ordering of the archival set is as fundamentally important as the internal structure of any member of the set. The meaning of any record depends on its relationship to other records. For Example, purchase order for a computer might be assumed to be an acquisition record, but that would be true only if the PO were found in the case files of the acquisition office. If the same document were found in a different case file, such as the investigatory files of an agency's Inspector General, it would be a record of an investigation of possible fraud, waste or abuse. If it were found in an investigatory file of the FBI it would be a record of an investigation of some significant criminal activity.

Often related records may be heterogeneous. Textual documents, databases, email, GIS, digital and hard copy. Archival systems must articulate and support access to such diverse sets.

Interoperability for Discovery and Delivery

The third interoperability challenge concerns discovery and delivery of preserved electronic records. This entails being able to deliver any type of preserved electronic records to anyone who wants them without imposing burdensome constraints on the technology they use to receive them. It also requires support for support discovery and delivery on unknown, even uninvented future systems.

To do this, we have to have a channel for communication with the future which is itself dynamic, or as we expressed it in the Request for Proposals for the ERA system, the archival system must be evolvable. It must be possible to replace any and all components of hardware and software used in the system with minimal impact on the functions of the system as a whole and negligible impact on the records preserved in it.

Obviously, communication with the future requires interoperability with future systems for storage, search, retrieval, delivery and use of preserved information assets.

One of the biggest challenges in supporting access to preserved records on unknown future systems is that is that the interests of future users are often independent of the purposes for which the records were kept in the first place. But challenges can be substantial even when the future recipient wants to use the records for the exact purpose for which they were created. E.g., CAD/CAM/CAE. We know nothing about what computer driven manufacturing system will be like in 25 year, other than that they will be very different from what we have today. Therefore, we cannot presume we will be able to use product data for necessary processes, such as manufacturing a replacement piece part for a ship or aircraft. NARA gets involved in issues like this because our statutory role includes assisting agencies to accomplish their mission. The Department of Defense and others maintain ships, aircraft and other physical systems for 40, 50 years or longer. We have an obligation to help them figure out how to preserve authentic digital records, including specification of both the products and the processes used to produce them. We are actively engaged with DoD and others on research in this very complex problem. We hope that, if we can help them find a way to preserve CAD/CAM/CAE records to meet their long term operational needs, we will be able to use the solution when such records are preserved in the National Archives.

Risks in preservation: getting the message into an adequate transmission channel in the first place; insuring that the connection is not broken; and effectively controlling the transmission process such that all essential attributes of the message persist.

By and large records are created to communicate within a specific process or activity where there is a high degree of shared information among the players. Whenever the process extends over a significant period of time, key information needed in completing it should be captured in records. Effective record keeping includes efficient retrieval of information needed at any point in the process and, therefore, should minimize the need for repeating information. For those who were not part of the process really understanding what was done requires access to the aggregate of the

records that were organically collected in the execution of the process. Thus records preservation – as distinguished from data preservation or preservation of publications – requires contextual richness. We must preserve not only records, but collections of records, and collections that are defined as ordered sets, where the order was determined as records were created and accumulated in the original process. We must also preserve contextual information about who created the records and for what purpose.

.

But purpose of archival preservation is to enable the recipient of the communication maximum latitude to use, repurpose or do whatever they want with an assumption of minimal corruption of the message in the interim.

Essence of archival interoperability is not an issue for interoperating IT systems. That is a necessary but far from sufficient condition. It is the ability to use information from the past to meet present objectives where the time lapse between the production of the information object and its subsequent use is ideally unbounded and the later use may be completely independent of the original purpose for which the record was created and kept.