ARKIVAL TECHNOLOGY CORPORATION

System Enhancement Study for Preserving and Validating Terabytes of Electronic Records transferred to NARA on Multiple Media Types or via FTP

FINAL REPORT

February 27, 2005

Contract #: NAMA-04-F-0055

ARKIVAL TECHNOLOGY CORPORATION
427 Amherst Street Suite 360
Nashua, NH 03063

Contact: Ronald D. Weiss
Telephone: 603 881 3322
Email: ron@arkival.com

(This page intentionally left blank)

Table of Contents

Preface

This final report summarizes ARkival Technology's Study under NARA contract NAMA-04-Q-0055 for the six month period beginning July, 2004. The subject matter entails a diversity of issues confronting NWME in their effort to make necessary changes in the preservation, validation and access processes.

This document classifies the work performed for the study into three different categories- though very inter-related in actual NWME operations. The three categories are Networks and Storage Devices, Electronic Transfers and Data Sharing and New Format accessions. Some of the categories required technical detail that may be of greater interest to a technical reader. For the purpose of satisfying all readers, the structure of this report has the corresponding technical detail in the APPENDIX of this document.

System Enhancement Study for Preserving and Validating Terabytes of Electronic Records transferred to NARA on Multiple Media Types or via FTP.

1. Executive Summary

1.1 Recommendations

This study resulted in four (4) primary recommendations:

Networks and Storage Devices

1. The implementation of a TAPE FARM STAR Network with appropriate enrichment to the APS software.

Electronic Transfers and Data Sharing

2. A new, secure and user-friendly "push- method" for Electronic Transfers

3. A Router Cluster-based design for Secure Data Sharing between a closed network (APS) and open net work (NARA net).

New Format accessions

4. Use of multiple COTS software for verification of new "rich" file formats with a phased approach to a longer-term solution embodying promising new "back-bone" software (JHOVE).

These recommendations…

- Address the need for processing files with DLT devices including DLT tape libraries.
- Eliminate multiple data copy processes and some manual operations.
- Provide sufficient capacity and hardware plans to meet forecasted demands.
- Provide a focused direction for accessioning files in new formats.

1.2 Executive Discussion of the Four Recommendations

1.2.1 Networks, DLT Storage Devices, Implementation & Constraints

ARkival's initial review of the current NWME operation was completed with the objective of integrating high density storage devices (DLT's) and recommending procedures for automating preservation and verification; including a comprehensive network review for future activities. These activities also focussed on scaling capacity by 100 to 1000 fold. The following information was determined:

- The currently configured APS (LAN) client-server network includes 3480 drives; some with autoloaders, 9 track tape drives, DLT tape drives, NAS devices (Iomega SNAP servers) and database servers along with a compliment of workstations with other peripherals as well. The devices in place combined with the forthcoming addition of more DLT drives and/or M1500 DLT Tape Libraries will exceed the network bandwidth during certain applications and use.

- The present LAN bandwidth is 100 Mbit per second (Fast Ethernet). In considering new DLT additions to the LAN, the present bandwidth will limit the number of DLT simultaneously operating on the network.

- The DLT's being introduced have a storage capacity 200X that of the IBM 3480 and are 2X faster.

- The use of full data DLT cartridges will significantly increase processing time in several NWME operations (file searches, tar-ing files, etc.) if there are no changes made to the APS software, processes and methods in use.

- The ability of the present network and operation to satisfy the forecasted demand prior to the arrival of the ERA is seriously in question at this time. Projections of increased quantities of files combined with greater file sizes and the replacement of 3480 preservation copies will stress the present system beyond its capability.

- Improving network efficiency alone will not enable NWME to handle projected increases in capacity because many of the limiting factors are not related to network performance. Many of these factors are more methods and systems related than hardware related.

Improvements to be proposed from this study are extensions of the present technology and are directed to improve operations and better prepare NWME to handle large volume and newer formats until ERA system is operational.

1.2.2 Electronic Transfers

The present NWME/NARA scheme for FTP transfers has not been well utilized by participating Agencies. ARkival believes this issue is important for the processing of increasing quantities of data in NWME and to provide an electronic transfer path for the ERA.

Successful electronic transfers usually take place when the system is readily available for the sender. Agencies responsible for transfers look to avoid work interruptions and special conditions and times to transfer files. The NWME system is not particularly user-friendly and is inconvenient for attracting greater electronic transfer volume.

In the overall scheme of future data transfers to NARA, the method of secure electronic transfers must be resolved. The benefits of new electronic transfer methods to NWME are the simplification of the ingestion process, the capability of providing greater security and improved compliance to archivist requirements.

A newer streamlined process can readily accept electronic records in a secure media-less transfer while automatically verifying the integrity of the records being sent using application software and controls. The method could make data transfers faster, easier for the sender and able to support automated integration with existing systems at the NWME receiving center.

The file sizes and frequency of submissions from the Agencies has some bearing on the practicality of secure electronic transfers to NARA. One could make the case that all-electronic accessions should be a significant part of future NARA operations. In today's technology not all submissions can be reasonably transferred because of existing bandwidth limitations. It seems reasonable therefore those smaller and more frequent transmissions could employ an updated and more user-friendly transfer process.

The technology and simplicity of sending/receiving secure electronic documents is not beyond reach but certainly not the preferred method of accession transfers to NARA from its participating agencies. Discussion at NWME indicated the present FTP process of electronic transfers is technically possible but clearly with complications for its use both to NWME operations and to its customer agencies.

The present process of electronic transfers at NWME can best be described as a "pull process" compared to a "push process" like that used for emails. In the "push-process" case, the sending party sends documents at a time and manner that best suits the sender. In a user-friendly environment, the "push-process" receiving system is designed to accept the secure transfer with ease and simplicity.

The limitations of present NWME/NARA FTP receiving process[4] has created a situation whereby participating agencies prefer accession transfers via physical transport or other non-electronic means. If NWME is to make secure electronic transfers part of its daily operation, changes at different levels will be required; including the sending agencies.

The more significant problems that need to be addressed involve bandwidth issues, file sizes, the frequency of agency transmissions and receipt of transferred files on NARAnet. A more user-friendly process for smaller and more frequently sent files should be considered as it will lay the groundwork for future submissions of larger and less frequently sent files and make the ERA transition less complex.

1.2.3 Secure Data Sharing

A review of NWME data flow suggested that a secure data sharing capability could improve certain operations in the accession and reference copy process. The network segregation currently in place for security is also an impediment to data sharing. Current technologies however, can provide a secure way to share data on different networks.

Inherent in this proposal is a Router Cluster design approach to network security whereby neither network affects the internal policies of the other. A solution to this problem will provide a single metadata entry to be shared by the multiple databases in NWME operations.

Although not critical to data flow, a secure method of data sharing will improve operational efficiency, increase throughput and provide ease of access to files.

1.2.4 6+ New Formats

This portion of the study identified software tools to validate the integrity of the e-records transferred to NARA in six new and different formats. In its conclusion, the study proposes methods/software to ensure that volume, content and structure are consistent with the NARA specifications for the transfer of e-records in the six formats.

ARkival has identified several commercial off the shelf (COTS) software, shareware, and Open Source product possibilities that can be used to validate the new format specifications and preserve the records with the present/proposed system design.

The proposed COTS software alternatives were evaluated for the method of checking the integrity of the record formats; the various conversions possible for the six different formats and the ability to address existing standards.

[4] Complexity results from bandwidth, times of transfers and file data sizes

Networks and Storage Devices

2. Networks and Storage Devices

2.1 Network & Hardware Devices

The APS LAN is considered a simple Client-Server network. The actual network is somewhat more complex and a complete hardware profiling of the network was necessary to identify all the hardware components and their associated IP addresses. The full listing is provided in APPENDIX A.1 of this document. In summary, the network (at the time of the trial) included seventeen (17) computer systems, one (1) server, four (4) Network Attached Storage (NAS) appliances, eleven (11) workstations- some with different storage devices attached, and one (1) print server.

2.2 Network Trials & Measurements

(This section summarizes much of the detail reported in the APS/ LAN NETWORK REPORT & ANALYSIS in APPENDIX A.1. In addition to the report and the summary charts provided, there exists some 14GB of actual data recorded during the trials. Software routines have been developed to study the data for future questions about design and component performance).

The following table summarizes the LAN network trials performed on the APS/LAN. The trials, separately and combined involved data transmission between the more important network devices for the accession and reference copy process. These network measurements were required to determine a baseline reference for data storage devices being used today and more so in the near term future. The resulting baseline data will be used to examine workload expansion and the network requirements for additional storage devices.

The following list describes the critical baseline operations/hardware used for the data transfer trials:

- 3480 to NAS
- CD to NAS
- DLT 8000 to NAS
- NAS to local workstation

[5] ARkival Network Analysis Report: NARA/NWME APS/ LAN NETWORK REPORT & ANALYSIS - *see Appendix A.1 of this document*.

In addition to the trial network activity itself there was some network overhead recorded simultaneously; that data is included in each trial summary and in most cases was considered minor[3].

The detailed network activity reports were collected and baseline data for two (2) primary tape storage drives were isolated; they include the 3480 transfer to the NAS and DLT transfer to the NAS. Typical network data for these primary storage devices was obtained from single time intervals (snippets) or multiple snippets and thereafter isolated for reporting purposes. *See Figure below*

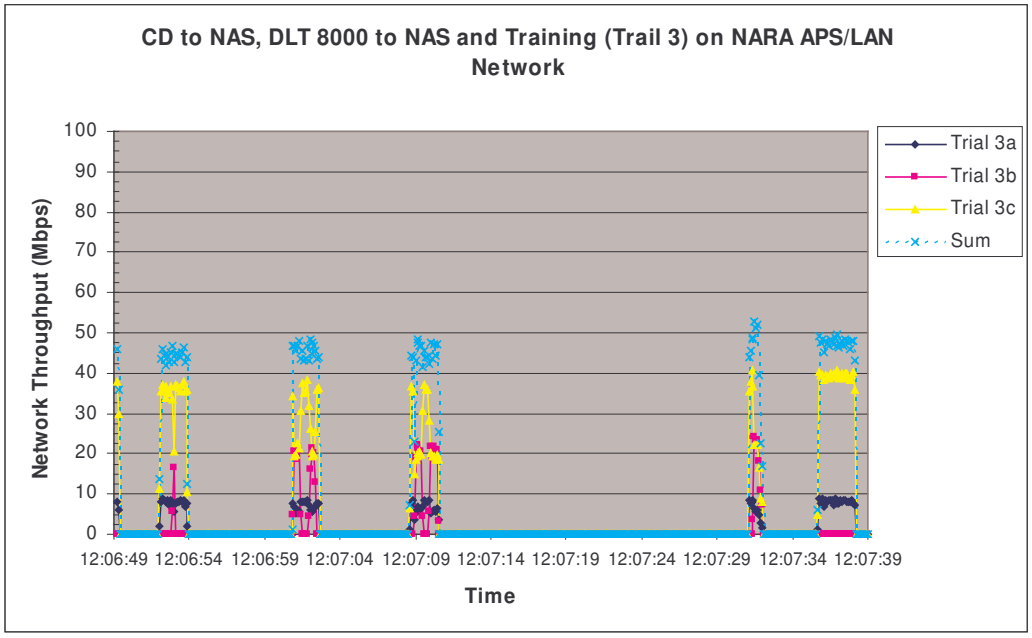


Figure 5. Representative time Snippets for simultaneous CD to NAS (3a), DLT 8000 to NAS (3b) and training (3c) data transfers to the NAS (trial 3). Note the total contribution of all the different data traffic occasionally exceeding 50% of the available throughput[4] for short periods of time.

In summary, present network performance may degrade during certain applications and the result can be somewhat misleading when one studies average network activity without observing peak performance times within the average. The periods of peak

[3] The only notable overhead was the contribution from a training session in NWME taking place during one of the ARkival trials. Training sessions are considered normal part of APS/LAN operations and its occurrence was helpful in the network activity measurement.

[4] "Average Data" is subject to interpretation via both the measurement technique and the CSMA/cd protocol. The data collected required detailed introspection in that the analyzer reports of "average data transmitted" was not always an indicator of potential network problems. Instead peak activity instances within the average, did indeed support a potential basis for conflicts in application and use.

network activity contained in device-essential operations, more of the same type of operations performed simultaneously with additional hardware devices combined with increased volumes of data (10 to 100 fold), larger data files and more preservation/reference copying are obvious reasons to question the future performance, reliability and application of the current network design for the projected need.

2.2 LAN Network Analysis *(See Network Schematic in APPENDIX A1)*

Hardware

The forecasted growth and subsequent network demands between client-attached storage and network-attached storage will impact future designs and performance. A network analysis was made on the basis of individual components presently making up the APS net. Most of the components are essential to future designs and any limitations need to be understood and designed-around for scaling, improved operation and inter-networking activity.

Industry specifications on standard network configurations, presently networked hardware and connectivity[5]considerations were the basis to make the following observations of the LAN and capacity planning for same:

- The NAS devices are used mostly in a "Write once, Read many times" style and may not be optimized for such use, both by purchase and configuration. In addition, the NAS devices should not be configured as RAID 1, as this configuration will slow read times.

- Network measurements for specified operations are helpful to confirm NAS performance

- The DLT tape drives are far slower[6] than either the network or the HDD's

- The underlying 100Mb/s (12.5MB/s) of the (Fast Ethernet) network may represent a serious bottleneck when transferring data between workstations and a fast NAS and/or by adding DLT devices.

- A bandwidth upgrade to a switched 1 Gb/s (125 MB/s) GigE or Fiber will likely be required to support multiple accessions simultaneously.

---

[5] Typically specified operations and speeds:
Copy data to/from a workstation HDD (max speed 3 MB/sec)
Copy data to/from a local HDD to/from NAS (max speed 12.5 MB/sec)

[6] DLT 8000 drives specify a 6 MB/s uncompressed Average Sustained Transfer Rates (ASTR). *(See OEM specs* http://www.quantum.com/NR/rdonlyres/7225D7FE-D6EF-4387-B865-FC4A77992E06/0/DLT8000DS618.pdf *and* http://www-1.ibm.com/ibm/history/exhibits/storage/storage_3480c.html)

- Should the additional data capacity and the processing of new format types require the use of 100-300 GB tape drives in the 40MB/s class, NAS or SANs in the 600+ MB/s class, and 10Gb/s networking, the LAN should have the necessary hardware and cabling to handle the volume.

Software, Processes & Operations

Hardware is only part of a successful solution. In order for the above hardware to be effective, both software and procedural issues must also be addressed[7].

With relatively large amounts of data being forecasted, accessed and manipulated, indexing, stacking[8] and other kinds of volume management software will be required.

During the current study, in particular during application demonstrations and the network utility trials there was a low utilization of both hardware and software resources, too many manual processes and too few automated processes. To that end, the following APS recommendations are being made. In some manner the APS software and application must be enhanced to…

- Accommodate network-addressable devices (for Ethernet and/or Fiber channel)
- Accommodate Tape Libraries
- Accommodate Batch processing of files using "check-point" software (e.g. allow job re-starting from point of failure and not back to the first tape in a series of tapes)
- Replace/upgrade the present APS-TAR to new TAR technology that saves files and has indexing capability (e.g., uses check sum capability for writing and verifying files and employs INDEXING capability for locating a single file in a large group of files)

2.2.1 AERIC & NARAnet Metrics

The majority of AERIC network activity takes place on the large, high performance NARAnet[9]. The NARAnet supports all NWME operations involving AERIC applications and few, if any critical components will affect forecasted

[7] The acquisition of hardware should be matched with needed software and procedural improvements.
[8] Consolidation issues (stacking) must be also addressed for determining a design for transferring archived records from 3480 cartridges to DLT cartridges.

[9] See NARA Enterprise Architecture Vers. 2.0 (Sept 1, 2003).

capacity and scaling to a 100 fold increase of data, including archiving activities related to some six (6) new format types.

2.3 New DLT Storage Devices and LAN application

DLT introduction to NWME operations can have a substantial impact on productivity and throughput. The DLT's being introduced have a storage capacity 200X that of the IBM 3480's and are 2X faster. There exist however several issues in APS operations that need addressing to benefit from DLT optimum performance and utilization (e.g. APS software, tape handling operations, work scheduling, etc.)

ARkival's review of currently used tape-oriented operations, procedures and processes in NWME suggests the introduction of faster, higher capacity DLT tape drives, tape libraries and other hardware will require optimization of all hardware devices with firmware and software to enable gains in throughput and performance. Although greater capacity DLT's will allow for larger files to be transferred in a single media cartridge, increased cartridge capacity will require changes in the present methods used for many tape operations.

There is opportunity to improve DLT applications and Tape Library use as follows:

- Tape processes will be faster and consume less physical space with DLT's
- New files brought into NWME can be stored on tapes along with discs providing easily accessed reference files
- Large files with multiple tapes and large file sizes (>100GB & TB) can be processed faster.
- For faster TARing & re-TARing of files
- For faster file searches in collections of any size
- Files being transferred to DLT's (migration).
- Addressing incoming file receipts using electronic transfers (FTP) to NARA
- DLT's and Libraries can be implemented for non-attended tape operations.

In summary, the current network configuration in the APS area is structured towards temporary large file storage onto Network Attached Storage (NAS) devices, permitting multiple users access to the data from a central location. In most small network configurations large amounts of data transfer is not the primary use of the network, but more random database access. Due to the large amounts of data being stored on the NAS's the capacity of these devices is limited from time to time.

Workflow optimization geared towards getting data into a disc-based storage system as early as possible as well as minimizing the number of data transfers should be addressed.

The NAS's on the LAN are the central storage container for APS network reference data and for assembling large accessions. Given that these particular devices are not designed for high volume throughput they will quickly become a production bottleneck as the number of concurrent users accessing these devices increases. To this end, a more suitable network design should be considered.

The feasibility of employing optical fiber network capability with the DLT device introduction onto the APS LAN was also considered as an option to Ethernet. A notable factor was that the Quantum DLT 8000 tape drives and other APS LAN storage devices do not natively support Fiber. These legacy SCSI devices (DLT 8000, IBM 3480, 9-track, et. al.) will require Fiber channel bridges (FC Bridges) for connectivity. The Quantum Tape library series (M1500) is Fiber-supported and library units configured with DLT 8000 products will require more costly SCSI to Fiber Channel bridges. In difference to the additional costs for making the legacy devices fiber-ready, Fiber networking remains an alternative option.

2.4 A proposed TAPE FARM STAR NETWORK

2.4.1 NETWORK DESIGN & BASIS

ARkival has studied, measured and reviewed the present APS/LAN network. Although functional for most of today's operations, it exhibits occasional inconsistencies and certain component and design limitations for anticipated network demands and future workloads. The complications and impediments to optimize the present network performance can be addressed. However when considering the increased data volumes forecasted[10], the likelihood of considerably larger accession files, the potential impact of the 6 new formats (particularly large GIS files), the likelihood of more EFT transfers and the direction to new DLT storage devices, NARA/NWME interests would be best served by considering a new network design.

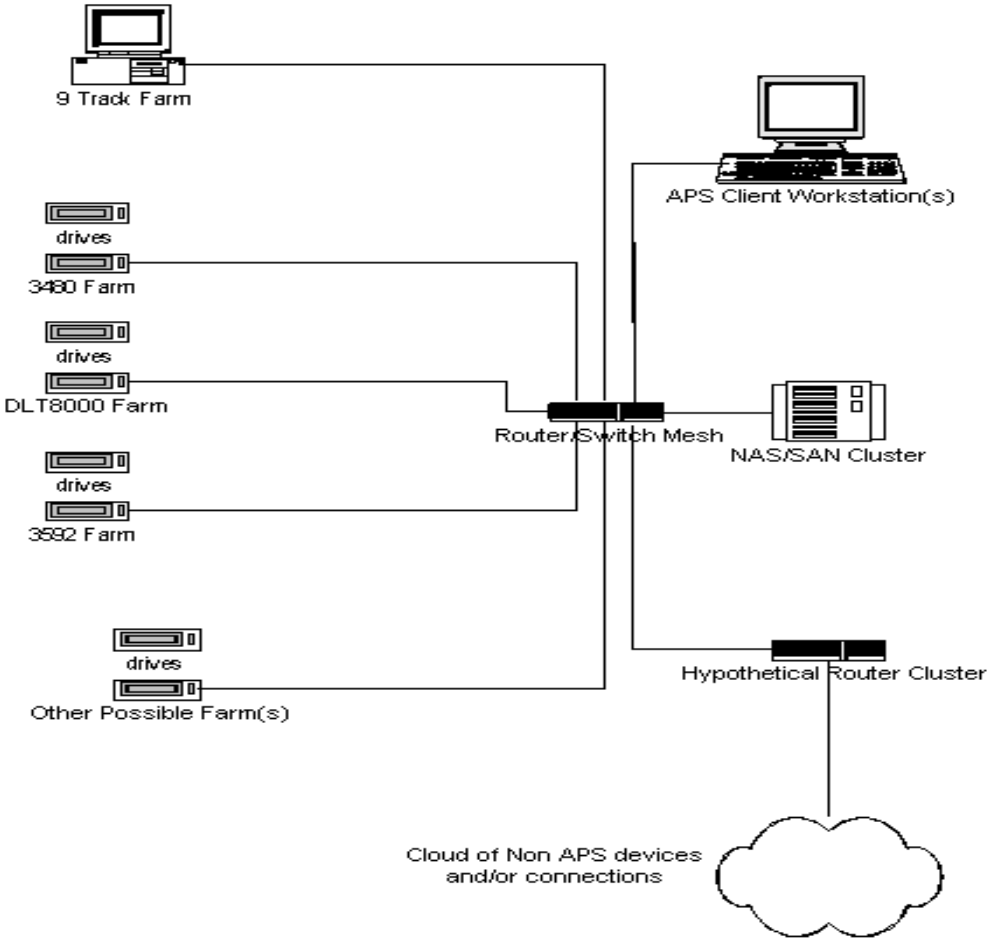
Proposed is a network design for safe processing accessions that are several hundred Gigabytes and Terabytes in size and also includes concepts for automating preservation and verification work with DLT and other storage devices in the same environment.

This proposed design incorporates the input of numerous technical and operational meetings, network tests and measurements, discussion with operations personnel and with operators performing key APS & AERIC functions and the actual observation of functionality and use of the APS/LAN.

The design concept embodies many facets of NARA operations today and also utilizes much of the existing LAN hardware to minimize cost without sacrificing performance. It incorporates and optimizes the performance of new DLT hardware and the M1500 tape libraries on the network, improves effectiveness of tape storage applications for NWME operations and optimizes tape related functionality.

[10] See Deliverable 4.2 September 9, 2004- FORECASTED NWME CAPACITY

THE PROPOSED DESIGN - ***A "Multiple Tape Farm" Star Network***



The network design is best described as a "Multiple Tape Farm Star Network" - *See Figure above.* The emphasis on tape results from the obvious dependence on tape as a storage means at NARA and the predominance of tape from past and likely large future accessions. Whereas many large storage applications are "Disk to Disk to Tape"-based, the NARA/ NWME operations is predominantly "Tape to Disk to Disk to Tape"-based. This distinction requires that all types of tape services be readily accessible, prioritized and not a detriment to either the network or the timely access to data.

Because of the tape dependence at NARA and the variety of tape formats in use and likely to be used in the future, the network concept had to embody the tape varieties as well as the use of other storage media such as CD's, DVD's, et. al. Also of equal importance in the design consideration was the fact that many of the APS operations are still tape and disc based and that the relatively lower speeds of tape-based operations should minimally affect disk and network speeds and traffic.

Each tape farm in the proposed design is a host or host-cluster dedicated to supporting a specific kind of tape device. Thus there is a 3480 farm, a DLT8000 farm, a 3590 farm, etc. This device focus and network locality is currently an IT "Best Practice" for handling multiple types of mass storage devices on a single network. The design embodies a NAS/SAN mesh in the center of a "star" network topology. The points on the star can be the LAN devices (APS workstations, tape farms, local servers, etc) and, indirectly a router cluster to WAN devices (connections to AERIC, AMIS, NARA-net,)

On the LAN level:

(Device specific tape farm)----**NAS/SAN Mesh**----(other local devices, including other device specific tape farms)

On the WAN level:

(Local device or device farm)----**NAS/SAN Mesh**----Router Cluster----(non local devices)

The NAS/SAN Mesh[11] is the heart of the communications infrastructure for the proposed network design. At this level, it is architectural concept, not a reference to a specific device. In this application, the NAS's (Network Attached Storage) are basically dedicated File Servers where the operating system and their internal hardware (usually RAID for the HD's, multiple fast network interface cards (NICs) etc) are completely focused on user-based file level I/O's. The SANs (Storage Area Networks) are more specialized application devices and are not user-based. The SAN devices have layers of software and hardware support removed and are designed for performance.

One or more interconnected NAS systems can provide some level of performance improvement. If greater performance is required than NAS technology can provide a SAN fabric made of one or more SAN systems would be added to the NAS/SAN mesh

[11] A disc-based device (NAS or SAN) will improve application performance because the communication with/within it have been optimized. Many SAN's today are used in automating and centralizing disk servers including tape backup operations. The actual backup process is usually shorter since channel performance is optimized for moving large amounts of data. Using a disc-dedicated server, data is sent to the storage medium (tape) through the server's front end. Productivity is improved because applications are not sharing bandwidth with the tape backup data stream. It is an important hardware consideration for improved APS/LAN performance.

and the NAS systems would become "NAS gateways" between the user level access layer and the SAN fabric[12].

This ability to scale by either "dimension", NAS or SAN or more devices of a particular category, is what makes this architectural concept building-block-based and a "Best Practice" by current IT industry standards.

The Router/Switch mesh can also be either a 1 GbE or Fiber channel device.

2.4.2 SAN Usage and APS Application

SAN devices[13] are usually designed for different kinds of applications than those used by NARA in their business operations. NARA's requirements involving preservation, validation and access require very high throughputs and not particularly low latencies. A SAN tailored for NARA's business processes would have lower than necessary latencies with added capability to meet the high throughput requirements. It is also unlikely that a low end/low cost SAN would be a COTS solution for the APS LAN without the costly addition of throughput capacity (e.g. for processing of large TB files and TB file sets).

SAN's tend to be fiber-oriented, and for APS application a new Fiber Channel SAN fabric would need to be installed. Legacy SCSI devices need to be connected to Fiber Channel via SCSI bridges. Existing Workstations would require Fiber Channel Host Bus Adapters and appropriate software drivers added. The final step would be to add Fiber Channel compatible disk storage to the network. An additional benefit of the SAN approach is that it also allows data sharing between workstations regardless of the Ethernet networks they are attached to. Tapes could be read into the SAN via AERIC and read by APS. Thus, Fiber Channel is another option for the proposed Tape Farm application.

iSCSI technology embeds SCSI commands into TCP/IP packets and uses some variation of Ethernet (100MbE, 1GbE, 10GbE, etc) as if it were a disk drive cable. SCSI commands inside IP packets have far greater overhead than a native SCSI bus. The resulting increased latency can be a problem for OLTP-like applications and the reduced

[12] The SAN's architecture works in a way that makes all storage devices available to all computers on the network. The hardware that connects computers to storage devices in a SAN is referred to as a "fabric." The SAN fabric enables any-computer-to-any-storage device connectivity through the use of Fiber Channel switching technology.

[13] SAN's are specialty devices for specific purposes and mainly reliant on fiber, either as Fiber Channel or iSCSI. Typical SAN applications are on-line transaction processes (OLTP) that require very low latencies. In general, SANs tend to have high cost of ownership due to the maintenance contracts and specialized hardware associated with them.

effective bandwidth is somewhat a problem for most applications. iSCSI is also a contender for the NARA application.

2.4.3 Design Benefits

The "*Multiple Tape Farm" design…*

- … Clearly differentiates physical services from application services
- … Gives consistency to the APS/LAN and addresses the complexity of different types of media storage required.
- … Optimizes tape utilization, has provision for different tape and disc formats, and provides for uniformity and growth prior to ERA functionality.
- … Can be scaled to need and does not embody the typical "stovepipe" processes.
- … Provides relative simplicity and is based upon decomposing the complex issues of the network and particularly its utilization.
- … Modernizes NWME network operations, improving reliability and performance at a reasonable cost.
- … Can be implemented in stages to minimize its impact and possible downtime.

… and the design will provide for a simpler, less complex conversion with high speed network inter-connectivity when required.

2.4.4 Operational Benefits

- Very few of the present APS "functional" operations are likely to be affected.
- The present APS/LAN room and space will be adequate for the proposed change but may require supplemental HVAC consideration.
- Supporting new devices, workstations, etc becomes much simpler by making the network architecture modular in this way
- Device and resource utilization can be higher than present since all devices become part of a common pool rather than being locally attached to dedicated resources.
- This scheme makes it easier to design automated processes and procedures for common NARA tasks like creating tapes, accepting submissions, etc.

- This design makes it easier to have automated 24x7 un-attended operations where desired.

2.4.5 Design Limitations and Considerations

1. The *Multiple Tape Farm" Star Network Design* is substantially different than the present network. It will affect hardware, operating systems and methods of operation.

2. The design embodies both a newer and proven concept that will change some tape handling operations.

3. The operational change will require a greater emphasis on planning and scheduling by users.

4. The design may have some high future costs should expansion be required from 10/100 to gigabyte capacity as well as some larger I/O caches.

5. Devices on the network must be upgraded to support 1GbE or Fiber channel to take advantage of this architecture.

6. This new design may require a replacement of the existing network wiring and router structure and relocates many of the SCSI devices currently on APS workstations to an alternate location.

2.4.6 Justification and Direction

The APS/LAN requires a network capability that can utilize new DLT devices (drives & library systems) along with 3480, 9 track tapes and other presently utilized tape and disc storage devices. The network design guideline used by ARkival implies that it should not be negatively affected by the introduction of new devices introduced or by the processing of large files to be introduced via DLT's and possibly FTP transfers.

The present APS/LAN workstations have demonstrated some inconsistent timing that may likely be affected by the presence of different types of storage devices. Processing delays may result from different drives and hardware drivers, different firmware levels and conflicts and revisions in the APS software.

The new design will make APS workstations more robust, reliable, predictable, and maintainable if their commonly "cloned" peripheral devices are gathered together in drive-specific tape farm rather than being scattered amongst the workstations. The tape devices themselves will also result in similar improvements because of their uniformity within each type device; the end result being a more efficient performance of tape devices in a drive-specific tape farm than in the current arrangement.

The NAS/SAN Mesh is the heart of the communications infrastructure and an architectural element that must be specified and interfaced with detail. These devices can move data via optimized channels to a localized tape farm with improved efficiency because of the communication with/within the device. It is an important hardware consideration for improved APS/LAN performance.

2.5 Network & Storage Device Recommendations

- Augment the APS workstations with tape device-specific tape farms
- Install disc-based storage devices (NAS or SAN's)
- Install an appropriate network (1 GbE with a Fiber channel bus)
- Use software, firmware and application software changes that optimize hardware performance
- Use appropriate Policies and Procedures that address the needs of the client-Agencies, Researchers, NWME personnel and public users.

Infrastructure alone cannot and will not provide the performance required in this most specific application. In conjunction with the proposed hardware changes, ARkival is

recommending the following changes to the APS and operations…

- Incorporate the use of Network-addressable devices
- Incorporate the use of Tape Libraries
- Provide for batch processing of files using "check-point" software (e.g. allow job re-starting from point of failure and not back to the first tape in a series of tapes)
- Upgrade APS-TAR to new TAR technology that uses check sum capability for writing and verifying files
- Upgrade APS with file management software to fast-find files in a large storage medium Software to employ INDEXING capability for also locating a single file in a large group of files, providing fast access to files and for toggling between files.

The incorporation of such changes will impact ease of record processing and improve processing times for preservation, validation and access.

(This page intentionally left blank)

Electronic Transfers
&
A Mechanism for Data Interchange on 2 Networks

3. Electronic Transfers

3.1 Present Process and Constraints

The present NWME/NARA method for FTP (file transfer protocol) transfers has not been well utilized by participating Agencies. ARkival believes this issue is important for the processing of increasing quantities of data in NWME and to provide an electronic transfer path for the ERA.

Description

The current method uses FTP running on a dedicated server that is connected to the internet but not to the APS network. This server is used only for conducting FTP file operations. Because this method uses FTP it is not possible to conduct the transfer of files in a secure manner. The process also requires significant amounts of time be spent by personnel both at the submitting agency and at NARA in order to coordinate the transfer operation. This complex and difficult process has the effect of discouraging use of the FTP system and the low number of electronic transfers received in past years supports this premise.

Successful electronic transfers usually take place when the system is readily available for the sender. Agencies responsible for transfers look to avoid work interruptions and special conditions and times to transfer files. The NWME system is not particularly user-friendly and is inconvenient for attracting a greater volume of electronic transfers.

In the overall scheme of future data transfers to NARA, the method of secure electronic transfers must be resolved. The benefits of electronic transfers to NWME will simplify the ingestion process and be capable of providing greater security and improved compliance to archivist requirements.

A newer streamlined process can readily accept electronic records in a secure media-less transfer while automatically verifying the integrity of the records being sent using application software and controls. The method could make data transfers faster, easier for the sender and able to support automated integration with existing systems at the NWME receiving center.

Constraints

The file sizes and frequency of submissions from Agencies have some bearing on the practicality of electronic transfers to NARA. One could make the case that all-electronic accessions should be a significant part of future NARA operations. In

today's technology not all submissions can be reasonably transferred because of existing bandwidth limitations. It seems reasonable therefore those smaller, more frequent transmissions could make best use of an updated and more user-friendly transfer process.

The present process of electronic transfers to NWME can best be described as a "pull process" compared to a "push process" *(e.g., like that used for emails)*. In a "push-process", the sending party sends documents at a time and manner that best suits them. In a user-friendly environment, the "push-process" receiving system is designed to accept the secure transfer with ease and simplicity.

The present NWME/NARA FTP receiving process[14] has created a situation whereby participating agencies prefer accession transfers via Federal Express or other non-electronic means. If NWME is to make secure electronic transfers part of its daily operation, changes at different levels will be required; including the sending agencies.

The more significant problems that need to be addressed involve bandwidth issues, file sizes, the frequency of agency transmissions and receipt of transferred files on NARAnet. A more user-friendly process for smaller and more frequently sent files should be considered as it will lay the groundwork for future submissions of larger and less frequently sent files and make the ERA transition less complex.

Addressing Security

Security considerations had to be included from the very beginning of the new design for an electronic submission process. It has been addressed at different levels from the sending-Agency, the receiving server on the NARAnet and the proposed APS/LAN Tape Farm network[15].

3.2 ESV- A Proposed Process for secure electronic transfers

A newer and more user-friendly secure way of accepting electronic transferred files has been reviewed and proposed. This new concept provides a NARA/client-agency electronic transfer process that is practical, appropriately secure and convenient for the agency and NARA and also believed to be ERA-friendly.

Today, there are multiple possible ways to replace the current FTP mechanism with a simpler, easier, and more secure system that provides all of the benefits outlined

[14] Complexity results from bandwidth, times of transfers and file data sizes
[15] See Security and Router-Cluster discussions in Section 4. and Section 2.

above. The use of encryption certificates and authentication certificates, for example, provide the foundation for secure file transfers- all mature, well defined, technologies that are in common use.

3.2.1 An Electronic Submission and Verification (ESV) process.

This ESV concept provides a NARA/client-agency electronic transfer process that is practical, appropriately secure and convenient for the agency and NARA. This process will require procedural changes for NARA and approvals from NARAnet Security (NH).

This proposed process would accept certain electronic records in secure media-less transfers while automatically verifying the format of the records being sent by using application software and controls. The method will make smaller sized data transfers faster, easier for the sender and incorporates automated integration at the NWME receiving location. Electronic submissions would be described as a *NARA controlled-Agency initiated* process.

ARkival is proposing that NWME consider the automation of the manual process in place today. The automated process would provide agencies a NARA form (website-accessible) that includes details for accession compliance. The system would reside on the NARAnet (or a location of NARA's choice) on a separate server and integrated into www.NARA.gov (e.g. ESVtransfer.NARA.gov).[16] The receiving system and process would employ a secure and authorized agency log-in that is user-friendly and provides automated access to NARA from an authorized client-Agency. The system would perform a secure, scheduled, unattended accession retrieval to NARA (via NARAnet) to a dedicated server. There would be no Agency-client communication other than the submittal forms and the subsequent automated retrieval (by NARA) of the agency accession.

Accessions would be later transferred from the NARAnet secure server to the NAS on the APS/LAN via a secure router-cluster and screened for compliance in the same manner used for manual transmissions received today. Compliant documents could be processed and readily moved through the APS operations for archival storage.

[16] Alternatively a direct connection to any Agency can also be created to support automated transfers of data files.

3.2.2 Process properties

A new and secure system should be implemented with the following characteristics and features:

- Authentication certificates for both ends of the transfer.
- Use of a separate data source to store information about security, authentication, and connection details.
- Use of encryption certificates and encrypted data channels.
- An accession process that allows for both agency initiation and NWME approval/acceptance to trigger automated data transfers.
- Uses commonly available technologies that are operating system independent on the agency end.
- Uses an extensible design that allows integration of data validation processes such as recommended in Section 5 discussing six new formats.

The integration of the ESV system with other NARA/NWME systems will require that new security hardware and software be put into place

3.3 Electronic File Transfer Recommendations

- Implement a new, secure, media-less electronic transfer process.
- Replace the current FTP process with a user-friendly process based upon web-driven automation.
- Integrate the new, secure file transfer process with other NWME software and processes.
- Optimize the process for file sizes that are practical for electronic transfers.

Together these recommendations will…

- Provide a contemporary, secure means for electronic transfers
- Computer-automate many of the manual operations involved with data transfers and reduce the steps in the incoming accession process
- Provide a simpler, more user-friendly process for the client-Agencies
- Make for fewer physical media types to process and fewer customized secure deliveries and receipts
- Reduce the paperwork and approval processes
- Improve compliance to archivist requirements
- Provide a direction for the future with a process that is more ERA-likely

4. Mechanism for Data Interchange between Two Different Networks[17]

4.1 Data Interchange Discussion

One additional consideration of the network redesign is that of Secure data sharing between two different networks. A solution to this problem will provide a single metadata entry to be shared by the multiple databases in NWME operations. It will allow authorized personnel to process data from their own desks and have full and secure access to NWME system-wide data.

Furthermore and perhaps of greater significance is that a redesign proposal for data sharing embodies almost all elements of this project. A secure network interconnect will improve operational efficiency, increase throughput, provide ease of access to files and make the ERA transition less complex.

4.2 Constraints of Present Operations

A flow chart of NWME operations and discussions with NWME personnel indicated that current methods require certain repeated data entries for application databases and result in process delays and room for error. The network segregation is clearly the primary roadblock. Current technologies can however provide a secure way to integrate the networks making data-sharing possible. The following Router Cluster Approach to Network Security may likely resolve the major issue of 'NH Security & Approval'

4.3 Benefits of the Router-Cluster Design

A Router Cluster connected to the new network architecture can serve as a secure gateway to and from closed (APS LAN) and open networks (NARAnet). This new architecture also makes it easier to design more user-friendly electronic transfers that are appropriately secure and convenient for the user-Agency as well as NARA.

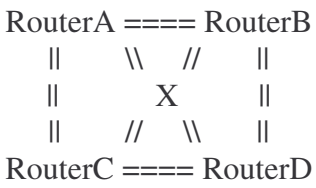
Security considerations are dealt with from the very beginning of the new design starting at the physical layer. Thus, sharing common data between a closed network (APS) and open net work (NARA net) and any other organization simply becomes a matter of

[17] Data sharing was not a part of this contract SOW. Throughout this Study, secure data sharing was identified and highlighted as an important issue for all NWME operations. It seemed obvious to include this issue and its solution in the study and the network/communication re-design.

passing (secure) electronic messages from one to another rather than the any of the ad hoc methods, often in the "sneaker net" category, currently used.

4.4 The Router-Cluster Design

ARkival believes many of the NARA/NWME security concerns can be addressed by the Router Cluster design. This revised network strategy is being pursued as an IT industry "Best Practice" for securely connecting two networks:



Routers A and B are administered by Network 1. Routers C and D are administered by Network 2. This means neither network affects the internal policies of the other.

The dual connections between routers and the cross connect make sure there is no single point of failure. For best results each router would be on a separate electrical circuit, and there will be a spare router available for rapid swap should one of the routers fail. The connections between the AB pair and the CD pair would be sized to guarantee appropriate bandwidth between the two networks. In extremis, the two networks can even be physically isolated from each other by simply pulling the AB to CD connections. The Router Cluster design includes security-based software that involves routing tables. Because routers are being used as the connection means different security measures have been considered[18].

A Router Cluster connected to this network architecture can serve as a secure gateway to APS for electronic transfers received on NARAnet and linked to the re-designed APS/LAN network.

Addressing Security

Because routers are being used as the connection, network layer 1-3 security policies can be implemented that are resistant to malware such as user datagram protocol (UDP)

[18] Network layer 1-3 security policies can be implemented that are resistant to malware such as User Datagram Protocol (UDP)-based worms and viruses that are impossible for network layer 4-7 or TCP based security measures to control. Given that variants of the UDP based Blaster worm are now considered "standard practice" in the Black Hat community, protecting against UDP based threats when designing network connectivity schemes is imperative.

based worms and viruses that are impossible for network layer 4-7 or TCP based security measures to control. Given that variants of the UDP based Blaster worm are now considered "standard practice" in the Black Hat community, protecting against UDP based threats when designing network connectivity schemes is imperative. In extremis, the two networks can even be physically isolated from each other by simply pulling the AB to CD connections. This level of security and access control is superior to any other known solution although other alternatives were also considered[19].

The dual connections between routers and the cross connect make sure there is no single point of failure in the set up. For best results each router should be on a separate electrical circuit, and there should be a spare router available for rapid swap in should one of the routers fail. The connections between the AB pair and the CD pair can also be sized to guarantee appropriate bandwidth between the two networks. The router-cluster for the application must be carefully specified and proofed with critical components and accompanying software to provide performance at peak efficiency.

4.5 Recommendation- Mechanism for Data Interchange between Two Different Networks[20]

- Establish a Router Cluster as the secure bridge between networks
- Determine methods to facilitate easy data sharing using the new connectivity

These recommendations will…

- Provide secure data sharing between a closed network (APS) and open net work (NARA net).
- Establish a single point where networks are cross-connected.
- Provide ease to administer the network
- Isolate the two networks to which a given workstation is connected.
- Reduce network traffic
- Provide security against viruses propagated within a UDP packet.

[19] See APPENDIX A.2 Discussion of Dual Network Card Design

[20] Secure data sharing was not a part of this contract SOW. A solution is included along with a proposed network/communication re-design because of its perceived applicability.

New Format Accessions

5.0 Overview

This part of the study evaluates available COTS[21] products and recommends those that can be used to accession agency records in the six new formats. ARkival also proposes a phased implementation that will result in a single NARA solution that can process the six new formats, future files that include audio and video data, as well as current flat files and database records.

5.1 COTS Products Evaluated

Email Messages with Attachments

- AdminSystem ANPOP POP3 Component
- Legato EmailXtender & EmailXaminer
- Weird Kid Software Emailchemy
- CompuSven
- Gens Software Ltd
- Wingra Technologies

Portable Document Format (PDF)

- JHOVE (JSTOR (Scholarly Journal Archive)/Harvard Object Validation Environment)
- *Enfocus PitStop Professional and PitStop Server
- Callas Pdfinspektor2, process|prepress Autopilot, and PDFengine
- Markzware Flight Check Professional
- *Adobe Acrobat 6.0

Scanned Images of Textual Records

- JHOVE (JSTOR (Scholarly Journal Archive)/Harvard Object Validation Environment).
- Aware Systems
- Trapeze from On Stream Systems
- *IrfanView
- Equilibrium DeBabelizer
- Global Image Viewer from Paragon Imaging
- *Stellent OutsideIn SDK
- Snowbound
- Shaffstall

[21] The term COTS in this study relates to COTS, open source and freeware.

* Indicates COTS Products used by NWME for test processing of documents in new formats

Digital Photographic Records

- JHOVE (JSTOR (Scholarly Journal Archive)/Harvard Object Validation Environment)
- *IrfanView
- *Adobe Photoshop Elements
- Aware Systems
- Trapeze from OnStream Systems
- *Stellent OutsideIn SDK
- Equilibrium DeBabelizer
- Snowbound
- Shaffstall

Digital Geospatial Data Records

- Autodesk Mapguide
- ∗Caris Easy-ENC 3.0
- *ESRI ArcGIS 9
- GE Energy Smallworld
- ∗Global Mapper (USGS Dlgv32pro)
- ∗Intergraph GeoMedia
- Leica ERDAS
- ∗MapInfo Professional
- Oracle Spatial and Locator
- ∗PCI Geomatica Freeview V9.1
- ∗Safe Software Feature Manipulation Engine Suite

Web Content Records

- W3C (World Wide Web Consortium) Markup Validator
- Cascading Style Sheet (CSS) Validator
- W3C XML (eXtensible Markup Language) Schema Validator
- The CPAN Module HTML::Tidy
- *W3C Link Checker
- *Xenu Link Sleuth

5.2 COTS Evaluation Data Tables

The following analysis by format type, evaluates selected COTS alternatives to their compliance with NARA guidelines and NARA defined specifications. A common set of notations has been used throughout the tabular reporting with exceptions footnoted.

Standards & Guidelines

The study's Statement of Work [22] (SOW) and the more recent web-published NARA Guidelines[23] provide specifications and guidelines for compliance. The latter dated web-published guidelines provided additional specifications for certain formats. These additional specifications have been included in the tabular evaluation. The web-published guidelines are referenced at the end of this report.

TABULAR COMMENTS *(for all 6 Formats)*

Y	Meets requirement out of the box
N	Does not meet requirement out of the box
U	Unknown if it meets requirement out of the box *(insufficient information available from developer)*
N/A	Not applicable
M1	Can meet requirement via available source code from software developer *(An open source code license may be required)*
M2	Can meet requirement with user programming of "tools" from the software developer. *(User-programming will be required to apply the tools to perform specified NARA verifications*
M3	Could meet requirements by modification of the source code if developer agrees to the modification and provides additional information.
M4	Could technically meet requirements by modification but unlikely that the source code developer will agrees to modifications regarding format verification. No 'M4" software is being recommended.
T	Meets requirement with manual review (for temporary use until automated processing software can be written)
S	Redirection (See description in another format section)

5.3 AERIC-like Functionality & the Six+ New Formats

The AERIC system is used to verify structured electronic data files. AERIC verifies that the data files received by NARA from Federal Agencies match the documentation sent by the Agencies.

[22] NARA contract NARA-04-0055; Section 3.2 pp5-8
[23] NARA website: http://www.archives.gov/records_management/initiatives/transfer_to_nara.html.

The present verification process consists of comparing the actual format and content of the data files received, to the description of the content as represented by the record layouts and codes provided by the client-agencies.[24] In effect, the AERIC system is used to verify structured electronic data files and is based upon an agency defined Table using a record layout provided by that same agency. AERIC uses this layout to create a new Table in the associated Oracle database. Historically, AERIC is operative on data files that have a similar metadata structure and can be classified as "flat files".

The introduction of the six new formats to NWME imposes a new challenge to the AERIC process because they have a "rich file" structure. Rich file structures are not as simple as the flat file structures and usually represent a collection of file information provided as embedded metadata and content.

Although a significant effort could be made to modify the AERIC software to accommodate rich file structures, such an investment is not justified because of the age of AERIC software and the complexity of enriching the software for these six new formats and others that are likely to follow.

Accessioning of rich file formats has necessitated that NARA/NWME establish new verification guidelines. New guidelines have been provided both in the Study's Statement of Work (S.O.W) and the more recent web Guidelines published on NARA's website.

5.4 New Format Metadata and Existing Databases

The "rich text" files, after being validated can be processed in the APS system just like flat files are processed today. The metadata forms submitted by the agencies can be entered- manually typed, if written or file-copied, if submitted electronically - into the APS catalog database.

In the future, the process can be facilitated via software. The JHOVE software architecture supports the use (and capture) of metadata while verifying the records. The metadata could also be electronically copied to the catalog database. The COTS software products could also be modified (with varying levels of complexity) to automatically capture corresponding metadata and enter same into existing databases.

[24] Archival Electronic Records Inspection and Control (AERIC) System- User's Manual June 07, 2004 Version 1.4

5.5 Requirements and COTS evaluations

Requirements and related evaluation criteria for applicable software alternatives were obtained from:

- The RFQ; NAMA-04-Q-0016; System Enhancement Study for Preserving and Validating Terabytes of Electronic Records Transferred to NARA on Multiple Media Types and via FTP.
- NWME COTS test results for PDF, web, photographic images, and GIS records.
- NARA Expanding Acceptable Transfer Requirements- web pages for all six new formats
- AERIC and APS user manuals
- Meetings at Archives II.

The identification, study and evaluation of COTS, Open Source, and Freeware products to the NARA requirements were performed via ARkival's subject-matter specialists. The identification of COTS alternatives for each format type was based upon working knowledge and/or experience with the formats, focused internet searches, vendor discussions and limited trials. The study also included searching web sites, reviewing of marketing documents, user and technical manuals and phone conversations and meetings with vendors. In some instances evaluation copies were downloaded and limited testing performed.

5.6 COTS Evaluations by Format Type

Summary of Applications' Ability to Validate Electronic Files and Compliance with NARA Guidance for Transfer of Six New Formats

5.6.1 Email Messages with Attachments

Requirement Enumeration	Requirements	1. AdminSystem ANPOP3 5.5	2. Legato EmailXtender & EmailXaminer	3. Weird Kid Emailchemy	4. CompuSven E-Mail Shuttle	5. Gens Software	6. Wingra
1a.	*Requirement 1a: Verify that a message in a transmitted file is in a standard markup language or in some native format.* **Test: Can determine message SML or native format**	Y	Y	Y	Y	Y	Y
1b.	*Requirement 1b: Verify that each attachment in a transmitted file is in a standard markup language or in some native format.* **Test: Can determine attachment SML or native format**	Y	Y	Y	Y	Y	Y
2a.	*Requirement 2a: Verify that messages and attachments form an identifiable, organized body of records.* **Test: Can determine identifiable, organized body of records**	Y	Y	Y	Y	Y	Y
2b.	*Requirement 2b: Messages and attachments must originate from a DOD 5015.2-STD RMA system or from an e-mail system.* **Test: Can determine origin of messages and attachments**	Y	Y	Y	Y	Y	Y
3a.	*Requirement 3a: Verify that e-mail content fields include labels providing the:* *i) Date* *ii) To & CC recipients* *iii) From* *iv) Subject* *v) Transmission showing time sent* *vi) Receipt showing time opened* *vii) Message size* *viii) File name* **Test: Can detect and report header information**	M3 1,2	M3 2	M3 1,2	M3 1,2	M3 1,2	M3 1,2
3b.	*Requirement 3b: Verify that messages and attachments are delimited at their beginning and end, and that attachments are properly separated from the message body.* **Test: Can detect delimiters for messages and attachments**	Y	Y	Y	Y	Y	Y
4.	*Requirement 4: Verify that each attachment is labeled with a filename and a file extension to indicate the proprietary software used to create the attachment. Report if that information is missing.* **Test: Can identify the attachment type using file extension or other information, reporting errors.**	Y	Y	Y	Y	Y	Y

Requirement Enumeration	Requirements	1. AdminSystem ANPOP3 5.5	2. Legato EmailXtender & EmailXaminer	3. Weird Kid Emailchemy	4. CompuSven E-Mail Shuttle	5. Gens Software	6. Wingra
5.	*Requirement 5: "COTS" or support exists for conversion of such.* **Test: Final product exists today.** **Test: Support for conversion exists.**	N N	N U	N Y	N Y	N Y	N U
6	*Requirement 6: Deals with mail messages in files after they have been received and possibly archived.* **Test: Examines entire mail files and archives.**	N	N	Y	Y	Y	Y
A.	*Additional feature A: Runs on Windows*	Y	Y	Y	Y	Y	Y
B.	*Additional feature B: Runs on UNIX/Linux*	N	N	Y	Y	U	N
C.	*Additional feature E: Provides batch processing capabilities*	N	N	M3	N	N	N
D.	*Additional feature F: Batch with summary reporting capabilities*	N	N	M3	N	N	N
E.	*Additional feature G: Ability to correct detected problems*	N	N	M3	N	N	N
F.	*Additional feature H: Open source software*	N	N	M3	N	N	N
G.	*Additional feature I: Freeware*	N	N	M3	N	N	N
H.	*Additional feature G: Ability to correct detected problems*	N	N	N	N	N	N
I.	*Additional feature H: Open source software*	N	N	N	N	N	N
J.	*Additional feature I: Freeware*	N	N	N	N	N	N

1. No user-readable reporting exists today, but could be added.

2. The application preserves all headers associated with both a message and the attachments to it.

3. Would need modifications to mail reader to capture information such as time message was first opened.

Recommendations

Emailchemy by Weird Kid Software is ARkival's recommendation for NARA's verification of e-mail files. It comes closest to satisfy the following criteria:

- Verifies that a message and each attachment in a transmitted file is in a standard markup language or in some native format.
- Verifies that messages and attachments form an identifiable, organized body of records.
- Verifies that messages and attachments originate from a DOD 5015.2-STD RMA system or from an e-mail system.
- Verifies that e-mail content fields include necessary labels

- Verifies that messages and attachments are delimited at their beginning and end, and that attachments are properly separated from the message body
- Verifies that each attachment is labeled with a filename and a file extension to indicate the proprietary software used to create the attachment. While missing information is detected today, Emailchemy would require modifications to create a report showing the list of such problems.

Discussion

Emailchemy requires some modification and customization to meet NARA's requirements. Emailchemy however, understands and interprets the greatest number of e-mail formats and makes it a likely alternative to process the different formats of e-mail files received by NARA, now and in the future.

Should it be necessary to verify mail from a DOD 5015.2-STD RMA system, Emailchemy will need additional modification.

The capture of the time when a message is first opened depends on a mail reader not the verification system. This requires adding to a mail message by the mail reader, i.e. the message itself would require modification if the information is passed to verification software. Interpretation and reporting of such times would require additional modification to Emailchemy.

Emailchemy runs on the Windows and UNIX operating systems.

5.6.2 Portable Document Format (PDF) Records

Requirement Enumeration	Requirements	1. JHOV	2. Enfocus PitStop	3. Callas PDFInspector2	4. Markzware Flight Check	5. Adobe Acrobat 6.0
1.	*Comply with PDF v 1.0 through v 1.4* **Test: Can determine version of PDF file.**	Y	Y	Y	Y	Y
2.	*Records must not contain security settings (e.g., self-sign, user passwords, and/or permissions) that prevent NARA from opening, viewing, viewing or printing the record.* **Test: Can determine status of security settings.**	Y	Y	Y	Y	Y
2.1	*Records created after April 1, 2004 must have all security settings deactivated (e.g., encryption, master passwords, and/or permissions) prior to transfer to NARA.* **Test: Can determine status of security settings.**	Y	Y	Y	Y	Y
3.	*Because of the complexities associated with certain PDF features, NARA will review PDF records containing special features on a case-by-case basis when the records are scheduled. Examples of special features include but are not limited to: digital signatures; links to other documents, files or sites; embedded files (including multimedia objects); form data; comments and/or annotations.* **Test: Can determine presence/absence of special features.**	Y	Y	Y	Y	Y
4.	*Electronic records that have been converted to PDF from their native electronic formats must include imbedded fonts to guarantee the visual reproduction of all text as created.* ***Test: Can identify converted PDF files and determine presence/absence of embedded fonts.*** *All fonts embedded in PDF records must be publicly identified as legally embeddable (i.e., font license permits embedding) in a file for unlimited, universal viewing and printing* ***Test: Identify fonts and check against a font license table.***	Y Y	Y Y	Y Y	Y Y	Y Y
5.	*PDF records that reference fonts other than the "base 14 fonts" must have those fonts referenced in the record (e.g., as a minimum, subsets of all referenced fonts) embedded within the PDF file.* **Test: Can identify fonts not "base 14" and presence/absence of full, or subsets of, embedded fonts.**	Y	Y	Y	Y	Y
6.	*PDF records created after April 1. 2004 must have all fonts referenced in the record, including the "base 14 fonts", embedded within the PDF file. This requirement is met by having, as a minimum, subsets of all referenced fonts embedded in the PDF file.* **Test: Can identify file creation date and presence/absence of full, or subsets of, embedded fonts.**	Y	Y	Y	Y	Y
7.	*Scanned images of textual paper records converted to PDF must adhere to the requirements in NWM 02.2003, MEMORANDUM TO AGENCY RECORDS OFFICERS: Expanding Acceptable Transfer Formats: Transfer Instructions for Scanned Images of Textual Records (Scanned Images Transfer Guidance), dated December 23, 2002.***See Section 3.c.**	S	S	S	S	S

Requirement Enumeration	Requirements	1. JHOV	2. Enfocus PitStop	3. Callas PDFInspector2	4. Markzware Flight Check	5. Adobe Acrobat 6.0
7.1	*Any agency that has PDF records that have not been scanned according to the minimum image quality specifications in the NWM 02.2003 guidance, should contact the NARA appraisal archivist assigned to that agency.* **Test: Can determine if file meets minimum image quality specifications. See Section 3.c.**	Y	Y	Y	Y	Y
8.	*PDF records that contain embedded searchable text based on Optical Character Recognition (OCR) must be identical in content and appearance to the source document. NARA will accept PDF records with uncorrected OCR'd text. It will not accept PDF records resulting from OCR processes that either alter the content or degrade the quality of the original bit-mapped image.* **Test: Can identify OCR'd text.**	Y	Y	Y	Y	Y
8.1	*NARA will accept PDF records that have been OCR'd using processes that do not alter the original bit-mapped image (e.g., Searchable Image – Exact).* **Test: Can identify text as PDF Searchable Image (Exact).**	Y	Y	Y	Y	Y
8.2	*NARA will not accept PDF records that have been OCR'd using processes that substitute OCR'd text for the original scanned text within the bit-mapped image. (e.g., Formatted Text and Graphics and PDF Normal).* **Test: Can identify text as Formatted Text and Graphics, and PDF Normal.**	Y	Y	Y	Y	Y
8.3	*NARA will not accept PDF records that have been OCR'd using processes that use lossy compression to reduce file size (e.g., JPEG, Searchable Image - Compact).* **Test: Can identify text as Searchable Image (Compact).**	Y	Y	Y	Y	Y
	Other Product Features					
9.	Provides batch processing capabilities	Y	Y	Y	Y	Y
9.1	Batch processing with summary reporting capabilities	Y	N	Y	Y	Y
10.	Optional ability to correct detected problems	N	Y	Y	N	N
11.	Application customization software tools	Y	Y	Y	N	N
12.	Maintenance service contract available	N	Y	Y	Y	Y
13.	Verifies file formats, field sizes, specific values, and ranges of values	Y	N	N	N	N
14.	Provides an image viewer	Y	Y	Y	Y	Y
15.	Supports statistical sampling	M1	N	N	N	N

Requirement Enumeration	Requirements	1. JHOV	2. Enfocus PitStop	3. Callas PDFInspector2	4. Markzware Flight Check	5. Adobe Acrobat 6.0
16.	Metadata review capability	Y	Y	Y	Y	Y
17.	Runs on Windows systems	Y	Y	Y	Y	Y
18.	Runs on UNIX/Linux systems	Y	N	Y[26]	N	Y
19.	Open source software	Y	N	N	N	N

Recommendations

ARkival recommends the use of **Adobe Acrobat V6.0** for validating PDF documents. Adobe recently released V6.0 with an embedded "preflight" module that offers competitive functionality to all the other COTS products in the table above. In addition it provides batch processing with both file-level and summary reporting capability.

None of the COTS alternatives validate file structure, as does AERIC. Because PDF documents are scheduled to be accessioned, ARkival recommends that NARA use the metadata reporting capabilities of Adobe Acrobat along with the image viewer to process submissions.

Discussion

The batch processing reporting features of Adobe Acrobat V6.0 are not available in the PitStop products tested by NWME. Additionally, Adobe is preferred because the Callas and Markzware products are Adobe plug-ins, requiring Adobe Acrobat which contains the same capabilities.

[26] Only Non-GUI (Graphic User Interface) Library modules are available for Solaris and Linux

5.6.3 Scanned Images of Textual Records

Requirement Enumeration	Requirement	1. JHOVE	2. IrfanView	3. Global Image Viewer	4. Aware Systems	5. Trapeze	6. OutsideIn	7. DeBabelizer	8. Snowbound	9. Shaffstall
1.	Verify scanned images of textual records: **Test: Verifies internal consistency contents of metadata fields present in the scanned image files for the file types listed below.**									
2.	**Ascertain the file format transferred**:									
2.1	TIFF	Y	T	U	M2	M3	T	M3	M2	N
2.2	GIF	Y	T	U	M2	M3	T	M3	M2	N
2.3	BIIF	M1	N	T	N	N	N	N	N	N
2.4	PNG	M1	T	U	M2	M3	T	M3	M2	N
3.	Ascertain that scanned images of textual records satisfy the minimum requirements related to image resolution and pixel (bit) depth: **Test: range check on each metadata value**	M1	T	M4	M2	M3	T	M4	M2	N
4.	Image Viewer (Implied Requirement) **Test: ability to view images listed below**									
4.1	TIFF	M1	Y	Y	M2	Y	Y	Y	M2	N
4.2	GIF	M1	Y	Y	M2	Y	Y	Y	M2	N
4.3	BIIF	M1	N	Y	N	N	N	Y	M3	N
4.4	PNG	M1	Y	U	M2	Y	Y	Y	M2	N
5.	**Recommendations**									
5.1	Recommended	M1	Y	N	N	N	N	N	N	N
5.2	Recommended for software viewing array for high priority extra-guidelines documents.	Y	Y	Y	Y	Y	Y	Y	Y	Y
6.	**Other Product Features**									
6.1	Additional feature A: Runs on Windows	Y	Y	Y	Y	Y	Y	Y	Y	Y
6.2	Additional feature B: Runs on UNIX/Linux	Y	N	N	N	N	Y	N	Y	N

Recommendations

JHOVE is the most complete software for Scanned Textual Records and can eventually provide NARA automated verification and batch processing with software modifications. It is not being recommended at this time because of its pre-release status.

IrfanView is the preferred COTS application for the manual viewing of statistically sampled Scanned Textual records for the purpose of verifying the record format, verifying that the metadata values for resolution and bit depth, and for visually verifying the scanned textual records.

If required, **Global Image Viewer** can be used for viewing Basic Interchange (BIIF) Format Standard 12087-5 (ISO) Part 5 (Dec. 1998).

Discussion

ARkival is recommending IrfanView for viewing images and for checking the metadata in the files, including resolution and bit depth until automated testing can be provided.

JHOVE is believed to be the easiest to modify, it is also open source and focused on archival applications. JHOVE is also likely to incorporate changes made for NARA into its 'official' product and to maintain such changes in the future. JHOVE's architecture already supports batch processing and file format validation. Adding metadata parameter testing (for resolution and bit depth) and statistically sampled viewing is easier than adding file validation to the commercial products that do image viewing.

ARkival is also recommending that image viewing of scanned textual images be maintained as a validation tool because direct viewing can identify problems that digital imaging technology may have caused. The COTS Snowbound application provides a "tool box" that can be software-implemented to perform automated verification. For documents received in unknown format and when off the shelf viewers are insufficient to view documents, ARkival recommends considering the tool box enhancements of Snowbound.

5.6.4 Digital Photographic Records

Requirement Enumeration	Requirement	1. JHOVE	2. IrfanView	3. Photoshop Elements	4. Aware Systems	5. Trapeze	6. OutsideIn	7. DeBabelizer	8. Snowbound	9. Shaffstall
1.	Verify image file format of digital photographic images **Test: Verify TIFF and JPEG file formats**									
1.1	TIFF	Y	T	T	M2	T	T	M4	M2	N
1.2	JPEG	Y	T	T	M2	T	T	M4	M2	N
2.	1b: Verify grayscale images are 8-bit (8 bit image) or 16-bit (16 bit image) per channel and color images (including red-green-blue images) are 8 bits (24 bit image) or 16-bit (48 bit image) per channel. **Test: Verifies metadata values.**									
2.1	TIFF	M1	T	T	M2	T	T	M4	M2	N
2.2	JPEG	M1	T	T	M2	T	T	M4	M2	N
3.	For Tagged Image File Format (TIFF) files: confirm that file extensions include .TIFF and .TI format, versions 4.0, 5.0, and 6.0 4. **Test: verifies TIFF format versions.**	Y	T	T	M2	T	T	M4	M2	N
4.1	Confirm that JFIF and JPEG images are compliant with (ISO/IEC) standard 10918-1 (1994): **Test: Verifies JFIF and JPEG format versions.**	Y	T	T	M2	M4	T	M4	M2	N
4.2	Verify that Default file extensions include .JPEG, .JFIF, and .JPG: **Test: Verifies JPEG, JFIF, and JPG file extensions.**	Y	T	T	M2	M4	T	M4	M2	N
5.	**Test: Verifies metadata values** (megapixels and array size)	M1	T	T	M2	T	T	M4	M2	N
6.	**Recommendation**									
6.1	Recommended	M1	Y	Y	N	N	N	N	N	N
6.2	Recommended for software viewing array for high priority extra-guidelines documents.	Y	Y	Y	Y	Y	Y	Y	Y	Y
7.	**Other Product Features**									
7.1	Additional feature A: Runs on Windows	Y	Y	Y	Y	Y	Y	Y	Y	Y
7.2	Additional feature B: Runs on UNIX/Linux	Y	N	N	N	N	Y	N	Y	N

Recommendations

JHOVE is also the most complete software for Scanned Photographic Records and can eventually provide NARA automated verification and batch processing with software modifications. It is not being recommended at this time because of its pre-release status.

IrfanView is the preferred COTS application for the manual viewing of Scanned Photographic records. Adobe Photoshop Elements can also be used for manually browsing the photographic images.

Discussion

ARkival is recommending IrfanView for viewing photographic images and for checking the metadata in the files (e.g., resolution, bit depth, caption information, et.al.) until automated testing can be provided.

JHOVE is believed to be the easiest to modify, it is also open source and focused on archival applications. JHOVE is also likely to incorporate changes made for NARA into its 'official' product and to maintain such changes in the future. JHOVE's architecture already supports batch processing and file format validation. Adding metadata verification for resolution and bit depth and statistically sampled viewing is easier than adding file validation to the commercial products that do image viewing[27]. Exports of metadata can also be structured to adhere to the ASCII-oriented, non-proprietary standards that the CFR requires for databases.

ARkival is also recommending that image viewing of photographic images be maintained as a validation tool because direct viewing can identify problems that digital imaging technology may have caused. The COTS Snowbound application provides a "tool box" that can be software-implemented to perform automated verification. For documents received in unknown format and when off the shelf viewers are insufficient to view the photographic images, ARkival recommends considering the tool box enhancements of Snowbound.

[27] The significance of Caption information has been recognized and can likely be incorporated into a recommended COTS alternative for processing metadata in both *EXIF and IPTC.*

5.6.5 Digital Geospatial Data Records

Requirement Enumeration	GIS Requirements	1. ESRI ArcGIS	2. Intergraph	3. Autodesk	4. GE Energy	5. Leica	6. MapInfo	7. Caris	8. Global Mapper	9. Oracle	10. PCI	11. Safe Software
1.	**GML Support**											
1.1	Read	Y	Y	N	N	N	N	N	N	N	Y	Y
1.2	Import	Y	Y	N	Y	N	Y	N	N	N	Y	Y
1.3	Export	Y	Y	N	Y	N	N	N	N	N	Y	Y
2.	**SDTS Support**											
2.1	Read	Y	Y	N	N	N	N	N	Y	N	Y	Y
2.2	Import	Y	Y	N	Y	Y	Y	N	Y	N	Y	Y
2.3	Export	N	N	N	N	N	N	N	N	N	N	N
3.	**MIL-STD-2407**[28]											
3.1	Read	Y	U	U	U	U	U	U	U	U	U	Y
3.2	Import	N	U	U	U	U	U	U	U	U	U	Y
3.3	Export	N	U	U	U	U	U	U	U	U	U	Y
4.	**MIL-STD-2411**[29]											
4.1	Read	Y	U	U	U	U	U	U	U	U	U	N
4.2	Import	N	U	U	U	U	U	U	U	U	U	N
4.3	Export	N	U	U	U	U	U	U	U	U	U	N
5.	**SDSFIE**[30]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
6.	**Developer Tools**	Y	Y	Y	Y	Y	Y	N	N	Y	Y	Y
7.	**Provides batch processing capabilities**	N	N	N	N	N	N	N	N	N	N	N
7.1	**Batch processing with summary reporting capabilities**	N	N	N	N	N	N	N	N	N	N	N
8.	**Optional ability to correct detected problems**	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
9.	**Application**	Y	Y	Y	Y	Y	Y	N	N	Y	Y	Y

[28] Vector Product Format (VPF) data is stored in a structure described in the Military Standard, Vector Product Format, MIL-STD-2407. It describes the structure and format conventions which must be met for a dataset to be considered a VPF data. It would best be described as an "architecture" specification, rather than a specific format. The specs first came out in 1992. There are a number of formats that follow the VPF specification, such as the Digital Chart of the World (DCW), Digital Nautical Chart (DNC), VMap Level 0, VMap Level 1, and UVMap database standards. ESRI can read some of these formats; Safe Software products can read and write some of these formats.

[29] Raster Product Format (RPF) data is stored in a structure described in the MIL-STD-2411. Specs for this came out in 1994. Format examples include Compressed ARC Digitized Raster Graphics [CADRG] and Controlled Image Base [CIB]). ESRI can read some of these formats; Safe Software products are anticipated to support raster formats in 2005.

[30] The SDSFIE standard was developed by the CADD/GIS Technology Center for Facilities, Infrastructure, and Environment (formerly the Tri-Service CADD/GIS Technology Center). The standard is not a data format, but a data ***content*** standard designed for use with the predominate COTS GIS, CADD, and relational database software. (Ex.: Naming conventions for tables and attribute fields.) The CADD/GIS Technology Center has detailed documents explaining how to implement the standard in ArcView, Autodesk, and others. ESRI and Autodesk data, for example, can both be SDS compliant.

Requirement Enumeration	GIS Requirements	1. ESRI ArcGIS	2. Intergraph	3. Autodesk	4. GE Energy	5. Leica	6. MapInfo	7. Caris	8. Global Mapper	9. Oracle	10. PCI	11. Safe Software
	customization software tools											
10.	**Maintenance service contract available**	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
11.	**Verifies file formats, field sizes, specific values, and ranges of values**	N	N	N	N	N	N	N	N	N	N	N
12.	**Provides an image viewer**	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
13.	**Metadata review capability**	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
14.	**Runs on Windows systems**	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
15.	**Runs on UNIX/Linux systems**	Y	N	Y	Y	Y	N	N	N	Y	Y	Y
16.	**Open source software**	N	N	N	N	N	N	N	N	N	N	N

Recommendations

Two COTS alternatives are recommended: **ESRI ArcGIS** and **Safe Software FME**.

For GIS software, **ESRI ArcGIS** is the tool of choice. The product captures 35% of overall market share[31] and is a major leader in numerous sectors. ESRI has just released a new product called the ArcGIS Data Interoperability extension. This toolset enables ArcGIS users to directly read and import more than 70 spatial data formats and export more than 50 formats. Users also have the flexibility to define custom data formats within an interactive visual diagramming environment.

As the leading ETL [Extract, Transform, Load] vendor with plug-ins to all the major GIS vendors, **Safe Software** is ARkival's choice for providing additional ETL tools. The FME Viewer, while a bit like some of the other examined free viewers for GIS data, goes farther. It includes a deeper view into the file, displaying coordinate information and tools to do queries (filters). This viewer is designed for those involved in translations with results that can be graphically examined during processing. Safe Software has recently released a new add-on called FME 2004 ICE. This product contains over 600 user-requested enhancements, as well as several new and updated formats that bring the total number of FME supported formats to over 130. FME 2004 ICE is also able to plug into and extend a number of other vendor GIS products.

[31] Public Sector 58%, Transportation 49%, Education 72%, Marketing & Sales 50%, Private Industry 37%, and Unix Workstation 56%

ARkival believes the recommended ESRI ArcGIS application with the ETL product, creates a COTS GIS solution for NARA.

Discussion

ARkival's study recommendations will allow NARA to verify GIS accessions. Determining what geospatial information is valuable will require an understanding of the submitting agency's GIS system and dataset.

A fundamental goal of any data transfer standard is to accomplish the transfer without losing information. For spatial data, information is composed of greater detail than just spatial objects and descriptive attributes. It also includes a description of the processing steps, data extraction guidelines used, positional accuracy reports, and many other similar items. The data producer must provide complete information – a subjective responsibility that can vary depending on how the data was and will be used.

Of the geospatial data standards acceptable under current NARA guidelines, GML is the recommended choice for at least the foreseeable future. GML is emerging as the industry and government standard of choice. Vendor applications support for GML is increasing. There will be some agencies with geospatial data based on the Military Standards (2407 and 2411) and other "non-proprietary, published, open standard maintained by or for a Federal, national or international standards organization." If submitted to NARA they may be validated using ArcGIS or FME. If they are among the ones that are not supported, the accessioning architect will have to determine the best method of validating the data.

The vast majority of data used by submitting agencies will not however, be in any standards-based format. To comply with NARA guidance, the submitting agency must take responsibility for proper conversion of its native datasets to GML Translation of data from one format to another can introduce error. The agency must know what issues are involved in the translation process, what must be checked or monitored, and what must be supplied by the agency. By comparing its native datasets (shapefiles, coverages, dwg, dgn, tab, etc) to the outputted GML files (perhaps using recommended software), the submitting agency analyst can evaluate whether the conversion was a success.

In addition to the conversion to GML, it is recommended that the original GIS format be preserved along with the GML. This recommendation is being made to maintain an alignment with the archival practice of endeavoring to change accessioned records as little as possible.

5.6.6 Web Content Records

Summary of Applications' Ability to Validate Electronic Files and Compliance with NARA Guidance for Transfer of Web Content Records

Requirement Enumeration	Requirement	W3C Markup Validator	W3C CSS Validator	W3C XML Schema Validator	W3C Link Checker	Xenu Link Sleuth	Perl CPAN module HTML::Tidy
1.	Verify HTML content	M2	M2	N/A	N/A	N/A	M3
2.	Verify XML content	N/A	N/A	M2	N/A	N/A	N/A
3.	Verify Links	N/A	N/A	N/A	Y	Y	N/A
A.	Additional Feature A: Runs on Windows	Y	Y	Y	Y	Y	Y
B.	Additional Feature B: Runs on Linux/UNIX	Y	Y	Y	Y	Y	Y

Recommendations

ARkival recommends **W3C (World Wide Web Consortium) utilities** as the best COTS alternative for validating web content. The W3C offers[32] a set of validation utilities for markup validation, cascading style sheet (CSS) validation, and Extensible Markup Language (XML) validation that taken together are capable of addressing NARA's requirements and guidelines.

The W3C utilities are currently designed for a user-interactive session to perform validation on a single URL or file[33]. The validation performed is very high quality and takes into account a wide variety of different technical details that are possible to encounter in web content. The research determined that modification of the code to facilitate the use of a batch oriented processing model is also feasible. Therefore, ARkival suggests a new user interface to the base code be developed that provides batch processing capability.

[32] http://validator.w3.org
[33] When the source code is obtained and built locally, the default configuration expects the code to run in the context of a web server process that validates a single URL/file at a time.

While not a validation process per se, ARkival also recommends the inclusion of a 'link check' process as a means of determining that the web content records have no omissions and that all external links have actually been removed in accordance with the NARA accession guidelines for web content. ARkival reviewed the Xenu Link Sleuth and the W3C LinkChecker applications. While they are comparable in capability we are recommending the use of the W3C LinkChecker because it is not constrained for use on Windows systems only while the Xenu product is. The W3C utility is also available as source code and can be modified as necessary to provide additional functionality as may be required by NARA.

Discussion

None of the COTS alternatives provided Web Content Record validation capabilities only. Validation is always bundled into web creation tools as a utility function to be used after creation or modification of web content. The use of a web creation product for validation only would increase the cost, complexity, use and the time required by NARA personnel to perform validation tasks. The COTS alternatives studied were designed for interactive use and provides no means for batch-processing of files.

The new user interface developed for batch processing could also be submitted to the W3C for possible inclusion in their standard distributions of the validator source code packages. This will help to insure that maintenance of the batch user interface will not be dependent on a single source as the source code will be available via the open source licensing model.

ARkival also investigated the use of the Perl CPAN module in HTML::Tidy as a simpler method of doing validation but were unable to establish a working instance of the module on our test configuration. The functionality of the code shows that the W3C modules are far more capable and more of a finished product.

Given the general guidance and requirements from NARA, ARkival's goals were to…

- make allowances for the very broad spectrum of content from which modern web sites are composed
- determine the submitted web content is complete, within a range of probabilities,
- determine the data type of a given component element matches the stated data type,
- identify the member components that make up a single body of web content

5.7 Recommendations of Software Products for Checking Integrity of Six New Electronic Record Formats

In the study findings none of the COTS products (for any of the six formats) will completely satisfy the NARA verification Guidelines without some modification. The selected software alternatives (for each format) were compared to the NARA guidelines and rated accordingly.

The validation requirements generally fall into two groups: (1) record format validations similar to that performed by the AERIC system (e.g. tabular comparisons, statistical and visual samplings) and (2) requirements published in NARA Expanding Acceptable Transfer Requirements for each of the six new formats. In general, the recommended software products have several common elements…

- The COTS products identified were developed for non-archival applications (usually format conversions) but can be utilized to verify compliance with NARA Expanding Acceptable Transfer Requirements

- The recommended COTS products do not provide AERIC-level automated file format verification or statistical sampling. They can, with varying levels of complexity, be modified to do so. The selected products have the potential to verify the six new formats records meet NARA Expanding Acceptable Transfer Requirements, but will require changes to do so.

The COTS products are for the most part, manual, non-batch applications. File and imageviewer products are recommended with sampling to provide record format validations.

In most cases, ARkival believes that the recommended COTS software for each of the formats could be modified to satisfy the major NARA guidelines and requirements. In some COTS applications, contact has been made with the software developers and they agreed with our assessments. It is also recommend that the AERIC-like statistical visual sampling be performed with the six formats[34] and consideration has been included in the software alternative analysis.

Each of the selected software products can also be modified for batch processing of incoming files[35], statistical sampling, metadata review and visual inspection. When commercial software is modified however, assurance[36] is needed that the modification will be included in all future releases in an unchanged form.

---

[34] Primarily for record/file format validation.
[35] Only PDF applications are presently capable of performing the batch processing of records without modification.
[36] In general, smaller commercial companies are more likely to make requested modifications but larger commercial companies are likely to have better business staying power.

5.8 JHOVE – JSTOR[37] Harvard Object Validation Environment Software

In the course of this study ARkival discovered promising software for potential application as a generic solution for all six (and more) formats[38]. This software is open source and will be available for customized modification for all six formats and the batch processing of files. JHOVE is currently in its 4th and last beta release. Production release is expected in early 2005. JHOVE is software worthy of consideration for present and future rich file formats. Specifically, JHOVE will require NARA/NWME profiles for each format verification and development of new modules for document viewing, batch processing, and batch reporting capabilities.

JHOVE presently provides functions to perform format-specific identification, validation, and characterization of digital objects. The initial JHOVE distribution includes the following standard modules…

AIFF	PDF
ASCII	TIFF
BYTESTREAM	UTF
GIF	WAVE
JPEG	XML
JPEG 2000	(S)HTML

Although commercial software could be modified to perform most format verifications, JHOVE was written specifically for this purpose.

5.9 Implementation

ARkival recommends that the six new format accessions to NWME be received and ingested at the re-designed APS/LAN network (e.g. minimizes redundant data entries). The new network re-design integrated with the router cluster application can provide secure shared data via a disc-based NAS[39] throughout NWME. The six new formats can be processed and verified utilizing host software developed to support one or more of the COTS alternatives for each software recommendation. Provision can be made to statistically sample, visually check and verify incoming files. Verification should include both the record's metadata and a visual inspection of the viewable version of the records-similar to current AERIC processing for flat files.

[37] http://www.jstor.org/
[38] The JHOVE software product is the result of a JSTOR and Harvard University Library collaboration project to develop an extensible framework for format validation (funded in part by the Andrew W. Mellon Foundation through a grant to JSTOR for an Electronic-Archiving Initiative).
[39] Network Attached Storage- see earlier deliverable 4.2.2(a)

The implementation of the different COTS applications for each format will involve a staging process. The first phase of the process will do format verification with certain limitations based upon the present capability of each software product. Subsequent phases will include adaptations of COTS alternatives as they are customized for the NARA application.

5. 10 Recommendations of Software Products for Checking Integrity of Six New Electronic Record Formats & Implementation

5.10.1 Software Recommendation by Format File Type

- **Email Messages with Attachments**
  Weird Kid Software Emailchemy
- **Portable Document Format (PDF)**
  Adobe Acrobat V6
- **Scanned Images**
  Irfanview & JHOVE
- **Digital Photographic Records**
  IrfanView & JHOVE
- **Digital Geospatial Data Records**
  ESRI & Safe Software FME
- **Web Content**
  W3C

5.10.2 Recommendation of common 'backbone' software for ALL New & Future Format accessions (JHOVE).

5.10.3 Recommended Implementation: 4 Stages:

1. Format verification based upon present capabilities of each COTS product
2. Customize COTS products for NARA application
3. JHOVE[40] - existing modules with new NARA profiles
4. JHOVE – develop new modules for remaining

[40] In the event JHOVE capabilities are not available as planned or do not meet NARA requirements, custom software will need to be developed along with continued modifications to the COTS products used in Stage 1.

APPENDIX

A.1 Network Trials, Measurements & Analysis

(This section includes the APS/ LAN NETWORK STUDY (original report) and ANALYSIS). In addition to the report and the summary charts included herein, there exists some 14GB of actual data recorded during the trials. Software routines have been developed to study the data for future questions about design and component performance).

APS/ LAN NETWORK Trials and Measurements

Test:
This test was conducted on October 4th and 5th , 2004 at the National Archives and Records Administration facility located at 8600 Adelphi Ave College Park MD.

Purpose:
The purpose of this test is to establish network throughput on the APS network. During this test files were transferred from workstations to Network Attached Storage (NAS) devices. Source files originated from several formats, including 3480 tape, DLT8000 tape, CD-ROM and local disk drives. In addition to the analysis of data throughput, extraneous packets were analyzed and are summarized in the recommendations section of this report.

Scope:
The analysis consisted of transfers from workstations to NAS's. To facilitate capturing packet analysis a hub was connected between the NAS's and the 3Com switch, the packet collection computer was also connected to the hub permitting all traffic to and from the NAS's would be collected.

Transfer One consisted of a 491MB (megabyte) transfer from **APS04 to SNAP3**, the source data was from a 3480 tape. During this transfer only APS04 was utilizing the network so that a baseline could be established.

Transfer Two consisted of a 650MB (megabyte) transfer from **APS08 to SNAP3**, the source data was from a CD-ROM. During this transfer packets from the first transfer had not completed and allowed for concurrent transfers to SNAP2 from both APS04 and APS08.

Transfer Three consisted of a 40GB (gigabyte) transfer from APS19 to SNAP3; the data source was from a DLT8000 tape. During this transfer training sessions where being conducted at other APS workstations.

Transfer Four consisted of a transfer from SNAP3 to APS8. The data source was the same set of files copied in transfer two.

Network Configuration:

The APS network is an isolated network comprised of sixteen (16) computer system, including one (1) server, four 4 Network Attached Storage (NAS) appliances, eleven (11) workstations and two(2) print servers. The physical network consists of; Category 5 100BaseT wiring interconnected via a 3Com 3C16465C Ethernet switch. All network cards are 10/100 integrated network interface cards, the type of interface for the printer server is unknown. *(A listing of all network devices and their IP addresses is included at the end of this report.)*

Server Operating System:

The server is a Windows 2003 Server configured as a Domain Controller in an Active Directory (AD) network. The domain name is NARA; the server name is APS-Finity with an IP address of 192.168.26.55.

This configuration requires TCP/IP be installed, as well as, DNS that is compatible with Windows Active Directory. Workstation must register with AD in order to function properly in the domain. Users log into the domain for authentication, which provides their rights and permissions necessary to access resources.

The some of the benefits of logging into a domain are as follows:

1) Centralized authentication of all logins.
2) Centralized user account database.
3) Centralized assignment of user rights and permissions.
4) Centralized roaming user profiles, which permits users desktop settings to follow them from system to system.

To fully take advantage of these benefits all systems in this network must be AD aware and log into the domain. The administration of the domain should be conducted by in-house staff since frequent changes require updates to user rights and permissions. Currently the domain is administered by a contractor, who was not present during our visit

Packet Analysis:

During testing a standard assessment of the types of packets broadcast on the network was conducted. As seen in the figure below 99.98% of all traffic is TCP related traffic. And less than 0.02% of all traffic is broadcast traffic, normally considered overhead and should be avoided if possible. This network configuration appears to be optimal and we would only recommend the removal of Service Advertisement Protocol (SAP), which is normally associated with Novell Netware.

Protocol	% Packets	Packets	Bytes	Mbit/s	End Packets	End Bytes	End Mbit/s
Frame	100.00%	64749	51397359	6.698	0	0	0.000
Ethernet	100.00%	64749	51397359	6.698	0	0	0.000
Internet Protocol	99.98%	64738	51396659	6.698	0	0	0.000
Transmission Control Protocol	99.98%	64734	51395927	6.698	39635	39568470	5.157
NetBIOS Session Service	38.76%	25099	11827457	1.541	30	41236	0.005
SMB (Server Message Block Protocol)	38.72%	25068	11784707	1.536	25068	11784707	1.536
Data	0.00%	1	1514	0.000	1	1514	0.000
User Datagram Protocol	0.01%	4	732	0.000	0	0	0.000
Simple Network Management Protocol	0.00%	2	241	0.000	2	241	0.000
NetBIOS Datagram Service	0.00%	2	491	0.000	0	0	0.000
SMB (Server Message Block Protocol)	0.00%	2	491	0.000	0	0	0.000
SMB MailSlot Protocol	0.00%	2	491	0.000	0	0	0.000
Microsoft Windows Browser Protocol	0.00%	2	491	0.000	2	491	0.000
Address Resolution Protocol	0.01%	4	222	0.000	4	222	0.000
Internetwork Packet eXchange	0.00%	3	180	0.000	0	0	0.000
Service Advertisement Protocol	0.00%	3	180	0.000	3	180	0.000
Logical-Link Control	0.01%	4	298	0.000	0	0	0.000
Internetwork Packet eXchange	0.01%	4	298	0.000	0	0	0.000
Service Advertisement Protocol	0.01%	4	298	0.000	4	298	0.000

During Test2, a sixty second snippet illustrates the number of packets and bytes transmitted and received from each computer. As shown in the figure below, SNAP3 (192.168.26.29) received a total of 27,005,057 bytes of data

Address	Packets	Bytes	Tx Packets	Tx Bytes	Rx Packets	Rx Bytes
192.168.26.16	35171	28008107	22838	26879805	12333	1128302
192.168.26.29	38706	31163023	14447	4157966	24259	27005057
192.168.26.25	25981	20229416	13313	8427178	12668	11802238
192.168.26.30	25988	20229808	12672	11802464	13316	8427344
192.168.26.22	3534	3154884	1421	125266	2113	3029618
192.168.26.55	25	1947	13	1095	12	852
192.168.26.28	21	1608	9	714	12	894
192.168.26.27	19	1488	8	654	11	834
192.168.26.24	12	1096	8	668	4	428
192.168.26.18	6	360	3	180	3	180
192.168.26.19	6	360	3	180	3	180
192.168.26.32	2	241	1	119	1	122
192.168.26.20	2	241	1	122	1	119
192.168.26.10	1	248	1	248	0	0
192.168.26.255	2	491	0	0	2	491

Detailed summary results, as depicted above were further analyzed in greater detail for the Throughput Analysis.

Recommendations:

The current network configuration in the APS area is structured towards temporary large file storage onto Network Attached Storage (NAS) devices, permitting multiple users access to the resource from a central location. In most small network configurations large amounts of data transfer is not the primary use of the network, but more random database access. Due to the large amounts of data being stored on the NAS's the capacity of these devices is limited from time to time. For example, SNAP3 has a total capacity of 336GB and 198GB of free space, EOPSNAP has a total capacity 224GB and a total of 200MB of free space. Due the large number of changes made to the data store, selected in-house staff personnel should have the ability to make changes to rights and permissions on the storage devices.

The NAS's are the central storage container for the APS network reference data and for assembling large accessions. Given that these devices are not designed for high volume throughput they will quickly become a production bottleneck as the number of concurrent users accessing these devices increases.

References: http://www.ieee802.org/3/

APS/ LAN NETWORK Analysis

The following table summarizes the LAN network trials performed on the APS/LAN. The trials, separately and combined involved data transmission between the more important network devices. These network measurements were required to determine a baseline reference for these central and higher priority devices being used today and more so in the near term future. The resulting baseline data can be used to evaluate effects of workload expansion and the implications of additional storage devices to present and future networks.

The following list describes the critical baseline operations/hardware measured for the trials:

- 3480 to NAS
- CD to NAS
- DLT 8000 to NAS
- NAS to local workstation

In addition to the trial network activity itself there is always some network overhead that was recorded simultaneously; that data is included in each trial summary and in most cases was considered minor. The only notable overhead was the contribution from a training session in NWME taking place during one of the ARkival trials. Training sessions are considered normal part of APS/LAN operations and its occurrence was helpful in the network activity measurement.

TRIAL			AVG. DATA* (mb/sec)	Comment	Reference
Description	Trial Reference #	Single (S) trial or multiple (M) trial			
3480 to NAS/Bkgnd	1	S	4.9	*Critical data*	*See Figure 3*
CD to NAS/ DLT 8000 to NAS/Bkgnd	2 (2a, 2b, 2c)	M	1.4	*Critical data: DLT 8000 to NAS-ONLY*	*See Figure 4*
CD to NAS/ DLT 8000 to NAS/NWME Training/Bkgnd	3 (3a, 3b, 3c)	M			*See Figure 5*
NAS to Workstation	4	S			*See Figure 6*

* "Average Data" is subject to error via both the measurement technique and the CSMA/cd protocol.

The detailed network activity reports were collected and baseline data for two (2) primary tape storage drives were isolated; they include the 3480 transfer to the NAS and DLT transfer to the NAS. The network data for theses primary storage devices was obtained from single time intervals (snippets) or multiple snippets and thereafter isolated for reporting purposes. *See Figures 3-6 below*. The data collected required detailed introspection in that the analyzer reports of "average data transmitted" is not always an indicator of potential network problems. Instead peak activity instances within the average, did indeed support a potential basis for conflicts in application and use- *see Figure 5*. Increased network usage therefore could likely account for poor performance and unexplained network slowdowns. Presently, the primary focus for application impairment hinges on data flow to the NAS* with complications resulting from relatively slow tape drive data rates.

* The NAS's today are used more as large temporary storage means than as network performance devices. The NAS's are central focus points for large data storage and applications can be a bottleneck for access as well their use for tape drive applications. The newer design emphasizes more switching operations of the NAS/SAN mesh and should provide greater efficiency and speed in peak application periods.

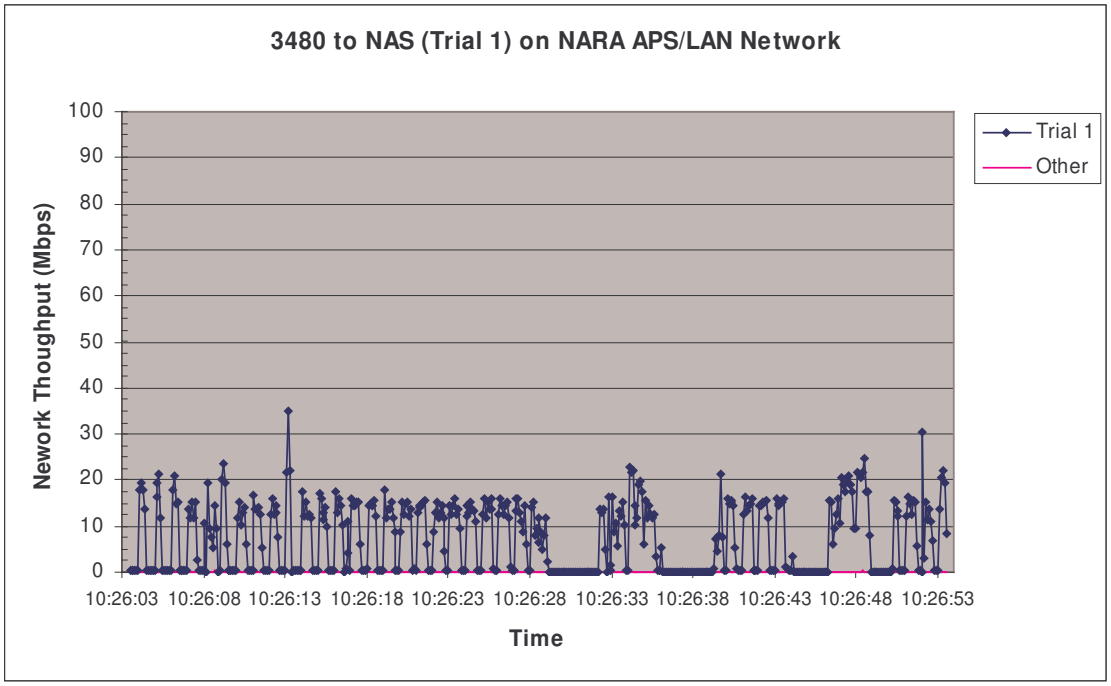


Figure 3. Representative time snippets for a single 3480 data transfer to the NAS (trial 1).

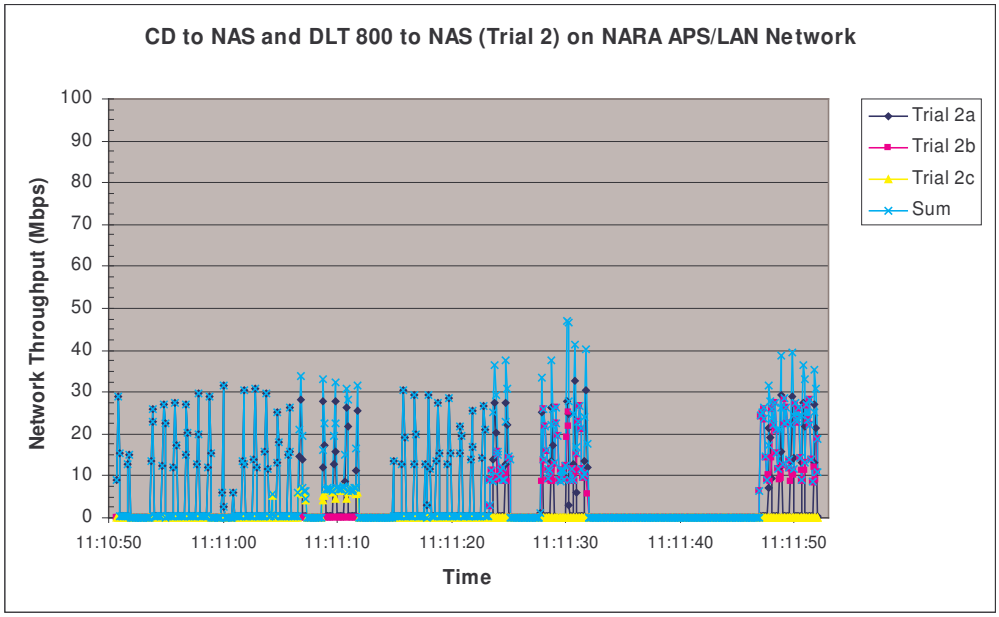


Figure 4. Representative time Snippets for simultaneous CD to NAS (2a), DLT 8000 to NAS (2b) and background traffic (2c) data transfers to the NAS (trial 2).

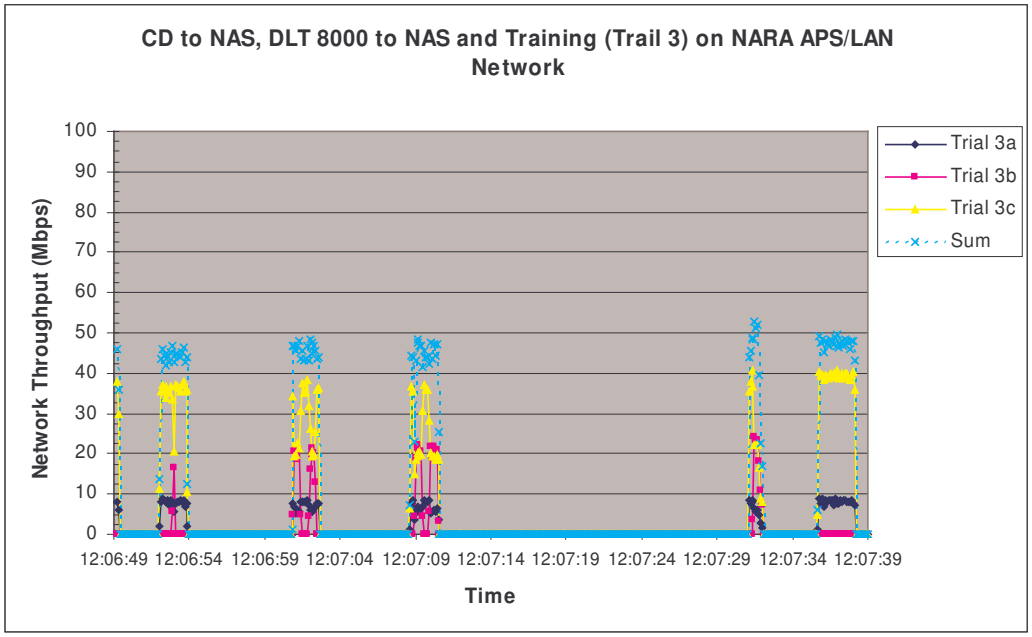


Figure 5. Representative time Snippets for simultaneous CD to NAS (3a), DLT 8000 to NAS (3b) and training (3c) data transfers to the NAS (trial 3).

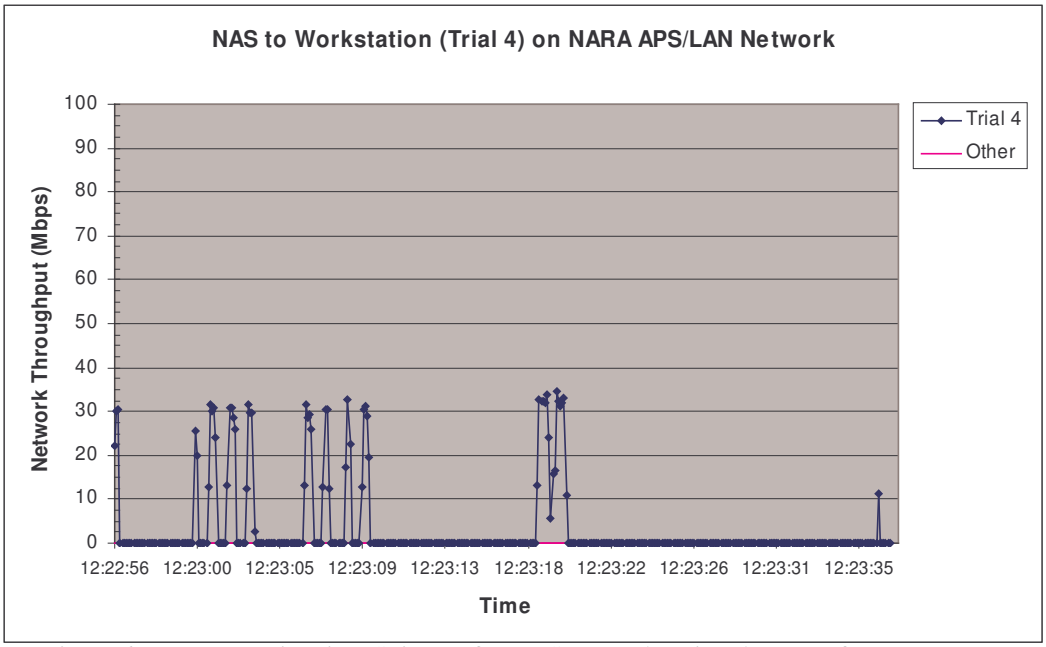


Figure 6. Representative time Snippets for NAS to Workstation data transfer.

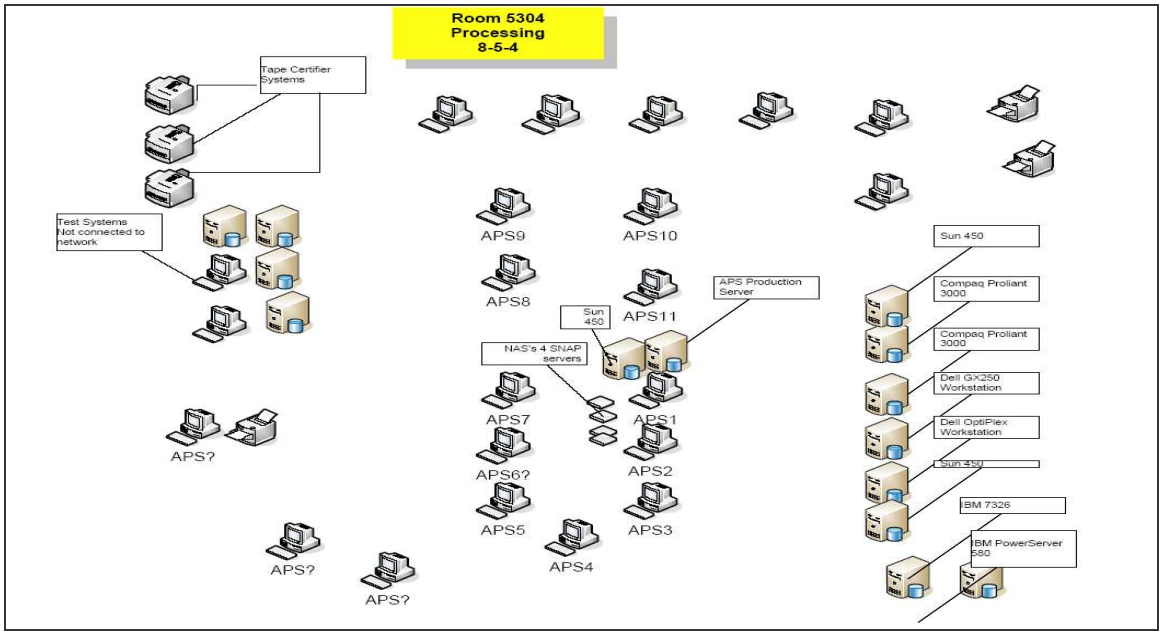
In summary, present network usage appears more susceptible to potential problems during certain applications and the result can be somewhat misleading when one studies average network activity. The detailed analysis above indicates the total contribution of all the different data traffic to occasionally exceed 50% of the available throughput. Any substantial increase in simultaneous workstation usage and/or capacity (10X to 100X) will likely elevate the environment traffic to be problematical and network performance unreliable.

The data also indicates that the present 100Mb/sec network, capable of delivering 9-10MB/sec bandwidth per "switched" leg is performing considerably less than specification. Although the basis for under-performance may likely result from several different causes the effort to analyze specific causes was not justified when considering the network ramifications for expansion (10X to 100X)[**]. The increased demand can not be met with the present network and its components.

The increased volumes of data, larger data files, more preservation/reference copying and more of the same operations performed with more hardware combined with observed periods of peak network activity contained in the 'average" are obvious reasons to question the performance, reliability and application of the current network design.

[**] Anticipated accessions 10X to 100X in terms of both average file size and total number of files. Contained in the requirement are 10+TB files and 1+TB file sets.

APS (LAN)



Network Hardware:

File Server:
Hostname: **APS-FINITY**
Operating System: **Windows 2003 Server**
IP Address: **192.168.26.55**
MAC address: **00-50-8B-B9-35-16**

Workstations:
APS01
IP Address: **192.168.26.14**
MAC address: **00-50-DA-BC-8E-C2**
Operating System: **Windows 98**
Attached Storage Devices: **One 3480 tape drives**

APS02
IP Address: **192.168.26.15**
MAC address: **00-50-DA-D6-82-30**
Operating System: **Windows 98**
Attached Storage Devices: **Two 3480 tape drives, two 9-track drives**

APS-3
IP Address: **192.168.26.17**
MAC address: **00-50-DA-D6-7F-12**
Operating System: **Windows 98**
Attached Storage Devices: **Two 3480 tape drives, two 9-track drives**

APS-4
IP Address: **192.168.26.16**
MAC address: **00-09-6B-E4-8E-B4**
Operating System: **Windows XP**
Attached Storage Devices: **Two 3480 tape drives, one DVD-RW**

APS-5

Attached Storage Devices: **Two 3480 tape drives, two DLT Tape drives**

APS-6

Attached Storage Devices: **Three 3480 tape drives**

APS07
IP Address: **192.168.26.22**
MAC address: **00-09-6B-E4-8E-F1**
Operating System: **Windows XP**
Attached Storage Devices: **Two 3480 tape drives, one DLT tape drive**

APS08
IP Address: **192.168.26.24**
MAC address: **00-09-6B-64-F4-70**
Operating System: **Windows XP**
Attached Storage Devices: **Two 3480 tape drives, two DLT tape drives**

APS09
IP Address: **192.168.26.11**
MAC address: **00-09-6B-E4-CC-9C**
Operating System: **Windows XP**
Attached Storage Devices: **Two 3480 tape drives**

APS10

Attached Storage Devices: **Two 3480 tape drives**

APS11
IP Address: **192.168.26.10**
MAC address: **00-09-6B-E4-B1-65**
Operating System: **Windows XP**
Attached Storage Devices: **Two 3480 tape drives, two DLT tape drives**

Network Attached Storage Devices:

FTP-SERVER
IP Address: **192.168.26.28**
MAC address: **00-C0-B6-04-B2-EA**
Operating System: **Quantum Snap Server**
Disk Configuration: RAID5 (4 disks)
Total MB: 84,181
Free MB: 10,655
Total Files: 390,434

SNAP2
IP Address: **192.168.26.27**
MAC address: **00-C0-B6-08-80-9D**
Operating System: **Quantum Snap Server**
Disk Configuration: RAID5 (4 disks)
Total MB: 280,377
Free MB: 100,029
Total Files: 24,498

SNAP3
IP Address: **192.168.26.29**
MAC address: **00-C0-B6-08-92-5E**
Operating System: **Quantum Snap Server**
Disk Configuration:
RAID5 (4 disks)
Total MB: 336,543
Free MB: 198,581
Total Files: 43,633

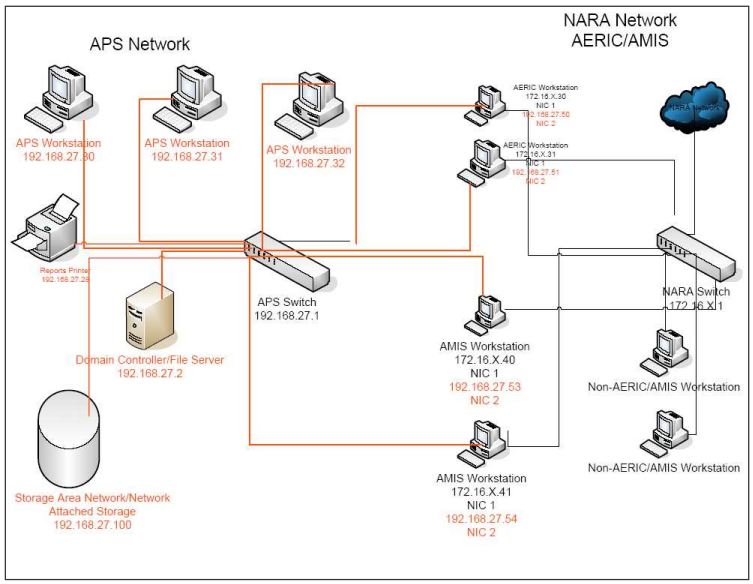
EOPSNAP
IP Address: **192.168.26.30**
MAC address: **00-C0-B6-08-AB-C0**
Operating System: **Quantum Snap Server**
Disk Configuration: RAID5 (4 disks)
Total MB: 224,311
Free MB: 202
Total Files: 749

Test equipment:
The process of collecting information was conducted using a Panasonic CF-27 notebook and an IBM T-21. Ethereal version 0.10.7 with WinCap version 3.0 alpha was used to collect Ethernet frames and GFI LanScan version 8.0 was used to collect networking information from the workstations

A.2 "Dual network card" design *(see Figure below)*

The first approach ARkival suggested as an inter-network communication means had the following advantages and disadvantages:

- It is a relatively inexpensive solution whereas a firewall/router is more expensive.
- It is fairly easy to implement.
- It significantly increases the complexity of the networks by establishing multiple points at which the networks are cross-connected. As a result of this complexity the difficulty in administering the network will increase substantially.

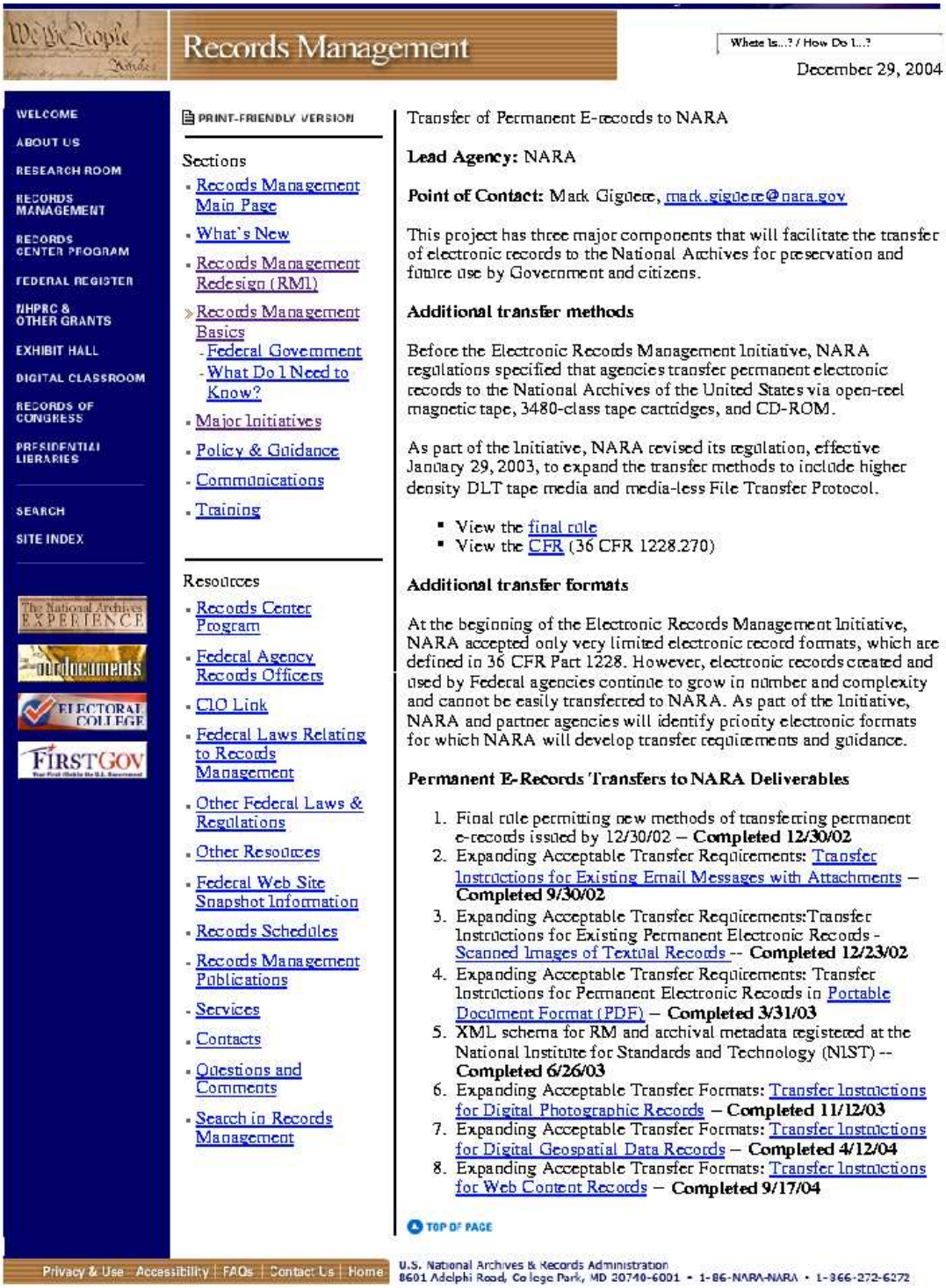


- The dual card approach does not provide true isolation of the two networks to which a given workstation is connected. There is the possibility that a poorly configured, or modified, workstation will allow network traffic from one card to be seen on the other, thus artificially inflating network traffic with network noise.

- Security- There are security concerns particularly with regard to the fact that there are now viruses capable of being propagated within a UDP packet. Therefore conventional TCP level filtering solutions, which a dual card workstation would normally use to filter traffic between the two networks, is inadequate to assure the security of the networks.

In summary, this "dual network card" approach alone will not solve all the major issues. It will not solve the need to share data between APS, AERIC, AMIS, and other systems directly and also has security limitations.

A.3 NARA recommendations:



We the People

Records Management

Where Is...? / How Do I...?

December 29, 2004

WELCOME
ABOUT US
RESEARCH ROOM
RECORDS MANAGEMENT
RECORDS CENTER PROGRAM
FEDERAL REGISTER
NHPRC & OTHER GRANTS
EXHIBIT HALL
DIGITAL CLASSROOM
RECORDS OF CONGRESS
PRESIDENTIAL LIBRARIES
SEARCH
SITE INDEX

The National Archives EXPERIENCE
ourdocuments
ELECTORAL COLLEGE
FIRSTGOV

PRINT-FRIENDLY VERSION

Sections

- Records Management Main Page
- What's New
- Records Management Redesign (RM1)
- Records Management Basics
  - Federal Government
  - What Do I Need to Know?
- Major Initiatives
- Policy & Guidance
- Communications
- Training

Resources

- Records Center Program
- Federal Agency Records Officers
- CIO Link
- Federal Laws Relating to Records Management
- Other Federal Laws & Regulations
- Other Resources
- Federal Web Site Snapshot Information
- Records Schedules
- Records Management Publications
- Services
- Contacts
- Questions and Comments
- Search in Records Management

Transfer of Permanent E-records to NARA

Lead Agency: NARA

Point of Contact: Mark Giguere, mark.giguere@nara.gov

This project has three major components that will facilitate the transfer of electronic records to the National Archives for preservation and future use by Government and citizens.

Additional transfer methods

Before the Electronic Records Management Initiative, NARA regulations specified that agencies transfer permanent electronic records to the National Archives of the United States via open-reel magnetic tape, 3480-class tape cartridges, and CD-ROM.

As part of the Initiative, NARA revised its regulation, effective January 29, 2003, to expand the transfer methods to include higher density DLT tape media and media-less File Transfer Protocol.

- View the final rule
- View the CFR (36 CFR 1228.270)

Additional transfer formats

At the beginning of the Electronic Records Management Initiative, NARA accepted only very limited electronic record formats, which are defined in 36 CFR Part 1228. However, electronic records created and used by Federal agencies continue to grow in number and complexity and cannot be easily transferred to NARA. As part of the Initiative, NARA and partner agencies will identify priority electronic formats for which NARA will develop transfer requirements and guidance.

Permanent E-Records Transfers to NARA Deliverables

1. Final rule permitting new methods of transferring permanent e-records issued by 12/30/02 – **Completed 12/30/02**
2. Expanding Acceptable Transfer Requirements: Transfer Instructions for Existing Email Messages with Attachments – **Completed 9/30/02**
3. Expanding Acceptable Transfer Requirements:Transfer Instructions for Existing Permanent Electronic Records - Scanned Images of Textual Records -- **Completed 12/23/02**
4. Expanding Acceptable Transfer Requirements: Transfer Instructions for Permanent Electronic Records in Portable Document Format (PDF) – **Completed 3/31/03**
5. XML schema for RM and archival metadata registered at the National Institute for Standards and Technology (NIST) -- **Completed 6/26/03**
6. Expanding Acceptable Transfer Formats: Transfer Instructions for Digital Photographic Records – **Completed 11/12/03**
7. Expanding Acceptable Transfer Formats: Transfer Instructions for Digital Geospatial Data Records – **Completed 4/12/04**
8. Expanding Acceptable Transfer Formats: Transfer Instructions for Web Content Records – **Completed 9/17/04**

TOP OF PAGE

Privacy & Use | Accessibility | FAQs | Contact Us | Home

U.S. National Archives & Records Administration
8601 Adelphi Road, College Park, MD 20740-6001 • 1-86-NARA-NARA • 1-366-272-6272

Transfer of Permanent E-records to NARA

Lead Agency: NARA

Point of Contact: Mark Giguere, mark.giguere@nara.gov

This project has three major components that will facilitate the transfer of electronic records to the National Archives for preservation and future use by Government and citizens.

Additional transfer methods

Before the Electronic Records Management Initiative, NARA regulations specified that agencies transfer permanent electronic records to the National Archives of the United States via open-reel magnetic tape, 3480-class tape cartridges, and CD-ROM.

As part of the Initiative, NARA revised its regulation, effective January 29, 2003, to expand the transfer methods to include higher density DLT tape media and media-less File Transfer Protocol.

- View the final rule
- View the CFR (36 CFR 1228.270)

Additional transfer formats

At the beginning of the Electronic Records Management Initiative, NARA accepted only very limited electronic record formats, which are defined in 36 CFR Part 1228. However, electronic records created and used by Federal agencies continue to grow in number and complexity and cannot be easily transferred to NARA. As part of the Initiative, NARA and partner agencies will identify priority electronic formats for which NARA will develop transfer requirements and guidance.

Permanent E-Records Transfers to NARA Deliverables

1. Final rule permitting new methods of transferring permanent e-records issued by 12/30/02 -- **Completed 12/30/02**
2. Expanding Acceptable Transfer Requirements: Transfer Instructions for Existing Email Messages with Attachments -- **Completed 9/30/02**
3. Expanding Acceptable Transfer Requirements: Transfer Instructions for Existing Permanent Electronic Records - Scanned Images of Textual Records -- **Completed 12/23/02**
4. Expanding Acceptable Transfer Requirements: Transfer Instructions for Permanent Electronic Records in Portable Document Format (PDF) -- **Completed 3/31/03**
5. XML schema for RM and archival metadata registered at the National Institute for Standards and Technology (NIST) -- **Completed 6/26/03**
6. Expanding Acceptable Transfer Formats: Transfer Instructions for Digital Photographic Records -- **Completed 11/12/03**
7. Expanding Acceptable Transfer Formats: Transfer Instructions for Digital Geospatial Data Records -- **Completed 4/12/04**
8. Expanding Acceptable Transfer Formats: Transfer Instructions for Web Content Records -- **Completed 9/17/04**

ACKNOWLEDGEMENTS

The author recognizes the technical contributions of the following individuals

Brian Alley	Dennis McMann
Dan Coutu	Ron Peacetree
John Fox	Jim Richardson
Steven Gilheany	Bob Schmidt
Richard Holstein	

Appreciation is extended to Dr.Vivek Navale of the National Archives & Records Administration for his guidance and technical advice throughout the Study. Special thanks are also extended to all the staff members of NWME who assisted ARkival in many aspects of the study and the many visits to NARA.