
JOHN F. KENNEDY PRESIDENTIAL LIBRARY & MUSEUM

**Beyond Bricks-and-Mortar:
Innovations in System Architecture for the
John F. Kennedy Presidential Library
Digital Archive**

September 27, 2010

Prepared by:

Raytheon

Raytheon Company Global Headquarters
870 Winter Street
Waltham, MA 02451

In conjunction with:

**The John F. Kennedy Presidential Library and Museum
The John F. Kennedy Library Foundation
AT&T
EMC
Iron Mountain**

Beyond Bricks-and-Mortar: Innovations in System Architecture for the John F. Kennedy Presidential Library Digital Archive

Executive Summary

The U.S. Presidential Libraries house a national treasure of historical documents, photographs, and audio/visual recordings that preserve not only the legacy of the Presidents but also the culture, history, and values of the times in which they lived. At present, many of these priceless artifacts have been available only by special appointment and a visit to the library itself. However, the John F. Kennedy Presidential Library and Museum is forging a path through this bricks-and-mortar barrier. Through its innovative Digital Archive Project, the Library is undertaking the digitization, description, and electronic archiving of millions of its holdings to enable world-wide access 24/7 via the internet. Students, researchers, journalists, and the public at large will have the ability to search, query, and retrieve one-of-a-kind handwritten notes, recorded interviews and debates, and other items that will enable them to re-experience those unique moments in our nation's history. The Digital Archive's debut is set for January 10, 2011, the 50th anniversary of Kennedy's inauguration. This initiative is being conducted under the direction of the National Archives and Records Administration (NARA), supported by the Kennedy Library Foundation.

The digital path has not been an easy one: The design for the Digital Archive must accommodate factors as diverse as the fragility of the artifacts, fluctuating budgets, limited facilities, data security, and incremental growth over a long period of time in a world of rapid technological change. To meet the challenges, the John F. Kennedy Library Foundation has partnered with leaders from across the industry in information technology and systems engineering. AT&T, EMC, Iron Mountain, and Raytheon are each donating time, expertise, and resources to ensure that the project is a success.

At the heart of the Project is a ground breaking system architecture, designed under the engineering leadership of Raytheon in close coordination with the JFK Library and Technology Partners. It combines the latest in information management technology with design elements for economy and modularity to ensure that the Digital Archive continues to be affordable and adaptable over its long lifetime. As shown in Figure 1, the architecture for the Digital Archive consists of three nodes which are distributed across three locations to take advantage of state-of-the-art facilities and resources available at those sites to perform key functions.

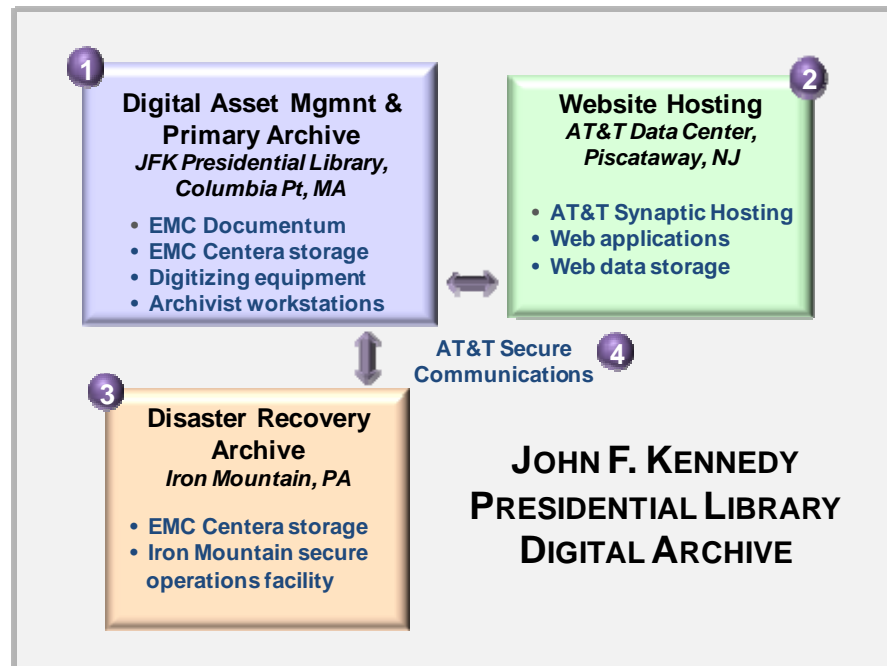


Figure 1: Digital Archive Distributed Architecture

At the JFK Presidential Library (❶ in the figure), NARA archivists use EMC's Documentum system to ingest, describe, organize, and research the digitized items in the Primary Archive which is built on EMC's Centera storage system. Users can access selected holdings from the archive via the internet by means of Web applications hosted at an AT&T data center in Piscataway, NJ utilizing AT&T Synaptic Hosting ServicesSM (❷). At the same time that NARA archivists at the JFK Library are adding data to the Primary Archive, it is replicated electronically to a backup Disaster Recovery Archive located at the Iron Mountain secure operations center in Boyers, PA (❸). Secure transmittal of data is provided among the sites by AT&T's communications network (❹). The result is an architecture that is highly modular, utilizing state-of-the-art technology that enables different components of the Digital Archive to be placed in facilities best suited for the particular function, security needs, and funding plan. The design modularity also readily supports the addition of storage and resources as the archive grows over time, as well as the ability to readily insert newer, more advanced technology as it becomes available.

Detailed Description

System Architecture

Figure 2 provides an expanded view of the Digital Archive system level architecture, illustrating the major components at each node. The following sections describe how each of these components contributes to the operation and use of the system.

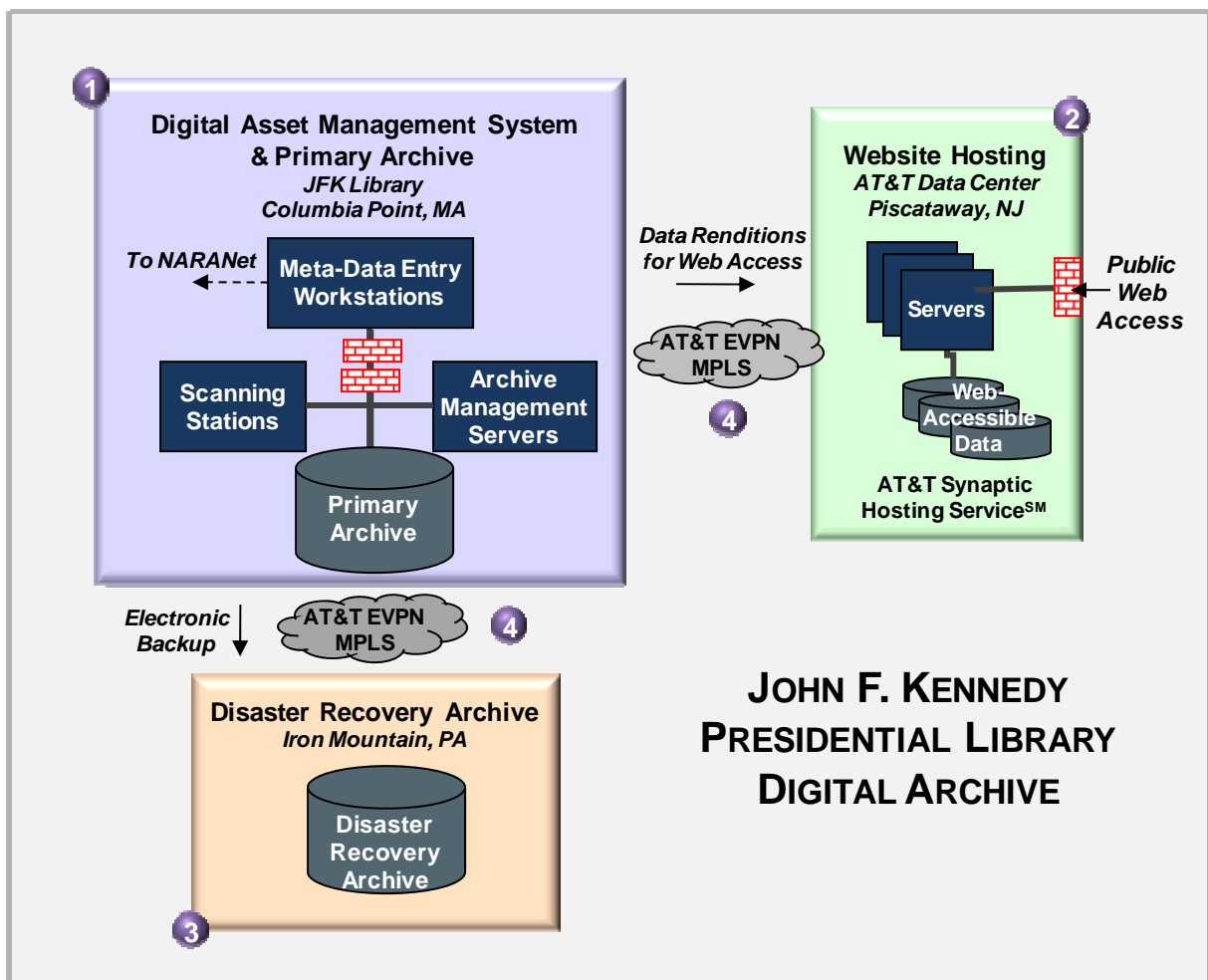


Figure 2: Digital Archive System Level Architecture

Primary Archive (1)

The Digital Archive collection is stored in the Primary Archive, based on EMC's Centera storage system located at the JFK Library (1). Digitization of documents and photographs for the collection is performed using high resolution scanners at the JFK Library location. Audio/visual material is digitized by Xepa, an Iron Mountain subsidiary, and delivered on hard drives to the Library. The digital files of all materials are ingested into EMC's Documentum enterprise content management system, where Library archivists enter descriptive meta-data for each digital scan, including historical records and keywords to aid in research. The 50th Anniversary debut of the Digital Archive will feature priority items from the Kennedy Library collections as shown in Table 1. Items will continue to be added over the years, building an ever-increasing, readily accessible knowledge base of President Kennedy's remarkable legacy.

Table 1: Digital Archive 50th Anniversary Demonstration Collection

Items	Pages (#)	Photos (#)	Audio (hours)	Moving Images (hours)
President's Office Files	160K	256		
JFK Personal Papers	60K			
JFK White House Correspondence	16K			
White House Photos		35K		
White House Audio			125	
CBS Video Footage				30

The historical documents, photographs, and audio and video recordings are digitized at very high resolution to preserve fidelity, thus resulting in large file sizes for each item. One standard document page consumes about 100 MB, and the size of the archive, inclusive of all media, is expected to increase to approximately 152 TB by 2016. To make the most of the available storage, EMC's Captiva software is used to apply LZW lossless compression to the digitized document files. This compression method perfectly preserves the high fidelity resolution but uses 25-30% less storage than the uncompressed files, thus reducing the five-year storage requirement to approximately 117 TB. To ensure the integrity of the stored data, EMC's Centera storage system employs parity checking which allows automatic detection of errors and reconstruction of data if a drive in the disk array fails. Additionally, as the archive grows beyond 2016, the system's modular design supports the future addition of newer, even more efficient storage technologies.

Disaster Recovery Archive (2)

Digitizing the records is a labor-intensive activity, with each page of each document or photograph carefully placed by hand on a special scanner. Special handling and digitization equipment are also required for the often fragile audio/visual materials, many recorded on analog media long since rendered obsolete in today's digital world. To ensure that this significant investment of time, labor, and money is not lost through accident, disaster, or malicious intent, a complete backup of the Digital Archive is maintained electronically via transmittal on an AT&T secure communications network (2) to the Disaster Recovery Archive (2). This backup archive, also built on the EMC Centera storage system, is located in Iron Mountain's secure underground operations facility in Boyers, PA. This facility provides unsurpassed protection for the archive data, including:

- Tier 3 Mission Critical Data Center certification
- Redundant commercial power feeds
- Full backup power for up to 7 days
- OSHA certified fire company
- 24 hour armed security
- 24 hour maintenance
- 24 x 7 service operation

In the event that data in the Primary Archive is lost or compromised, the Disaster Recovery archive provides the capability to readily restore the full data holdings, including all associated meta-data.

Website Hosting (2)

The public's experience of the Digital Archive is created through Web applications hosted in the AT&T data center (2). The AT&T Synaptic HostingSM environment at the data center provides a state-of-the-art computer architecture that can automatically adjust processing power depending on load. It consists of a large grid of computers and storage which can run multiple applications as "virtual machines", dynamically re-allocating processors and storage among the applications as demand goes up or down – an ideal architecture for a public-facing Website. This is especially important for the Digital Archive, where surges in demand can be expected at significant anniversaries, e.g., the first moon landing. Using any standard Web browser, users at schools, universities, other libraries, news agencies, and homes can search, query, and retrieve selected documents, photos, and streaming audio/visual content in compressed formats optimized for Web presentation. The Web-ready renditions for documents and photographs are generated by the Documentum system at the JFK Library at the time the full resolution digitized files are produced and ingested. The renditions are then transmitted via a secure AT&T network (4) to the data center where they are stored on the synaptic host storage system, ready to be discovered and accessed. Renditions of audio/visual material are produced by Xepa as part of the full resolution digitization process, and hosted for optimized Web streaming via a third party content delivery network provider, Ooyala. In addition to direct Web access, users can also special order larger files to be delivered on CD, DVD, or through download to an ftp account.

Information Security (all nodes)

Given the significance of the historical data that the John F. Kennedy Presidential Library Digital Archive contains, security plays a critical role in the design, with an aim toward protecting the integrity of the archive contents at all times. Raytheon certified security engineers have worked with NARA representatives to design and incorporate required safeguards for compliance with Federal Information Processing Standards and other IT security policies. To this end, data in the Primary Archive (1) can only be accessed by authorized Library personnel through the Documentum system. All public access is through the Website to data stored at the AT&T data center (2), which uses configurable virtual firewalls to protect the Website itself from outside hacking or denial of service attacks. Raytheon is providing independent testing of the Website firewall to identify and resolve any vulnerabilities. The Digital Archive security architecture also features two firewalls at the JFK Library (1) between the meta-data entry workstations, which are connected to NARANet, and the rest of the Digital Archive. These firewalls ensure that the U.S. Government's NARA network is protected from unauthorized access in accordance with Homeland Security regulations, while allowing the Library archivists to access both the Digital Archive and the full range of NARANet-based historical resources. All electronic transmittance of data between the Digital Archive sites occurs via an AT&T secure network (4).

Network Communication (4)

The AT&T secure network (4) is the inner backbone of the Digital Archive, linking the Primary Archive at the JFK Library with the Disaster Recovery Archive at Iron Mountain and with the Web hosting site at the AT&T data center. This secure network employs AT&T's Enhanced Virtual Private Network (EVPN) Multi-Protocol Label Switching (MPLS) technology. The EVPN service is a fully managed, fully bundled network-based internet protocol (IP) VPN service that ensures tight control and protection of data transactions. Use of the MPLS protocol augments protection and efficiency of data transmittance using a technology known as "label switching". A label is appended to each data packet which uniquely identifies that packet as belonging to a specific destination port within the customer's network. When the data packet reaches its destination, the label is removed automatically. The combination of the EVPN and MPLS technologies provides absolute data segregation from other traffic on the network and source and route assurance. It ensures that connections between the sites can only be established and data transmitted by authorized personnel and processes.

System Administration (all nodes)

Although the parts of the Digital Archive are located in widely separated geographic areas, the staff at the Library has insight into the health, status, and usage levels of all equipment, processing, and Web operations through advanced system administration and monitoring tools. Additionally, EMC's Centera storage system also features a "Call Home" service which issues an automatic alert when a problem is detected. Both Iron Mountain and AT&T provide 24/7 operations centers at their respective sites.

Conclusion

By incorporating modularity throughout the design, the Digital Archive architecture provides significant flexibility to adapt and grow as both the amount of data increases and technology continues to advance. At the highest level of modularity, the architecture supports the placement of people, processors, storage, and applications in the facilities which can best support them from both a capabilities and a funding plan perspective. Within each site, the design readily accommodates upgrades in technology or addition of storage without significant re-design or impact to the other parts of the archive. These characteristics ensure that the Digital Archive will remain current and cost-effective to serve its invaluable historical and educational mission for many generations to come.