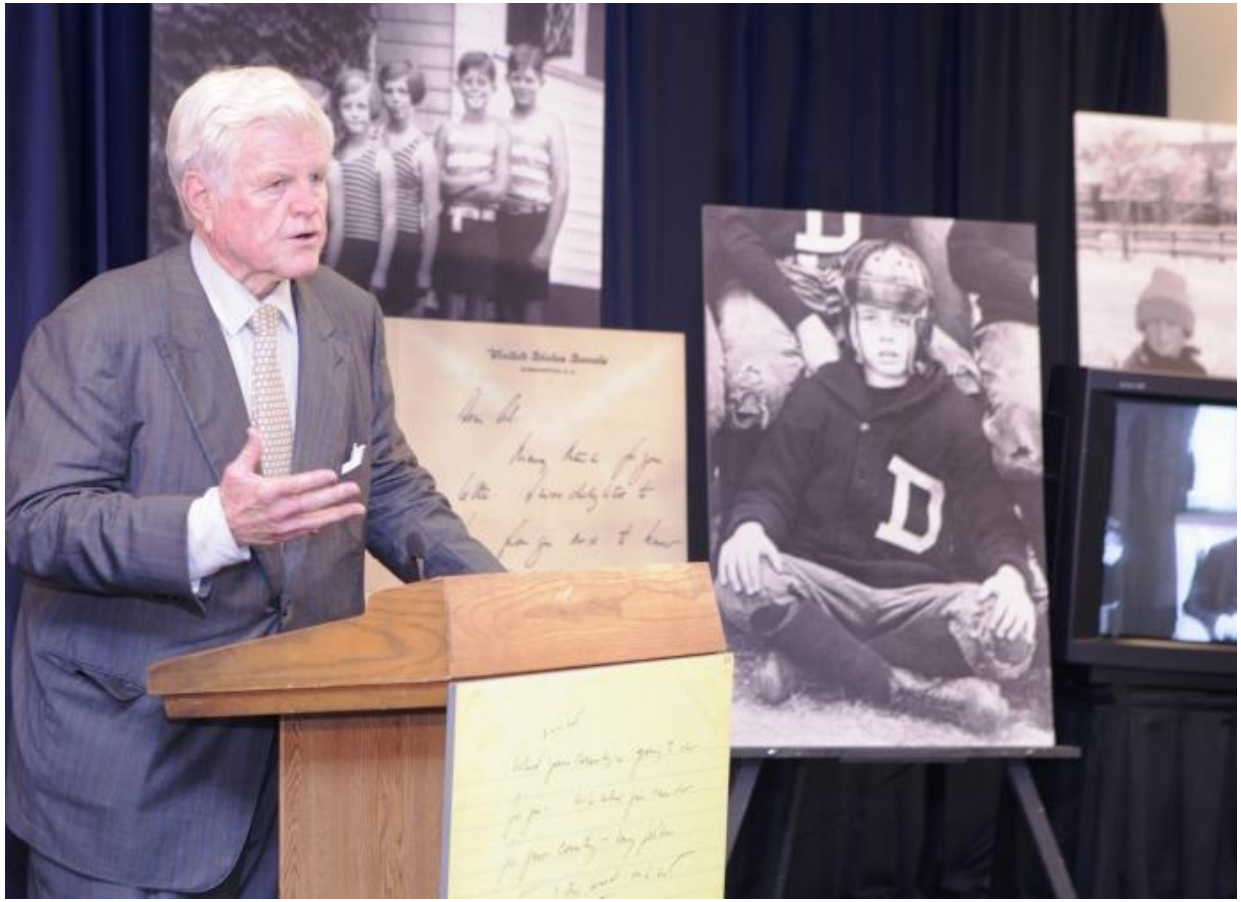
# Managing Large (Digitization) Collections

## and What Smaller Institutions Can Learn

JAMES ROTH AND ERICA BOUDREAU
JFK PRESIDENTIAL LIBRARY
DIGITAL COMMONWEALTH
APRIL 26, 2011

*The groundbreaking initiative we are announcing today is the first of its kind in the nation…Because of this historical initiative, millions of documents, miles of film, and hundreds of thousands of photographs from President Kennedy's administration will be scanned, digitized, indexed and permanently preserved. More importantly, they will be available to all citizens of the world – not just the scholars and researchers who make the journey to Boston.*

Senator Edward M. Kennedy



# June 2006: The Digital Archives initiative is announced

*The Kennedy Library's Access to a Legacy project is a service to our nation. The digitization of archival records is becoming an essential means to allow the public greater access to our national treasures via websites, social media or the growing area of mobile applications," said David Ferriero, Archivist of the United States. "This initiative will open new areas of learning and discovery through the library's archives and will preserve precious documents on digital media for future generations.*

David Ferriero, Archivist of the United States


**January 10, 2011 – the launch of the Digital Archives is announced by Caroline Kennedy and Archivist of the United States at the National Archives in Washington, DC**

# As of January 2011 launch

- Since 2006, Kennedy Library staff have digitized, described, and made available several archival collections in their entirety:
  - President's Office Files
  - White House Central Chronological Files
  - John F. Kennedy Personal Papers
  - White House Audio Collection
- As well as portions of:
  - White House Photograph Collection
  - State Gifts Collection (museum artifacts)
  - John F. Kennedy and Robert F. Kennedy Oral History Collections
  - White House Central Subject Files
  - Several moving image collections

# As of January 2011 launch

- All of that added up to approximately:
  - Over 200,000 pages of textual documents
  - 300 reels of audio tape, containing more than 1,245 individual recordings of telephone calls, speeches and meetings
  - Almost 300 museum artifacts
  - 72 reels of film
  - 1,500 photographs

# Creation of a Digital Archives

## Definition of a digital archives vs. an online exhibit

- Digital Archives
  *Organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities.* – based on *American Digital Library Federation's* definition of Digital Libraries
- Online Exhibits
  *An exhibition that generally includes materials such as artworks, documents, or objects that have been selected and ordered so that their interaction demonstrates an idea or theme for cultural or educational purposes. The selection of materials for an exhibit is sometimes called curation, and the individual responsible a curator.* – based on *SAA's Glossary of Archival and Records Terminology*

# Things to consider before digitizing

- Know **why** you want to digitize
- Know **what** you want to digitize
- Know **how much** you want to digitize
- Know **when** you will digitize a particular collection, series, or item

# Things to consider before digitizing (con't)

- Know **what information or *metadata*** you want to capture about your digital assets
- Know **who** will do the work
- Know **how you will make your digitized assets available** to your users
- Know **how you will pay for** your digitization initiative

# Know **why** you want to digitize

- Your reasons for digitizing will inform decisions about everything else. The two main drivers of digitization are ***access*** and ***preservation***.

# Know **why** you want to digitize

- Access
  - Generally most important benefit of digitization
    - Anyone with internet connection can view materials online
  - File quality will depend on audience
    - Casual users who just need to see or read content can be presented with low resolution renditions
    - Other users may require higher resolution renditions for print or documentary purposes
    - Internal users might need higher resolution renditions for exhibits or other work

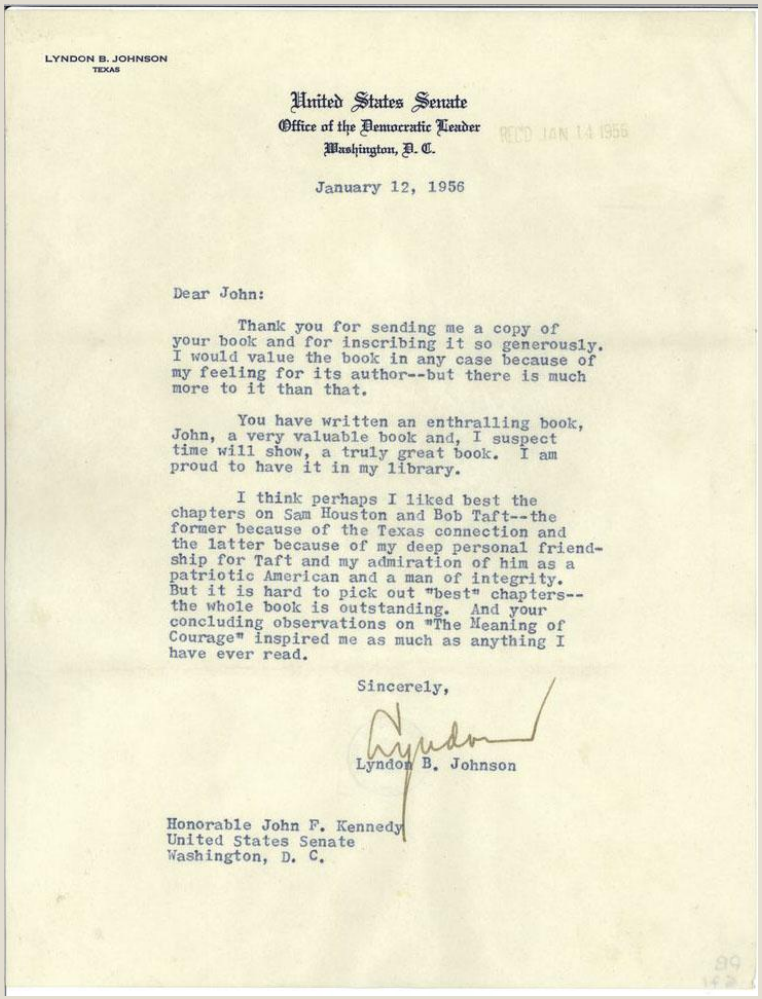
# Know **why** you want to digitize

- Preservation
  - Digitization is not true preservation
  - Still need to preserve original documents
  - However, digitization can reduce wear and tear
  - Disaster recovery - intellectual content reconstructed
  - AV materials, reformatting could be considered preservation

# Know **what** you want to digitize


- Knowing what types of material you plan to digitize will help you make decisions about equipment and storage needs.

# Know **what** you want to digitize



LYNDON B. JOHNSON
TEXAS

United States Senate
Office of the Democratic Leader
Washington, D. C.

RECD JAN 14 1956

January 12, 1956

Dear John:

Thank you for sending me a copy of your book and for inscribing it so generously. I would value the book in any case because of my feeling for its author--but there is much more to it than that.

You have written an enthralling book, John, a very valuable book and, I suspect time will show, a truly great book. I am proud to have it in my library.

I think perhaps I liked best the chapters on Sam Houston and Bob Taft--the former because of the Texas connection and the latter because of my deep personal friendship for Taft and my admiration of him as a patriotic American and a man of integrity. But it is hard to pick out "best" chapters--the whole book is outstanding. And your concluding observations on "The Meaning of Courage" inspired me as much as anything I have ever read.

Sincerely,

Lyndon B. Johnson

Honorable John F. Kennedy
United States Senate
Washington, D. C.

- Textual documents
  - Flatbed scanners (avoid automatic document feeder)
  - TIFF files with LZW lossless compression (average 60MB or so per page) as preservation master
  - Several JPEG renditions for internal and web use
  - Text file containing OCR data, generated from TIFF, allows documents to be searched

# Know **what** you want to digitize

- Photographs
  - Flatbed scanners and a negative scanner (Hasselblad)
  - Uncompressed TIFF files (average 35-100MB, depending on color or black and white)
  - JPEG renditions for internal, reference, and web use

# Know **what** you want to digitize


- Audio recordings
  - Digital audio workstation or pay to have assets digitized off site
  - .WAV file, 96 kHz/24 bits as preservation master
  - .WAV file, 44 kHz/16 bits as reference rendition
  - .mp3 file for web delivery

# Know **what** you want to digitize

- Moving images
  - Digital video workstation or have assets digitized off site
  - Lossless JPEG 2000 frames wrapped in MXF, 720 x 486 pixels, 30 frames per second as preservation/master
  - MPEG2 files for reference renditions
  - MPEG4 files for web delivery

# Know **what** you want to digitize

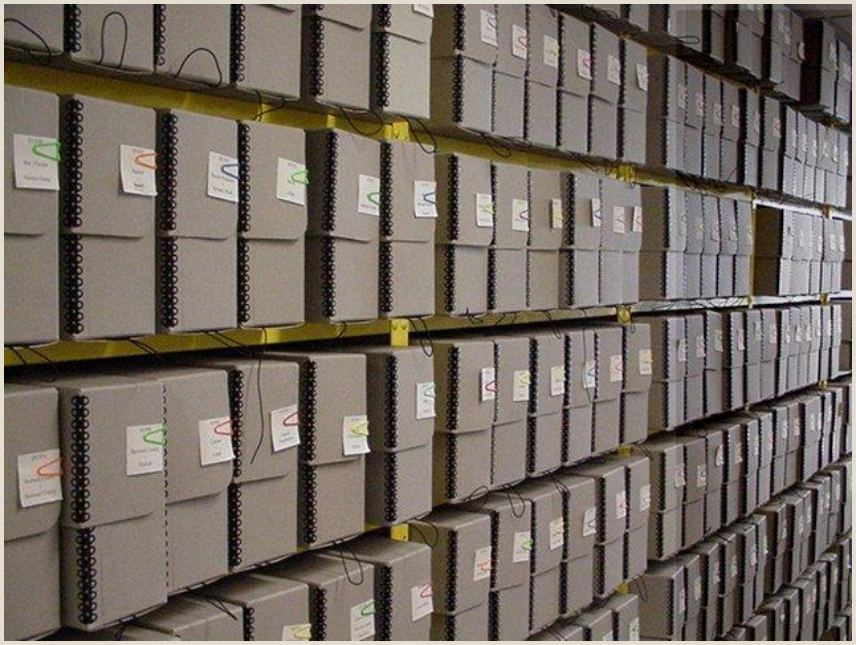
- Museum artifacts
  - TIFF images of all sides of an object, taken with digital camera
  - High tech, 3D ways of presenting objects

# Know **how much** you want to digitize

- Important workflow consideration: level at which you will digitize
  - Entire collections
  - Individual series
  - Specific items

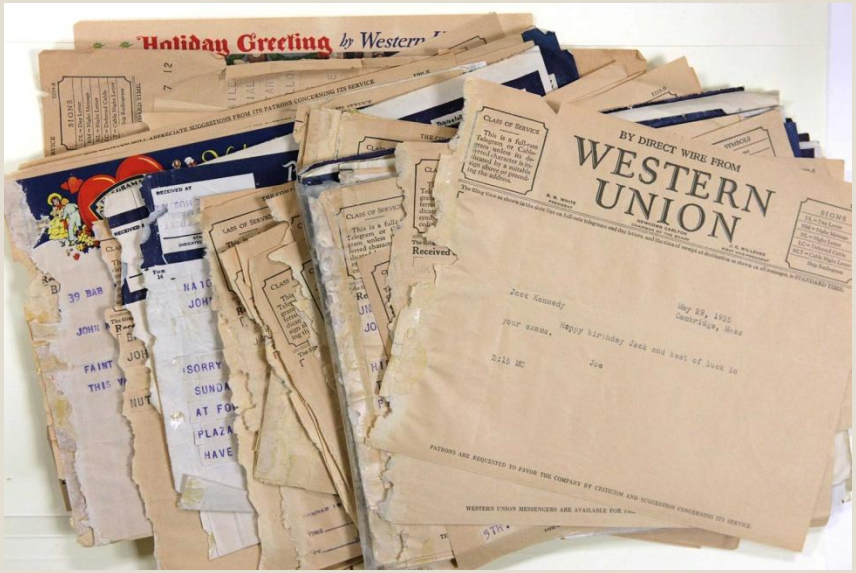
# Know **how much** you want to digitize

- Entire collections
  - JFK: digitizing whole collections, regardless of size
  - Avoids any judgment by archivist
  - Time consuming
  - Really need to commit to the idea



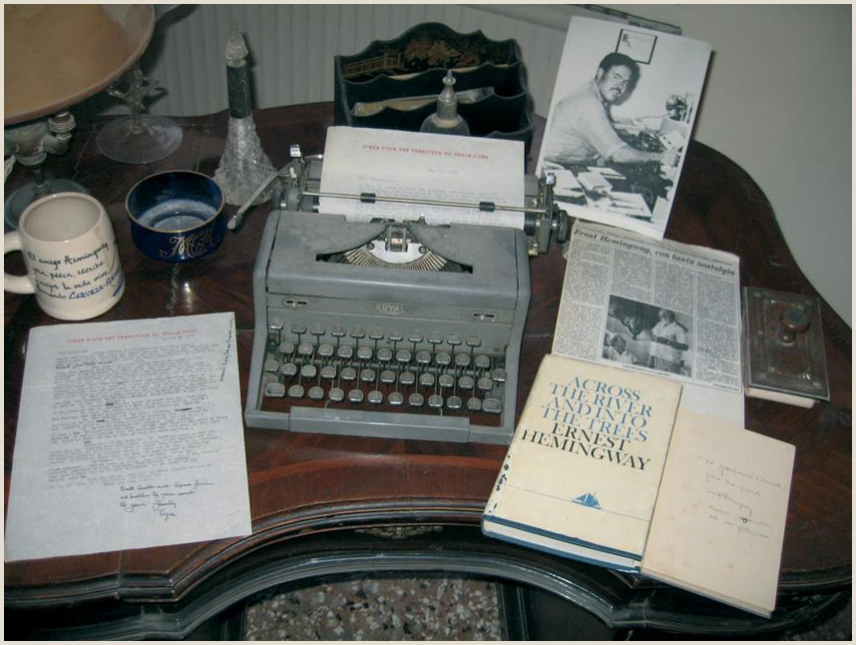
# Know **how much** you want to digitize

- Specific series
  - Some selection by the archivist
  - Entire series allows users to decide what is relevant
  - Single series might be marketable to donors or administrators



# Know **how much** you want to digitize

- Highlights of single or multiple collections
  - More like an online exhibit than digital archives
  - Extremely important to retain contextual information
  - May require more descriptive information



# Know **when** you want to digitize


- Under what circumstances will you digitize a particular collection, series, or item? Will digitization at your institution be systematic, opportunistic, on demand, or a combination of all or some of the above?

# Know **when** you want to digitize

- Systematic digitization
  - Based on priorities set by the institution
    - Core mission
    - Core collections
    - Researcher interest
    - Internal needs
  - Clear beginning and end to each project
  - Requires short and long term planning
  - Pace can be sped up or slowed down based on staffing and funding levels

# Know **when** you want to digitize

- Opportunistic digitization
  - Donor driven projects
  - Internal projects: exhibits or fundraising campaigns
  - Keep sight of institutional priorities and general archival principles

# Know **when** you want to digitize

- Digitization on demand
  - Driven by reference requests
  - Reflects portion of holdings of interest to researchers
  - Consider digitizing entire folders as part of the process
  - Consider charging a lower reproduction fee for digital files
    - National Archives of Australia – overcoming the “tyranny of distance”

# Know **when** you want to digitize

- You don't have to choose just one of these strategies
  - Systematic digitization a constant activity
  - When opportunity or demand arises
  - We currently employ all of these methods
    - Systematic (POF, Human Rights series of White House Central Subject Files)
    - Opportunistic (White House Photographs)
    - On demand (reference requests – mostly AV)

# Know what metadata you want to capture


- Metadata, or data about data, is the key to the success of your digitization program.
  - Without it: an unmanageable jumble of orphaned files
  - With good metadata: use and reuse your digital assets

# Know what **metadata** you want to capture

- What established standards will you adopt?
  - Administrative, descriptive, technical, and preservation metadata
  - Crucial to document standards
    - Data Dictionaries for each object type
  - Descriptive standards used by JFK include
    - Describing Archives: A Content Standard
    - NARA Lifecycle Guidelines

# Data Dictionary Sample Entry

| | |
|---|---|
| Documentum Object Name: | jfk_collection_abbr |
| Dublin Core Element: | source |
| Modifier: | abbreviation |
| Scheme: | (none) |
| ARC Crosswalk: | (none) |
| Mandatory: | Yes |
| Repeatable: | No |
| Controlled Vocabulary Source: | NLJFK |
| Definition: | A unique identifier manually created during the digitization process to coordinate the physical materials with their digital representations on the collection level. |
| Guidelines: | Before digitization occurs, each collection is assigned a **Collection Abbreviation,** which is based on the Collection Name and forms the root of the Digital Identifier for each file unit within the collection. |
| Examples: | JFKPOF |
| | RSSPP |

# Know what **metadata** you want to capture

- What established standards will you adopt?
  - Descriptive standards
  - Metadata scheme standards
    - Dublin Core
    - MARC (Machine Readable Cataloging)
    - MODS/METS (Metadata Object Description Schema/Metadata Encoding and Transmission Standard)
    - Encoded Archival Description (EAD)
    - Encoded Archival Context (EAC-CPF)

# XML record containing textual folder metadata

<?xml version="1.0" encoding="iso-8859-1" ?>
- <rdf:RDF xmlns:rdf="**http://www.w3.org/1999/02/22-rdf-syntax-ns#**" xmlns:dc="**http://purl.org/dc/elements/1.1/**">
- <rdf:Description>
<dc:type modifier="**dcmtobject**">**Textual Folder**</dc:type>
<dc:source modifier="**abbreviation**">**JFKPOF**</dc:source>
<dc:identifier>**JFKPOF-087-005**</dc:identifier>
<dc:source>**John F. Kennedy Personal Papers**</dc:source>
<dc:source modifier="**number**">**3**</dc:source>
<dc:relation modified="**isPartOfSeries**">**Departments and Agencies.**</dc:relation>
<dc:relation modified="**isPartOfSn**">**7.**</dc:relation>
<dc:title>**Public Printer**</dc:title>
<dc:title modifier="**alternative**">**Papers of John F. Kennedy. Presidential Papers. President's Office Files. Departments and Agencies. Public Printer**</dc:title>
<dc:publisher modifier="**contact**">**John F. Kennedy Library (NLJFK), Columbia Point, Boston, MA 02125-3398 Phone: 617-514-1600, Fax: 617-514-1652, Email: kennrdy.library@nara.gov**</dc:publisher>
<dc:publisher>**John F. Kennedy Library (Boston, MA)**</dc:publisher>
<dc:date>**3 May 1962, undated**</dc:date>
<dc:date modifier="**start**">**1962-05-03**</dc:date>
<dc:date modifier="**end**">**1962-05-03**</dc:date>
<dc:creator scheme="**ARC**">**President (1961-1963:Kennedy). Office of the Personal Secretary., 1961-1963**</dc:creator>
<dc:creator modifier="**type**">**MOST RECENT**</dc:creator>
<dc:contributor />
<dc:type scheme="**ARC**">**Textual Records**</dc:type>
<dc:format modifier="**media_type**" scheme="**ARC**">**PAPER**</dc:format>
<dc:description modifier="**type**">**FILE UNIT**</dc:description>
<dc:type modifier="**content**" />
<dc:description>**This folder contains material collected by the office of President John F. Kennedy's secretary, Evelyn Lincoln, concerning the Public Printer. Materials concern the publication and distribution of the book "Public Papers of the Presidents of the United States- John F. Kennedy, 1961."**</dc:description>
<dc:format>**4 digital pages, including 1 photograph**</dc:format>
<dc:format modifier="**digitized**">**4**</dc:format>
<dc:language schema="**ISO639-2**">**eng**</dc:language>
<dc:subject>**United States government departments and agencies**</dc:subject>
<dc:subject scheme="**person**“>**Kennedy, John F. (John Fitzgerald), 1917-1963**</dc:subject>
<dc:subject scheme="**organization**">**Government Printing Office. (1861-)**</dc:subject>
<dc:subject scheme="**geog**“>**Boston (MA)**</dc:subject>

# Know what **metadata** you want to capture

- What established standards will you adopt?
  - Descriptive standards
  - Metadata schema standards
  - Authority control for subjects, personal names, organizational names, and geographic terms
    - Library of Congress authority headings
    - Getty Art and Architecture Thesaurus
    - Local authority lists

# Know what **metadata** you want to capture

- Does this information already exist for your materials, or will it have to be created?
  - Reuse data from finding aids or other descriptive resources
    - How reliable is existing description?
  - And/or add new metadata
    - Do you have the resources to do this?

# Know what **metadata** you want to capture

- At what level will you describe your assets?
  - Collection level
    - The most basic level of description; appropriate for a record in a library catalog
  - Series level
    - The most common level of archival description found in finding aids
    - Each folder is identified with a title, but no other descriptive information
  - File unit level
    - Level of description the JFK Library chose for its textual materials
    - Practical compromise between broad and granular level description
  - Item level
    - Level at which JFK Library describes audiovisual and museum assets

# Know what **metadata** you want to capture

- How will you name your files?
  - A systematic means of identifying assets within a collection, decided before digitization
  - Digital identifiers: include collection abbreviation, followed by further identifying information
    - JFKPOF-001-001 (Folder 1 in Box 1 of the President's Office Files)
    - JFKPP-054-003 (Folder 3 in Box 54 of the John F. Kennedy Personal Papers)

# Know **who** will do the work

- The choices are generally existing staff, new staff, outside vendors, interns, and/or volunteers

# Know **who** will do the work

- The level of description and the scale of project affect level of staffing
  - Use existing descriptive resources, may not need new staff
  - If need to create new descriptive metadata, hire professional archivists
  - If plan to digitize large scale, will need many scanners

# Know **who** will do the work

- We have a combination:
  - Existing staff manage the project
  - Six new staff members as metadata catalogers
  - Iron Mountain Studios provide digitized audio and video assets
  - Interns (both paid and unpaid) are used to scan or digitize

# Know how you will **present** your digital assets


- Digitizing your material isn't enough
- Make digital assets accessible to the public
- Consider before beginning a large project
- Affects renditions and metadata

# Know how you will **present** your digital assets

- Do you have an existing website that can handle digitized content, or will you need to develop an entirely new site?
  - JFK Library website underwent a redesign to incorporate the new Digital Archives
  - Good and bad – too much focus on the site as a whole can take away from the design of the Digital Archives specifically
  - On the other hand, an old site will not be able to support a large number of digital files or their metadata

# Know how you will **present** your digital assets

- Will you share your content with other archival sites/search engines/consortia? If so, how?
  - May provide an alternative to using your site to host digital assets
  - Need to meet established imaging and metadata standards
  - Collaborative sites can increase the visibility of your records, making them available to a wider audience

# Know how you will **present** your digital assets

- Will you charge users to view and/or download your content, or will you offer them free of charge?
  - If you charge, how will that exchange occur?
  - Can you accept credit card payments on website?
  - Are your users willing to pay?
  - JFK Library: available to the public free of charge, but charge for the delivery of high resolution images and AV files

# Know how you will **present** your digital assets

- What kind of search tool will you use?
  - A robust search engine will make the most of your metadata
    - Allows for robust searching by keyword and/or browsing by controlled lists of vocabulary
    - Expensive and time consuming to configure
  - Online finding aids can provide access to digitized assets if a search engine is too expensive
    - Reflects traditional means of discovery
    - Retains contextual information that search results can sometimes lack
    - Encoded Archival Description (EAD), <dao> tag can be used
    - Titles can be linked to digital content using simple hyperlinks if finding aids are HTML or PDF documents

# Know how you will **present** your digital assets

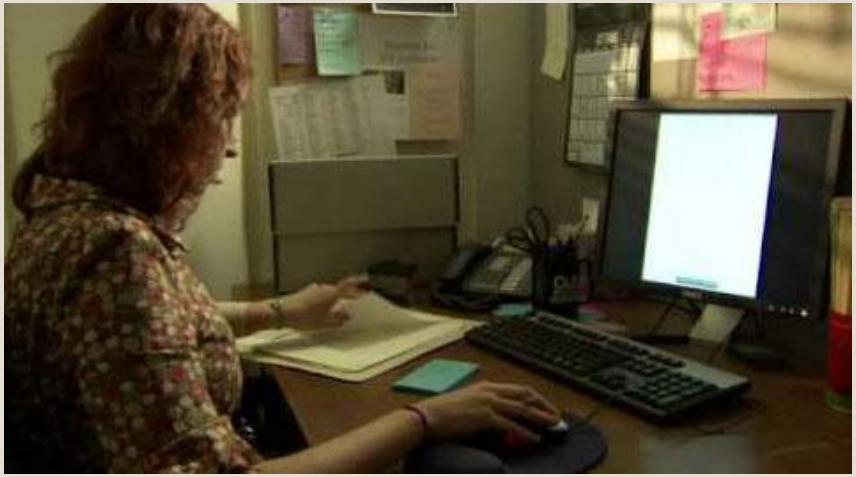
- Are there copyright concerns in your holdings?
  - Can either avoid digitizing copyrighted material or digitize but not make available on public site
  - Depending on volume of digitized material, contact copyright holders and request permission to make their material available online
  - Fair Use is not clear cut when it comes to digital archives
- Similarly, are there privacy concerns in your holdings?
  - Be careful what you post – a personal letter buried in a collection in an archives is one thing, but that letter posted on the internet is something else entirely!

# Know **how you will pay** for digitization


- Digitization can be extremely costly in terms of staff time, equipment (hardware and software), and storage

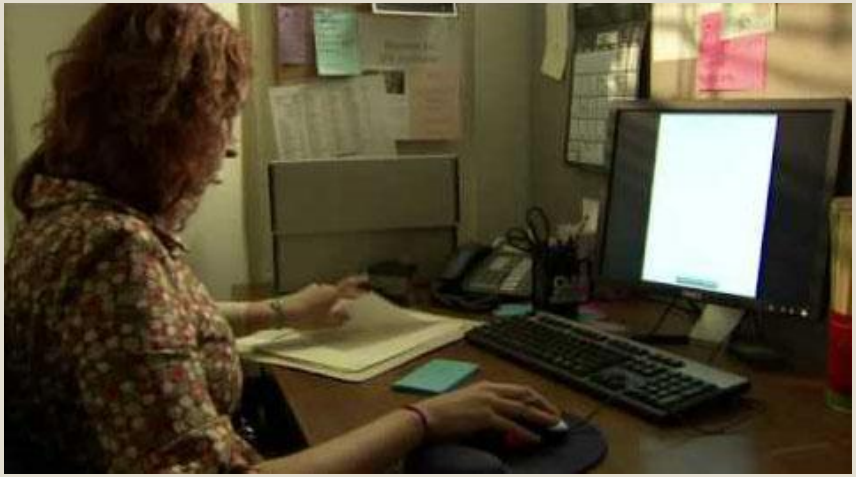
# Know **how you will pay** for digitization

- Staffing costs
  - Institutional commitment to new professional staff
    - Not necessarily sustainable in the long term
    - Flexible planning allows for raising or lowering production based on varying staffing levels
  - Apply for grants
  - Train existing staff
  - Partner with local universities
  - Volunteers



# Know **how you will pay** for digitization

- Staffing costs (con't)
  - Description is most time consuming
    - Scanning happens twice as fast as cataloging (textual and photographic prints)
    - Can create a backlog of digitized material

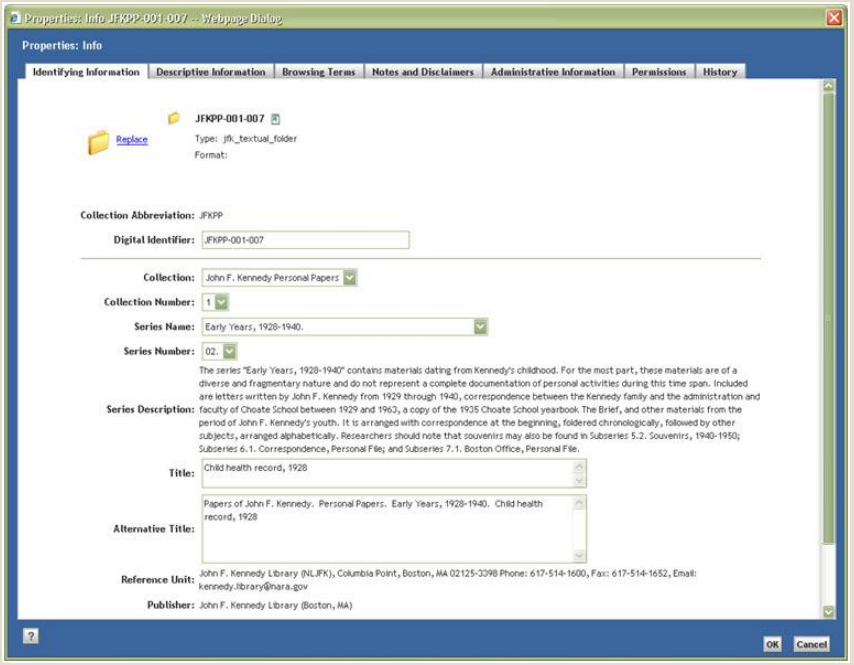


# Know **how you will pay** for digitization

- Hardware costs
  - The more scanners, the more material scanned
  - Flatbed scanners for documents, photographs, and photo negatives (with proper holders)
  - Audio and moving image digitization require more complex equipment

# Know **how you will pay** for digitization

- Software costs
  - Ideally, use a cohesive digital asset management system (DAMS) to manage digital assets
  - There are a range of proprietary and open source options
  - We use EMC's Documentum, a proprietary system – it was donated to us



# Know **how you will pay** for digitization

- Storage costs
  - Creating high quality digital assets quickly fill up storage space
  - JFK's EMC Centera storage system: currently 80 TB, will expand to 300 TB over 5 years
  - Removable hard drives, flash drives, CDs, DVDs - risk of failure
  - Disaster recovery system
    - ALWAYS maintain a back up copy of your files

# Know **how you will pay** for digitization

- A few more things to consider:
  - If soliciting donations, how will donor's interests affect your work?
  - A realistic sustainable permanent digitization program, or a project-based approach?
  - Digital files require long term preservation plan – will files be accessible 10 years from now? 50? 100?

# Questions?

# THANK YOU FOR YOUR TIME

**Contact Information:**

James Roth, Deputy Director

james.roth@nara.gov

Erica Boudreau, Digital Archivist

erica.boudreau@nara.gov

Digital Archives URL:

www.jfklibrary.org/Research/Search-the-Digital-Archives.aspx