NARA Datasets for Artificial Intelligence / Machine Learning

Use Cases for AI/ML

Self-describing Records

- Agencies transfer digitized and electronic records, for which archivists currently have to manually describe for the Catalog
- NARA seeks to process raw archival objects like PDFs, images, and emails (aka “digital objects”) to automatically produce descriptive metadata for public access with minimal archivist intervention

Use Cases for AI/ML

Personalized Search

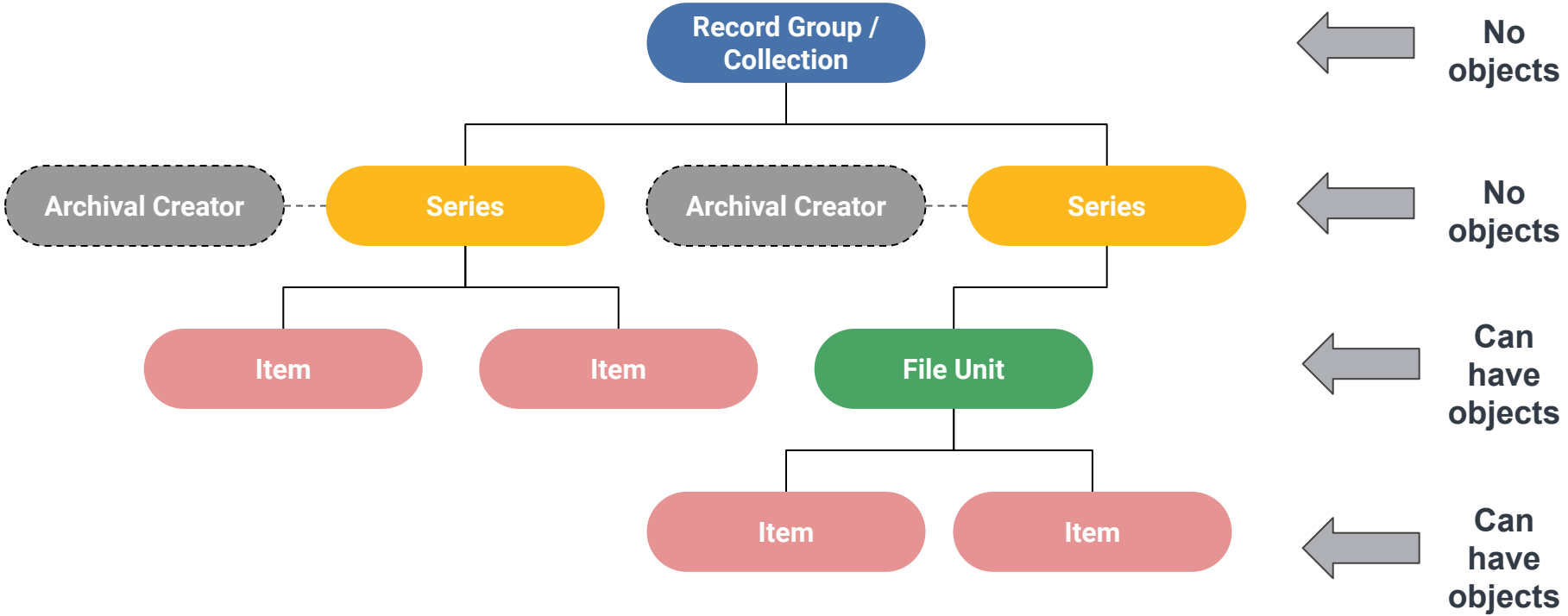
- As the volume of content available for research in our Catalog grows, it becomes increasingly difficult for users to find what they seek
- NARA seeks to provide users of the National Archives Catalog with a personalized search experience that provides relevant search results and makes recommendations for related records based on the user's behavior and interests

Use Cases for AI/ML

Facial Recognition

- NARA holds many photographs of notable people in history, however in many situations these people are not identified in the metadata created by the originating agency or by archivists
- NARA seeks to explore facial recognition technology for it's photographic holdings to assist users with finding relevant photographs of notable people in history

NARA's Archival Hierarchy



Criteria for Datasets

1. **<100,000 objects**
   a. Manageable, but scalable
2. **Textual records**
   a. Majority of NARA's holdings (>90%) are textual records
3. **Object-level metadata**

Datasets with Object-Level Metadata (<100,000 objects)

Record Group / Collection	Series	# of objects	# of descriptions	OCR data?	Citizen archivist data?	National Archives Catalog API query
RG 109 - War Department Collection of Confederate Records	Record Books of Executive, Legislative, and Judicial Offices of the Confederate Government, 1874 - 1899	94,063	1,401	Yes	Yes	link
RG 92 - Office of the Quartermaster General	Correspondence, Reports, Telegrams, Applications, and Other Papers Relating to Burials of Service Personnel, 1915-1939	47,100	982	Yes	Yes	link
RG 75 - Bureau of Indian Affairs	Correspondence Relating to Reindeer Herds in Alaska, 1911-1960	21,426	636	No	Yes	link
RG 120 - Records of the American Expeditionary Forces (World War I)	Records of Divisions, 1918-1942	9,120	2,405	Yes	Yes	link

Datasets with Object-Level Metadata (<10,000 objects)

Record Group / Collection	Series	# of objects	# of descriptions	OCR data?	Citizen archivist data?	National Archives Catalog API query
RG 276 - U.S. Court of Appeals	Case Files, 1891 - 1997	8,238	27	Yes	Yes	link
Collection LBJ-PCTJWHD	Lady Bird Johnson's Daily Diary, 12/1963 - 1/31/1969	4,017	1,604	Yes	Yes	link
RG 412 - Environmental Protection Agency	Program Development Files on Seabrook Nuclear Power Plant, 1/1/1973 - 12/31/1979	3,547	29	Yes	Yes	link
RG 341 - U.S. Air Force	Reports Regarding Proposed Air Force Academy Site Selection, 1950 - 1950	2,872	43	Yes	No	link

Datasets without Object-level Metadata

Record Group / Collection	Series	# of objects	# of descriptions	OCR data?	Citizen archivist data?	National Archives Catalog API query
RG 181- Naval Districts and Shore Establishments	Shipyard Logs, 1888 - 1958	25,694	48	No	No	link
RG 472 - U.S. Forces in Southeast Asia	General Records, 1965 - 1972	23,959	260	No	No	link
RG 60 - Department of Justice	Files of Associate Deputy General Merrick B. Garland, 1994 - 1997	23,552	447	No	No	link
RG 22 - U.S. Fish and Wildlife Service	Endangered Species Delisting Files, 1975 - 2000	7,144	42	No	No	link

Datasets of Electronic Records

Record Group / Collection	Series	# of objects	# of descriptions	National Archives Catalog API query
RG 541- Assassination Records Review Board	Electronic Records Relating to John F. Kennedy Assassination Research, 4/1/1994 - 9/30/1998 [includes emails]	83	83	link
RG 330 - Office of the Secretary of Defense	Defense Casualty Analysis System (DCAS) Files, ca. 2001 - 3/16/2009	4	4	link

Datasets with Photographic Records of People

Record Group / Collection	Series	# of objects	# of descriptions	National Archives Catalog API query
BHO-WHPO - Records of the White House Photo Office (Obama Administration)	Presidential Photographs, 1/20/2009 - 1/20/2017	8,010	8,010	link
WJC-WHPO - Photographs of the White House Photograph Office (Clinton Administration)	Photographs Relating to the Clinton Administration, 1/20/1993 - 1/20/2001	499	499	link
FL - Frank W. Legg Photographic Collection of Portraits of Nineteenth Century Notables	Portraits, 1862 - 1884	91	91	link

NARA Resources

- Information for the NARA - Virginia Tech AI Conference, April-May 2021
- National Archives Catalog
- National Archives Catalog API documentation

Questions? Contact Jason Clingerman (jason.clingerman@nara.gov), the Digital Engagement Division Director in NARA's Office of Innovation