

NEW OFFICE OF SCIENTIFIC RESEARCH AND DEVELOPMENT

1530 P STREET, N.W. · WASHINGTON 25, D.C.

PROJECT MEMEX

AMERICA'S HARDTECH HERITAGE

MEMORANDUM FOR: THE DIRECTOR, NATIONAL ARCHIVES

SUBJECT: Industrial Digitization of the Nation's Technical Records for the
Training of Thinking Machines

DATE: September 13, 2026



1. PURPOSE. This office proposes an industrial effort to digitize the one percent of the National Archives' holdings which constitute America's Hardtech Heritage, its key science and technology records, in order to create a new corpus of pre-training data for the next generation of frontier models. The work is to be accomplished within 180 days, commencing 1 January 2027.
2. BACKGROUND. The Archives hold more than 13 billion pages: paper, microfilm, parchment, and everything in between. The 2022-2026 Strategic Plan set a goal of 500 million pages in the Catalog by 30 September 2026, but this represents less than four percent of the holdings. At the current pace the last box will be scanned sometime in the 24th century, even if the Archives never received another page.

Meanwhile, every frontier laboratory has exhausted the open internet and is now training on synthetic text, the internal correspondence of defunct companies, scraped web forums, and obscure, destroyed books. The single largest reservoir of high-quality, provenance-tracked, public-domain English prose on the planet is sitting in Archives II. The taxpayer paid tens of billions of dollars for decades of engineering studies, laboratory notebooks, and test results from NASA, the Tennessee Valley Authority, the Bureau of Ordnance, and the Geological Survey. It is time to put them back to work.

The tools now exist. Vision-language models read cursive, damaged film, and marginalia at scale. The bottleneck is no longer the technology; it is the decision.

Archives II is the La Brea tar pit of technical data: a paleontological wonder and an oil well to power a new century of American invention. Beyond final reports it preserves the process of discovery itself, the unsuccessful configurations, iterations, measurements, and operating conditions which give the final report its meaning. These data are multi-modal: drawings, schematics, hand calculations, tables, plots, and photographs, all bound to the text that explains them. Grounded engineering data of this kind is nearly nonexistent in open corpora. Its release would accelerate the frontier laboratories and the startups alike, and could start a new wave of American open-weights models.

3. THE DOCUMENTS. We propose to select the one percent of the holdings most essential to training a machine capable of meaningful engineering: 50,000 Hollinger boxes, 150 million pages, 5.4 million cubic feet of technical history. The forty record groups proposed include the NASA research center experiments and mission engineering files; Army Corps of Engineers civil works; the wartime files of this Office; Atomic Energy Commission and Department of Energy technical records; the Bureaus of Aeronautics, Ships, and Ordnance; the early ARPA files; the Office of Science and Technology Policy; the Geological Survey; and the Agricultural Research Service laboratory reports, among much more.
4. THE PEOPLE. The Willow Run plant built 6,972 B-24 bombers during the war, peaking at 428 in the single month of April 1944, better than fourteen a day around the clock. The same industrial methods can be put to work digitizing our technical heritage. One trained person can capture about 2,500 pages per day; that is 60,000 workdays for 150 million pages. Three shifts across 195 capture stations, some 942 capture personnel and 330 in support, can digitize the entire one percent in six months, with a forty percent margin built in for efficiency losses. All personnel will be trained to NARA handling standards, including the handling of photographs, fragile paper, and bound volumes.
5. THE PROCESS.
 - a. Confirm the Targets. Coordinate with NARA and the principal laboratories to ensure the 50,000 boxes selected are of the highest value for pre-training.
 - b. Manifest and Pull. Assign container and batch identifiers, reconcile the requested record groups to the pull manifest, record condition, track exceptions, and stage cleared batches, with full traceability throughout.
 - c. Transfer and Receive. Use approved, trained shippers with dispatch and receipt reconciliation, along documented routes, under pre-approved procedures. A ten-day buffer of roughly 3,600 box equivalents limits the number of originals out of storage at any one time while keeping the stations supplied. The 50,000 boxes require 63 round trips at 800 boxes per trip, about one every three days.
 - d. Prepare and Capture. Stage receiving, preparation, capture, quality review, and return lines. Annotate every document with metadata before capture, and use commodity scanning equipment at every station, with specialty equipment reserved for oversize, bound, and fragile material.
 - e. Accept and Upload. Confirm integrity and image quality by a control process. Upload the scanned files with their metadata (page sequence and container number, checksum, and an audit log naming the responsible personnel and the NARA authorization) to multiple commercial cloud environments.
 - f. Return and Reconcile. Originals are returned as received, in documented order and in approved boxes. Each container is closed only after NARA personnel confirm that all records have been returned, and record their approval by digital receipt. All exceptions are tracked and routed back for resolution.

6. THE FACILITY. The Archives were not built for industrial digitization. We propose to truck the selected material to a nearby site of not less than 32,000 square feet of floor space and there build a digitization factory. Multiple candidate warehouses, industrial buildings, and former retail spaces lie within sixty minutes of College Park and can be brought to the 36 CFR 1234 standard for archival facilities by retrofit of heating, ventilation, and fire suppression. Federal and university partners may also have space at hand.
7. THE PRECEDENT. The Archives' partnership with Ancestry.com, the largest of many such arrangements, authorized large-scale, off-site digitization under strict handling requirements and opened the genealogical record to hundreds of millions of Americans. This project proposes to do the same, at an accelerated pace.
8. THE COST. The total estimated cost is \$50.8 million, negligible beside the cost of a single data center, and an investment in the country's future. The figure comprises some \$25 million of direct capture labor at \$35 per hour loaded, with a fifteen percent relief allowance; a further \$9 million for logistics, preparation, and quality assurance; \$750,000 for NARA oversight; \$3.5 million for compute and storage for optical character recognition and extraction, at one cent per page equivalent, with preservation-quality storage for the first year; \$513,000 for shipping at \$8 per box and \$1,800 per vehicle round trip; \$314,500 for capture equipment at \$1,000 per station with ten percent spares and \$100,000 of specialty equipment; and \$830,000 for the facility at \$24 per square foot per year, \$6 per square foot for utilities and security, six months of occupancy, and \$350,000 of fit-out. A contingency of \$10.2 million (twenty-five percent) is held against retrofit, shipping, equipment, or labor overages.

Capture labor	\$25,806,690
Preparation, quality assurance, and floor support	9,032,342
Mobilization and archival oversight	750,000
Facility, fit-out, and utilities	830,000
Capture equipment	314,500
Pulls, manifests, and return transport	513,400
Optical character recognition and structured extraction	1,500,000
Preservation storage, first year	1,944,000
Contingency (25%)	10,172,733
TOTAL	\$50,863,664

9. THE SCHEDULE. To meet the urgent demand for training data, this program proposes to do years of work in 180 days, commencing 1 January 2027.

Days 1-30: Facility fit-out and testing; train staff; run a representative pilot to validate process and quality.

Days 31-167: Three eight-hour shifts per day of capture.

Days 168-181: Complete reconciliation, test recovery, and delivery.

10. THE DELIVERABLE.

a. Unprecedented Data. No public dataset of this kind exists today. The result is an enduring repository which unlocks new companies focused on the difficult, grimy work of hardtech engineering, strengthens the frontier laboratories' pre-training, and makes possible new American open-weights models trained upon it.

b. Clean Title. No copyright litigation, no licensing fees, no foreign jurisdiction. The American people already own it; the results become a common good.

c. A New Process. A facility, and a method, for the rapid mass digitization of technical data, applicable thereafter to further Archives holdings and to legacy industrial records of every kind.

11. RECOMMENDATION. That the Director approve Project MEMEX; make the selected record groups available for off-site capture under the partnership authorities already in use; and support the request for the funds set out in paragraph 8, so that work may begin on 1 January 2027 and the corpus be delivered to American laboratories before the following summer.

The country paid for this data once. It is respectfully submitted that we make it work a second time.

APPROVED:

Director, National Archives

APPROVED · SEP 13 2026

Enclosures:

1. Schedule of Record Groups (40)
2. Facility Survey, College Park Vicinity
3. Cost Estimate, Detail

Distribution:

The President
Director, Office of Science and Technology Policy
File