Computational Archival Science (CAS) and National Funding Priorities

Richard Marciano

UMD iSchool

“Finding New Knowledge: Archival Records in the Age of Big Data”
COMPUTATIONAL ARCHIVAL SCIENCE:

An interdisciplinary field concerned with the application of computational methods and resources to records/archives processing, storage, long-term preservation, and access:

> *To improve efficiency, productivity and precision in helping archivists make appraisal, arrangement and description, preservation and access decisions through large-scale analysis of archives.*

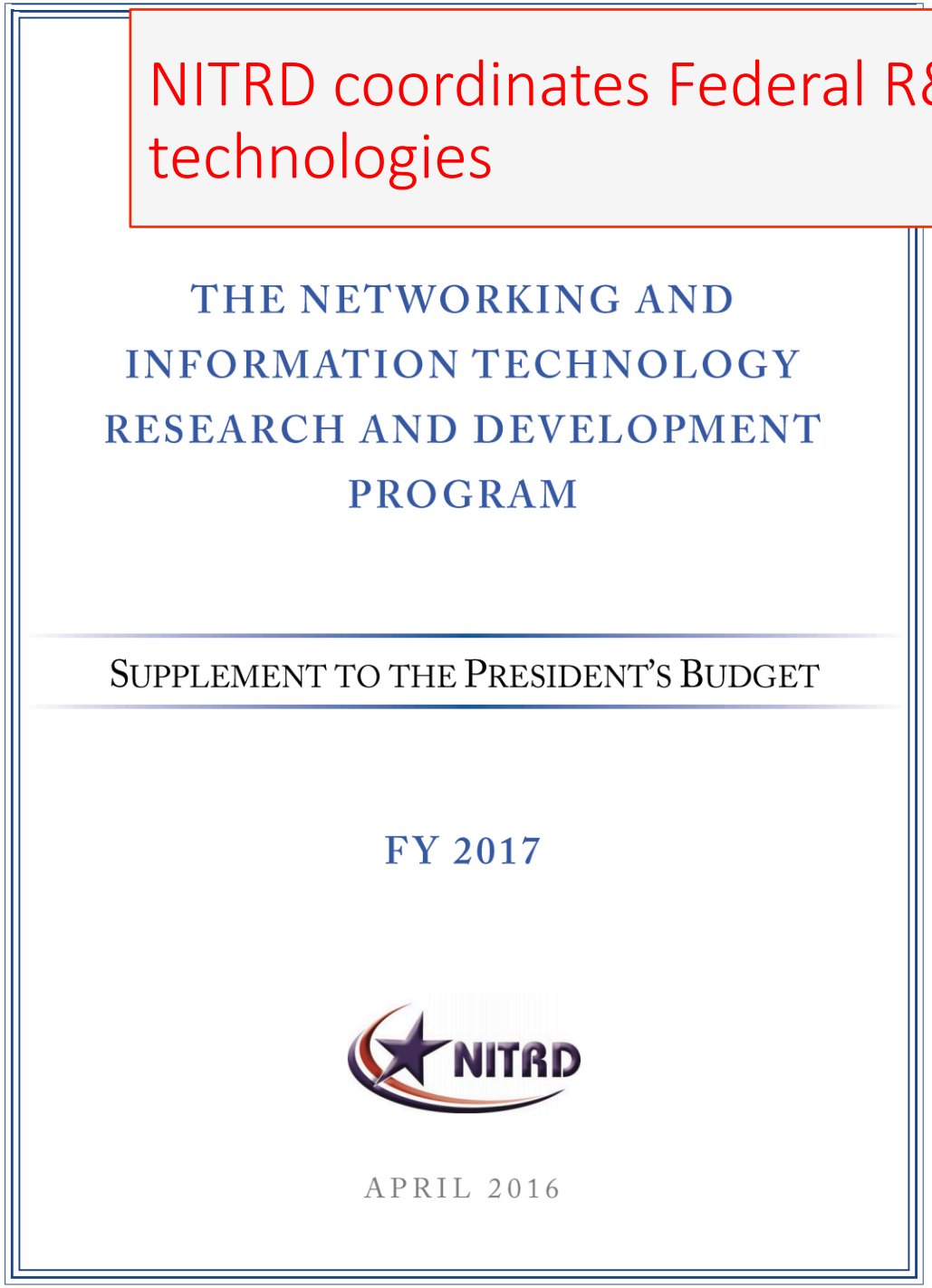
This suggests that computational archival science is also a blend of *computational thinking with archival thinking.*

Some of the challenges might arguably be (part of the discussion):

- We need a way to better convey to funders the principles and challenges of big archives.
- We need a way to better educate archivists about the principles and challenges of big archives.

10:00 - 11:15	Federal Agency Panel: *Diane Travis (MC)* • **NITRD NCO** (Keith Marzullo) o Director of the *Networking and Information Technology Research and Development* (NITRD) National Coordination Office (NCO) • **NSF** (Robert Chadduck) o Program Director, *Data & Cyberinfrastructure* • **NIH** (Jennie Larkin) o Senior Advisor, *Extramural Programs & Strategic Development, Office of the Associate Director for Data Science* • **IMLS** (Trevor Owens) o *Senior Library Program Officer at the Office of Library Services Discretionary Programs* • **NEH:** (Brett Bobley) o *Director, Office of Digital Humanities*

NITRD coordinates Federal R&D investments in advanced digital technologies



- NITRD:
  - https://www.nitrd.gov/
- MEMBER AGENCIES:
  - https://www.nitrd.gov/SUBCOMMITTEE/nitrd_agencies/agency_contacts.aspx
- Topics directly relevant to archivists and records managers

Subcommittee on Networking and Information Technology Research (NITRD)

- Dept. of Commerce: NIST, NOAA
- Dept. of Defense: DARPA, NSA, OSD, Service Research Organizations (Air Force, Army, Navy)
- Dept. of Energy: DOE/NNSA, DOE/OE, DOE/SC
- Dept. of Health and Human Services: AHRQ, NIH, ONC
- Dept. of Justice: DOJ
- Dept. of Homeland Security: DHS
- Independent Agencies: EPA, NASA, NARA, NRO, NSF, OMB, OSTP, NITRD/NCO

Program Component Areas (PCAs) for FY 2017:

1. Cyber Security and Information Assurance (CSIA)
2. Enabling-R&D for High-Capacity Computing Systems (EHCS)
3. High-Capability Computing Systems Infrastructure and Applications (HCSIA)
4. High Confidence Software and Systems (HCSS)
5. **Human Computer Interaction and Information Management (HCI&IM)**
6. **Large-Scale Data Management and Analysis (LSDMA)**
7. Large Scale Networking (LSN)
8. Robotics and Intelligent Systems (RIS)
9. Software Design and Productivity (SDP)
10. Social, Economic and Workforce Implications of IT and IT Workforce Development (SEW)

NARA's programmatic interests:

- Global-scale, open source, next-generation technologies, architectures, and services enabling effective, sustainable management, intellectual control, and access to nationally distributed billion-file-and-larger scale, complex digital object collections.

5. HCI & Info. Mgt.: AHRQ, DHS, DoD Service Research Orgs, EPA, NARA, NIH, NIST, NASA, NOAA, NSF

- *Transforming data to knowledge* (tools to accelerate scientific discovery and productivity from heterogeneous data stores, and development of innovative multidimensional approaches to highly complex data).
- *Human engagement and decision-making* (research to design effective HCI mechanisms for distributed collaboration, knowledge management, virtual organizations, and visual environments).
- *Effective stewardship of science and engineering data* (federation, preservation, and analysis of large, heterogeneous collections of scientific data, information, and records).
- *Information integration, accessibility, and management* (tools for optimized, scalable ingest and processing for high-capacity data integration -- GIS, management, exploitation, modeling, and analysis)
- *Health information technologies* (clinical decision-support systems, electronic health records).
- *Information search and retrieval* (legal discovery, domain-specific search and machine reading of records).
- *Cognitive, adaptive, and intelligent systems* (trustworthiness and reliability of automated systems)
- *Multimodal language recognition and translation* (document summarization/distillation, automatic content extraction)

6. Large-Scale Data Management and Analysis (LSDMA): DARPA, DHS, DOE/NNSA, DOE/SC, EPA, NARA, NASA, NIH, NIST, NOAA, NSA, NSF, OSD

- *Next-generation capabilities and improved trustworthiness of data for decision making* (when dealing with large-scale and complex data, new ways to enable trustworthy and intuitive visualization of data as well as effective analytical tools for decision makers).
- *Cyberinfrastructure* (improve big data and big compute to develop a capable exascale computing system that support both simulation and data analysis at scale).
- *Data capture, curation, management, and access* (digital resource discovery and indexing, data and metadata standards, and sustainability – storage of data, reference datasets to enable new tools).
- *Data privacy, security, and ethics* (guidelines and best practices to enable privacy protection).
- *Education and training*
- *Collaboration*

NARA: Short List of Research Partners

- University of Maryland Institute for Advanced Computer Studies (UMIACS): Producer-Archive Workflow Network (PAWN) for electronic records transfer, data management at scale.
- DARPA: DOCT project – early research (circa 1998) in the use of high performance computing (supercomputers and related technologies) for long term preservation of archival collections of electronic records.
- San Diego Supercomputer Center – later University of North Carolina at Chapel Hill - now the University of Maryland (DAKS/DICE/CI-BER/DCIC): data management at scale (Storage Resource Broker (SRB), integrated Rule-Oriented Data System (iRODS), metadata extraction, arrangement, description, and visualization of large collections of electronic records, email ingest, etc.
- Georgia Tech Research Institute (GTRI): Advanced tools for processing Presidential Records (Presidential Electronic Records PilOt System (PERPOS)), metadata extraction, content summarization, automated description, file type identification – including substantial contributions to PRONOM and DROID, assistive technologies for review/redaction of sensitive information, auto-categorization of email, grammars for file type specification/parsing/validation.
- National Center for Supercomputing Applications (NCSA): Data format description language parsers, hierarchical data format (HDF), email ingest, automated content comparison for versioning and de-duplication, Conversion Software Registry (CSR), reconstruction of computer-aided decision-making, searchable access to digital images of handwritten Census forms, Polyglot for automated file format conversion.
- Texas Advanced Computing Center (TACC): Visualization, metadata extraction, content summarization for very large collections.
- **PDES, Inc.:** Long term preservation and reuse of electronic engineering records.
- **Army Research Lab:** Foreign language records auto-translation, cybersecurity.
- **Networking and Information Technology Research and Development (NITRD) Program:** NARA's research efforts have been coordinated through this program. The NITRD Program provides a framework in which many Federal agencies come together to coordinate their networking and information technology (IT) research and development (R&D) efforts. The Program operates under the aegis of the NITRD Subcommittee of the National Science and Technology Council's (NSTC) Committee on Technology. The Subcommittee, made up of representatives from each of NITRD's member agencies, provides overall coordination for NITRD activities.