COLLABORATION IS THE THING

MARK CONRAD

ARCHIVES SPECIALIST

NATIONAL ARCHIVES AND RECORDS ADMINISTRATION

COMPUTATIONAL ARCHIVAL SCIENCE WORKSHOP

IEEE BIG DATA 2016

8 DECEMBER 2016

WASHINGTON, DC

CAVEAT EMPTOR!

All statements, assertions, conjectures, etc are my own unless otherwise specifically attributed.

I'M AN ARCHIVIST…

BIG DATA IS A BIG DEAL FOR ARCHIVISTS

- Got more electronic records than you can easily manage, preserve, and provide access to?
- Got records in so many formats your head spins?
- Got more text than you could read in several lifetimes?
- Can't figure out how to redact all the sensitive information?

YOU'VE GOT BIG DATA!

PEN AND PAPER METHODS DON'T WORK

- For decades archivists have tried to pound a square peg into a round hole
- When you have a new hammer everything looks like a square peg
- We often grasp at straws and think they are hammers

A CHANGE OF PERSPECTIVE

- We need to be exposed to new ideas/methods
- Realize that other disciplines have similar challenges
- Collaborate to “kick the tires” on new technology
- Get our hands dirty
- Don’t lose sight of our unique requirements
- Look for quick wins, but appreciate the losses

COLLABORATION IN ACTION: A FEW EXAMPLES

- Collaborators: NARA, ARL, GTRI
- Collaborations:
  - Language of Records Disposition
  - PERPOS
  - Information Extraction
  - Content Summarization
  - File Type Identification
  - Many more!
- http://perpos.gtri.gatech.edu/

COLLABORATION IN ACTION: A FEW EXAMPLES

- Collaborators: NARA, DARPA, NSF, SDSC ->UNC-CH -> UMD
- Collaborations:
  - Feasibility study
  - Data management at scale – SRB -> iRODS -> DRAS-TIC
  - Metadata extraction and automated description
  - Management of Archival Collections at scale
  - Many more!

COLLABORATION IN ACTION: A FEW EXAMPLES

- Collaborators: NHPRC, Archivists, SDSC -> UNC-CH, GTRI
- Collaborations:
  - Persistent Archives Testbed
  - InterPARES
  - Distributed Archival Custodial Preservation Environments for electronic records (DCAPE)
  - Many more!

COLLABORATION IN ACTION: A FEW EXAMPLES

- Collaborators: TACC, NARA, NSF
- Collaborations:
  - The Embedded Archivist
  - Visualization and Archival Collections
  - Data Mining for “Big Archives” Analysis
  - Integrating Multi-touch in High-Resolution Display Environments
  - Content Clustering
  - Many More!

COLLABORATION IN ACTION: A FEW EXAMPLES

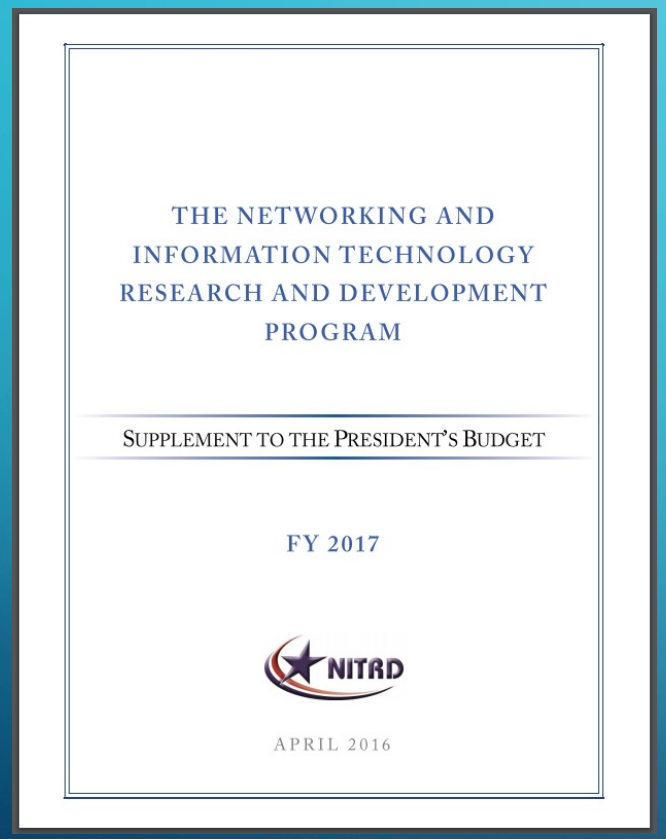
- Collaborators: NARA, CCSDS, NASA, Other Space Agencies
- Collaborations:
  - OAIS – ISO 14721
  - Trustworthy Digital Repositories – ISO 16363
  - Requirements for bodies providing audit and certification of candidate trustworthy digital repositories – ISO 16919
  - Many more!

MANY MORE COLLABORATIONS!

- NCSA
- NIST
- UMIACS
- Internet2
- NAVSEA
- PDES
- WVU
- NNSA
- Pittsburgh SC
- LOTAR

https://www.archives.gov/applied-research/papers/publications.html

COLLABORATION ACROSS THE U.S. GOVERNMENT



https://www.nitrd.gov/pubs/2017supplement/FY2017NITRDSupplement.pdf

NITRD FY 2017 PRIORITIES
HCI & INFO. MGT.:

- *Transforming data to knowledge*
  - Tools to accelerate scientific discovery and productivity from heterogeneous data stores,
  - Development of innovative multidimensional approaches to highly complex data
- *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 – esp. GIS
  - Management, exploitation, modeling, and analysis
- *Information search and retrieval*
  - Legal discovery, domain-specific search, recognition of opinion, and machine reading of records.
- *Multimodal language recognition and translation*
  - Document summarization/distillation, automatic content extraction

NITRD FY 17 PRIORITIES, CONT'D
LARGE-SCALE DATA MANAGEMENT AND ANALYSIS (LSDMA):

- *Next-generation capabilities and improved trustworthiness of data for decision making*
  - Enable trustworthy and intuitive visualization of data
  - Effective analytical tools for decision makers
- *Data capture, curation, management, and access*
  - Digital resource discovery and indexing
  - Sustainability
  - Reference datasets to enable new tools
- *Data privacy, security, and ethics*

NITRD FY 17 PRIORITIES, CONT'D

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.

COMMON ELEMENTS

- Practical problems
- Multiple perspectives
- Something for all parties
- Scalable, evolvable, extensible
- Many involved students
- 1990s to 2012 and still learning lessons
- Needs to continue!

POSSIBLE IMPLICATIONS

- More faculty participation in such collaborations at scale
- Ditto for students
- Cross-discipline team teaching
- No more pen and paper methods for electronic records

QUESTIONS?

- Comments?
- Rotten tomatoes?