PARTICIPANT NAME.	**GEORGE O. STRAWN**
TITLE.	DIRECTOR
INSTITUTION/ AFFILIATION(S).	NITRD/NCO
EMAIL ADDRESS.	GSTRAWN@NITRD.GOV

PRIMARY RESEARCH OR PRACTICE AREA(S):

- COORDINATION OF FEDERAL PROGRAMS IN NETWORKING AND IT R&D

PREVIOUS EXPERIENCE

- IOWA STATE UNIVERSITY COMPUTER SCIENCE DEPT AND COMPUTATION CENTER; NSF CISE AND NSF CIO

RELATED WORK

- PROJECT NAME & URL

WWW.NITRD.GOV/NITRDGROUPS/INDEX.PHP?TITLE=BIG_DATA_(BD_SSG)#TITLE

DASPOS Data and Software Preservation for Open Science

SURVEY OF COMMONALITY WITH OTHER DISCIPLINES

WORKSHOP 2 – JULY 25, 2013
INDIANAPOLIS, INDIANA

Some US Data to Knowledge Matters

George O. Strawn
NITRD.gov

Caveat auditor

The opinions expressed in this talk are those of the speaker, not the US government

Outline

- NITRD
- US Big Data research initiative
- Open Access to US govt data and Public Access to science results supported by the US govt
- Semantic Medline

NITRD Interagency Committee (Networking and IT R&D)

- Reports to the White House Office of Science and Technology Policy (OSTP)
- A 22-year-old program to enhance coordination and collaboration among the Federal agencies that perform and support IT R&D

NITRD Member Agencies

Department of Commerce (2)

Department of Defense (5)

Department of Energy ((3)

Department of Health and Human Services (3)

Department of Homeland Security

Environmental Protection Agency

National Archives and Records Administration

National Aeronautics and Space Agency

National Reconnaissance Office

National Science Foundation

National Security Agency

NITRD program component areas

- Cyber Security and Information Assurance
- High Confidence Software and Systems
- High-End Computing
- Human Computer Interaction and Info Mgmt
- Large Scale Networking
- Social, Economic, and Workforce Implications
- Software Design and Productivity

NITRD senior steering groups

- Big Data
- CyberPhysical Systems
- Cybersecurity
- Health IT R&D
- Wireless Spectrum Efficiency

Big Data

- A term applied to data whose size, velocity or complexity is beyond the ability of commonly used software tools to capture, manage, and/or process within a tolerable elapsed time.
- volume, velocity, *variety*, etc

NITRD's
Big Data Initiative

- Core Technologies
- Domain Research Data
- Challenges/Competitions
- Workforce Development

Core Technologies

- Collection, Storage and Management of Big Data
- Data Analytics
- Data Sharing and Collaboration

Domain Research Data

- NSF projects such as DataOne, DataNet
- Earth Observation Systems
- Astronomy, Virtual Observatory
- Genomics
- Nano S&T, Nanohub
- Materials Genome
- Particle Physics, LHC
- data.gov

Challenges/Competitions

- Engage a broader public

Workforce Development

- Data Science, Big Data degrees

USG and Data

- *Open Access* to usg data becomes the default (http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf)
- *Public Access* to Federally funded science journal articles *and* science data required of all agencies funding more than $100M per year (http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf)

Questions for USG data

- Where do you put it? In the cloud?
- How do you find it? Browsing? Search, Semantic search?
- How do you use it? Web service? APIs? Semantics?

Semantic Medline

An example of the automatic extraction and integration of biomedical information

Medline & UMLS

- Developed and Maintained by HHS/NIH/NLM
- Medline is a database of the titles and abstracts of ~20 million biomed research articles
- UMLS (Unified Medical Language System) is set of biomedical vocabularies

Semantic Medline

- A knowledge base of ~60 million "key sentences" from Medline

- A key sentence is of the form subject-verb-object (an RDF triple)

- The key sentences are derived from the Medline titles and abstracts by linguistic analysis and are normalized by a controlled vocabulary derived from UMLS

Semantic Medline provides

- A graphic view of a specified portion of the Semantic Medline graph (the subject and object nouns label the graph nodes and verbs label the arcs)

- The graphic view supports browsing and recall of the articles containing the graph link (aka key sentence)

- A sparql query capability

A new mode of discovery

- Why do older men have more sleep problems?
- What connections exist between cancer, obesity and circadian rhythms?
- Who would use a new mode of discovery?

Semantic Medline + Nanopublications?

- Semantic Medline addresses the past
- Nanopublications address the future, assuming investigators are educable
- Nanopublication author burden could be lowered by concurrent Semantic Medline processing and dialog with author

In the long run?

- All science disciplines develop UMLS-like vocabularies
- All science disciplines develop Semantic Medline-like knowledge bases
- All disciplinary knowledge bases are interoperable to facilitate interdisciplinary discovery