APLU Council on Research, Summer Meeting, June 2015

Steven Daley-Laursen
University of Idaho
Northwest Knowledge Network
USDA NAREEE Data WG

Research Data Management—Routes Around Potholes
Many Types of Data to be Managed by Universities

- Research and Outreach Data & Computing
  - On-Line Teaching
  - University Business Analytics
  - Student Data
Multiple Concepts in Play

Big Data  Open Data  Publications
Data-Enabled Science

Data Management Services
Practitioner & Professional Applications
Questions for you to answer... today

• Why are Big/Open Data and Data Intensive Science important to you as a university officer?
  • Challenges it presents? Opportunities it presents?

• Who can you partners with to position your university?
  • research faculty, graduate students and undergraduates?
  • other university officers?
  • external partners?
The ‘main’ reason we need to be onto this topic....
Reasons Why You/I Need to Care

✓ Enables research/ers and education/ors
✓ Makes faculty and institution more competitive
✓ Increases effectiveness of federal AND state-funded research, education and outreach
✓ Enables more informed/skilled decision-making by managers and practitioners served by research
✓ There are costs of doing something AND doing nothing.
✓ Government is acting; Universities must foster partnerships with government.
Drivers of the Data Phenomenon
Big data growth

Big data market is estimated to grow 45% annually to reach $25 billion by 2015

Growth of Global data - Zettabytes

Zettabyte = one million petabytes

2010 Stored data* - Petabytes

Petabyte - one quadrillion (short scale) bytes

*greater than
Sources: Nasscom -CRISIL GR&A analysis
The **FOUR V’s of Big Data**

**Volume**

It’s estimated that 2.5 quintillion bytes (2.3 trillion gigabytes) of data are created each day.

- **6 billion people** have cell phones.
- **World population**: 7 billion

**Velocity**

Modern cars have close to 100 sensors that monitor items such as fuel level and tire pressure.

By 2015, there will be 18.9 billion network connections — almost 2.5 connections per person on earth.

**Variety**

30 billion pieces of content are shared on Facebook every month.

1 in 3 business leaders don’t trust the information they use to make decisions.

**Veracity**

27% of respondents in one survey were unsure of how much of their data was inaccurate.

By 2014, it’s anticipated there will be 420 million wearable, wireless health monitors.

4 billion hours of video are watched on YouTube each month.

400 million tweets are sent per day by about 200 million monthly active users.

As of 2011, the global size of data in healthcare was estimated to be 150 exabytes (1.67 trillion gigabytes).

By 2020, it is estimated that 40 zettabytes (4 trillion terabytes) of data will be created by 2020, an increase of 30 times from 2005.

**The New York Stock Exchange captures 1 TB of trade information during each trading session.**

**IBM**

Sources: McKinsey Global Institute, Twitter, Cisco, EMC, SADS, IBM, MEPTEC.
2. Emergence of a Fourth Research Paradigm

Thousand years ago – **Experimental Science**
- Description of natural phenomena

Last few hundred years – **Theoretical Science**
- Newton’s Laws, Maxwell’s Equations…

Last few decades – **Computational Science**
- Simulation of complex phenomena

Today – **Data-Intensive Science**
- Scientists overwhelmed with data sets from many different sources
  - Captured by instruments
  - Generated by simulations
  - Generated by sensor networks

\[
\left( \frac{a}{a} \right)^2 = \frac{4\pi G \rho}{3} - K \frac{c^2}{a^2}
\]

**eScience** is the set of tools and technologies to support data federation and collaboration
- For analysis and data mining
- For data visualization and exploration
- For scholarly communication and dissemination

*(With thanks to Jim Gray)*
“...everything about science is changing because of the impact of information technology. Experimental, theoretical, computational science are all being affected by the data deluge, and a fourth, data intensive science paradigm is emerging.

The goal is to have a world in which all of the science literature is online, all of the science data is online, and they interoperate with each other.

Lots of new tools are needed to make this happen.”

- Jim Gray, Microsoft Research
Reduced Time to Insight

Real-world Data

Computational Modeling

Persistent Distributed Data

Workflow, Data Mining & Algorithms

Interpretation & Insight

(Thanks to Craig Mundie)
3. Issues and solutions are enabled by applications, everywhere.
4. Feds Make Big Data a Big Deal

• White House Office of Science and Technology
  • 03/12: Big Data Research and Development Initiative
  • 02/13: All data collected with public funds must be public
  • 5/13: OSTP/OMB Require DM Plans (Big, Open) from each Department

• National Science Foundation (NSF, NIH) and so on....
  • Data acquired with public funds belongs to the public.
  • Data management plans required for all grant applications
  • Data may now be listed in NSF biosketches
Executive Branch Mandates

...executive offices and funding agencies requiring data accessibility and encouraging data intensive use...

2-2013

OSTP Policy: “Increasing Access to the Results of Federally Funded Scientific Research” Requires a plan to support increased public access to the results of research (scholarly publications and science data) funded by the Federal Government

5-2013

OMB: “Open Data Policy—Managing Information as an Asset”

· May 9: WH Executive Order: “Making Open and Machine Readable the New Default for Government Information”
I’m Lonely and Unsure
Who Else is Doing This That I Need to Connect with at On/Off Campus???
Complying with Mandates and Enabling Data-Intensive Activity Dictates Cooperators

+ Quality data and metadata throughout lifecycle
+ Data management policies
+ Data/Pub cataloging, serving, application tool services
+ Centralized IT, access to HPC, pipelines
+ Research; Interoperability (TEK-BioP-Social) and Virtualization
+ Workforce development; domain and software

= RESEARCH – LIBRARY – ITS – ACADEMICS – GCOUNSEL
STATE-REGIONAL-GLOBAL NETWORKS
2009 to 2015

NORTHWEST KNOWLEDGE NETWORK

UNIVERSITY OF IDAHO AND COOPERATORS

www.northwestknowledge.net
NKN Mission

Enable research teams to address complex societal problems by facilitating quality metadata, and the storage, discovery and dynamic analysis of data as long term, dependable assets.

Advance research and education in support of data intensive science.
NKN Basic Beliefs

- Data must be captured and stored -- Redundancy
- Uncurated data will be lost – Metadata, Links, Provenance
- Data must be discoverable and accessible.
- Data will have value beyond its original intent -- DOI
Northwest Knowledge Network

- Lifecycle management for heterogeneous research data
  - Tiered, distributed data storage
  - Metadata Tools, Standards
  - Data discovery and retrieval
  - Data-centric researcher collaboration tools
  - Interoperability across scale, time, data discipline (incl TEK, Social)

- NKN Big Data Functions
  - Capture
  - Storage
  - Curation
  - Search
  - Sharing
    - Analysis
    - Visualization
Advancements in CI research and CI training to enhance NKN architecture, services and application.

Tools and services enable interdisciplinary research between geographically distributed researchers and visualization of results for policy makers.

Policies and procedures facilitate data sharing and preservation between academic institutions and with state and federal agencies.
BIG/OPEN DATA NETWORKS AND TEAMS

Agencies: USDoI, USDoE, NSF, USDA

Academic Units

CI Research and Workforce Development

Customer Relations and Communications

Technology and Project Management (Service Center)

Partnerships

Library

Global-Local Networks (DataONE, IRON)

ITS

LEGAL

INL HPC
Staff funded, hired and located in collaboration with Library, Research, ITS, and “member” Agencies
Welcome to NKN

The Northwest Knowledge Network (NKN) supports researchers with data management services and is focused on building and maintaining a catalog of research data and standards-compliant metadata. Learn more about NKN.

Search NKN’s Database (enter keywords) Go! Reset

HELP: Enter keyword(s) above and/or draw a search box on the map below (results appear below map). Use of data provided on this website constitutes acceptance of our TERMS OF SERVICE.

Limit bounding box search: Intersecting Within

Want to Upload Data?

Contact NKN
Northwest Knowledge Network
University of Idaho
Moscow, ID 83844-2358
Office: Library 416A
Building the Infrastructure

- Staffing for Services, Research, Training and Cooperation
- Hardware, Software, Licenses
- Data Policies, Procedures and Governance Across Institutions and Sectors
- Connectedness/Access to Distributed Data and Computational Capacity
Funding Model Stage One – Establishment

• Collaborative regional and national fiscal partnerships around common interest in data curation and access. Membership Grants”. USDOI-USGS, NSF-EPSCoR, USDA-Forest Service Research.

• Grants from federal agencies to support regional partnerships on critical issues (climate) with significant data management activity.

• Pressure on agencies to fund “directly” in grants, programs, joint infrastructure development.
NKN Leverage

Western Region CI Strategic Partners

NW Regional CI-Data Clients

National/International Common Policies, Node, Sharing

Statewide CI Strategic Partner

Failover, Network, HPC Partners

NKN Leverage

How 5 Years of Strategic Agency-University CI Partnerships Have Supported Data-Enabled Research

2005-08, **UI Bio-Informatics; USDA FS Fire Research Mgmt Info System**

2008, CI Capacity Priority in **EPSCoR Trk 1 Climate**;

2009, UI Strategic Plan for Research Data Management and CI

2010, Northwest Knowledge Network, **EPSCoR CI Coord, CI Faculty**

2010-12, Regional **USDI/USDA/NSF** Projects w/ NKN Leading Data/CI

2013, **EPSCoR** Funds Faculty & Tri-State CI, State CI Strategic Plan
MISSION

To coordinate the expertise of federal and university scientists to provide scientific information and tools that may help address federal, state, and tribal resource managers’ priorities in response to a changing climate.
DOI-USGS Climate Science Centers
and $800K in supplemental grants to NKN to develop DM tools of benefit to the national CSC network.

Goal 1: Establish a significant and effective leadership presence to strengthen the region’s ability to plan and implement a coordinated climate science portfolio.

Goal 2: Develop resources and programs to enhance climate science literacy, and give regional audiences the necessary tools and information to promote climate change awareness.

Information Subset

NKN Required Fields
(minimum for data discovery)

General Information:
- File Identifier
- Metadata Language
- Metadata Data Stamp
- Organization
- Organization Role

Identification Information:
- Title
- Dataset Publication Date
- Dataset Language
- Abstract
- Data Type

Browse Graphic:
- Browse Graphic URL
- Browse Graphic Caption
- Browse Graphic Type

Data Theme:
List of various theme topics

Reference System Information:
- Reference System URN
- Reference System Organization ID

Spatial Domain:
- Lat/Long bounding coordinates

Time Period of Content:
- Beginning Date
- Ending Date

Data Quality Information:
- Scope
- Data History

Contact Information:
- Organization
- Contact Position
- Address
- City
- State
- Postal Code
- Country
- Email
- Contact website
- Telephone
- Fax

Distribution Information:
- Format Name
- Format Version
- Distribution Link
- Distribution Link Function
- Download Size

Metadata Reference Information:
- Metadata Standard Name
- Metadata Standard Version
DM “Member” Partnership with USGS

* Provide data-metadata support, services, tools for researcher teams, and education for teams of climate researchers-students-resource managers,

* Store, secure and serve climate data, analytical functions and interpretive services to enhance science and application.

* Establish National DM WG in USGS CSC Network re: joint interests.

* 2012 -- DOI decision to use centralized repository for all CSC research data. And, Membership grants from USGS ending.
CC-NIE Networking Infrastructure: Enhancing Network Capabilities to Foster Big Data Science at the University of Idaho

NSF Org: AGI DIV Of Advanced Cyberinfrastructure

Initial Amendment Date: September 11, 2013
Latest Amendment Date: September 11, 2013
Award Number: 1341040
Award Instrument: Standard Grant
Program Manager: Kevin L. Thompson
CSE Direct For Computer & Info Sci & Engnr

Start Date: October 1, 2013
End Date: September 30, 2015 (Estimated)
Awarded Amount to Date: $447,969.00

Investigator(s):
- Paul Gessler paul@uidaho.edu (Principal Investigator)
- James Foster (Co-Principal Investigator)
- Daniel Swart (Co-Principal Investigator)
- Lucas Shememan (Co-Principal Investigator)

Sponsor: University of Idaho
Develop our capacity to create new knowledge about ecosystem services, landscape change, and socio-economic systems, and develop our ability to provide science-based decision support to manage for sustainable ecosystem services during urban growth.
*New Capacity in Research Data Interoperability*

**MILES — NKN Cyberinfrastructure plan:**

- Supports data storage, mapping, modeling, and visualization tools,
- Enables sharing of large data collections and HPC access,
- Sponsors interoperability research fostering data accessibility across domains, formats, and spatio-temporal scales.
- Invests in faculty positions and technical support to expand Idaho’s CI-related infrastructure.

*Facilitates integrated modeling across 35 faculty in ecology, economics, physical and social geography, hydrology, public policy, planning, computer sciences, visualization and communications.*
Advancing Data Sharing and Application
Researchers, Students, Citizens, Policy Makers, Managers

* New Virtualization-Visualization Tools communicate results of multi-disciplinary scenario modeling in vulnerable ecosystems and generate new feedback data.

  * promoting intellectual exchange across disciplines;
  * improving communication to communities about Social-Ecological Systems;
  * providing decision makers with data/knowledge for sustainable urban growth.
*Data Management and Services CI*

A new *Virtual Watershed Platform* to link scientific modeling, visualization and data management, for:

* a) streamlined access to local-remote data,
* b) analytic scenario development through interactive visualization,
* c) rapid assimilation of model outputs into DM system for use by other models.
Funding Model Stage One – Establishment

• Collaborative regional and national fiscal partnerships around common interest in data curation and access. Membership Grants™. USDOI-USGS, NSF-EPSCoR, USDA-Forest Service Research.

• Grants from federal agencies to support regional partnerships on critical issues (climate) with significant data management activity.

• Pressure on agencies to fund “directly” in grants, programs, joint infrastructure development.
Seeking a Sustainable Fiscal Model

Northwest Knowledge Network FY2013 through FY2020 Budget Plan

<table>
<thead>
<tr>
<th>Revenue</th>
<th>FY13</th>
<th>FY14</th>
<th>FY15</th>
<th>FY16</th>
<th>FY17</th>
<th>FY18</th>
<th>FY19</th>
<th>FY20</th>
</tr>
</thead>
<tbody>
<tr>
<td>VP Funds</td>
<td>102,413</td>
<td>133,609</td>
<td>242,460</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PI F&amp;A Return</td>
<td>1,728</td>
<td>23,069</td>
<td>26,500</td>
<td>15,000</td>
<td>5,000</td>
<td>15,000</td>
<td>15,000</td>
<td>15,000</td>
</tr>
<tr>
<td>UI Central FA</td>
<td>-</td>
<td>-</td>
<td>100,000</td>
<td>156,000</td>
<td>179,400</td>
<td>206,310</td>
<td>237,257</td>
<td>272,846</td>
</tr>
<tr>
<td>Service Center</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>EPSCoR</td>
<td>39,720</td>
<td>78,727</td>
<td>172,631</td>
<td>164,154</td>
<td>84,284</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>USGS Grants</td>
<td>215,868</td>
<td>301,521</td>
<td>212,861</td>
<td>42,558</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Misc Grants/Dept Funds</td>
<td>61,628</td>
<td>63,915</td>
<td>63,742</td>
<td>7,191</td>
<td>7,188</td>
<td>7,332</td>
<td>7,478</td>
<td>7,628</td>
</tr>
<tr>
<td>New Grant Funds</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>New Equipment Funding</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Revenue Total</td>
<td>421,357</td>
<td>600,841</td>
<td>818,195</td>
<td>384,903</td>
<td>475,872</td>
<td>528,642</td>
<td>559,735</td>
<td>595,473</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Expense</th>
<th>FY13</th>
<th>FY14</th>
<th>FY15</th>
<th>FY16</th>
<th>FY17</th>
<th>FY18</th>
<th>FY19</th>
<th>FY20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Payroll</td>
<td>402,116</td>
<td>563,468</td>
<td>798,734</td>
<td>785,220</td>
<td>788,737</td>
<td>804,328</td>
<td>820,229</td>
<td>836,450</td>
</tr>
<tr>
<td>Operating</td>
<td>19,058</td>
<td>18,861</td>
<td>46,017</td>
<td>38,517</td>
<td>38,517</td>
<td>38,517</td>
<td>38,120</td>
<td>38,517</td>
</tr>
<tr>
<td>Computer Equipment</td>
<td>-</td>
<td>89,203</td>
<td>45,000</td>
<td>40,000</td>
<td>40,000</td>
<td>140,000</td>
<td>180,000</td>
<td>180,000</td>
</tr>
<tr>
<td>Office Furniture/Equip</td>
<td>183</td>
<td>120</td>
<td>20,000</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Expense Total</td>
<td>421,357</td>
<td>671,661</td>
<td>909,751</td>
<td>863,737</td>
<td>867,254</td>
<td>982,845</td>
<td>1,038,349</td>
<td>1,054,967</td>
</tr>
</tbody>
</table>

Net FY Balance     | -     | (70,811) | (91,556) | (478,834) | (394,383) | (454,204) | (478,614) | (459,494) |
Need To Grow

University-Agency Communication, Cooperation and Collaboration
## Federal Agencies Under OSTP Mandate

<table>
<thead>
<tr>
<th>Agency</th>
</tr>
</thead>
<tbody>
<tr>
<td>AHRQ</td>
</tr>
<tr>
<td>ASPR</td>
</tr>
<tr>
<td>CDC</td>
</tr>
<tr>
<td>DHS</td>
</tr>
<tr>
<td>DOC</td>
</tr>
<tr>
<td>DOD</td>
</tr>
<tr>
<td>DOE</td>
</tr>
<tr>
<td>DOI</td>
</tr>
<tr>
<td>DOT</td>
</tr>
<tr>
<td>ED</td>
</tr>
<tr>
<td>EPA</td>
</tr>
<tr>
<td>FDA</td>
</tr>
<tr>
<td>NASA</td>
</tr>
<tr>
<td>NIH</td>
</tr>
<tr>
<td>NIST</td>
</tr>
<tr>
<td>NOAA</td>
</tr>
<tr>
<td>NSF</td>
</tr>
<tr>
<td>ODNI</td>
</tr>
<tr>
<td>SI</td>
</tr>
<tr>
<td>USAID</td>
</tr>
<tr>
<td>USDA</td>
</tr>
<tr>
<td>USGS</td>
</tr>
<tr>
<td>VA</td>
</tr>
</tbody>
</table>
Federal Agency Responses to OSTP Report Mandate

Valen, Dan; Blanchat, Kelly (2015): http://dx.doi.org/10.6084/m9.figshare.1367165

### Overview of OSTP Responses

<table>
<thead>
<tr>
<th>Funder</th>
<th>Policy Coverage</th>
<th>Policy Stipulations</th>
<th>Support Provided</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Published Outputs</td>
<td>Data</td>
<td>Time Limits</td>
</tr>
<tr>
<td>AHRQ</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>ASPR</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>CDC</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>DOD</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>DOE</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>DOT</td>
<td>○</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>FDA</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>NASA</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>NIH</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>NIST</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>NOAA</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>NSF</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>USDA</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>USAID</td>
<td>○</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>USGS</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>VA</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>
February 22\textsuperscript{nd}, 2014 OSTP Memo

- Each agency shall submit its draft plan to OSTP within six months of publication of this memorandum (August 22\textsuperscript{nd}, 2013).
- OSTP, in coordination with the Office of Management and Budget (OMB), will review the draft agency plans and provide guidance to facilitate the development of final plans that are consistent with the objectives of this memorandum and, where possible, compatible with the plans of other Federal agencies subject to this memorandum (guidance provided March 2014).
- Final USDA plan resubmitted November 7\textsuperscript{th}
  - December OSTP approval
  - January NAREEEB Data Management Working Group consultation
  - February 7\textsuperscript{th} Publication on USDA Open Government website
USDA Plan: Digital Data

- Publish Plan
- Listen
- Implement Plan
  - Policy Development
  - Business Processes, Procedures, and Resources Development
  - Information Systems and Automation Development
  - Outreach, Education, and Training Development
Phased Approach

• Phase 1 (Balance of 2015)
  – Listen
  – Form Scientific Output Access Policy Council
  – Inventory and Pilot
• Phase 2 (January 1\textsuperscript{st}, 2016)
  – Pre Implementation
• Phase 3 (January 1\textsuperscript{st}, 2017 & Beyond)
  – Full Implementation & Refinement
Continuum of Data Sharing

- Interoperable set of complete department based RDS systems
- Complete non-interoperable coverage with a highly distributed RDS systems
- Complete & interoperable coverage with a federal-academic-private hybrid RDS systems
- Incomplete coverage with highly distributed RDS systems, some interoperability
- Single perfect federal RDS systems
- Highly distributed RDS systems congealing around academic disciplines
- Spotty RDS systems with little or no coordination
- Few or no research data sharing (RDS) requirements, few or no RDS systems

More Desirable

Less Desirable
Adopt a Distributed & Virtual Solution

Data Publishers

Data Users

USDA Data Repository

- Web Portal/Interface
- USDA Scientific Data Catalog
- APIs
- Data Analytic
What you can expect to see in the rest of FY 2015

• Broader stakeholder feedback opportunities as plans gel and USDA develops policy
• Broad policy framework development by SOAP council
• Pilot NIFA grant programs requesting a two page data management plans
• Rule making process for scholarly publications, likely in collaboration with other federal agencies outside of USDA
• USDA Digital Data Commons emerging at NAL
USDA Digital Data Commons; 5/2014

- ARS investing $3 million in establishment of new DM infrastructure in four of their several regions (e.g., Oakland CA). Minimal university consult.

- National Ag Library setting up Ag Data Commons to be repository for all USDA extramural and intramural data in cooperation with distributed repositories. General trend of feds centralizing data: DOI, DOE, NSF...

- Using LTAR network as pilot test for organizing data (e.g., NKN-WA LTAR becoming node of NAL to host data, assist with data management; not unlike the NSF DataONE model of nodes).

- USDA Ag Data Commons  https://data.nal.usda.gov
NAREEE DM WG Established, 5/2014

• **NAREEE Initial Concerns:**
  
  • *Stakeholder involvement.* Producers, ECOP, ESCOP, universities
  
  • *Cost models.* Data management services require financing.
  
  • *Responsibility for compliance.* Institution or an individual researcher.
  
  • *Privacy.* Handling of personal identifiable information.

**NAREEE Action:** The Board decided to form a working group to discuss the aforementioned issues surrounding the Big and Open Data Initiatives and to develop recommendations for the USDA.
1. Realistic Mandates and Business Plans.

2. Focus. a) storage, curation, quality, access, interoperability of data, b) curating publications/data; c) public access, d) proprietary/intellectual aspects of data affecting industries, universities.


4. Intra-Agency Collaboration. Maximize collaborative planning and implementation of data management activity on major issues.

5. Stakeholder Interaction and Consultation. reach out discover solutions and innovations in universities, industry, other agencies.

6. Inter-Agency Collaboration.

7. Scales. importance of local, state, regional, national and global scales.
NAREEEE WG USDA Briefing; DC 9/2014

• Office of Undersecretary; Dr. Cathy Woteki, Undersecretary

• Office of Chief Scientist; Open Data/GODAN; Dr. Jaime Adams, Senior Advisor

• Agricultural Research Service, ARS; Dr. Steven Shafer, Associate Administrator; Paul Gibson, Chief Information Officer; Dr. Doreen Ware, Scientific Information Officer; Tom Houston, Information Technology Specialist

• National Institute of Food and Agriculture, NIFA; Dr. William Hoffman, C of S

• Economic Research Service, ERS; Dr. Mitch Morehart, Farm Income Team Ldr

• Office of the Chief Information Officer, USDA; Joyce Hunter, Deputy CIO, Policy and Planning; Bobby Jones, Sr. Advisor to the Deputy CIO; Acting Chief Data Officer; Ray Payton, Acting Assoc CIO, Policy and Directives Chief Privacy Officer

• National Agricultural Statistics Service NASS; Christy Meyer, Head Census Data Section Donald Buysse, Head Census Section
Share Planning Information:

• Provide NAREEE with USDA’s Open Data, Big Data and data management planning documents at the earliest possible time, including both departmental and agency level plans and reports, those internal to USDA and those submitted to OMB and OSTP.

• Provide briefings to the NDMWG on the activities and plans of the US Open Data Council and other interagency and inter-departmental groups focused on Open Data, Big Data and data management.
Engage Stakeholders:

- Sponsor additional stakeholder engagement on the topics of Open Data, Big Data and data management, especially with regulatory agencies, and agricultural producers, support industries, and consultants.

- Launch a USDA REEE Data Management website to communicate with universities and other stakeholders about USDA initiatives and plans. Include information about the value, best practices and benefits of actively managing data, feature USDA initiatives in Big Data, Open Data and data management, and encourage collaboration between government and universities on the build out of data management infrastructure.

- Develop a fact sheet defining Open Data and Big Data, the requirements for complying with mandates for each, and a glossary of terms and definitions relating to each. This sheet can be posted on the USDA Data Management website.
Expand Inter-Department and Inter-Agency Collaboration

• Encourage all USDA agencies to consult with the ERS on best practices for data management. ERS has a long history of developing, collecting, curating, serving and publishing data with an emphasis on stakeholder engagement and service, and has refined protocols for how to train, certify and allow access to data with potentially private or sensitive information.

• Actively expand interagency, inter-departmental and university-agency data management collaboration on topics where data-intensive science and education can have an immediate, positive effect on major, complex issues of the day (e.g., climate change, food security, and genomics). Provide NAREEE and NDMWG with briefings on collaborative efforts.
Initiate Collaborations with Universities

- Encourage ARS leadership, at both regional and national levels, to consult with and conduct joint planning and implementation exercises with land grant universities on the provision of data management infrastructure and policies throughout the nation. Data management infrastructure will be needed to support major research facilities and activities between universities and agencies throughout the nation and beyond.

- Encourage REEE agencies to work with universities on the planning and implementation of symposia where all can share their plans for the build-out of data management infrastructure, and identify areas for cooperation, collaboration and complementarity. Consider an annual conference and regional symposia.
• NIFA to formalize data management requirements and support in all appropriate RFAs/RFPs, including: a) requirement of data management plans, b) requirement of metrics for evaluation of data management efforts and processes for accountability; and c) instructions to include data management expenditures in direct costs.

• Work with universities to eventually include capacity funded programs (Hatch, Smith-Lever, McIntire-Stennis, Evans-Allen, Animal Health, Renewable Resources (RREA), 1890, and Tribal) under the umbrella of Open Data. Include Experiment Station and Cooperative Extension directors in planning conversations.

• Advise universities on the development of curricula for the preparation of a next generation workforce with both domain science and software development expertise.

• Convene conversation with OSTP, OMB, U Leaders to pursue new grant administration policy (authorization language) enable federal funding of infrastructure development.
DM Plans Require:

- types of data, samples, physical collections, software, curriculum materials to be produced in the project;
- standards used for data and metadata format and content;
- policies for access and sharing including provisions for protection of privacy, confidentiality, security, intellectual property, or other rights;
- Policies-provisions for re-use, re-distribution, production of derivatives;
- plans for archiving data, samples, and other research products, and for preservation of access to them.
NSF Big Data Innovation Hubs


• To augment OSTP Big Data Initiative and ignite public-private partnerships, NSF CISE establishing National Network of *Big Data Regional Innovation Hubs*.

• Four geographic Hubs NE, M-West, South, and West. Consortia of academia, industry, government. Support the breadth of interested local stakeholders, and achieve common Big Data goals not possible for independent members.

• Multi-phase process to develop a National Network of BD Hubs. 1) sets up governance structure to ensure cross-hub collaboration and sustainability over long term. 2) focuses on building out various sectors of particular interest to each BD Hub (transportation, smart cities, health, energy, public safety, education). 3) focuses on connecting BD Hubs and regional sectors as US innovation ecosystem.

• Charrettes held in each region Spring 2015. HUBzero community portal at [http://bdhub.info](http://bdhub.info) to communicate with members in/between their region.
Collaborative Strategies for Env Sci Data

• How to meet long term needs for environmental research data management by examining the original goals and objectives of single repositories, identifying efficiencies of scale, and assessing cost-benefit and sustainability of centralized/coordinated services and collaborations.

• Drawing on collective experience of data curators from environmental research fields, data aggregators, tool developers, computer scientists and environmental scientists in managing data and repositories. Seeking collaborative, cost-effective strategies for managing infrastructure, defining standards, and developing tools.

Case Study in Regional Data Management

• Startup venture; partner (member) and institutional funding.

• Learned critical level of staffing hardware, software to sustain core services.

• Demand for services quickly exceeded capacity. Ventures ending.

• Established Service Center.

• Need to keep sharing stories about our respective efforts at U’s.

• Need dependable revenue flow for data services, and more sophisticated, coordinated, strategic partnership between universities and federal government, on behalf of the PI’s (a triangle of value propositions).
Seek a Sustainable Business Model via University-Agency Cooperation

Need activity-interaction on all three sides of a the value triangle; Federal agencies, PIs and universities must relate to each other.

• Agency require PI’s to do DM planning;

• Agencies/Universities require PI’s to dedicate direct costs for DM;

• Universities provide PI’s with essential DM services;

• Universities/agencies convene national workshop on joint sustainable data management; common priorities, policies, protocols, complementary infrastructure. Shared strategy, cost.

• Engage Congress in authorization conversations.
Thank you!!

stevendl@uidaho.edu (USDA, USDI, NKN);
paulg@uidaho.edu (NSF, NASA, NOAA, NKN)
jmciver@uidaho.edu (Institutional Leadership)
IRON Network 2012

- Core Nodes in Boise, Pullman and IDF
- Aggregation in Spokane, SLC, Clarkston and Lewiston.
- 1GE/10GE Backbone and Access connections
- Connectivity to neighboring RONs (PNWGP, UEN)
Facets/Elements of the Data Mgmt

• Supporting research faculty and their data/science
• University reputation and attractiveness
• Workforce development, graduate and undergraduate
• Internal and external partnerships
• Triangle of federal executive/legislative, researchers, universities
In 2013 the White House Office of Science & Technology Policy (OSTP) released a memo directing certain federal agencies (those with more than $100M in R&D expenditures) to "develop plans to make the published results of federally funded research freely available to the public within one year of publication and requiring researchers to better account for and manage the digital data resulting from federally funded scientific research." See full memo titled "Increasing Access to the Results of Federally Funded Scientific Research" here (PDF). Many funding agencies now require the inclusion of a data management plan in grant proposal submissions.
Future

• Continue investments to build-out IRON, NKN and similar models across the state – formalized governance structure

• Adapt and evolve our state CI for supporting science while making all research proposals more competitive as a result of our strategic investments

Broader Impacts

• All Idaho researchers and institutions become more competitive

• New science is facilitated as a result of interoperability initiatives that enable new query and exploratory analysis to do cross- and multidisciplinary discovery across space-time scales with disciplinary data in separate repositories

• Truly support outreach to industry and stakeholder communities in Idaho and the region, improved tech-transfer economic development
National LambdaRail

- NLR provides the national backbone
- An NLR member or associate organization coordinates the regional component
- An individual institution coordinates the local connection

NLR offers a variety of connectivity options to meet your high-performance network requirements:

- **PacketNet**: NLR’s shared routed service connects educational institutions with one another, as well as with the commodity Internet and with major commercial sites.
- **FrameNet**: provides users with shared or dedicated point-to-point or multipoint transport at multi-megabit, 1- and 10-gigabit Ethernet (GE) data rates for high-volume data flows.
- **WaveNet**: provides users with dedicated point-to-point, high-capacity 10G LAN-PHY or OC-192 network connections, enabling users to control their own end-to-end pathways.
Three Separate But Related Issues

• Big Data
  – Industry Sole Proprietorships, Corporate, Academic, Government

• Open (Government) Data
  – May 2013 Executive Order
  – Federal Government Data

• Open Access to Federally Funded Research Outputs
  – February 22\textsuperscript{nd} OSTP Memo
  – Access to Federally Funded Data and Scholarly Publications
Open Access to Federally Funded Research Outputs

- February 22\textsuperscript{nd}, 2014 OSTP Memo
- Activities To Date
  - Authorship of the Plan
  - OSTP Plan Approval
- Need to Develop
  - Policy
  - Business Processes, Procedures and Resources
  - Information Systems & Automation
  - Outreach, Education & Training
- Time to Listen Again
February 22\textsuperscript{nd}, 2013 OSTP Memo

- The Office of Science and Technology Policy (OSTP) hereby directs each Federal agency with over $100 million in annual conduct of research and development expenditures to develop a plan to support increased public access to the results of research funded by the Federal Government.

- Each agency plan must be consistent with the objectives set out in this memorandum. These objectives were developed with input from the National Science and Technology Council and public consultation in compliance with the America COMPETES Reauthorization Act of 2010 (P.L. 111-358).
OSTP Memo - Availability of Research Datasets

- Digitally formatted scientific data resulting from unclassified research supported wholly or in part by Federal funding should be stored and publicly accessible to search, retrieve, and analyze.
- For purposes of this memorandum, data is defined, consistent with OMB circular A-110, as the digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings including data sets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as laboratory specimens.
USDA Plan: Scholarly Publications & Digital Data

- Publish Plan
- Listen
- Implement Plan
  - Policy Development
  - Business Processes, Procedures, and Resources Development
  - Information Systems and Automation Development
  - Outreach, Education, and Training Development
The Case of Capacity Programs

Should Open Data mandate apply to Hatch, Smith-Lever, McIntire-Stennis, Evans-Allen, Animal Health, Renewable Resources (RREA), 1890, and Tribal?

A. $.5 billion in applied, regional and demonstration research programs and their data may be important;

B. Could be cumbersome, questionably effective and time consuming for data to be organized and called for from this community.

Capacity leaders need to provide input on whether to be included, and if so, how would they help design an approach that will work.
Extension Delivery and Application

• Help producers, managers and policy makers with the application of data to scenario building, modeling, visualization....

• Pursue cooperative arrangements between industry, producers and universities on the collection, storage, access, use of data.

• Pursue RFPs with integrated Extension and Research in the data-intensive context.
USDA NAREEE Big/Open Data and Science

- USDA provide glossary of terms, more definition(s) about what is required, preferred
- USDA provide basics on the value, best practices, benefits of managing Open/Big Data
- USDA gather input from universities re: their capacity for providing Open/Big data services
- USDA engage Capacity programs as a special case; get input from leaders
- USDA incentivize researcher for data preparation (offering scrubbing and other services)
- USDA provide guidance to universities on how Open and Big Data mandates will be enforced
- ARS conduct joint planning exercises with land grant universities leading data management
- USDA RFPs explicitly require data management activity and hold accountable
- USDA RFPs instruct PIs to include data management expenses in direct costs
- USDA work with smaller/medium sized universities to minimize negative economies of scale
• Collaborative regional data partnerships (NIFA USDA, NW CSC USDol, INL USDoE, EPSCoR NSF, NW Climate Hub USDA, NW Forest Fire Science Center and Sustainable NW Dairies Center.

• Network of resources, services, and expertise
  • Policies, protocols, standards in support of effective data/metadata;
  • Systems admin, software development for data-intensive science;
  • Stable and enduring storage and access to data and metadata;
  • Hosting of virtual machines, applications, websites databases; and
  • Consulting/technical services for data and metadata management.
  • NSF DataONE, access to HPC and national high-speed data networks
Connecting real and synthetic data from user-centered visualization techniques, sensing networks, modeling and computer-enabled processing to create Virtual Watersheds and simulations to improve understanding of watershed dynamics.
Creating a Virtual Watershed Framework to formulate and integrate watershed models, including simulation of watershed dynamics, easy acquisition, integration and visualization of data, and identification of consequences of hydrologic change.
The Next Phase
The Online Data Observatory

- Greater than the sum of its parts.

- Enable investigators to visualize and inter-compare heterogeneous datasets without struggling with file formats, unit conversion, sub-setting, scales.

- New research with existing data.

- Important Components
  - Data representation/interoperability
  - New tools
  - Web service APIs