The Federal Big Data Initiative: Where it has been and where it is going

Brand Niemann
Director and Senior Data Scientist
Semantic Community
Fairfax, Virginia, United States of America
bniemann@cox.net

***Abstract*— *Abstract*—Since the White House announced the Big Data Initiative in 2010, there have been a series of activities for government agencies, academia, and industry to participate in to develop data scientists, perform research, and to develop applications, which this presentation will summarize.**

The work of the Federal Big Data Senior Steering Work Group, the NSF Big Data Funding Opportunities, and the Federal Big Data Working Group Meetup will be described and specific examples will be shown.

The roles of the Presidential Digital Government Strategy and Open Data / Open Government Policy, the new Congressional Data Act, and the Open Research Data Policy will be described and specific examples of their implementation will be given.

One should be able to see where they might participate in the Federal Big Data Initiative as a result of reading this paper.

Keywords—White House Big Data Initiative, Federal Big Data Senior Steering Work Group, NSF Big Data Funding Opportunities, Federal Big Data Working Group Meetup

I. BACKGROUND

Since the White House announced the Big Data Initiative in 2012, there have been a series of activities for government agencies, academia, and industry to participate in to develop data scientists, perform research, and to develop applications. The work of the Federal Big Data Senior Steering Work Group, the NSF Funding for Big Data and Data Science, and the Federal Big Data Working Group Meetup have all been important in carrying out these activities which are summarized in: TABLE 1: The Federal Big Data Initiative: Where it has been and where it is going, and in TABLE II. Federal Big Data Working Group Meetup: Data Science Data Publications in Data Browsers. Please note that because these are large linked data tables, they are in single column format at the end of the paper.

The context for this paper is shown in TABLE I. The federal government has been moving towards data publications in data browsers with the help of Semantic Community since 2009. Our purpose was and still is to support The Presidential Digital Government Strategy and Open Data / Open Government Policy, the new Congressional Data Act, and the Open Research Data Policy, because all essentially require Data Science Data Publications in Data Browsers.

In TABLE I., the so-called Prelude was January 2009 to the White House Announcement in March 29, 2012, and the Semantic Community Semantic Data Science Team started to work on Semantic Medline.

The Federal Digital Government Strategy has been interpreted as "treating all content as data", so big data = all your content. Thus even a small government agency or organization has Big Data if they utilize all of their content. Semantic Community says: "We make Big Data Small" using Semantics & Advanced Analytics.

TABLES I. and II. also show the Semantic Community Data Management Plan and commitment to the community-at-large to publicly preserve the reports, documents, meeting proceedings, data stories, data sets, and metadata for reuse.

II. SEMANTIC COMMUNITY

The Semantic Community Semantic Data Science Team pioneered a government big data application for the Federal Big Data Senior Steering Work Group called Semantic Medline on the YarcData Graph Appliance in which a massive medical publication data base (PubMed) was converted to a Semantic Web Graph Data Format (RDF) consisting of about 25 billion triples whose complex graph relationships are

instantaneously visualized for discovery of diseases and treatments by medical scientists and researchers. For more details see: Finding a Needle in a Digital haystack The Opinion Pages, Gartner on YarcData Urika, MEDLINE Solutions Brief, and Urika Product Brief.

Now the challenge is to apply this successful combination of collaboration and technology to other scientific subject matter and organizations.

The next opportunity for our Semantic Community Semantic Data Science Team was with the CODATA International Society for Digital Earth (ISDE) Workshop on Big Data for International Scientific Programmes: Challenges and Opportunities, June 8-9, Beijing, China. In preparation for this workshop presentation and tutorial, we prepared Big Data Science for CODATA and Digital Earth: Big Earth Data and Geospatial Analytics by mining their two principal scientific journals (Data Science and Digital Earth) to make both the journals and individual publications with data, scientific data publications in data browsers. It is obvious from the long list of science journals and their organizations that this process can be replicated many times and all the data results converted to the Semantic Web Graph Data Format (RDF) and visualized in the YarcData Graph Appliance for discovery of relationships between the individual disciplines and organizations. This is in fact the objective of the second year of the Big Data Initiative: to foster collaborations between the individual initiatives!

The Semantic Community Semantic Data Science Team also needed to expand to include those with scientific data publishing experience because NSF Assistant Director Farnam Jahanian said recently: "Implementation plans for public access (to scientific research data) could vary by discipline, and new business models for universities, libraries, publishers, and scholarly and professional societies could emerge." Our team needs to broker access to scientific research data publications like the Elsevier Research Data Services so that it creates a win-win for both the scientific community and the publisher.

III. THE FEDERAL BIG DATA WORKING GROUP MEETUP

The Federal Big Data Working Group Meetup is a broad community of participants focused on big data products for the Federal Big Data Initiative.

Our mission statement is as follows:

- Federal: Supports the Federal Big Data Initiative, but not endorsed by the Federal Government or its Agencies;
- Big Data: Supports the Federal Digital Government Strategy which is "treating all content as data", so big data = all your content;
- Working Group: Data Science Teams composed of Federal Government and Non-Federal Government experts producing big data products; and
- Meetup: The world's largest network of local groups to revitalize local community and help people around the world self-organize like MOOCs (Massive Open On-line Classes) being considered by the White House.

Our Framework is as follows:

- Leadership of the Semantic Data Science Team that produced Semantic Medline running on the Yarc Data Graph Appliance.
- Organize a Community of Data Scientists and Related Fields to focus on treating all of your content as "Big Data" by founding and co-organizing of the Federal Big Data Working Group Meetup.
- A graduate class prepared for GMU entitled "Practical Data Science for Data Scientists".
- Follow the Cross Industry Standard Process for Data Mining (CRISP-DM; Shearer, 2000) to build a Data Science Knowledge Base
- Mine prominent scientific journals for data policy, data bases, and data results that can be reused like Data Science and Digital Earth scientific journals for the CODATA International Workshop on Big Data for International Scientific Programmes, (June 8-9, in Beijing).
- Participation in the Data FAIRport (Findable, Accessible, Interoperable, and Reusable) with "Data Publication in Data Browsers".

- Obtain NSF funding for sustained data science for data publications work over a period of years
- Providing data stories that persuade and presentation materials for public education conferences like the COM.BigData Conference (August 4-6, in Washington, DC).

Our Meetup presentations focus on answering four essential questions:

- How was the data collected?
- Where is the data stored?
- What are the data results?
- Does the data story persuade?

Examples of the answers to these questions are given in the examples in the next section.

All are welcome to participate in our Meetups and learn big data science from tutorials and be mentored in their university and professional work and proposal writing.

During the past six months the Federal Big Data Working Group Meetup has focused on Federating Uses Cases, Data Publications, and Solutions & Technologies in the Meetups shown in TABLE II.

IV. Examples for Ten Senior Government People

Data Publications in Data Browsers have been created for at least the ten senior government officials as shown in TABLE III. as a way of educating and motivating them to task their staff and contractors to start doing the same.

Again please note that because these are large linked data tables, they are in single column format at the end of the paper.

V. Example of a Data Publication in a Data Browser

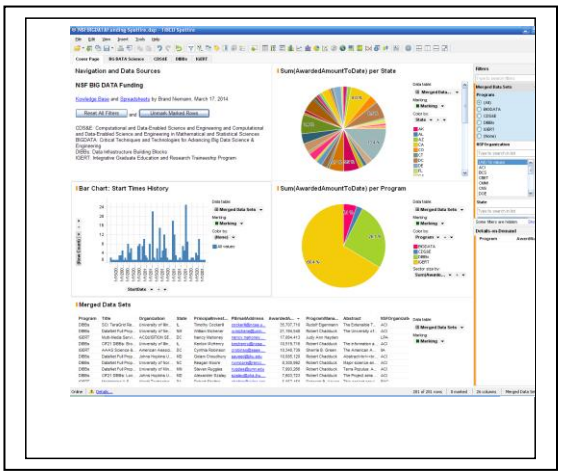
The NSF Grant Proposal Guide PDF file was converted to wiki format with structure and the grant proposal paper format was added with links to the appropriate sections of the Guide.

The index of the wiki page was captured to a spreadsheet as linked data in both graph and relational table formats for input to a Spotfire Dashboard along with 4 other spreadsheets of NSF Grants Awards data table for display in 4 tabs (Cover Page, BIG DATA Science, CDS&E, DIBBs, and IGERT) shown in Fig. 1.

The Cover Page contains the Merged Data Set in an overview interactive display where one can selected one or more Filters to the right and drill down and then select a graph element or row of data to see the details-on-demand.

Those looking to know more about the NSF BIG DATA Initiative and related programs can search these spreadsheets and Spotfire dashboards to identify projects they might want to partner with in their completions and/or new projects they might want to propose that do not duplicate existing projects.

Fig. 1 Spotfire Dashboard of NSF BIG DATA Funding

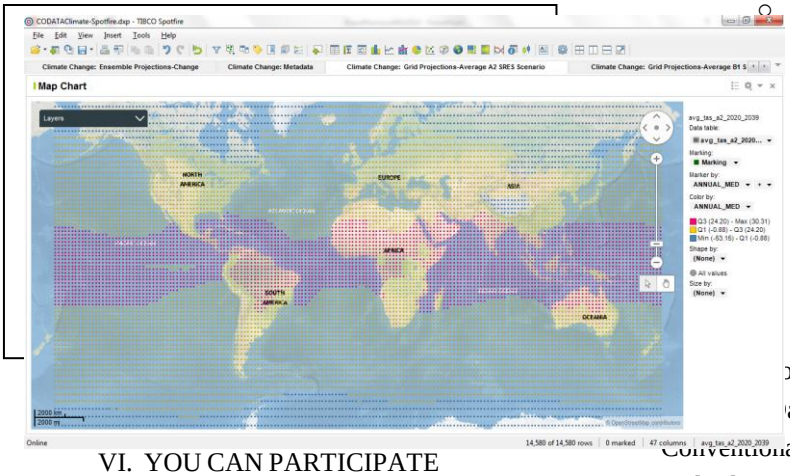


The entire (publically available) International Journal of Digital Earth has been copied into a wiki and structured so a data publication index can be built in a spreadsheet that can be used in Spotfire for content analytics and publication analytics along with Spotfire data analytics.

One significant highlight is the Fig. 2. Spotfire visualization of the Climate Change: Grid Projections - Average A2 SRES Scenario superimposed on the global geospatial infrastructure, which is still a work in progress to measure all the parameters of interest on a grid like this, or at enough locations to be confidently interpolated to such a grid.

Please note that the Spotfire Dashboards shown are available as both client-based and web-based data browsers.

Fig. 2. Spotfire Dashboard of Climate Change: Grid Projections-Average A2 SRES Scenario



VI. YOU CAN PARTICIPATE

You can participate in the following ways:

- Federal Big Data Senior Steering Work Group:
    - The Big Data Senior Steering Group (BD SSG) works to facilitate and further the goals of the White House Big Data R&D Initiative. The BD SSG strategic priorities include: Core technologies, Big data infrastructure, Workforce development, and Competitions and challenges.
        - Primarily government with some non-government invited presentations like Semantic Community.
- Faster Administration of Science and Technology Education and Research (FASTER):
    - FASTER's goal is to enhance collaboration and accelerate agencies' adoption of advanced IT capabilities developed by Government-sponsored IT research. FASTER hosts Expedition and Emerging Technology workshops as well as monthly meetings with invited guest speakers to achieve this goal.
        - **Open to public. Get on email list**.
- NSF Funding for Big Data and Data Science:
    - Recent Program Solicitation: Critical Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA).
        - The Semantic Community Data Science Team submitted a proposal you can see.
- Federal Big Data Working Group Meetup:
    - Mission Statement, What We Are Doing, and How We Are Doing It.
        - About 280 members now (government and non-government) with diverse employment and interests. **Open to public, just become a member at the web site.**

VII. SUMMARY AND CONCLUSIONS

old purpose of this paper (NSF BIG DATA Grant ata Publication in a Data Browser, and nal Paper Publication) has been presented and multiple examples of Data Publications in Data Browsers have been provided in three linked data tables (TABLES I, II, & -III.)

So the way forward with the Federal Big Data Initiative now is

- Implementing existing techniques and technologies
- Applying existing techniques (e.g., machine learning, statistical analysis) to specific data sets
- Developing databases to serve specific scientific communities using existing database techniques

Why? Because the rest of the world is racing ahead with all of this and doing the supporting research along the way with venture capital investment because the government research funding process has become too slow.

ACKNOWLEDGMENT

The author acknowledges with gratitude the influence of Dr. George Strawn, Director, National Coordination Office/NITRD (USA), for his leadership and wisdom during my years of federal service and since. He also acknowledges the many supervisors and colleagues that have supported his work, especially current members (about 290) of the Federal Big Data Working Group Meetup.

REFERENCES

[1] This paper uses web links

TABLE 1. The Federal Big Data Initiative: Where it has been and where it is going

Data Source: Excel Federal Big Data Initiative Tab

Date	Event	Comments
January 2009	Report of the Interagency Working Group on Digital Data to the National Science and Technology Council	Harnessing the Power of Digital Data for Science and Society
April 26, 2010	"OSTP in the Open" (R&D dashboard)	OSTP Open Government Plan
June 29 – July 1, 2010 (Published on March 31, 2011)	Scientific Data Management (SDM) for Government Agencies: Report from the Workshop to Improve SDM	Harnessing The Power Of Digital Data: Taking the Next Step
December 2010	The President's Council of Advisors on Science and Technology Report on Designing a Digital Future	Crosscutting Themes (Interoperable Interfaces). My Note: See Spotfire Dashboards below!
Early 2011	The Big Data Senior Steering Group (BD SSG) formed.	The Big Data Senior Steering Group (BD SSG) works to facilitate and further the goals of the White House Big Data R&D Initiative. The BD SSG strategic priorities include: Core technologies, Big data infrastructure, Workforce development, and Competitions and challenges.
March 21-22, 2011	Open Government Research & Development Summit, March 21-22, 2011	NITRD Dashboards in Spotfire (3)
March 29, 2012	Obama Administration's $200 million "National Big Data Research and Development Initiative"	The Big Data Initiative launch featured more than $200 million in new commitments from six Federal departments and agencies aiming to make the most of the explosion of Big Data and the tools needed to analyze it.
April 2012	Semantic Data Science Team Started to Work on Semantic Medline	"Both language and human thought are large, for feasibility we need to scale down the complexity of the process of semantic interpretation." Thomas C. Rindflesch, Ph.D., Lister Hill National Center for Biomedical Communication
April 4, 2012	Semantic Search (and Data Science Dashboards) for NSF Decision Making	Research.gov Dashboards in Spotfire
January	Presentation to BDSSG: Semantic	About a year ago, Dr. George Strawn challenged me to

24, 2013	Medline and Government Challenges with Big Data	pilot a new partnership to make NIH's Semantic Medline "the killer semantic web app for the government" and the keystone of workforce development for future data scientists.
May 3, 2013	White House Big Data Partners Workshop	The first workshop brought together representatives from industry, academia, and government to learn about existing BD partnerships, make connections with interested parties, and explore future possibilities.
May 29, 2013	Big Data Senior Steering Group (BDSSG) Workshop: Data Sharing and Metadata Curation: Obstacles and Strategies	Future strategies for managing scientific data and metadata for basic and applied research
June 20, 2013	Request for Two-Page Summary of Big Data Projects	Semantic Data Science Team Submission: Making the Most of Big Data
July 1, 2013	Free Online Version: Frontiers in Massive Data Analysis	Data Publication: Frontiers in Massive Data Analysis
September 10-11, 2013	AFEI Cloud: SOA, Semantics, and Data Science (15th SOA for eGov Conference)	First Live Semantic Medline-YarcData Graph Appliance Demos: YarcData Videos Demos: Schizo- 7 minutes, Cancer-21 minutes
November 12, 2013	Data to Knowledge to Action Event Takes Big Data Initiatives to Innovative Heights	Semantic Data Science Team Attends White House Big Data Event
January 7, 2014	Federal Big Data Working Group Meetup Kickoff My Note: See Table Below for More Detail	Semantic Big Data Science Application: Semantic Medline on the YarcData Graph Appliance for the Federal Big Data Senior Steering Work Group
March 3, 2014	White House Blog Post: "Privacy Workshop to Explore "Big Data" Opportunities, Challenges"	Big Data Privacy Workshop
March 4, 2014	Federal Big Data Working Group Meetup: Number 4	Joint NSF-NIH Biomedical Big Data Research: Euretos BRAIN
June 9, 2014	Critical Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA)	See NSF Funding Opportunities in Data Science

TABLE II. FEDERAL BIG DATA WORKING GROUP MEETUP

Data Source Excel Federal Big Data Working Group Meetup Tab

MindTouch	**Meetup.com**	**Story (s)**
Kick-off Meetup: Tuesday, January 7, 6:30 p.m.	Let's have our Kickoff in early January 2014.	Tutorials Start: Practical Data Science for Data Scientists and Semantic Big Data Science Application: Semantic Medline on the YarcData Graph Appliance for the Federal Big Data Senior Steering Work Group
Second Meetup: Tuesday, February 4, 6:30 p.m.	Second Meetup: Tuesday, February 4, 6:30 p.m.	Healthcare.gov Data Science and Be Informed Prototype Video
Third Meetup: Tuesday, February 18, 6:30 p.m.	Evolution of Semantic Technologies-The Value of Merging Smart Data With Big Data	Modus Operandi Semantic Knowledge Base
Fourth Meetup: Tuesday, March 4, 6:30 p.m.	Joint NSF-NIH Biomedical Big Data Research	NIST Data Science Symposium, Euretos BRAIN, and Data Culture at the NIH
Fifth Meetup: Tuesday March 18, 2014, 6:30 p.m.	Continue Data Science Tutorial and Learn About Bigdata SYSTAP	Bigdata SYSTAP Literature Survey of Graph Databases and Graph Databases
Sixth Meetup, Tuesday April 1, 2014, 6:30 p.m.	Marc Smith, Network Analytics, and Kate Goodier on Big Data Privacy	Data Science for VIVO, NodeXL and Sci2 for Data Science and Big Data Privacy Workshop
Seventh Meetup: Tuesday, April 15, 6:30 p.m.	Kate Goodier, Cognitive Metadata, and Cambridge Semantics, Insider Trading	Data Science for FIBO
Eight Meetup: Tuesday, May 6, 6:30 p.m.	EPA/NASA Climate-Environmental Data Analytics & A Redesigned, Open	Data Science for EPA Air Data, Chesapeake Bay Program, andNASA Big Data

	Data.gov	
Ninth Meetup: Tuesday, May 20, 6:30 p.m.	Data Science at GMU and Elsevier Research Data Services	A Data Science Big Mechanism for DARPA and Data Science for Climate Change Impacts
Tenth Meetup: Monday, June 2, 6:30 p.m.	Ontology Summit 2014 Postmortem and Reading & Reasoning with Semantic Insights for the DARPA Big Mechanism	Ontology for Big Data, Big Data Science for CODATA and Semantic Insights
Eleventh Meetup: Monday, June 30, 6:30 p.m.	MIT Big Data Initiative: Sam Madden, & Current Elephants: Michael Stonebraker	MIT Big Data Initiative: bigdata@CAIL and the new Intel Science and Technology Center for Big Data, Sam Madden and Why the current "elephants" are good at nothing, Data Tamer, and data integration issues, Michael Stonebraker Workshops on Extremely Large Databases
Twelveth Meetup: Monday, July 7, 6:30 p.m.	Data Science of White House Big Data Review and Brooke Aker: Big Data Lens on OpenFDA	Mary Galvin, AIC, HPCC Systems Academic Program and the Georgetown University McCourt School of Public Policy's Massive Data Institute, Katherine Goodier, Excelerate Solutions, Legislative Data and Transparency Conference and Chuck Rehberg: SIRA Part II, and Brooke Aker. Big Data Lens A Look at OpenFDA API and Big Data Design(s) Based on It.
Thirteenth Meetup: Monday, August 4, 9:00 a.m.	COM.BigData 2014: The 1st International Summit on Big Data Computing	Keynote and Panel

TABLE III. EXAMPLES OF DATA PUBLICATIONS IN DATA BROWSERS FOR SENIOR GOVERNMENT PEOPLE

DATA SOURCE: SEMANTIC COMMUNITY NSF BIG DATA PROPOSAL

Person	Interest	Data Publication in Data Browser	Example
Dr. John Holdren	Climate Change	Data Publication in Data Browser	Climate Change Assessment
Dr. George Strawn	Research Objects as Digital Objects	Data Publication in Data Browser	VIVO
Dr. Farnam Jahanian	NSF Big Data Publications	Data Publication in Data Browser	NSF Big Data
Dr. Phil Bourne	Data Culture at NIH	Data Publication in Data Browser	Bourne Research & NIH
Dan Kaufman and Paul Cohen	Big Mechanism for Cancer	Data Publication in Data Browser	DARPA Contract
Bryan Sivak	Hack-a-Thon	Data Publication in Data Browser	HHS IDEALAB
Todd Park	Code-a-Palooza	Data Publication in Data Browser	Health Datapalooza V
Brian Lee	Health United States 2013	Data Publication in Data Browser	Centers for Disease Control & Prevention Report
The Honorable Kathleen Sebelius	Dynamic Case Management	Data Publication in Data Browser	HealthCare.gov Web Site