Towards a shared vision for success and pilot projects for the National Data Service

> **Note:** *These session notes are from the morning and afternoon segments of the June 13, 2014 NDS workshop.*

Introductory Comments (Ed Seidel):

- Importance of collaboration across complementary initiatives
- Recognition that early pilot projects can serve as building blocks – provided they are well connected
- Think of the work at this stage of NDS as a matrix, with types of services on one dimension and communities being served on the other – pilot projects can reside in the various cells of the matrix
- It is important to engage the builders and it is important to be aligned with the communities

Initial framework for potential NDS pilot initiatives:

Services Stakeholder Pilots	**Data Collection**	**Publish**	**Link**	**Store**	**Discover**	???
Materials Genome						
Astronomy						
Humanities						
Civil Engineering						
Geoscience						
Biomedicine						
???						

Part I: If the NDS is fantastically successful and it is the year 2025 and we are celebrating success, what are we celebrating having accomplished?

- A long-tail repository where any discipline can add data and metadata with cultural context, and with security and privacy protections consistent with the initial input
- Dramatically accelerating research discovery
- Create a single place to go that is well known for service and resources for data – it may not all be provided but it is all visible and discoverable
- Enhanced data reuse
- If NDS is fantastically successful and it is the year 2025, what accomplishments would be celebrate?
    - All scientists know where to put their data and how to document it
        - It's as simple as creating a web page to document, publish, and preserve data
        - Data sharing is the expected, normal outcome of research
    - There is a national infrastructure for long-term preservation and access
        - Business model will be defined, accepted, and in place
        - The long-tail will no longer be a concern
        - Access through programming language/API of choice
    - NDS services are seen as essential for research, a routine part of doing research
- All scientists know where to put their data and to document it; it is as simple. .. a national infrastructure for long term access and storage . . . NDS services are essential for research and a routine part of doing research

- A culture change – it is clear that data are a critical part of the scientific process and the NDS is the vehicle for sharing, publishing, discovery, etc. – well understood, simple, part of the culture, every child uses it
- Have science driven science – informed by data – data is by default preserved, accessible, and used – the conversation is about exploiting the data better, not issues of storage, discovery, etc.
- Data is a first-class citizen across all fields and disciplines – just like publications – data has at least the same importance as HPC – data management practices are taught – a first-class discipline so that data are shared, discovered, reused through a system of connected repositories
- Data management is taken for granted – funding agencies fund data management plans as a matter of course – parallel to the need in the past to talk about exchange data, but that is taken for granted now – ubiquitous

Additional Comments:

- This is about exploiting the data better, not about storage/preservation, but we do need to address the initial infrastructure issues around storage/preservation
    - People are more focused on the "sexy stuff"
- Caution around what types of science discoveries will be enabled by NDS – hard to predict and important to manage expectations
    - Consider focusing more on the efficiency in the research process
- Talking about a place to put data and metadata in, and to get it out – that is a one-to-two year goal – get the infrastructure in
- We are talking a lot about the long-tail, but don't exclude the established repositories – these need to connect to the NDS as well
- NDS should think seriously about quantifiable measurable impact
- Don't forget about the existing array of campus infrastructures
- Making it easier for researchers to trust the data – provenance, intended use, etc.
- Be careful to not over promise – work from concrete use cases – builds interest
- Major data initiatives are complementary, collaborative, and sustainable
- What we put in place can't just be a demo – it has to be persistent cyberinfrastructure
- NDS influence provost office and tenure process so that data is recognized and tensure takes this into account

Part II: Focusing on potential next step action priorities, what would be an actionable pilot project that would be worth some level of discretionary effort on your part – either in a functional domain or in a disciplinary domain?

Potential Functional Domains:

1. Creating Data Collection (structure and unstructured data)
2. Publishing
3. Linkage
4. Access
5. Archiving, Preservation, Curation, Stewardship, Storage
6. Data Discovery, Notification
7. Analysis, Use, Visualization, Computation, Modeling, Analytics
8. Feedback, Credit, Impact
9. Coordination, Federation, Alignment among institutional actors

10. Education, Public Policy, Ethics, Diversity, Digital Divide

Potential Disciplinary Domains:

A. Materials Genome
B. Astronomy and Space Science
C. Humanities
D. Engineering
E. Geoscience
F. Biomedicine
G. Life Sciences
H. Social Science

Reporting Format:

- Motivation/Vision/Goal/Scope (2014-2015)
- Benefits/Risks
- Timing/Milestones/Resources

Note*: The same words mean different things to different people – be sure to ask "say more about what you mean by that . . ."*

Working Group Reports:

Creating Data Collection (structure and unstructured data)

- Focus on manifest file, integrating a number of existing services – authorship, readership
- Broad collection of ideas
- There are existing tools that can be leveraged
- First steps would be functional goals
- Tag line – as easy to share data as it is to create a web page
- Action items are identified

Comments:

- Would this allow collection several different types and bundled to have a new DOI
- Some attention to provenance
- There are important standard that are relevant – such as data conservancy

Publishing and Linkage

- Pilot project on linked open data repository
- Several publishers and several certified repositories
- All dedicated to this project
- Need volunteer data repositories for the pilot study
- ICPSR may be a candidate
- Materials initiative may be a candidate
- Astronomy may be a candidate

Comments:

- This is a nice pilot – demonstrating connects that can then be expanded

Archiving, Preservation, Curation, Stewardship, Storage

- A large domain – focus on archiving and storage initially
- Debate on curation and supercomputing orientations
- Initial phase is research own storage, followed by link to archive
- Goal is API to all storage – service using API
- First identify storage resources – identifying and profiling the current archiving landscape in the US – key is the profiling

Comments:

- This is a rate limiter
- NCSA SDSC are committed to some archival storage

Data Discovery, Notification and Access

- Inspiration from the video
- Have the NDS consortium work with the SHARE notification service to drive notification on new deposits into SHARE
- Issues of granularity to be addressed
- Ensuring SHARE can be used in this way for discovery in domain areas

Comments:

- Would that relate to archives?
- There are notifications for articles when published
- Central source that can be used for discovery
- A notification service could be integrated with the repositories
- There is an ESIP initiative – beyond a registry to a more distributed model
- Both push and pull models are relevant
- There is a red flag when we focus on particular solutions such as SHARE – take a step back and think of underlying protocols in the eco-system
- Potential for experiments with multiple services

Analysis, Use, Visualization, Computation, Modeling, Analytics

- This is the tail end of what everyone else is discovering
- Hard to move forward without seeing what emerges from others
- Still, want to identify two disparate fields that could play together
    - Earth Science/Atmospheric Science and Astronomy/Exo-Planet Science
- NDS standards or establish domain standards on tools and workflows
- Translational services, such as brown dog, combined together to illustrate usability interfaces – so disparate fields can work in cahoots with each other

Comments:

- Example of a storage system that could be connected to some analysis tools
- Dialogue on work flows, but don't want to constrain
- Astronomy group was proposing some analysis that would fit in this

Coordination, Federation, Alignment among institutional actors

- Important to have a mission and vision up front
    - A charter within the next year – rules of engagement
- A clear value proposition specified
    - Evolving as NDS evolves
- Clear ways to plug into NDS, such as through the pilot projects
    - Technical and social
    - People need to know how they can contribute

Comments:

- Will be on the agenda at the close out

Education, Public Policy, Ethics, Diversity, Digital Divide. . . Social Science . . . Humanities

- Abstract for a grant on loosely coupled systems in the development of a national data service
- Interaction is the relationship between loosely coupled systems and tightly structured disciplinary systems
    - Little research on this
    - Stakeholder alignment is key to NDS development
- Language has different meanings
- Inequities in funding across institutional actors
- Issues of competition
- Many candidate initiatives

Comments:

- Connection to the governance group
- Potential link to the RDA

Materials Genome

- Building on previous discussions on this
- Proposals to establish a materials data facility pilot
- A petabyte of storage is committed for this and other projects
- Data to be stored and shared readily at no cost initially
- Long term sustainability is a challenge for the community
- Hope to link to publications
- Computational data, crystal structure data, and other possibilities
- Clear strong commitment
- Can be announced at the third anniversary of the materials genome initiative

Comments:

- There is an RDA working group on materials interoperability

Astronomy and Space Science

- Very good structures in place in Astronomy – virtual observatory, etc.
- Looking for a promising niche – astrophysical simulations – archive with data in this space
    - Numerical relativity and cosmology communities

- Big data
- Simulation data to be discoverable
- Hosting at NCSA, SDSC
- Standards for metadata and simulation data
- Build on existing virtual observatory data
- YT for visualization
- Catalogue of derived data products available
- Data products are almost metadata

Comments:

- Each one of these topics maps onto different parts of the matrix – coordination with different groups – archiving, analysis, etc.
- Multi-messenger – gravitational waves, neutrinos, combined with electromagnetic

Engineering, Geoscience and GIS

- Focus on special data – adding civil engineering
- Issue with the distributed nature of the data
- Spatial data synthesis
- Integration of different data types – including different special coordinate systems
- Multi-level data – local, regional, national, international
- Semantic nature of integrated data across multiple sources
- Initial focus on a use case to integrate three or four sources for synthesis and aggregation

Comments:

- This can connect to GIS initiatives at NCSA

Health and Life Sciences

- Builds on two position papers
  - Data producer and data consumer
- Create a specification with health and life sciences reporting
- Begins with Autism data base and connect to different consumers in the health sciences
  - Data can't be pushed, but the meta data push has potential

Comments

- Could be a driver for different kinds discovery services
- There is the potential for VIVO to help provide context for research data – references to people, publications, etc.
- Potential extensions into many different health domains
- Highlighting investment options
- An extension of the archiving and storage group