National Data Service Consortium
Second Meeting

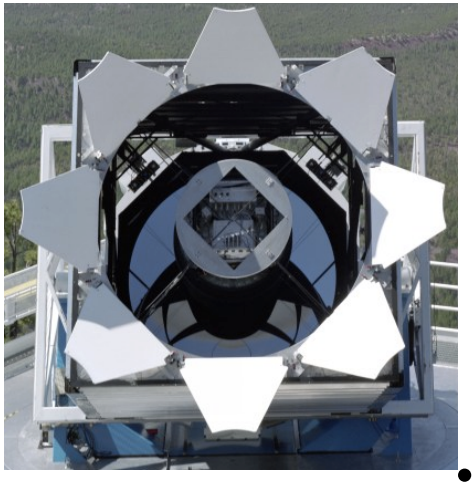
Ed Seidel

University of Illinois Urbana-Champaign

Data-enabled Transformation of Science

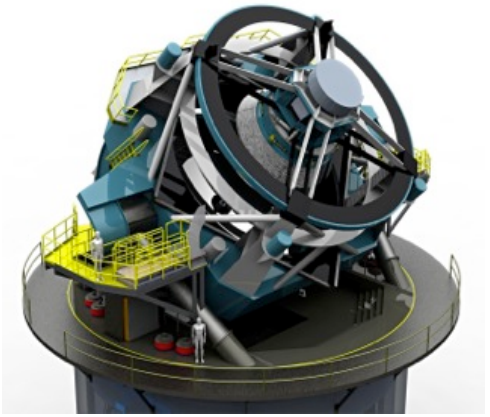
Astronomy 1500- 2000:

- Single scientist looks through telescope
- Record KB of data in notebook
- Require reproducibility



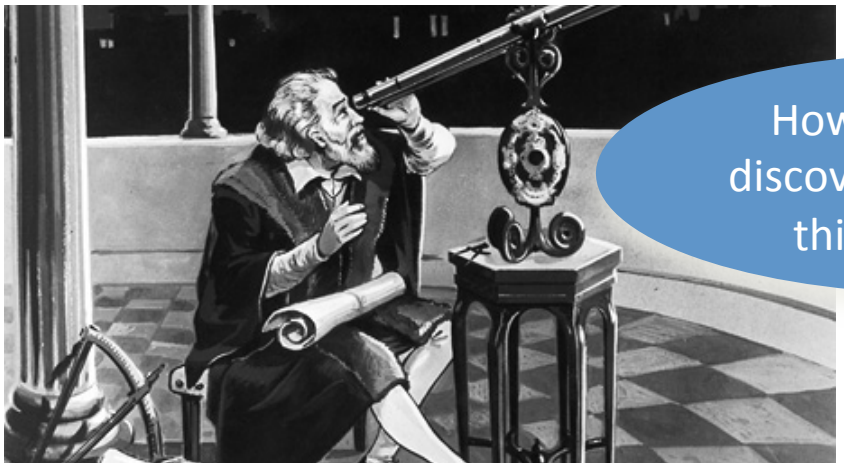
Sloan Digital Sky Survey 2000+

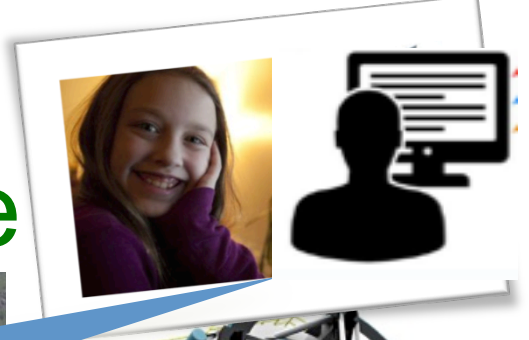
- Record data for decade (40TB)
- Serve to entire world
- Thousands of scientists work “together”



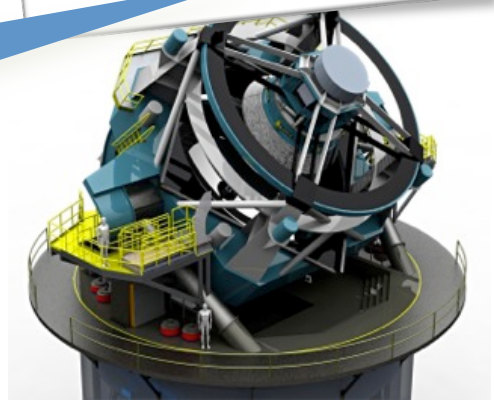
- DES (now)
  - 200GB/night
  - PB in decade
- LSST (6 years)
  - Record data for decade
  - SDSS/night!
  - 200 PB/decade

Data-enabled Transformation of Science





How can I publish, discover, verify data in this new world?



Astronomy 1500- 2000:

- Single scientist looks through telescope
- Record KB of data in notebook
- Require reproducibility

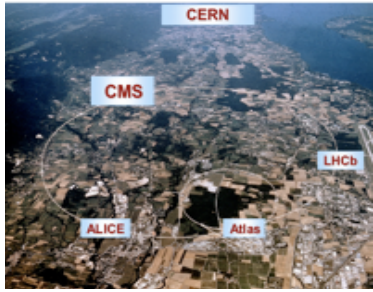
Sloan Digital Sky Survey 2000+

- Record data for decade (40TB)
- Serve to entire world
- Thousands of scientists work “together”

- DES (now)
  - 200GB/night
  - PB in decade
- LSST (6 years)
  - Record data for decade
  - SDSS/night!
  - 200 PB/decade

Big Data vs The Long Tail of Science

- Many “Big Data” projects are “special”
  - Highly organized, singular sources of data, professionally curated, a lot attention paid
- What about the “Long Tail” (the other 99%)?
  - 1000s of biologists sequencing communities of organisms
  - Thousands of chemists and materials scientists developing a “materials genome”
  - Characteristics:
    - Heterogeneous, perhaps hand generated
    - Not curated, reused, served, etc…

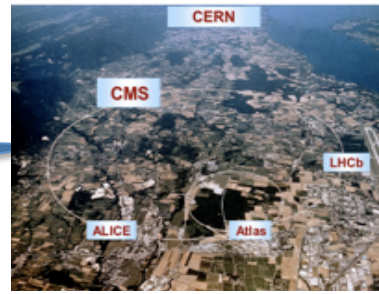


MATERIALS GENOME

Big Data vs The Long Tail of Science

- Many "Big Data" projects are "special"
  - High… …es of data, pr… …on paid
- What ab… …(the other 99%)?
  - 1000s of biologists sequencing communities of organisms
  - Thousands of chemists and materials scientists developing a "materials genome"
  - Characteristics:
    - Heterogeneous, perhaps hand generated
    - Not curated, reused, served, etc…

Fundamental Observation: Scientists communicate by sharing data…



MATERIALS GENOME

Basic Vision for Open Data and Publication Services

- Make it possible (easy) for anyone to:
  - Create a data collection and get an "identifier"…
  - Deposit it somewhere where it can be kept safe…
  - Provide services so others can find it, analyze it, repurpose it…
  - Link it to traditional (open, please!) publications…
    - OA aspects very important to this
- With these capabilities in place
  - Many important things will happen…

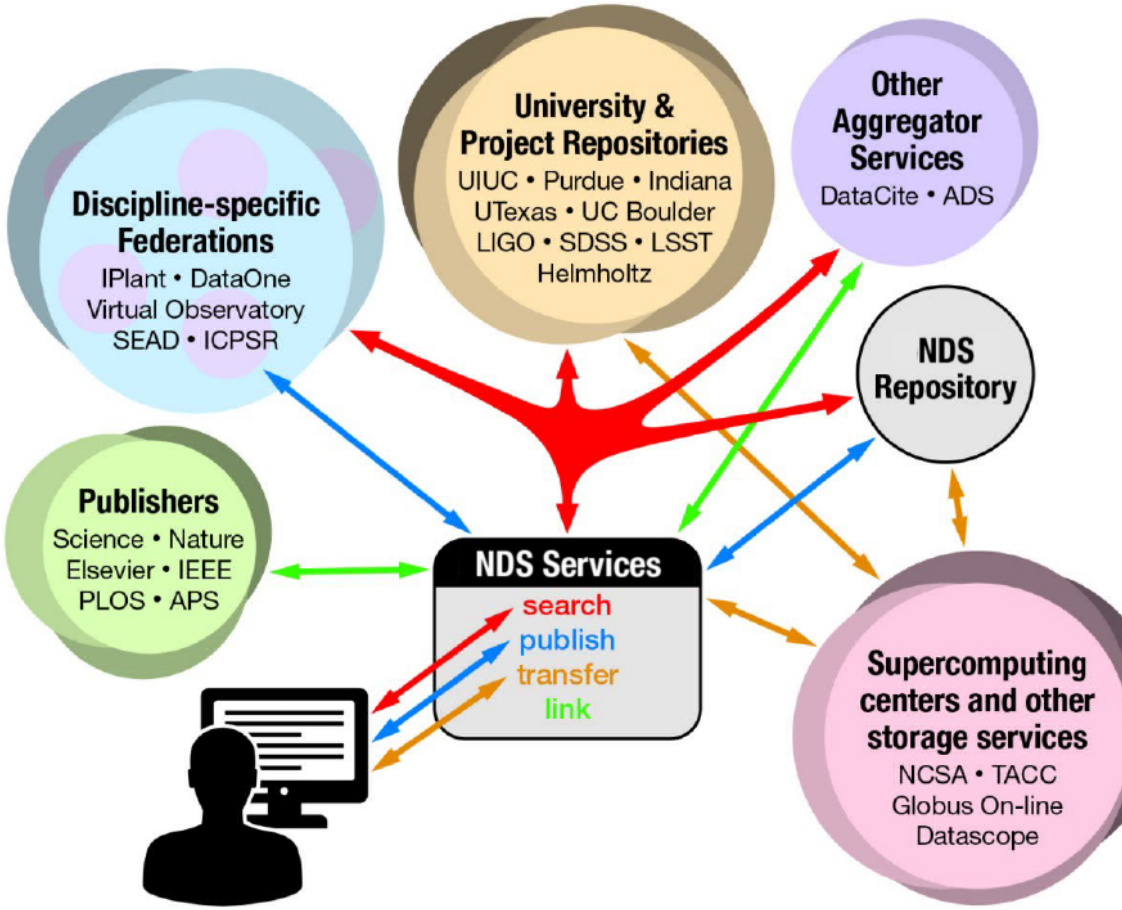
Basic Vision for O… Publica…

"We need to take steps to make scientific research data more liquid. The more we move towards open as the default for scientific research data, the more we will get out of the research enterprise. It is time to take deliberate steps to make that a reality." Mike Stebbins, White House OSTP

- Make it possible …
  - Create a data colle…
  - Deposit it somewhere …
  - Provide services so oth… …n find it, analyze it, repurpose it…
  - Link it to traditional (open, please!) publications…
    - OA aspects very important to this
- With these capabilities in place
  - Many important things will happen…

Why is this so critical for the future?

- *Reproducibility of a scientific result*: heart of science
  - Needed: access to complete state of a result…
- *Accelerating discovery*: faster, deeper dissemination of results to other researchers; *Repurposing data*
  - Needed: services to find, retrieve, analyze, describe…
- *Interdisciplinarity and complex problem solving*
  - Needed: ability to find, integrate results across communities
- *Public dissemination* of publicly funded research results
  - Needed: open, accessible results, searchable by public
- *Economic development*
  - Needed: availability of all the above to companies (MGI!)



A Builder's Consortium!

NATIONAL DATA SERVICE CONSORTIUM

NationalDataService.org

What should NDS do for researchers?

- Help researchers ***find data***
  - Cross-disciplinary searching: across federations, projects, archives, and other repositories
  - Find data related to a publication
  - Allow drill down to leverage specialized community-specific discovery
- Help researchers ***use data***
  - Download data, browse metadata, track provenance
  - Move data to processing platforms for specialized (re-)processing and analysis

What should NDS do for researchers?

- Help researchers ***share and publish data***
  - Engage researchers early in the publishing process
  - NDS and federated local/domain repositories
    - sharing privately with collaborators prior to publishing
    - tools to help organize the data for publishing
    - automatically ensure links to literature
    - assign DOIs, provide links to publishers, synchronize data publishing with papers
  - Recommend appropriate discipline/community repository for long-term preservation
    - NDS Repository as archive of last resort

What are we doing and why?

DEVELOPMENTS SINCE NDS-1 IN BOULDER

Activities

- First NDS meeting hosted by NCAR in Boulder
  - ~85 participants from all the above types of organizations
    - General services discussed and with much agreement emerging
    - Specific groups agreed to help pilot use cases for early services (MDF, Astro, other reports here)
- Since NDS-1 at NCAR…
  - Interim steering committee formed
  - OSTP and NDS announce Materials Data Facility (MGI + 3 years)
  - Hackathon last month at NCSA to explore prototype services, connecting together existing tools
  - NCSA, SDSC, TACC, ANL have agreed to create federated storage and capabilities for development of services
  - Demos at SC14; see booths at Globus, NCSA, SDSC, TACC…

NDS: an ecosystem in 3 parts

- The Portal

  Complete end-to-end set of vanilla national services for storing, sharing, publishing, finding and re-using data

- The Framework

  The system into which a community can plug specialized tools, portals, and services

- The Infrastructure

  Foundational storage, hosting environment and software that allow communities to build their own specialized data services

We're exploring all three of these with pilot activities now underway

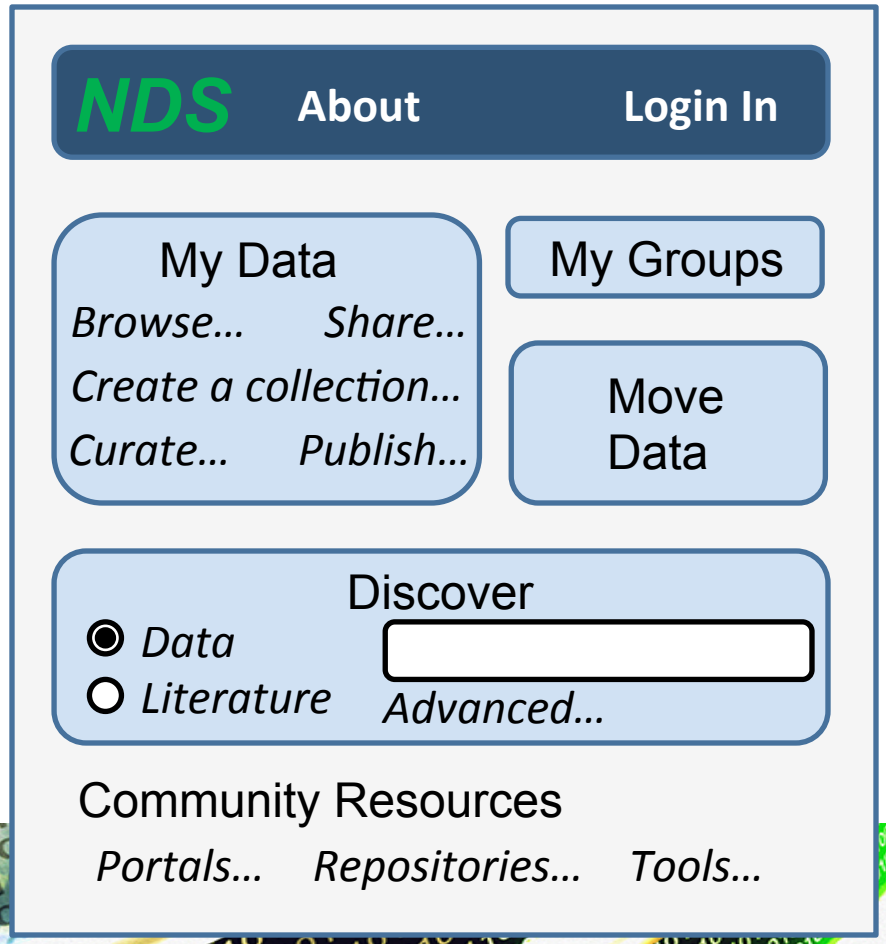
The NDS Ecosystem: The Portal

A complete end-to-end set of vanilla services for sharing, publishing, finding and re-using data

Imagine a portal that supports cross-disciplinary research

- Supports data of any discipline
- Enables private sharing, publishing, data discovery, data movement
- Connects researcher with community-specific resources where they exist

The Materials Data Facility (MDF) is a prototype for a generic NDS portal



The NDS Ecosystem: The Framework

The system into which a community can plug specialized tools, portals, and services

- The framework is how existing and new tools connect together
- Any generic NDS component should be replaceable with a community specific version
- User can leverage which ever tools work best for her research group.
- Based on the "Data Fabric" recommendations from the RDA

Imagine…

1. *A group begins sharing data informally using Dropbox, SciDrive, Figshare, or …*
2. *They move the data into SEAD to create a publishable collection*
3. *They combine it with a collection created in the NDS portal*
4. *The data is published into Dryad for long-term preservation*

We're currently exploring this through the Epiphyte Pilot and NDSShare

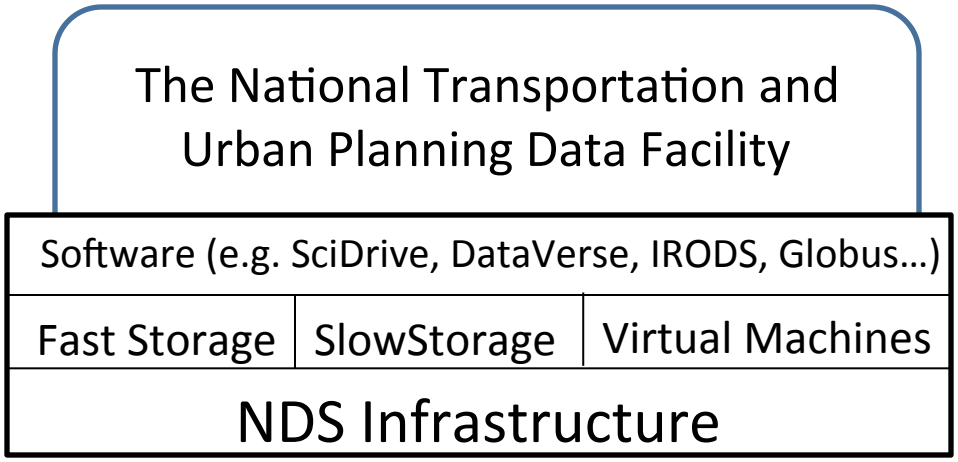
The NDS Ecosystem: Infrastructure

Foundational storage, hosting environment and software that allows communities to build their own specialized data services

Includes

- Nationally distributed storage
  - Replication services
  - Databases
  - Community repository tools
- Cloud-based Hosting environment
  - For hosting portals and services
- Common data software
  - For building community-specific capabilities

Imagine…



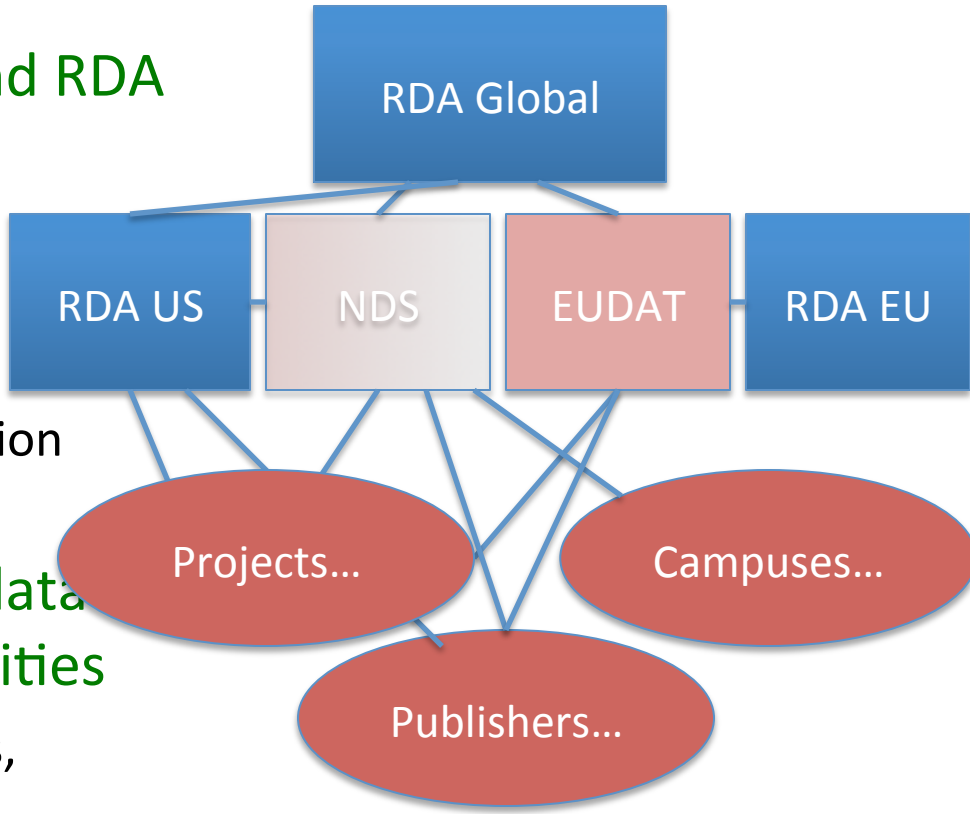
NDSLabs will provide an experimental prototype of NDS Infrastructure

NDS Lab and NDS Share

- NDS Lab
  - Target: friendly developers
  - A community support environment for developing, coordinating, deploying prototype service
  - Spinning disk, storage, virtual machines for developing and hosting services
  - Available to NDS community members
- NDS Share (or better name: Kalliope? Help us name it!)
  - Target: friendly scientists
  - Experimental platform for sharing data
    - Enable anyone to create data collections, store data, get DOI
  - Include installations of community data sharing applications
  - Will evolve over time
- Partnership between NCSA, ANL, TACC, and SDSC
  - Other interested partners?
- Look to make available by January 2015

How will NDS relate to other activities?

- NDS will complement and extend RDA
  - No attempt to duplicate
  - Working closely with RDA groups to implement
  - Progress on NDS as member of RDA Global and documented relation with RDA and other organizations
- NDS: operating framework for data services, on top of existing activities
  - Narrow focus on specific functions, leveraging existing capabilities
  - Create collections, identify, deposit, publish, link…



Vision for 6 months and 5 Years Hence

- Six Months
  - Clear membership, relationships, governance models
  - Numerous pilots progressing with coordination
  - Funding opportunities clearly identified, coordinated
- Five years
  - It is *routine*, and *part of culture*, to **store**, **publish**, **share**, **discover**, **link** data
  - National structures in place that connect campus, domain, and national services federated with local
  - Promise of data-enable science really begins!

NDS: *A Builders Consortium*

- NDS vision requires collaboration of many kinds of institutions
  - Compute and data services centers
    - Effort spearheaded by UIUC (NCSA, Library), UC/ANL, UT/TACC, UCSD/SDSC
  - Universities and project repositories
    - Internet2 and numerous members; ARL and members
    - LIGO, IceCube, LSST, DES, etc
  - Discipline-specific federations
    - E.g., SEAD, DataONE, iPlant, Virtual Observatory, SEAD, ICPSR, HASTAC, …
  - Publishers
    - Science, Nature, APS, PLOS, IEEE, Elsevier, JORS, et al…
- NDS Consortium to guide the building, governance of services
  - Coordinate separately funded efforts to build NDS components
    - Ensure interoperability, integrate existing tools and resources
  - Interim steering committee formed
  - Joel Cutcher-Gershenfeld to moderate governance discussions here

Activities

- First NDS meeting hosted by NCAR in Boulder
  - ~85 participants from all the above types of organizations
    - General services discussed and with much agreement emerging
    - Specific disciplines agreed to help pilot use cases for early services
- Since NDS-1 at NCAR…
  - OSTP and NDS announce Materials Data Facility (MGI + 3 years)
    - Stakeholder Map of 200+ groups in progress
  - Hackathon last month at NCSA to build out prototype services for data collections, storing, DOI minting, linking to publishers
  - NCSA, SDSC, TACC, ANL have agreed to create federated storage capabilities for development of services
  - Aiming for demos at SC14; see NCSA, TACC, SDSC, other booths

NDS ecosystem in 3 parts

Refine these elements here…

- Layer 1: Complete end-to-end set of vanilla *national* services for storing, sharing, publishing, finding and re-using data
  - Components for collaborative sharing, creating collections, archiving, re-use, and linking to literature
  - General-purpose search system discovers data across disciplines
- Layer 2: Framework into which a community can plug specialized components: Community-specific…
  - Federations integrate with NDS by plugging in, say, specific publishing tools that capture specialized metadata
  - Search tools can reach out to neighboring disciplines by accessing the generic search service
- Layer 3: Foundational infrastructure that allows communities to build their own specialize data services
  - Distributed storage, replication services, repository services
  - Cloud computing services for hosting portals and services

NDS ecosystem in 3 parts

Refine these elements here…

- Layer 1: Generic services for storing, discovering, linking…
  - Today and tomorrow we will see some prototype services for some of these functions
  - Still need to develop better definition
    - What services are needed
    - Consortium Governance model to determine what is to be supported
- Layer 2: Framework for community development
  - NCSA and partners will support *NDSLab*
  - Hosting environment for developing community
- Layer 3
  - NCSA and partners will support *NDSShare*
  - Storage and computing services

NDS Lab and NDS Share

- NDS Lab
  - A community support environment for developing, coordinating, deploying prototype service
  - Spinning disk, storage, virtual machines for developing and hosting services
  - Available to NDS community members
- NDS Share
  - Experimental platform for sharing data
    - Enable anyone to create data collections, store data, get DOI
  - Include NCSA-based installations of community data sharing applications
  - Will evolve over time
- Hoped for Dates
- Who else agrees to partner in advance

National, Federated Data Service(s)

Urgent need for national infrastructures for data

- Extensible, integrated national-scale services
  - Storing, sharing, finding, verifying, publishing, citing, reusing…
- Open and *federating* architecture
  - Building on the infrastructure currently at discipline/community level
  - Allow data providers to make data accessible in the national environment
  - Allow new and community-produced tools and resources to be plugged in