Green group:
http://bit.ly/1MA4ZNJ

Integration and collaboration of data repositories over a temporal scope - 5 - 10 years, spatial scope of global?

1. What results enabled?
    - People are able to (re-)use research data for any stage of the research cycle - not just discovery, but exploration and integration, etc.
    - Re-use of data outside of research cycle (e.g., decision-makers, policy makers, teachers in schools)
    - Serving a designated community and perhaps beyond the designated community
        - serving beyond the primary designated community may be further out - first focus should be on a designated community (community of individuals w/ shared/common knowledge).
        - Primary designated community: Earth and environmental science researchers
    - Enable open science
        - Reproducibility, open methods, interoperability
    - Facilitate data-intensive science
    - Repository/institution disappear into the background
    - Sustainable infrastructure
    - Scientists spend most time doing science, and a lot less 'wrangling' data
2. Three greatest limiting factors

1. Current Governance and Funding Structures
    a. Funding links to institutions inhibits sustainable cooperation/interoperability
    b. Programs @NSF and elsewhere are ephemeral, not support for long-term infrastructure
    c. Infrastructure dependent on benefactor rather than users

The current governance and funding structures for repositories (government agency-based, domain-based, institution, and publisher) inhibits sustainable cooperation/interoperability. Different types of repositories compete for limited funds, build and customize their own software, and create redundant services. NSF, NIH and and other government funding for research is short term and does not support development of sustainable infrastructure or long-term preservation. Infrastructure requirements are defined by available funding and goals of benefactors rather than scientific users.

2. Organization of repositories, workflow, and "data publication" model

    - Repositories mirror disciplinary silos, institutional silos, etc.
    - Data publication metaphor/model may be inadequate

        - Published snapshots are useful for fixed amounts of authoritative data associated with a published paper, but don't capture the full granularity of the data, and doesn't match the volume, heterogeneity, and recombination possibilities of data
3. Relationships Between data producers, consumers, and repositories
    - Lack of participation from data producers / incentives mismatch between producers/consumers
    - Repositories don't understand the re-use community
        - Unclear how people locate data
        - Unclear how data are re-used, why and how
    - Data heterogeneity necessitates varied management approaches
    - Redundant infrastructure development
    - Legacy investment in built infrastructure inhibits interoperability, slows change
    - Agency view that they are legally mandated to be the authoritative source for federally generated data
4. Changes needed in way env repos are managed to overcome limitations
    - Repositories should be redefined as a set of interoperable services (not end-to-end stacks); (e.g., data storage, metadata authoring and management, replication services); repositories provide a set of services that can be swapped out, interchanged, scaled to new levels, based on an evolving suite of standards
    - Repositories need to become selective of the data to be curated for preservation
    - Need to increase the number of data managers and curators in repositories by two orders of magnitude; and we need every scientist who produces data to be proficient at managing data in her own discipline (scientific metadata standards, analytical tools, credibility and integrity, etc.) from advisors/mentors through to graduate and undergraduate students
    - Need incentives for producers to participate/contribute; specific value
    - Repositories need to recognize the granularity, identification, and provenance issues inherent in data that differ from the traditional publication model
5. To engender collaboration, what changes are needed in how repositories conduct their work?
    - Long-term funding of repositories and repository services;
    - Funders need to evaluate infrastructure based on stability and sustainability, rather than novelty and experimentation (distinguish between infrastructure and research on infrastructure)
    - Initiate consistency in review criteria for data to justify initial and continuing curation investments (taxonomy of review criteria)

Green group Vision

Vision: We envision an interoperable, collaborative network of sustainable data repositories that supports open science for a designated earth and environmental science community over the next 5 - 10 years at global scales. This network will allow researchers to discover, access, and re-use data from other researchers, and will grow to support other designated communities, including decision-makers, policy makers, teachers, and students. Researchers will be able spend much more time doing science, and a lot less 'wrangling' data in pursuit of synthetic and collaborative studies.

Limitations: 1) The current governance and funding structures for repositories (government agency-based, domain-based, institution, and publisher) inhibit sustainable cooperation and interoperability. Different types of repositories compete for limited funds, build and customize their own software, and create redundant services. Government funding for research is short term and does not support development of sustainable infrastructure or long-term preservation. Infrastructure requirements are defined by available funding and goals of benefactors rather than scientific users. 2) The current organization of repositories, workflow, and the "data publication" model creates and/or reproduces disciplinary silos, institutional silos, and other divisions that make discovery and reuse overly dependent on repositories rather than scientific need. The data publication metaphor/model may be inadequate. Although published snapshots are useful for fixed amounts of authoritative data associated with a published paper, they don't capture the full granularity of the data, and do not match the volume, heterogeneity, and recombination possibilities of data. 3) The relationships between data producers, consumers, and repositories are compromised by the lack of participation from data producers, and a mismatch of incentives between producers/consumers. Repositories don't have a sufficient understanding of the re-use community.

Approaches: To overcome these limiting factors, repositories should be redefined as a set of interoperable services, not end-to-end stacks (e.g., data storage, metadata authoring and management, replication services), and in so doing, provide a set of services that can be swapped out, interchanged, scaled to new levels, and based on an evolving suite of standards. Repositories should become selective of the data to be curated for preservation. We must increase the number of data managers and curators in repositories by two orders of magnitude and ensure that every scientist who produces data is proficient at managing data in her own discipline. Repositories need to recognize the granularity, identification, and provenance issues inherent in data that differ from the traditional publication model.

Enabling Collaboration: In order to engender successful collaboration among repositories, funders must develop new models that support long-term funding of repositories and repository services, and that evaluates infrastructure based on stability and sustainability rather than novelty and experimentation. Repositories need consistency in data review criteria to justify initial deposit and sustained (re-)use of research data throughout the research cycle, with far greater efficiency than today.