Tech Vision

http://bit.ly/1HH5Yz7

Cyndy Parr, Dave Vieglas, Ken Casey, Margaret Hedstrom, Greg Janée, Mark Servilla, Kevin Brown, Matt Jones

How do you define this arena? li9i99k06 kll pł. Ll

1. Goal: sustainability of data, not repositories, more robust links in the face of dependencies (e.g. controlled vocabularies), better adherence to standards
2. Unified framework - clear architecture and understanding of projects and their responsibilities and where they fit into the framework
3. Goals, standards, APIs and service endpoints, may or may not involve common file storage or structures, provenance, version control
4. Scientists and science funding agencies pay for building as little of the infrastructure as possible
5. Some level of shared infrastructure
6. Interoperability of repositories
7. Sustainability - do we need more in common under the hood compared to just commonality at the level of interfaces
8. distributed infrastructure
9. Infrastructure needs to support heterogeneous needs and heterogeneous data

Why is this a critical arena for action?

1. Avoids inefficiencies of duplicated effort -- build on economies of scale
2. There's a lot of heterogeneity in solution implementations and some of that is for good reason, but there should be a framework that allows interoperability and collective goals to be met even if not all participants handle all parts of the life cycle
3. Make it easier to migrate to future platforms or across platforms

What might the top 3-4 top outcomes / results be within this arena? In what time frame might these be realistically achieved?

Many of these exist but need a bit more refinement to be useful, so could be done within a couple of years:
Next 3-5 years:

1. Best practices for researchers on how to package their data for broad use without losing all the relationships in the package
2. Guidance for repository developers and data producers on how to consistently apply standards like PROV in this domain (and encourage development of validation tools to help with that)
3. Common set of use cases (and anti-use cases) with examples in our domain of how to apply abstract standards)
4. Identification of appropriate standards based on similar recommendation efforts (US GEO, etc)

These will take a few more years:
Longer term:

1. An implementation plan (based on above recommendations) sitting above the new OAIS ref model, bringing to it current technologies, emphasizing interoperability in our domain (searching across many packages, computing across a repository)
2. Establishing a technical review board to advise on plans for repositories

What are the high level actions that would need to be taken to pursue these outcomes / results?

1. Convene the appropriate working group such as an ESIP cluster then working group
2. Write an NSF planning grant to bring more stakeholders to provide input to all of these things
3. Participate in DataONE outreach efforts

What capabilities / resources (those in existence and those yet to be developed) would be needed to engage in these actions?

1. Travel funds
2. Commitments from organizations for employee time
3. Community sharing infrastructure (e.g. Google Docs or a wiki, teleconferencing)

Ken Casey's vision:
Start with data assembly centers, make sure there are enough of them to cover the disciplines
Link them together with DataONE framework, to connect these to federal data structures for long term archiving
Mark Servilla:
Goal is to persist the data to make it sustainable

Functions:
Replication

Tech Vision Action Items

- Strengthen a technical community discussion of repository sustainability and interoperability features (inject DataONE Member node forum into ESIP cluster discussion --both Jan and July meetings)
    - entrain CDF as well
    - start by writing a cluster paragraph with Erin's input -- Matt & Cyndy -- Dec 15 (make sure there is remote participation)
- Articulate the objectives to be achieved via a federated approach (e.g., sustainability, interoperability, etc) (related to the ROI) (at Jan ESIP meeting 4-8 Jan)
    - For example, catalog interoperability
- Compile use case listing articulating shared vision of technical sustainability and interoperability across repositories on shared space (Github please)
    - Start with compiling list of use case links for existing work (Matt, …)
    - Review these at Jan meeting, discuss process for expansion/edit/improvements
- Articulate a high level conceptual framework (concept of operations) to satisfy the use cases --
    - Compile list of existing frameworks -- May 30 (Cyndy will coordinate)
    - Draft presented at July ESIP meeting -- decide who leads in January ESIP meeting (might be DataOne folks depending on the site review dates)

- Identify the services / infrastructure essential for implementing the conceptual framework and determine the cost of providing such (community wide) services (conceptual equivalent of DNS for internet).
    - Compile existing lists
    - Discuss this July ESIP
    - Assign responsibility in January
- Develop best practices for developers that include *good* validation tools and shared conventions and proposed changes in standards
    - Compile existing
    - Refine based on other work
    - Timing -- sometime in the future
- Discuss with Peter and EAGER time about write a proposal to further support the design process (planning grant) (Mark Servilla) (by the end of this year)
    - E.g. A month of funding each for 12 people, designated person (e.g. environmental sciences masters or recent graduate) to facilitate the process
    - supplement to EAGER award or some other award
    - ESIP fellows just got awarded, Erin will check to see if there's an option to get another ($2000 + travel). We want quite a bit more.