Considerations Associated with Building, Marketing, Maintaining and Managing a Big Data Repository

April 16, 2018

Helen Ray
William Savage
Adam Weeks

Focus – PREDICT and SAMHDA

- PREDICT – Department of Homeland Security
  - Protected Repository for Defense of Infrastructure Against Cyber Threats – later called “IMPACT”
- SAMHDA – Substance Abuse and Mental Health Services Administration (SAMHSA)
  - Substance Abuse and Mental Health Data Archive

Project History – Data Repository Building

- SAMHDA –Substance Abuse and Mental Health Data Archive
  - Offers data to researchers for analysis
  - Data types - public, semi-private, restricted
  - Consolidated model – one source for distribution of data
  - Free data resource

Project History – Data Repository Building (continued)

- PREDICT – Repository of Cyber Threat Data
  - Data for researchers to test and evaluate research prototypes
  - Government technology decision-makers evaluate competing cyber security tools and methods
  - Data types – semi-private, restricted
  - Distributed model – Multiple sources for distribution of data, one site for data request and access
  - Free resource

Construction of Data Repositories - Considerations

FOUR Questions - Associated Data Issues

1. What data do we include?
2. Where does the data come from?
3. How to ensure the accuracy and integrity of the data?
4. What data formats to offer?

Pre-construction

- Develop solid goals
- Understand user needs
- Determine timeline for implementation

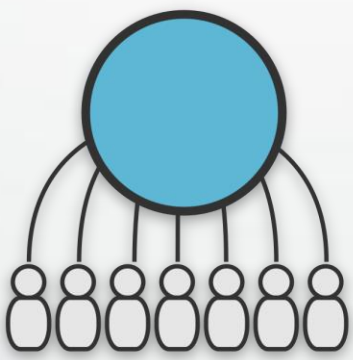
Construction of Data Repositories

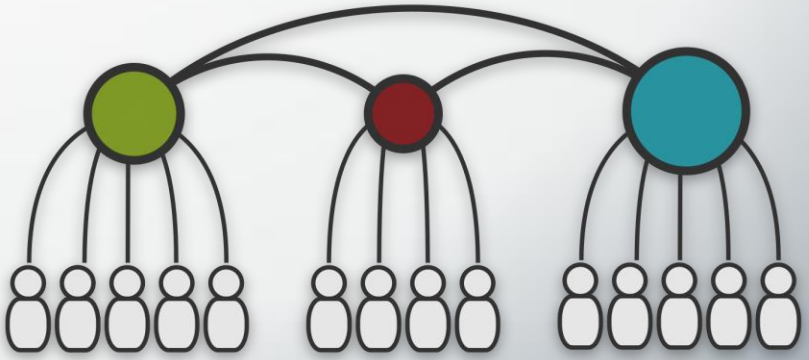
- Develop website
- Offer online analysis tools
- Provide Data Archive
- Managed releases

Models of Data Distribution

- Decision Points

- Centralized/Decentralized?
- Distributed?
  - One or small number of providers
  - Distributed - multiple data providers





Data Access Constraints

- ✓ Rights to collect data (contributors)
- ✓ Rights to share data (distributors)
- ✓ Data use agreements
- ✓ Data access levels – public, semi-private, private
- ✓ Monitoring and controlling access to data
- ✓ Timing limitations for data access – data destruction/return

Data Collection - Process

- Data Providers - standard formats vs random
- Data files moved into custody
- Delivery manifest files created to ensure traceability
- Data is placed in a secure directory
- Internal data delivery ticket created

Data Support and Maintenance

- Metadata to support the data
- Maintenance of datasets
- Statistical support
- Visualization tools
- Archival plans

Outreach

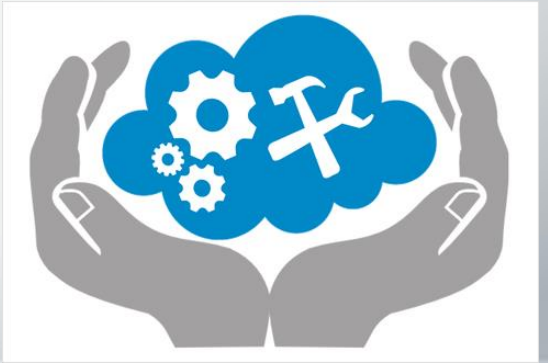
- Identifying the user community
- If you build it, will they come?
- Marketing and social media

Post Construction – Achieving Intended Goals

- Monitor use via Google Analytics and other tracking
- Serve the user community – informing
- Support for the site – takes resources!
- Database and hosting costs over time – locally hosted/public cloud/private cloud - GovCloud?
- Detailed plans for updates, maintenance
- Review by experts – focus groups to gather feedback
- Determine what's missing – gather feedback from users

Project Management – Methods and Tools

- MS Project schedules and Gannt chart process monitoring
- Tracking database for Help Desk inquiries
- Tracking database for internal development
- Vantage – proprietary software for performance monitoring – cost to complete
- IBM Cognos Analytics for financial and budget planning
- Confluence for stakeholder collaboration



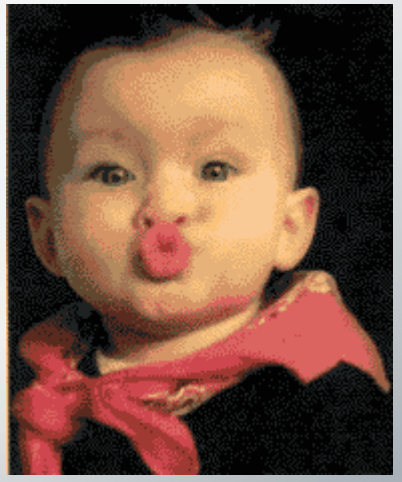
Summary

- Know your audience
- Understand data distribution model
- Develop solid project plan
- Provide analysis tools
- Monitor and expand usage

Questions?

?

Thank you



RTI International

delivering the promise of science
for global good

RTI INTERNATIONAL

Name: Helen Ray
Email: hmp@rti.org
Phone Number: 919-541-6954