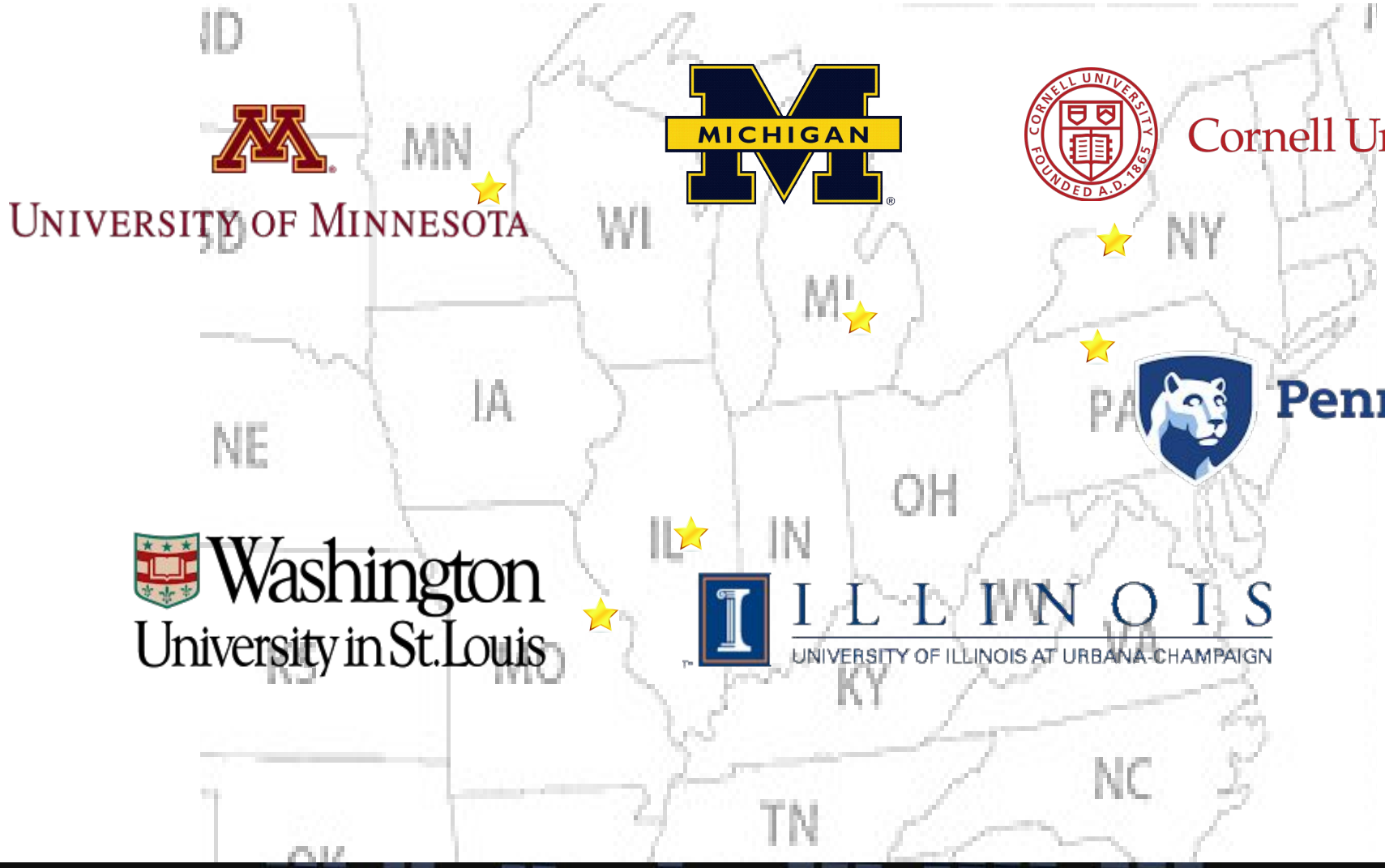
Data Curation Network Preliminary Results

Presented by Lisa Johnston to the SHARE User Group Meeting in Atlanta GA
January 24, 2017

https://sites.google.com/site/datacurationnetwork

Data Curation Network Partners



Challenge for Institutional Data Curation Services

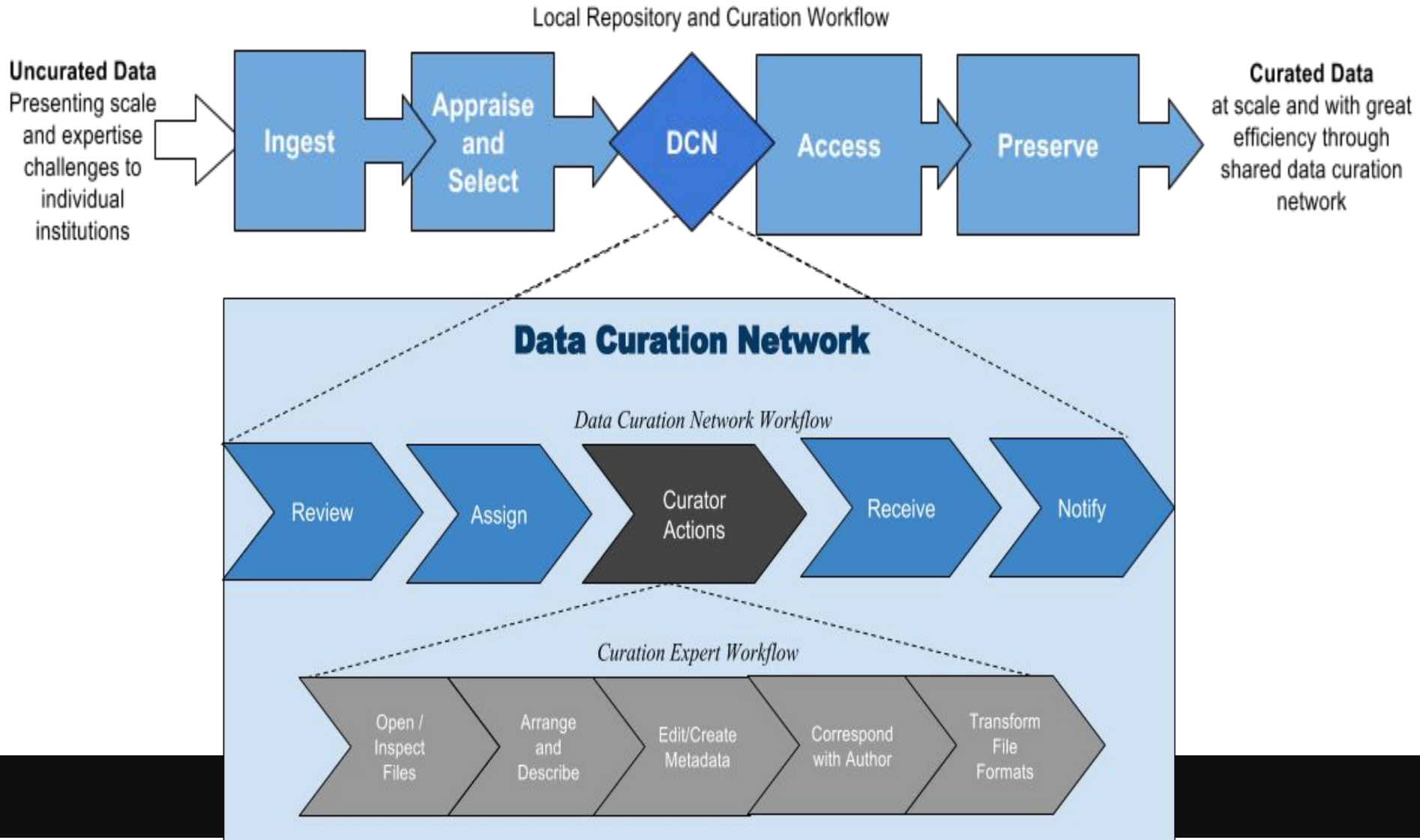
How to scale data curation services across all disciplines?

Multiple data curation experts are needed to effectively curate the diverse data types an institution typically generates.

Data curation expertise needed:

- File format-- GIS, spreadsheet/tabular, statistical/survey, software code, video/audio, images/3D, simulations...
- Discipline-specific-- genomic sequence, chemical spectra, biological image...
- Frequency-- Centers of excellence, departmental concentration

Draft Model for the Data Curation Network



Our Vision for the Next 3-5 Years

1. Develop standards-driven data curation techniques for all types of repository workflows and infrastructure.
2. Expand into a sustainable entity that grows beyond our initial six partner institutions.
3. Datasets curated by the Data Curation Network will be used to advance research and education in ways that are measurably of greater reuse value than non-curated data.
4. Build an innovative community that enriches capacities for data curation writ large.

Our Planning Phase activity to date

- ✓ **Summer** → Assessed infrastructure/policy/workflow differences and monitor the demand across institutions. Baseline report.
- ✓ **Fall 2016** → Seek input from researchers to better understand how data curation services fit into their research workflow (focus groups).
- **Underway Jan 2017** → ARL Spec Kit survey on library data curation activities.
- **Spring 2017** → Develop financial/governance models. Share our draft Data Curation Network model with stakeholders for feedback.

Data Curation Engagement: Card Rating Exer.

- Code review
- Contextualize
- Documentation
- Embargo
- File Format Transformations
- Persistent Identifier
- Quality Assurance
- Use Analytics
- Versioning
- Data Citation
- Deidentification
- File Audit
- File Inventory or Manifest
- File validation
- Metadata
- Metadata Brokerage
- Rights Management
- Risk Management
- Terms of Use
- Peer-review
- Technology Monitoring and Refresh

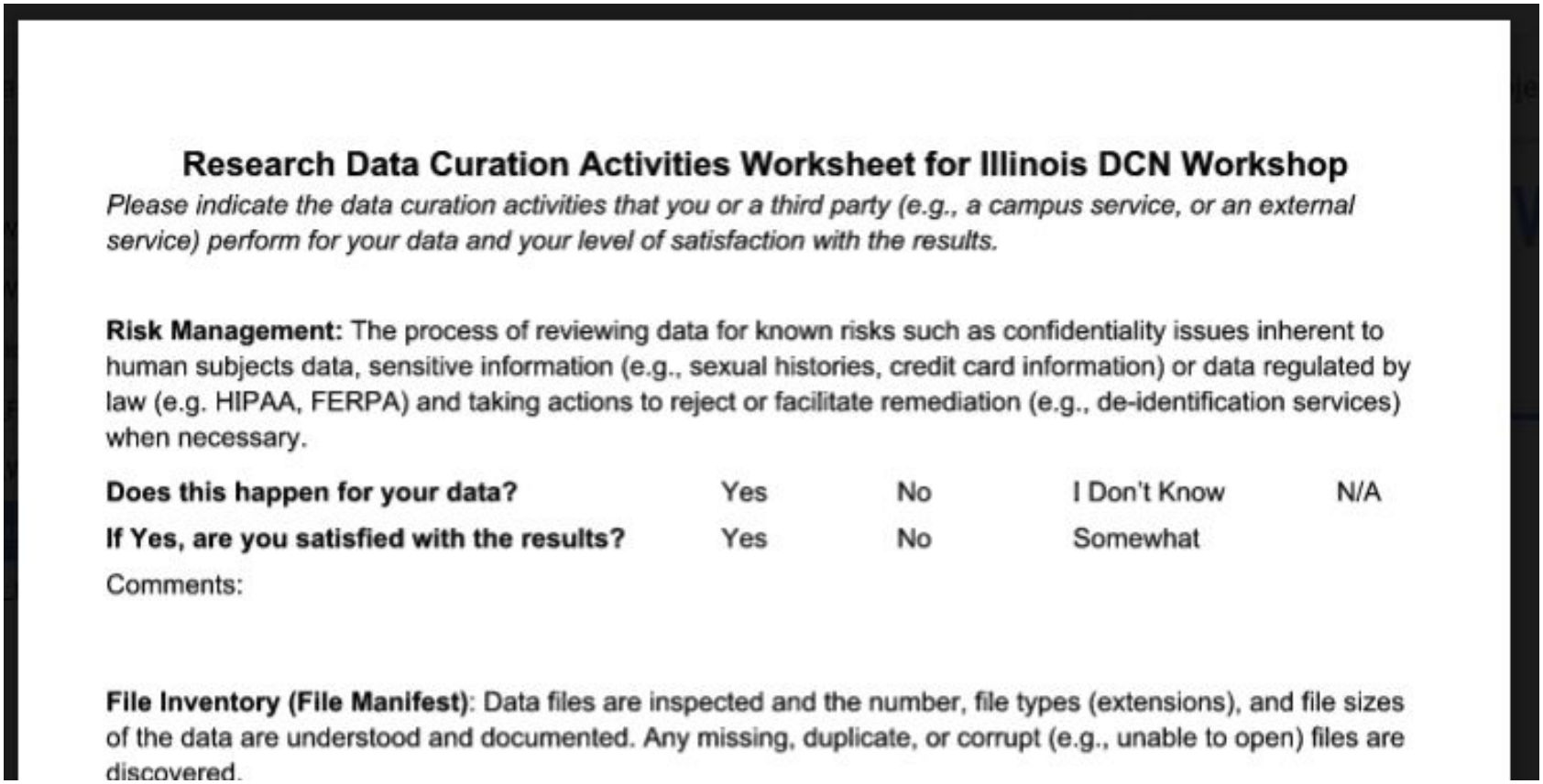
Activity #1 Results: Highest Ranking Activities

1. File Audit
2. Risk Management
3. Rights Management
4. Deposit Agreement
5. Persistent Identifier
6. Restricted Access (Data Enclave)

Activity #1 Results: Lowest Ranking Activities

1. Curation Log
2. Use Analytics
3. Technology Monitoring and Refresh
4. Peer Review of Data
5. Arrangement and Description
6. Authentication

Researcher Engagements



Research Data Curation Activities Worksheet for Illinois DCN Workshop

Please indicate the data curation activities that you or a third party (e.g., a campus service, or an external service) perform for your data and your level of satisfaction with the results.

Risk Management: The process of reviewing data for known risks such as confidentiality issues inherent to human subjects data, sensitive information (e.g., sexual histories, credit card information) or data regulated by law (e.g. HIPAA, FERPA) and taking actions to reject or facilitate remediation (e.g., de-identification services) when necessary.

Does this happen for your data?	Yes	No	I Don't Know	N/A
If Yes, are you satisfied with the results?	Yes	No	Somewhat	

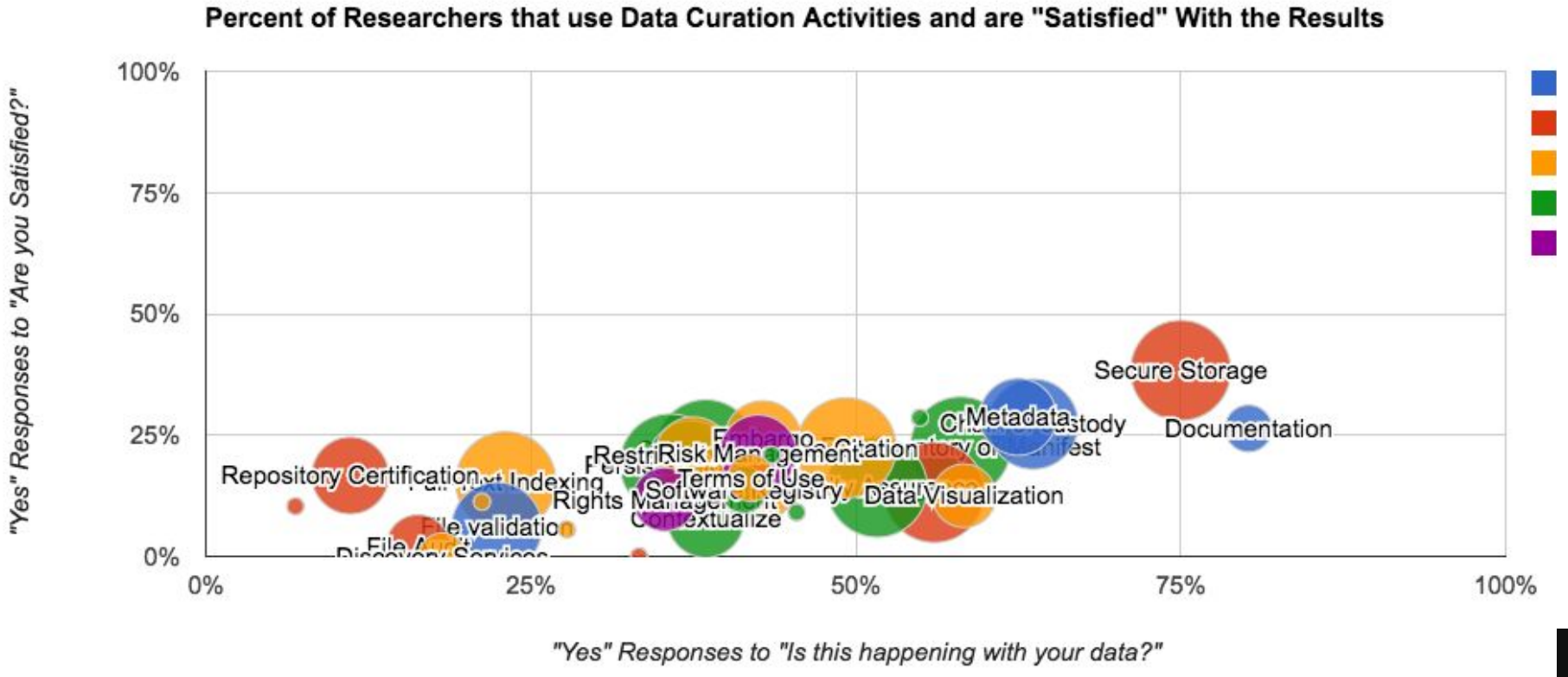
Comments:

File Inventory (File Manifest): Data files are inspected and the number, file types (extensions), and file sizes of the data are understood and documented. Any missing, duplicate, or corrupt (e.g., unable to open) files are discovered.

Results: Researcher Engagements

Methodology: Six focus groups with a total of 91 researchers

Findings: Most Data Curation Activities are Important (at least 3 out of 5)!
But….none are happening in a satisfactory way for a majority of participants!



Results: Researcher Engagements

Goal: Identify gaps in important curation activities that are either not happening/well.

Most Important Activities (4 out of 5)

- **(Create) Documentation (4.6)**
- **Chain of custody (4.5)**
- **Secure Storage (4.4)**
- **Quality Assurance (4.3)**
- **Persistent Identifier (4.3)**
- **Discovery Services (4.3)**
- **Tech Monitoring and Refresh (4.1)**
- **Software Registry (4.1)**
- **Data Visualization (4.0)**
- **File Audit (4.0)**
- **(Create) Metadata (4.0)**

Not Happening for Majority of Researchers

- **Persistent Identifier** (37% happens)
- **Discovery Services** (18% happens)
- **Technology Monitoring and Refresh** (33% happens)
- **Software Registry** (41% happens)
- **File Audit** (16% happens)

Results: Researcher Engagements

Goal: Identify gaps in important curation activities that are either not happening/well.

Most Important Activities (4 out of 5)

- **(Create) Documentation (4.6)**
- **Chain of custody (4.5)**
- **Secure Storage (4.4)**
- **Quality Assurance (4.3)**
- **Persistent Identifier (4.3)**
- **Discovery Services (4.3)**
- **Tech Monitoring and Refresh (4.1)**
- **Software Registry (4.1)**
- **Data Visualization (4.0)**
- **File Audit (4.0)**
- **(Create) Metadata (4.0)**

Not Happening for Majority of Researchers

- **Persistent Identifier** (37% happens)
- **Discovery Services** (18% happens)
- **Technology Monitoring and Refresh** (33% happens)
- **Software Registry** (41% happens)
- **File Audit** (16% happens)

Happening, but not satisfactorily

- **Documentation** (26% satisfied),
- **Chain of custody** (27% satisfied),
- **Secure storage** (38% satisfied),
- **Quality assurance** (14% satisfied),
- **Data Visualization (**12.5% satisfied),
- **Metadata** (29% satisfied)

Results: Researcher Engagements

Goal: Identify gaps in important curation activities that are either not happening/well.

Most Important Activities (4 out of 5)

- **(Create) Documentation (4.6)**
- **Chain of custody (4.5)**
- **Secure Storage (4.4)**
- **Quality Assurance (4.3)**
- **Persistent Identifier (4.3)**
- **Discovery Services (4.3)**
- **Tech Monitoring and Refresh (4.1)**
- **Software Registry (4.1)**
- **Data Visualization (4.0)**
- **File Audit (4.0)**
- **(Create) Metadata (4.0)**

Not Happening for Majority of Researchers

- **Persistent Identifier** (37% happens)
- **Discovery Services** (18% happens)
- **Technology Monitoring and Refresh** (33% happens)
- **Software Registry** (41% happens)
- **File Audit** (16% happens)

Happening, but not satisfactorily

- **Documentation** (26% satisfied),
- **Chain of custody** (27% satisfied),
- **Secure storage** (38% satisfied),
- **Quality assurance** (14% satisfied),
- **Data Visualization (**12.5% satisfied),
- **Metadata** (29% satisfied)

More Results Will be Posted to our Web site
https://sites.google.com/site/DataCurationNetwork

Twitter #DataCurationNetwork

Thanks for your participation!

Lisa Johnston (ljohnsto@umn.edu)