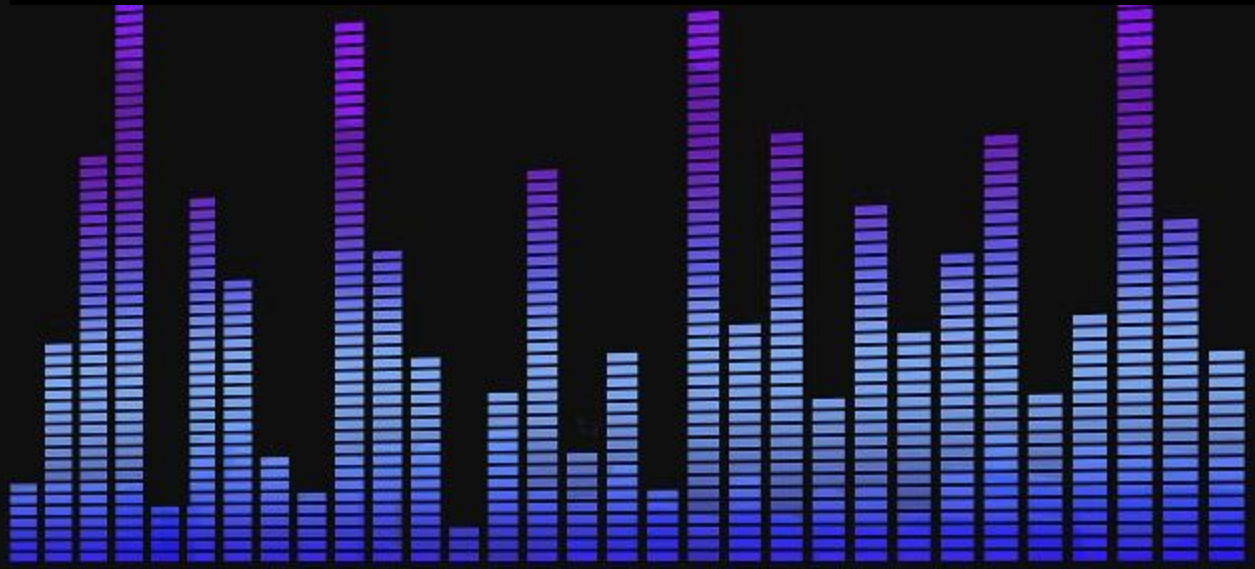
Data Curation Network



Planning a network of expertise model for curating research data in academic libraries

2016-2017

The Data Curation Network project is supported by a grant from the ALFRED P. SLOAN FOUNDATION.

Digital Library Forum

Data Curation Network

Lisa Johnston
University of Minnesota Libraries

11-8-2016

Rise of the Data Sharing Culture

Researchers are increasingly required/incentivised to share data

- Funder data sharing mandates
- Journal data sharing policies
- Disciplinary practices → emphasis on transparency and reproducibility

Data repositories: But...It's not enough to just keep the files!

Goal of data curation ⇒ Prepare and maintain research data in ways that make it findable, accessible, interoperable and reusable (FAIR),

Data curation = metadata, documentation, access, preservation, and more...

Data Curation Activities

- 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

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

Data Curation Network

The Data Curation Network will enable academic institutions to better support researchers that are faced with a growing number of requirements to ethically share their research data.

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

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.

Data Curation Network Partners

Planning the Data Curation Network

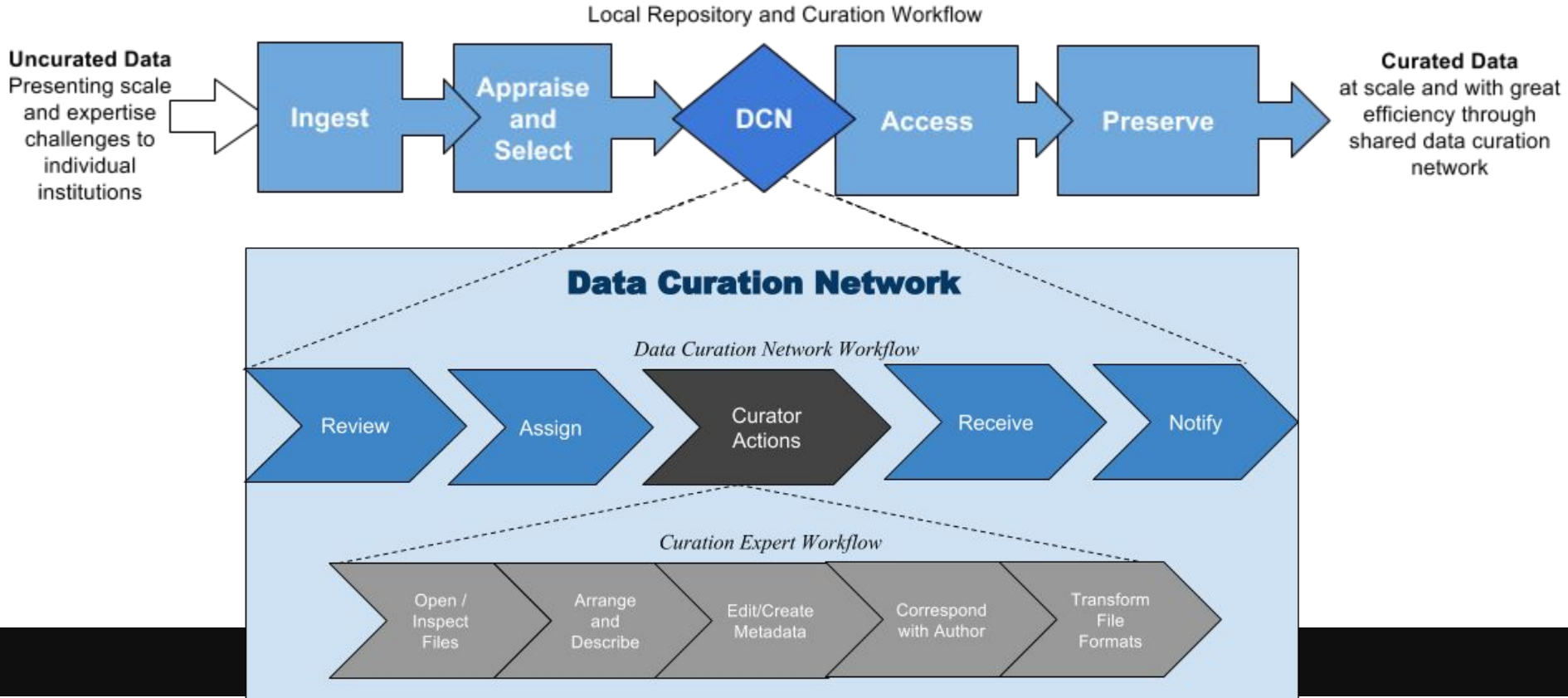
(Current) **Planning phase,** supported by the Alfred P. Sloan Foundation to:

- Develop a Data Curation Network ‘model of expertise’ for data curation staff that includes the projected staffing, costs, skills sets, and demand necessary for implementation.

(Future) **Pilot phase** will

- Test the model across our six institutions
- Plan for how to grow and sustain the Network

Draft Model for the Data Curation Network



Our Planning Phase activity to date

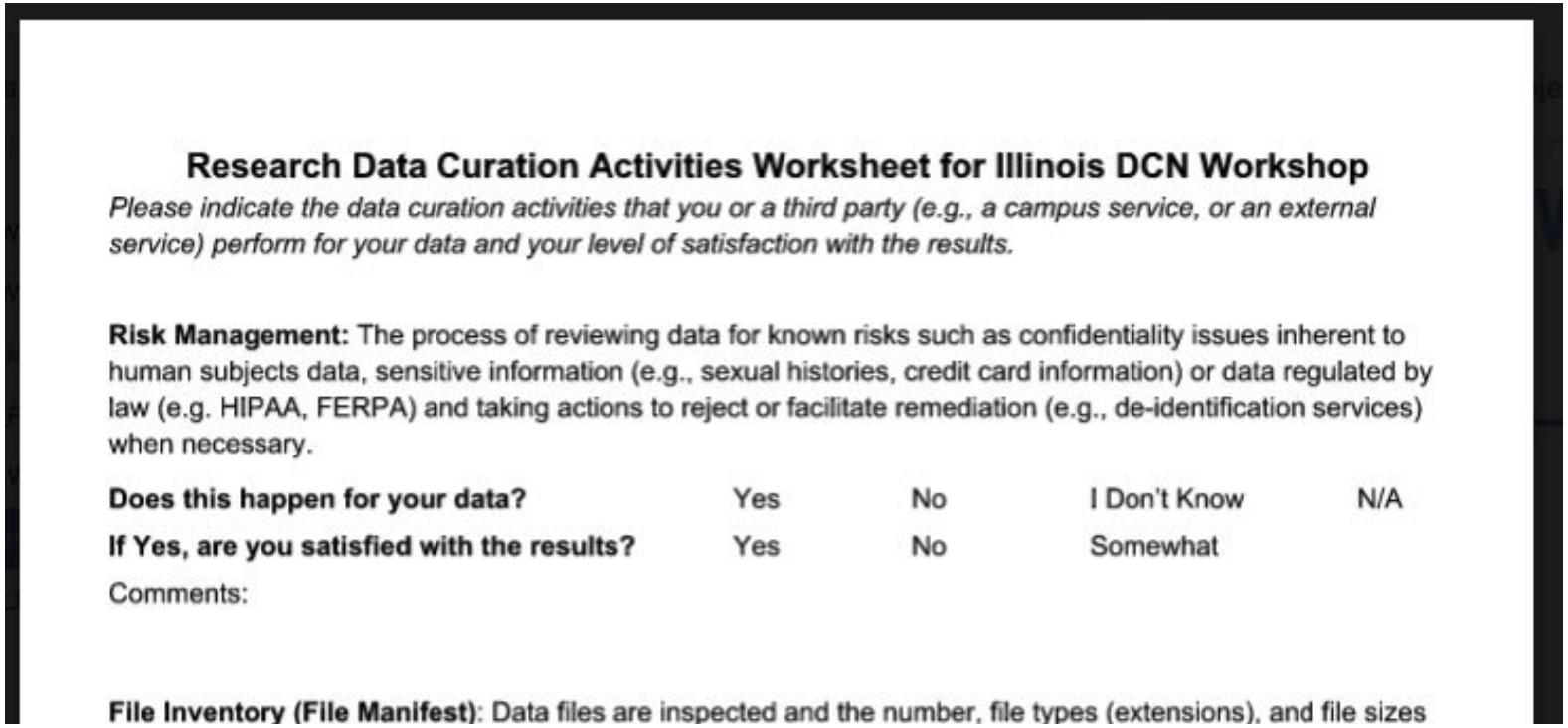
✓ **Summer** → Assessed infrastructure/policy/workflow differences and monitor the demand across institutions. Baseline report.

- **Underway Oct/Nov 2016** → Seek input from researchers to better understand how data curation services fit into their research workflow (focus groups).

- **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.

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:

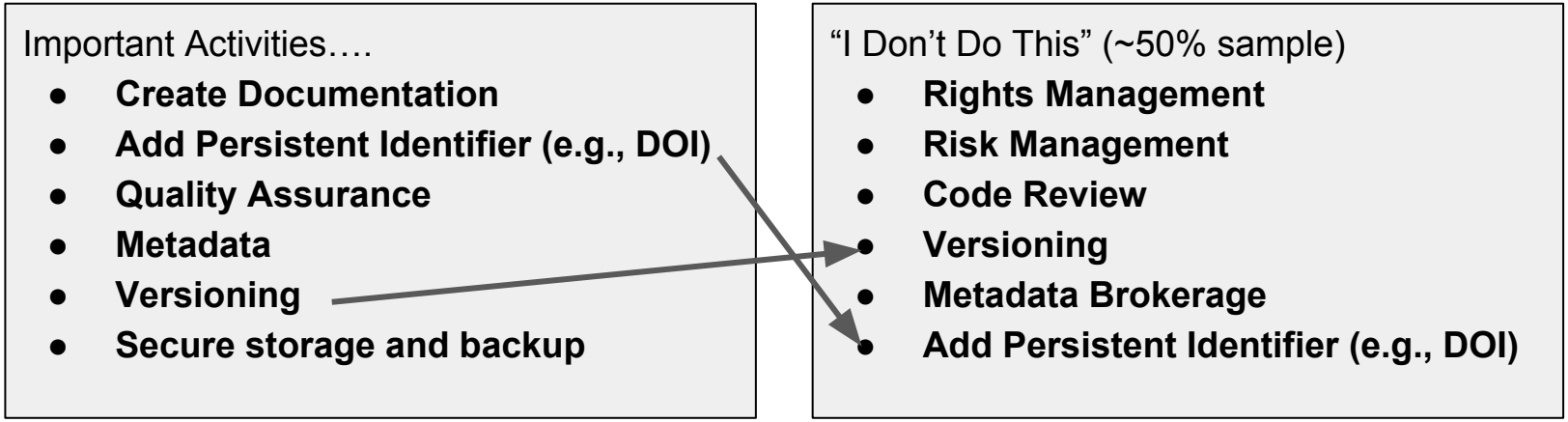
File Inventory (File Manifest): Data files are inspected and the number, file types (extensions), and file sizes

Results: Researcher Engagements

4 of 6 completed so far!

Goal: Identify value/importance placed on 40+ data curation activities in order to Identify gaps in important curation activities that are either not happening/well.

Preliminary findings…Gaps

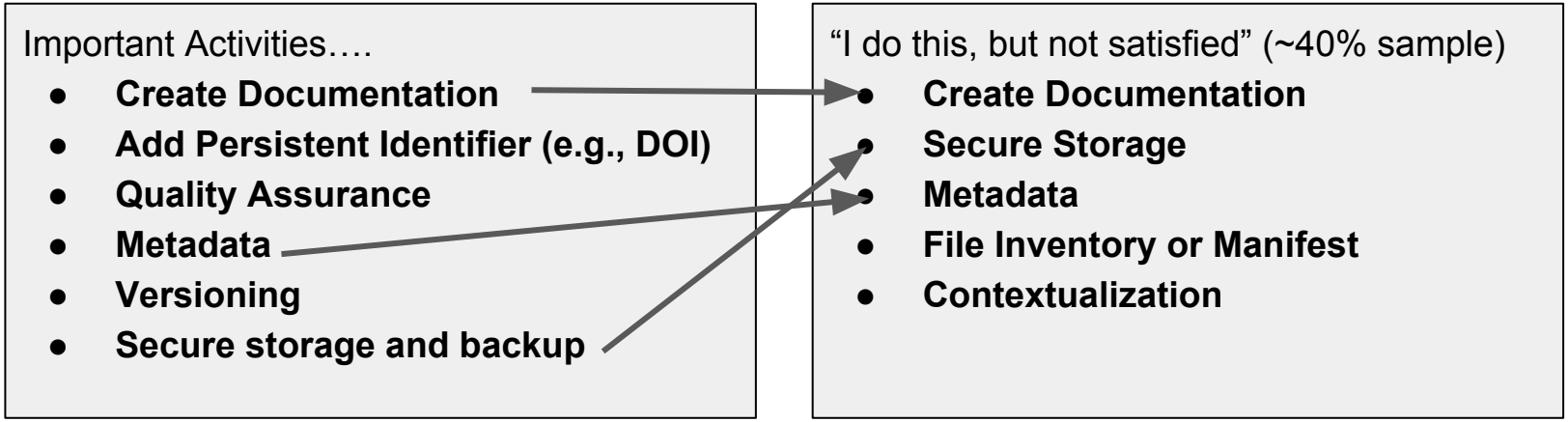


Results: Researcher Engagements

4 of 6 completed so far!

Goal: Identify value/importance placed on 40+ data curation activities (high!) in order to Identify gaps in important curation activities that are either not happening/well.

Preliminary findings…Gaps



Results: Repository Curation Workflows

	Repository Overview				Pre-ingest Curation?		Mediated vs Self-deposit?		Accept/Reject Stage?		Public	Post-ingest Curation?		
	Repository name	Repository URL	Date launched	Data holdings*	Consult only	Staging area for deposit	Mediated deposit	Self-deposit	Approval required to accept or reject	Auto accept	Go live here	Review metadata only	Review files and metadata	Add DOI
LIBRARIES University of Minnesota	Data Repository for the University of Minnesota (DRUM)	http://hdl.handle.net/11299/166578	Nov 2014	92	✓			✓	✓		✓		✓	✓
Cornell University Library	eCommons at Cornell	http://ecommons.cornell.edu	Fall 2002	108	✓		✓**	✓		✓	✓		✓†	✓**
	Illinois Data Bank	https://databank.illinois.edu	May 2016	11	✓			✓		✓	✓		✓†	✓
M Library	Deep Blue Data	https://deepblue.lib.umich.edu/data	Feb 2016	35	✓			✓		✓	✓		✓†	✓**
PennState University Libraries	Scholarsphere	https://scholarsphere.psu.edu	Fall 2012	439	✓			✓		✓	✓			
Washington University in St. Louis University Libraries	Digital Research Materials Repository	http://openscholarship.wustl.edu/data	Jan 2015	6	✓		✓	✓		✓	✓		✓	✓

* As of Sept 7, 2016 **On request †When possible

Results: Repository Technologies

	Technology Platform	Upload limits		Features						Service/Software Add Ons				Discovery Services						Metadata	
		Self-deposit	Other	Open Access	Versioning	Related Material Linking	API	OAI/PMH	Other	DOIs	Box.com Integration	Dropbox Integration	ORCID Integration	Web indexing: full text and metadata	Web indexing: metadata only	DataCite	SHARE	Data Citation Index (WoS)	re3data.org	Schema	Published?
LIBRARIES University of Minnesota	DSpace 5.5	2GB	Larger files must be mediated (up to 100GB per collection).	✓	✓*	✓*	✓	✓		✓				✓		✓	✓	✓	✓	Dublin Core	http://hdl.handle.net/11299/171761
Cornell University Library	DSpace 5.5	2GB	Larger files must be mediated. Total size per project per year is 10GB.	✓	✓*	✓		✓		✓*				✓					✓**	Dublin Core	Not yet published
	Custom-built Ruby on Rails webapp as a microservice to Medusa, a local preservation repository	15 GB (via Box.com)	Larger files may be ingested via a mediated mechanism.	✓*	✓*	✓*			Descriptive metadata editing	✓	✓		✓		✓	✓	✓**		✓	Compatible with DataCite Metadata Schema 3.1	https://www.ideals.illinois.edu/handle/2142/91019
LIBRARY	Hydra/Fedora Sufia 7	2GB	Larger files must be mediated. No defined limits. Exploring how to handle large data sets.	✓		✓**				✓	✓				✓					Dublin Core	Not yet published
PennState University Libraries	Hydra/Fedora (soon to be Sufia 7)	500 MB	Larger files via Dropbox (1.9 GB) or Box (5 GB). Up to 100 files and totaling less than 1 GB in size.	✓					Content can be private or restricted to Penn State only		✓	✓					✓		✓	Dublin Core	Not yet published
Washington University in St. Louis University Libraries	Digital Commons		Recommended 2 GB per file (not a hard limit - up to 10-20 GB).	✓	✓*	✓*		✓		✓				✓		✓	✓			Dublin Core	Not yet published

* Mediated **Forthcoming

Results: Repository Policies

	Features			Data Types			IR Policies				Data Licensing	
	Shibboleth Login	Embargoes	Preservation commitment	General	Private data	Real time or "streaming" data	Deposit license agreement	End-user terms of use	Disclaimer on data quality	Documentation required	Required / Optional	Licenses
LIBRARIES University of Minnesota	✓	✓	10 years via Rosetta	✓			✓	✓	✓	✓	Optional	Author-specified, CC0, CC-BY, CC-BY-NC
Cornell University Library	✓	✓	"...committed to preserving the binary form of the digital object..."	✓			✓				Optional	Author specified, CC0, CC-BY, CC-BY-ND, CC-BY-SA, CC- BY-NC, CC-BY-NC-ND, CC-By-NC-SA
		✓	Minimum of 5 years via the preservation repository, Medusa	✓			✓	✓	✓		Optional	CC0 and CC-BY encouraged, licence.txt allowed
LIBRARY			Minimum of 10 years with 3 tiers of commitment depending on format	✓			✓	✓	✓		Required	Author-specified, CC0, CC-BY, CC-BY-NC
PennState University Libraries	✓		For long-term preservation & access (no finite number of years expressed)	✓			✓	✓			Required	Default is CC-BY-NC-ND. Depositor may change to: CC-BY; CC-BY-SA; CC-BY-NC; CC-BY-ND; CC0
Washington University in St. Louis University Libraries		✓	Minimum 10 years followed by collection review in IR	✓			✓			✓	Optional	CC-BY, CC-BY-SA, CC-BY-NC, CC-BY-NC-SA, CC0

Thanks!

Web: https://sites.google.com/site/DataCurationNetwork

Twitter #DataCurationNetwork