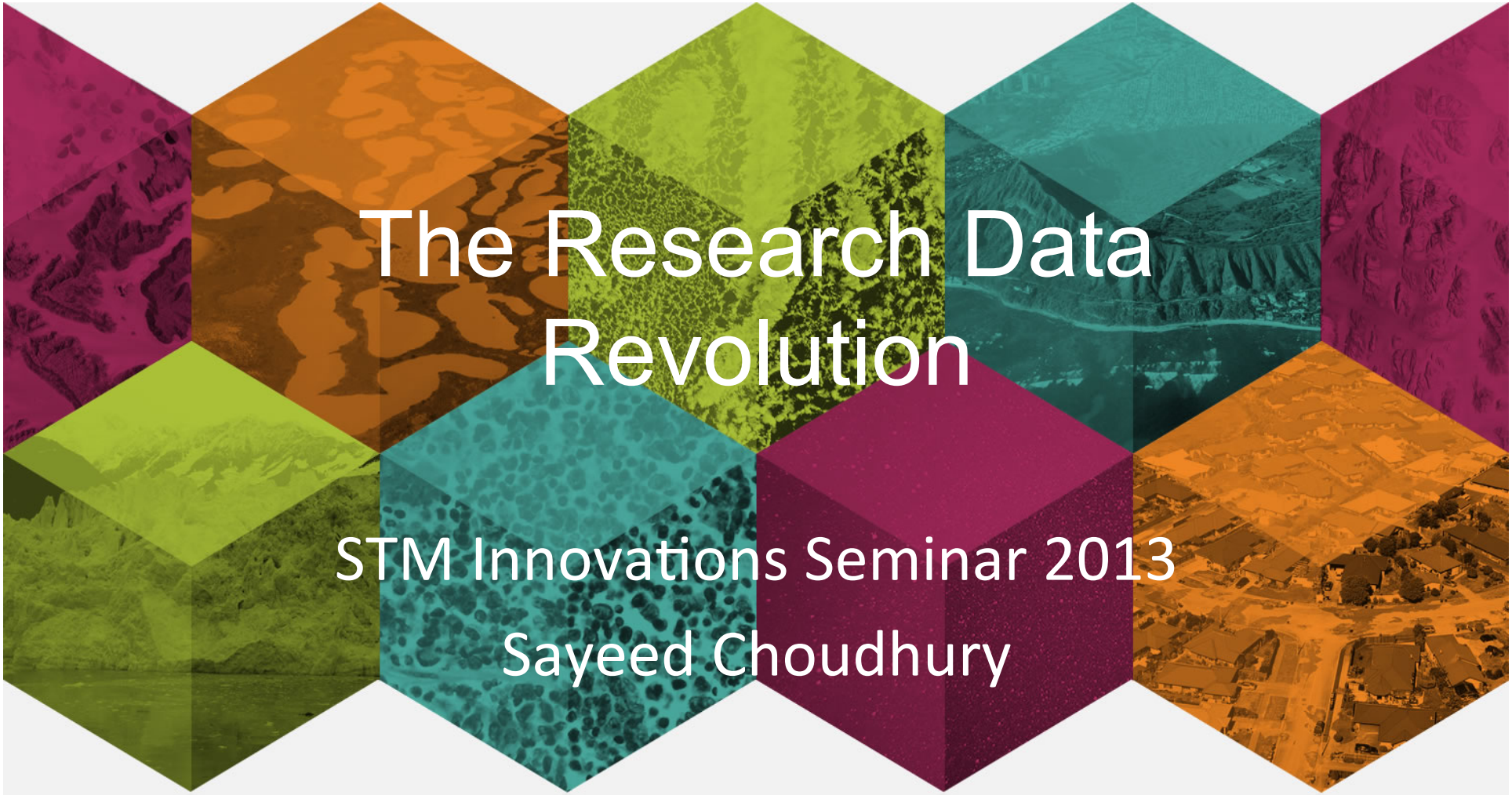


The Research Data Revolution

STM Innovations Seminar 2013

Sayeed Choudhury

DataConservancy

Data Conservancy (DC)

- One of five awards through US National Science Foundation's (NSF) DataNet program
- $10 million award to build national-scale data infrastructure
- Part of overall Cyberinfrastructure for the 21st Century program
- Culmination of over a decade of experience with Sloan Digital Sky Survey (SDSS) data

Data Conservancy Objectives

- Data Conservancy is a community that develops solutions for data preservation and sharing to promote cross-disciplinary re-use.

- Preserve – collect and take care of research data
- Share – reveal data's potential and possibilities
- Discover – promote re-use and new combinations

Is Data Really Different?

- “Data is the new oil” (stated in Qatar, European Commission, etc.)
- Data is the fourth factor of production (McKinsey)
- Todd Park estimates location sensitive apps generate $90 billion of value annually
- McKinsey estimates potential $3 trillion of economic value across seven sectors within US alone
- White House Office of Science and Technology Policy Executive Memorandum
- White House Open Government Initiative

Implications for Libraries

- Libraries are built on three pillars – collections, services and infrastructure."
  - Winston Tabb, Sheridan Dean of University Libraries, Johns Hopkins University

- Consider data and libraries (and publishers) from these three pillars

Collections

- Data are a new form of collections – though they are fundamentally different in nature
- Created or converted to digital format for processing by machines
- Entirely new methods are required
- New form of special collections

"Big Data"

- What is Big Data?
- There are definitions based on the "V's" of Big Data (e.g., volume, velocity, variety)
- What is clear is that it's different from "spreadsheet science" (or long-tail science)
- For me, if a community's ability to deal with data is overwhelmed, it's "Big Data" – it's more about "M's" (methods or lack thereof) than "V's"

Services

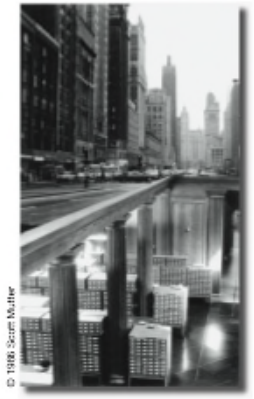
- There is a core of services that span across data from different disciplines, contexts, etc. – archiving is a good example
- If data collections are basically open, libraries may need to differentiate themselves by the services they offer
- Combination of machine and human mediated services
- There will be a set of services that only “experts” will be able to offer

Data Management Layers

Layers	Characteristics	Implication for PI	Implication relative to NSF
Curation	Adding value throughout life-cycle	• Feature Extraction • New query capabilities • Cross-disciplinary	• Competitive advantage • New opportunities
Preservation	Ensuring that data can be fully used and interpreted	• Ability to use own data in the future (e.g. 5 yrs) • Data sharing	• Satisfies NSF needs across directorates
Archiving	Data protection including fixity, identifiers	• Provides identifiers for sharing, references, etc.	• Could satisfy most NSF requirements

Storage	Bits on disk, tape, cloud, etc.
	Backup and restore
		• Responsible for:
		  • Restore
		  • Sharing
		  • Staffing
			• Could be enough for now but not near-term future

Understanding Infrastructure: Dynamics, Tensions, and Design



Report of a Workshop on "History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures"

Paul N. Edwards
Steven J. Jackson
Geoffrey C. Bowker
Cory P. Knobel

January 2007

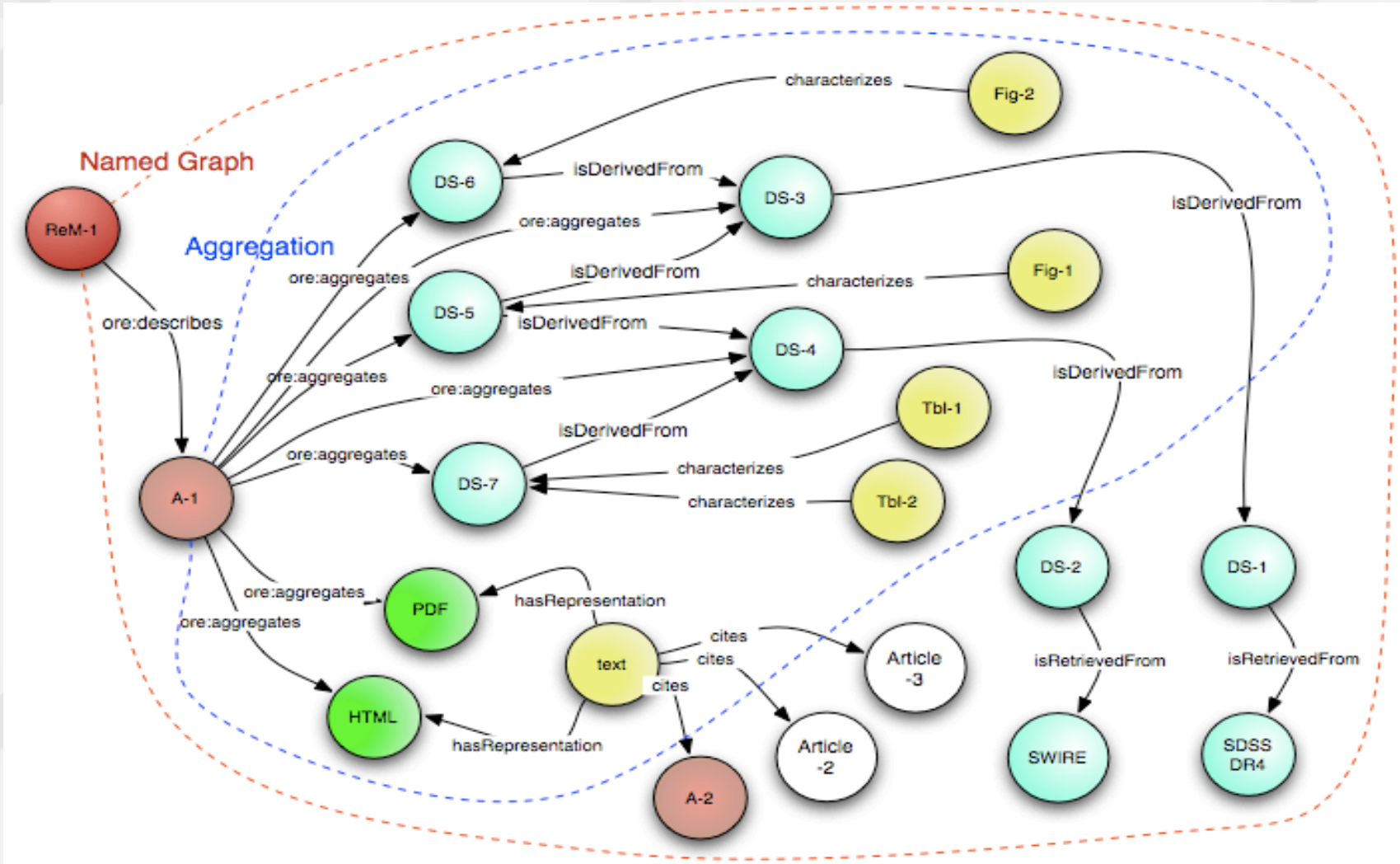
NSF Grant 0630263 • Human and Social Dynamics • Computer and Information Science and Engineering • Office of Cyberinfrastructure

…not a rigid road map but principles of navigation. There is no one way to design cyberinfrastructure, but there are tools we can teach the designers to help them appreciate the true size of the solution space – which is often much larger than they may think, if they are tied into technical fixes for all problems.

Infrastructure

- Data will require fundamentally new systems and infrastructure
- Institutional repositories can be useful gateways but not long-term solutions (particularly for “Big Data”)
- Libraries will need to operate at scale through an integrated, ecosystem approach to infrastructure
- Customized (“human mediated”) services most effective as interpretative layer on machine based services

Information graph using OAI-ORE



What about Publishers?

- "Publishing is about content, not format."
  - Wendy Queen, Associate Director of Project Muse, Johns Hopkins University Press
- *No one can claim a specific role or act with a sense of entitlement when it comes to data*
- The future of data curation is a competition between information graphs

Acknowledgements

- NSF Award OCI-0830976
- Sheridan Libraries and JHU financial support

- http://dataconservancy.org
- http://dmp.data.jhu.edu -- JHU DMS
- http://www.dlib.org/dlib/september12/mayernik/09mayernik.html -- DC blueprint document
- https://www.youtube.com/watch?v=F6iYXNvCRO4 -- data management layer stack model