Library Curation of Long-tail Science Data

October 27, 2010
CODATA
Cape Town, South Africa

P. Bryan Heidorn
Director
University of Arizona
School of Information Resources and Library Science

Thesis

- Large amounts of data remain uncurated
- Most of that data is from small data sets and is currently largely invisible – Dark Data
- This data should be curated locally but not by scientists

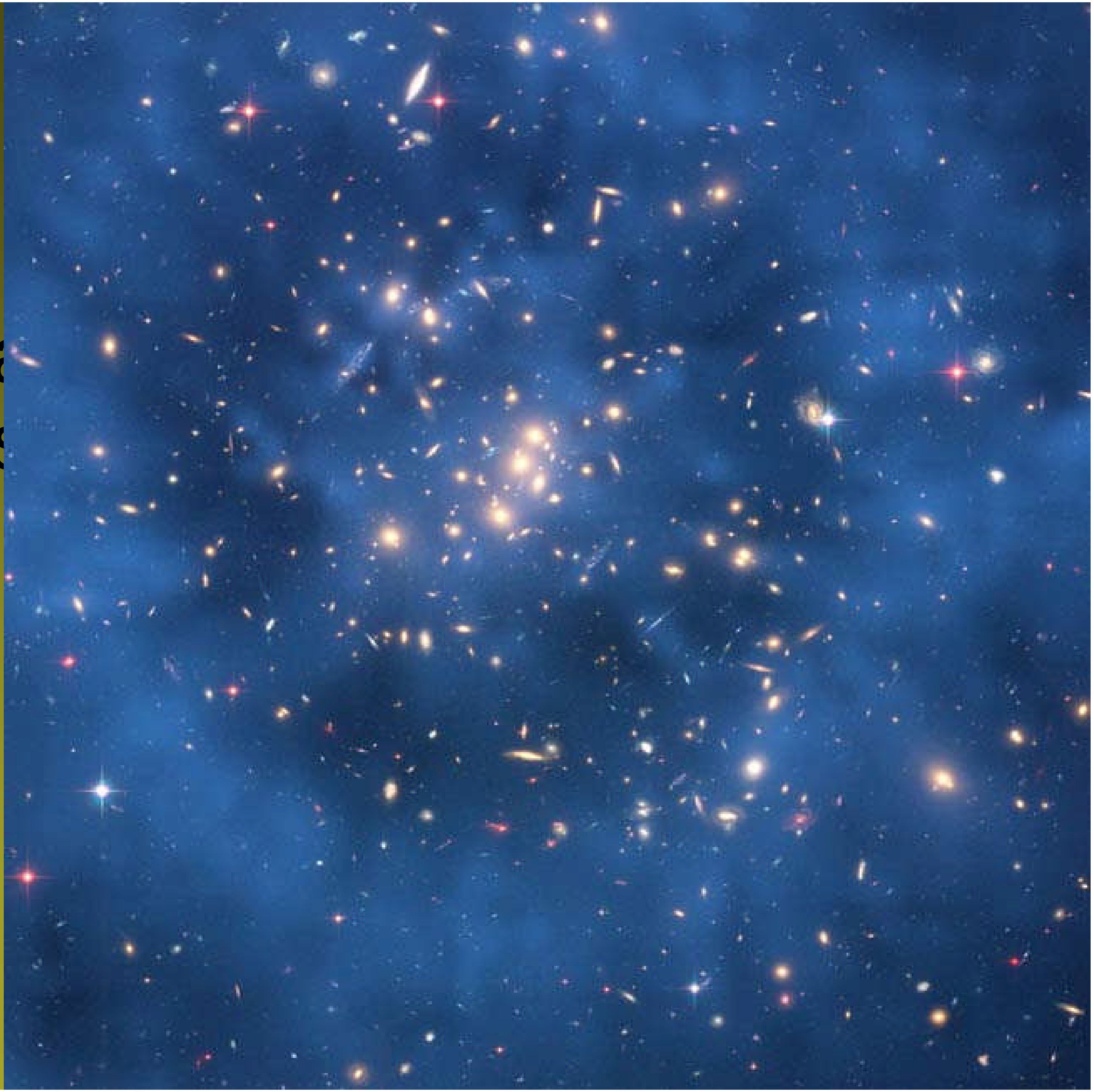
Why Libraries

- Long history of scholarly data management
- Skills overlap such a development of metadata standards, ontologies, controlled vocabularies, thesauri
- Long-lived institutions
- Overlap with museums and archives

The problem

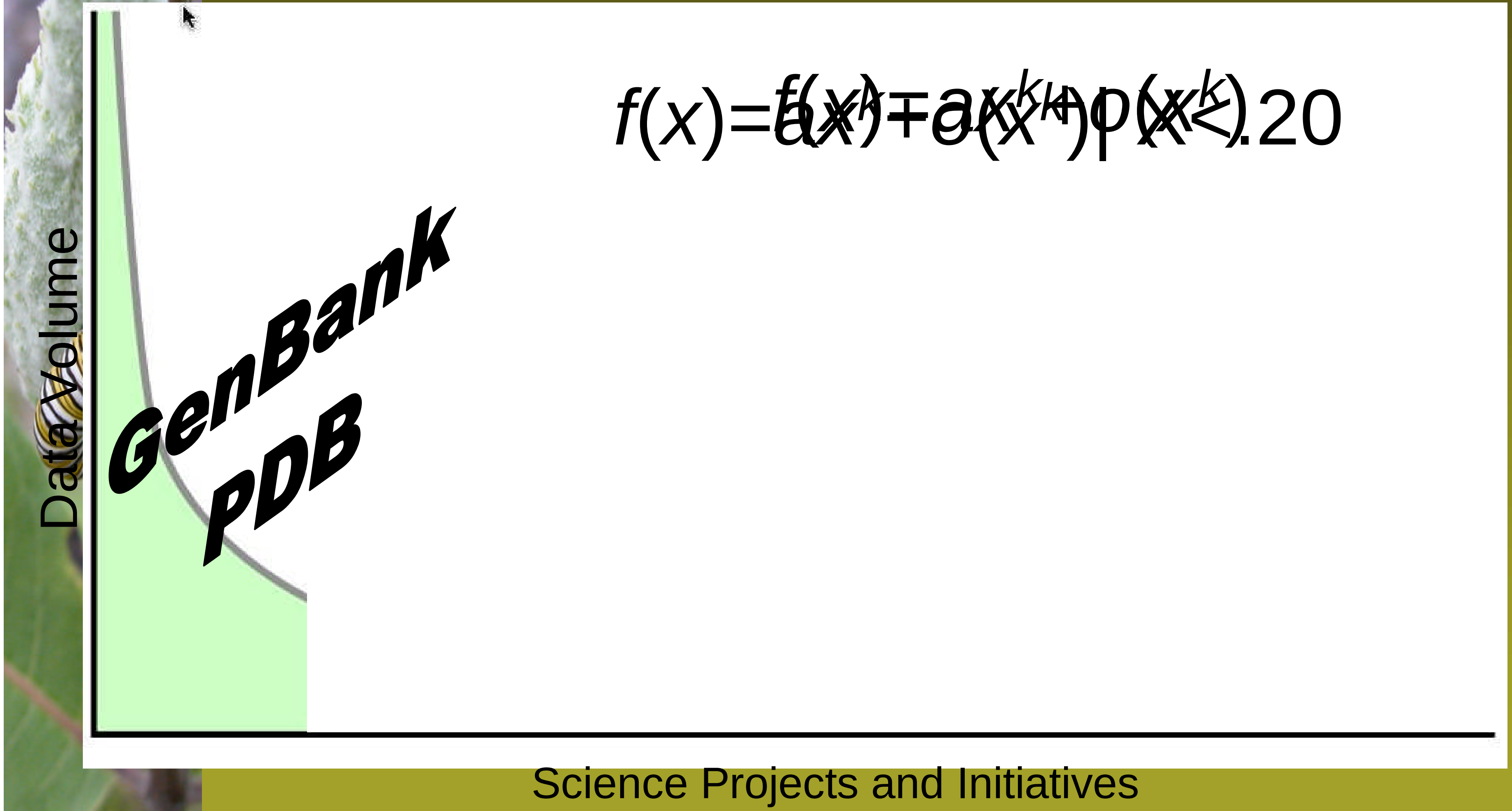
- Information is not in accessible format
- Computer Science, Information Science and Technology has not addressed the problem

Dark da
is/was



Hubble Space Telescope composite image "ring" of dark matter in the galaxy cluster Cl 0024+17

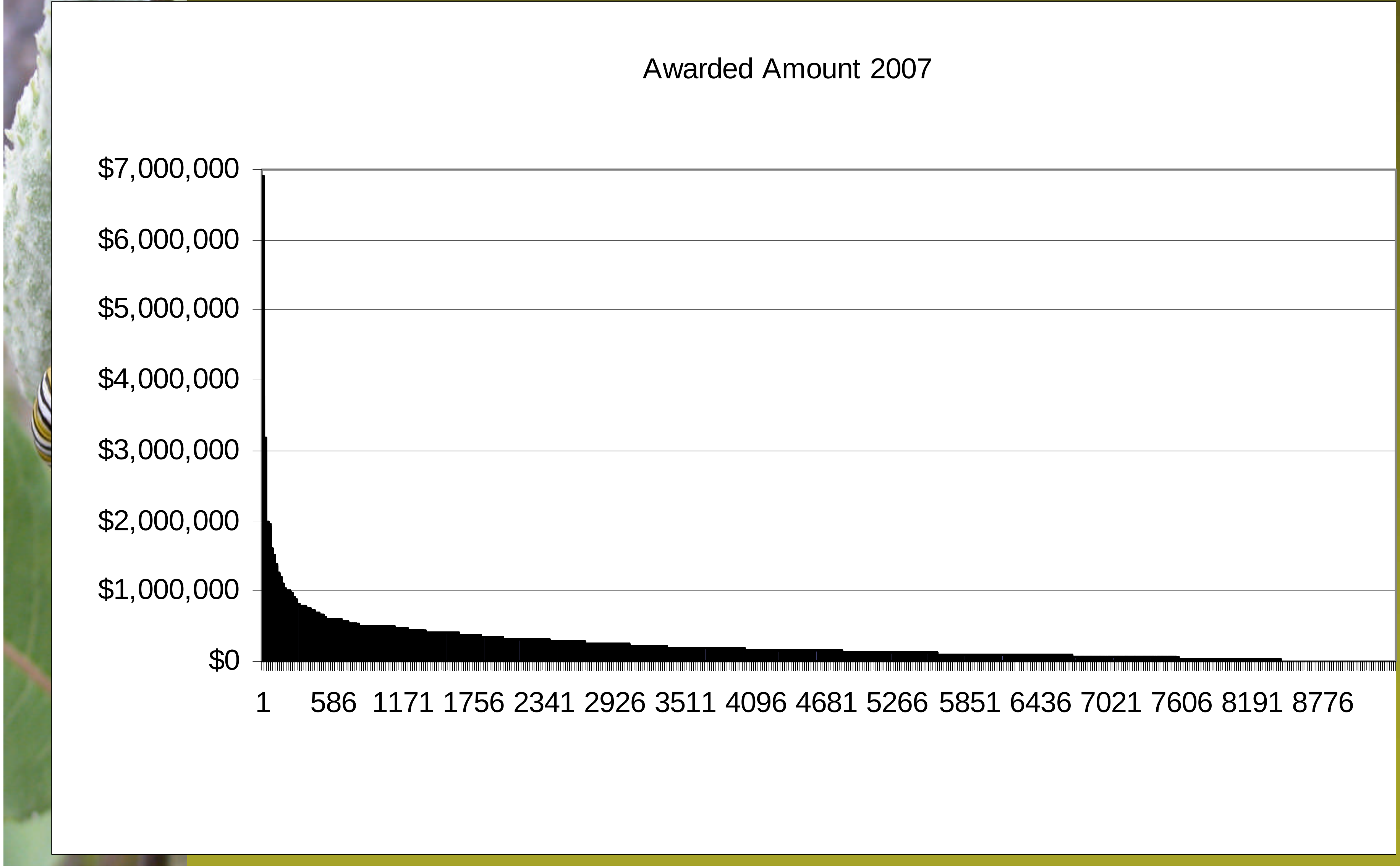
Power Law of Science Data



$$f(x)=ax^k+o(x^k)\mid x<.20$$

$$f(x)=ax^k+o(x^k)$$

Does NSF’s Data Follow the Power Law?



20-80
Rule The small are big!

Total Grants	9347 $2,137,636,716	
	20%	80%
Number Grants	1869	7478
Total Dollars	$1,199,088,125	$938,548,595
Range	$6,892,810-$350,000	$350,000-$831

Related Ideas

- John Porter:
  - Deep verses Wide databases
- Swanson:
  - Undiscovered Public Knowledge
- Science Commons:
  - Big Verses Small science

Small data is big science

- Because it is high volume
- Because it is information rich – high entropy
- While needs of large data are understood small data and integration are not understood
- Heidorn, P. Bryan (2008). Shedding Light on the Dark Data in the Long Tail of Science. Library Trends 57(2) Fall 2008 . Institutional Repositories: Institutional Repositories: Current State and Future. Edited by Sarah Sheeves and Melissa Cragin. (http://hdl.handle.net/2142/9127).

Where to find dark data

- Literature/Biodiversity Heritage Library
- Museum Specimens
- Field notes
- (Un)Experimental data sets
- Citizen Observations

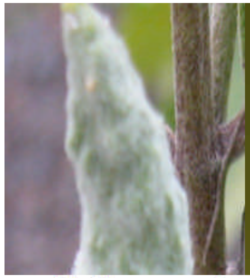
What is dark data good for?

- Ecological Niche Modeling
- Climate Change niche change prediction
- Taxonomic Name Resolution
- Literature Search Support
  - Taxonomic intelligence
  - Key-like – character searching
- Phenology and Phenology change
- Food-web / trophic level

New Information Disciplines

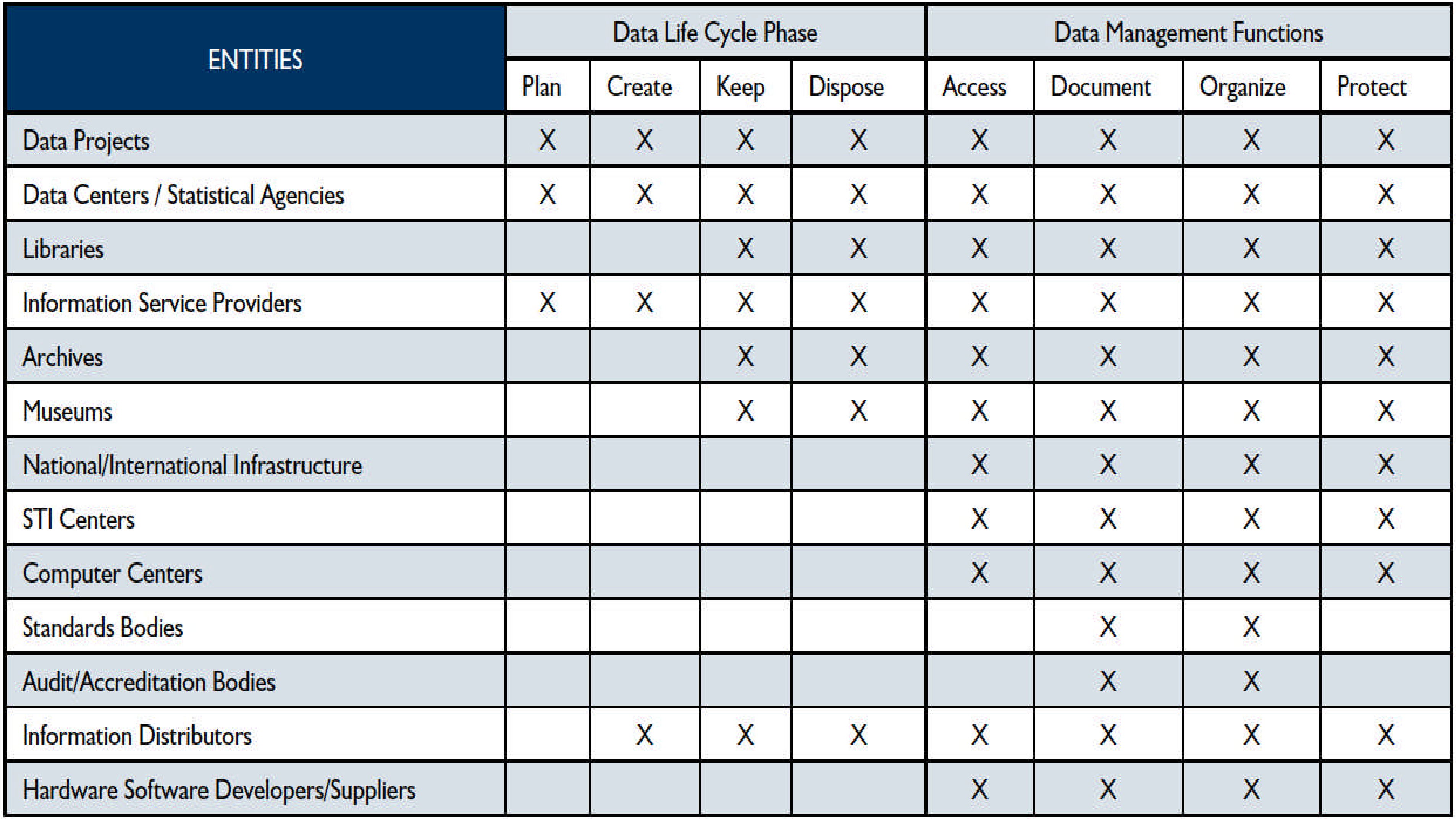
- **Digital Curator**: an expert knowledgeable of and with responsibility for the content of a digital collection(s)
- **Digital Archivist**: an expert competent to appraise, acquire, authenticate, preserve, and provide access to records in digital form
- **Data Scientists**: the information and computer scientists, database and software engineers and programmers, disciplinary experts, expert annotators, and others, who are crucial to the successful management of a digital data collection

(Long Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century, report of the National Science Board, September, 2005)

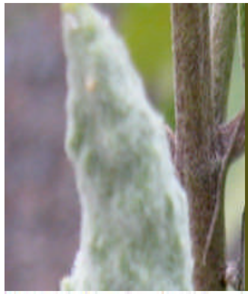


Library Roles

Exhibit C-6. Entities by Life Cycle Phase/Function

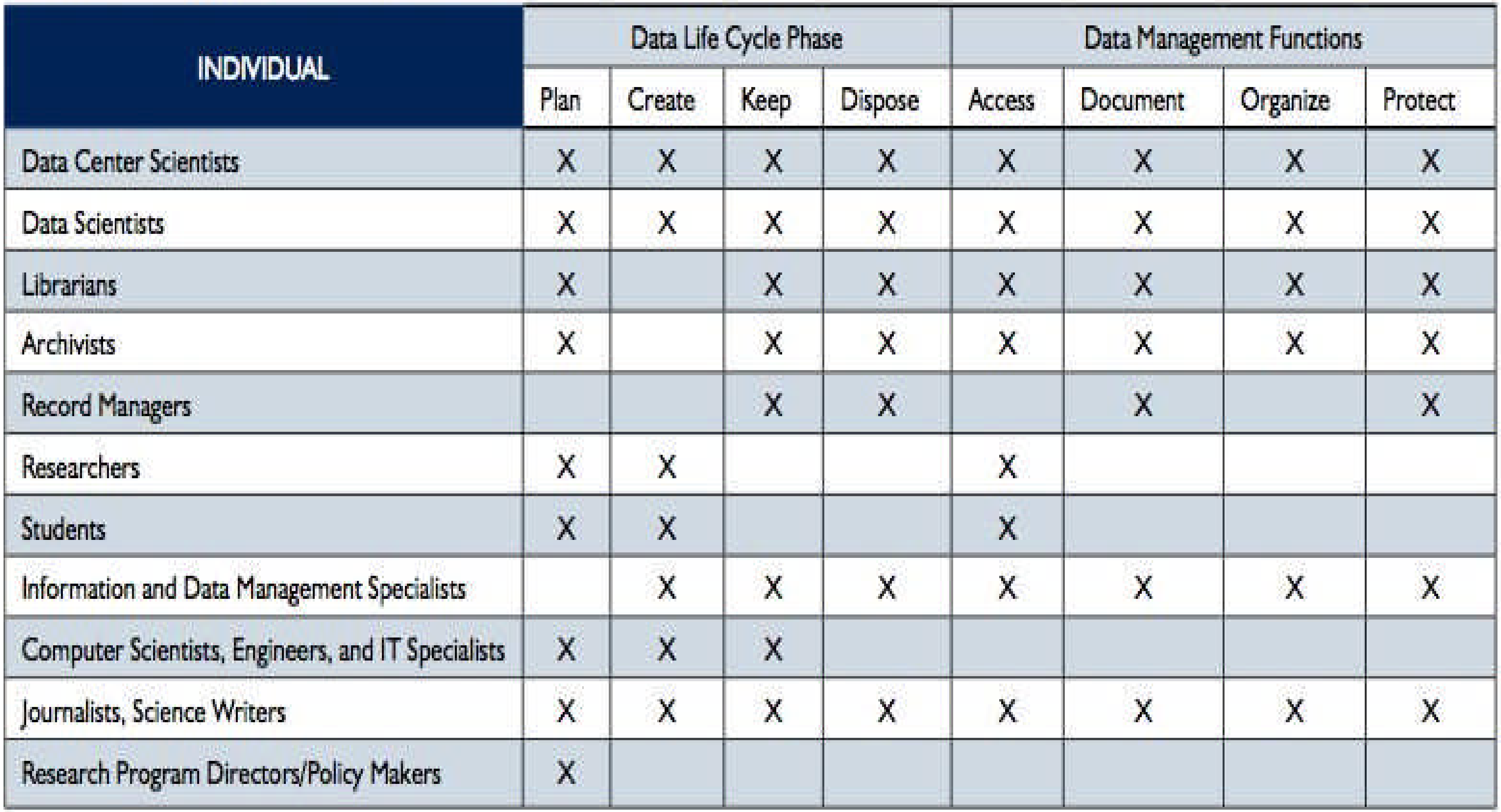


ENTITIES	Data Life Cycle Phase				Data Management Functions			
	Plan	Create	Keep	Dispose	Access	Document	Organize	Protect
Data Projects	X	X	X	X	X	X	X	X
Data Centers / Statistical Agencies	X	X	X	X	X	X	X	X
Libraries			X	X	X	X	X	X
Information Service Providers	X	X	X	X	X	X	X	X
Archives			X	X	X	X	X	X
Museums			X	X	X	X	X	X
National/International Infrastructure					X	X	X	X
STI Centers					X	X	X	X
Computer Centers					X	X	X	X
Standards Bodies						X	X	
Audit/Accreditation Bodies						X	X	
Information Distributors		X	X	X	X	X	X	X
Hardware Software Developers/Suppliers					X	X	X	X



Library Skills

Exhibit C-5. Individuals by Life Cycle Phase/Function



INDIVIDUAL	Data Life Cycle Phase				Data Management Functions			
	Plan	Create	Keep	Dispose	Access	Document	Organize	Protect
Data Center Scientists	X	X	X	X	X	X	X	X
Data Scientists	X	X	X	X	X	X	X	X
Librarians	X		X	X	X	X	X	X
Archivists	X		X	X	X	X	X	X
Record Managers			X	X		X		X
Researchers	X	X			X			
Students	X	X			X			
Information and Data Management Specialists		X	X	X	X	X	X	X
Computer Scientists, Engineers, and IT Specialists	X	X	X					
Journalists, Science Writers	X	X	X	X	X	X	X	X
Research Program Directors/Policy Makers	X							