A Perspective on the Structure of the EarthCube Community

Chaitan Baru
San Diego Supercomputer Center
UC San Diego

Stakeholder Alignment for EarthCube

- NSF-funded studies
  - Joel Cutcher-Gershenfeld, University of Illinois, Urbana-Champaign
  - Stakeholder Alignment in Socio-Technical Systems, NSF OCI RAPID 1229928,
  - Stakeholder Alignment for EarthCube, NSF SciSPR-STS-OCI-INSPIRE 1249607
  - Enabling Transformation in the Social Sciences, Geosciences, and Cyberinfrastructure

Respondent Profile	Earth-Scope Work-shop (n=21)	Plate Tec-tonics Work-shop (n=24)	Early Career Work-shop (n=37)	Earth-Cube Website (n=126)	Data Centers (n=576)
U.S. Institutional Affiliation	95.2%	100%	100%	88%	77.1%
International Institutional Affiliation	4.8%	0%	0%	12%	22.9%
Female	28.6%	0%	40.5%	26.1%	27.9%
Male	71.4%	100%	59.5%	73.9%	72.1%
Under 5 years of experience	4.8%	12.5%	5.4%	2.4%	12.9%
5-10 years of experience	19.0%	12.5%	37.0%	17.5%	20.5%
11-20 years of experience	14.3%	29.2%	56.8%	27.0%	28.5%
Over 20 years of experience	61.9%	45.8%	0%	53.2%	37.9%
Never heard of EarthCube	0%	12.5%	21.6%	14.3%	54.3%
Aware, but no direct experience	4.8%	37.5%	32.4%	21.4%	29.4%
Visited website	43.8%	12.5%	24.3%	12.7%	10.4%
Participated in discussions	31.3%	16.75	13.5%	15.1%	3.7%
Actively involved with EarthCube	6.3%	16.7%	8.1%	28.6%	1.7%
Leadership role in EarthCube	12.5%	4.2%	0%	7.1%	0.5%

Responses on Data Access, Use, and EarthCube (all responses normalized on a scale of zero to one, with one being most positive) Mean (s.d.)	Earth-Scope Work-shop (n=21)	Plate Tec-tonics Work-shop (n=24)	Early Career Work-shop (n=37)	Earth-Cube Web-site (n=126)	Data Cen-ters (n=576)
How IMPORTANT is it for you to find, access, and/or integrate multiple datasets, observations, visualization tools, and/or models in your field or discipline?	.94 (.10)	.91 (.14)	.89 (.19)	.89 (.18)	.87 (.20)
How EASY is it for you to find, access, and/or integrate multiple datasets, observations, visualization tools, and/or models in your field or discipline?	.42 (.25)	.35 (.23)	.33 (.30)	.41 (.25)	.42 (.24)
How IMPORTANT is it for you to find, access, and/or integrate multiple datasets, observations, visualization tools, and/or models that span different fields or disciplines ?	.84 (.23)	.74 (.27)	.77 (.31)	.79 (.24)	.73 (.27)
How EASY is it for you to find, access, and/or integrate multiple datasets, observations, visualization tools, and/or models that span different fields or disciplines?	.32 (.24)	.27 (.25)	.20 (.24)	.30 (.24)	.32 (.22)
Please use the scale ranging from "Inadequate" to "Adequate" to assess the present suite of publicly accessible datasets, data analysis tools, and modeling software – to what degree is it adequate for your research and education needs?	.40 (.20)	.32 (.17)	.40 (.26)	.42 (.24)	.49 (.26)

Responses on Data Access, Use, and EarthCube (all responses normalized on a scale of zero to one, with one being most positive) Mean (s.d.)	Earth-Scope Work-shop (n=21)	Plate Tec-tonics Work-shop (n=24)	Early Career Work-shop (n=37)	Earth-Cube Web-site (n=126)	Data Cen-ters (n=576)
In 5-7 years, I anticipate that EarthCube will result in substantially increased productivity for me and others whose work is similar to mine.	.60 (.28)	.73 (.21)	.67 (.22)	.65 (.22)	.62 (.25)
In 5-7 years, I anticipate that EarthCube will result in substantially expanded research opportunities for me and others whose work is similar to mine.	.63 (.30)	.73 (.21)	.73 (.20)	.69 (.22)	.65 (.25)
In 5-7 years, I anticipate that EarthCube will result in substantially expanded educational tools for me and others whose work is similar to mine.	.66 (.24)	.81 (.16)	.68 (.22)	.68 (.22)	.67 (.23)
In 5-7 years, I anticipate that EarthCube will result in substantially expanded capabilities to integrate multiple sources of data, datasets, observations, visualization, and models.	.64 (.28)	.80 (.18)	.75 (.19)	.73 (.20)	.69 (.24)

Responses on Data Access, Use, and EarthCube (all responses normalized on a scale of zero to one, with one being most positive) Mean (s.d.)	Earth-Scope Work-shop (n=21)	Plate Tec-tonics (n=24)	Early Career Work-shop (n=37)	Earth-Cube Web-site (n=126)	Data Cen-ters (n=576)
My employer/org. will most likely value and reward my efforts in the shaping and dev. of EarthCube.	.50 (.34)	.38 (.29)	.45 (.36)	.49 (.32)	.40 (.30)
My employer/org. will most likely see my participation in the shaping and dev. of EarthCube as an integral part of my job.	.45 (.33)	.35 (.29)	.43 (.34)	.43 (.32)	.34 (.29)
My contributions to the shaping and dev. of EarthCube will most likely be recognized and highly valued by colleagues in my field/domain.	.54 (.32)	.51 (.30)	.48 (.32)	.52 (.26)	.46 (.28)
There is currently a high degree of cooperation and sharing of data, models, and simulations among geoscientists.	.51 (.16)	.39 (.20)	.40 (.23)	.40 (.25)	.48 (.24)
There is currently sufficient communication and collaboration between geoscientists and those who develop cyberinfrastructure tools and approaches to advance the geosciences.	.31 (.22)	.25 (.17)	.26 (.22)	.29 (.22)	.34 (.23)
There is currently sufficient geoscience end-user knowledge and training so they can effectively use the present suite of cyber-infrastructure tools and train their students/colleagues in its use	.29 (.18)	.21 (.14)	.24 (.21)	.24 (.19)	.32 (.23)

Top Ten Barriers to Sharing Data (categories):

1. **No time/Needs too much QA/QC**
2. **No repository/No known repository**
3. **Inadequate standards/No standardized formats**
4. **Want to publish first/Don't want to be scooped**
5. **File size too large/Server size too small**
6. **Classified/proprietary/Agency or company restrictions**
7. **No credit/No incentive to share**
8. **Cost**
9. **Not sure what to do**
10. **Not sure anyone wants it**

Note: Approximately 45% of respondents did not respond to the open ended question "It is difficult to share my data because. . . " and another 6% said it was easy to share their data. The balance of responses were organized into the above categories; some individuals cited more than one reason (all of which were tabulated).

Responses on Data Access, Use, and EarthCube (all responses normalized on a scale of zero to one, with one being most positive) Mean (s.d.)	Earth-Scope Work-shop (n=21)	Plate Tec-tonics (n=24)	Early Career Work-shop (n=37)	Earth-Cube Web-site (n=126)	Data Cen-ters (n=576)
There are presently substantial unresolved issues around the access and use of geoscience data housed in federal government repositories.	.69 (.22)	.68 (.24)	.67 (.24)	.77 (.24)	.67 (.24)
There are presently substantial unresolved issues around the access and use of data held by invest. funded by NSF and other federal agencies.	.68 (.29)	.74 (.22)	.66 (.29)	.68 (.26)	.61 (.25)
There are presently substantial unresolved issues around the attribution/authorship of data in the use of data housed or retrieved by data aggregating systems like EarthCube.	.61 (.30)	.72 (.17)	.73 (.23)	.63 (.25)	.59 (.23)
The EarthCube initiative should specify guidelines so there is more interoperability and uniformity in discovering, accessing, sharing, and disseminating geoscience data.	.76 (.26)	.87 (.19)	.88 (.23)	.84 (.23)	.84 (.21)
The EarthCube initiative should specify guidelines so there is more interoperability and uniformity in geoscience data analysis tools, methods, & models.	.75 (.24)	.82 (.21)	.84 (.19)	.76 (.27)	.79 (.25)
The EarthCube initiative should specify guidelines so there is more interoperability and uniformity in geoscience visualization tools.	.75 (.23)	.84 (.19)	.81 (.20)	.75 (.26)	.78 (.25)

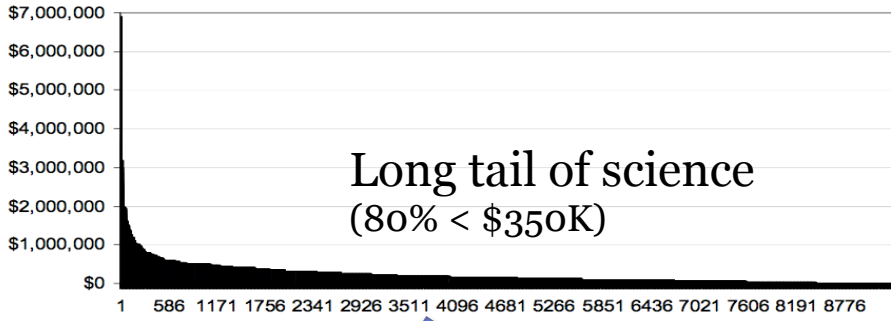
Long Tail Data

- Shedding Light on the Dark Data in the Long Tail of Science [B. Heidorn 2008]
- "Dark data"
  - "Data that is not carefully indexed and stored so it becomes nearly invisible to scientists and other potential users and therefore is more likely to remain underutilized and eventually lost."
- "Long tail": 80% of projects < $350K
  - Not much resources for data management

The EarthCube Data Challenge

"Big head" of science
- Established Data Archives, Databases, Data Repositories

- Projects with resources for data management, metadata creation, QA/QC processes, etc.



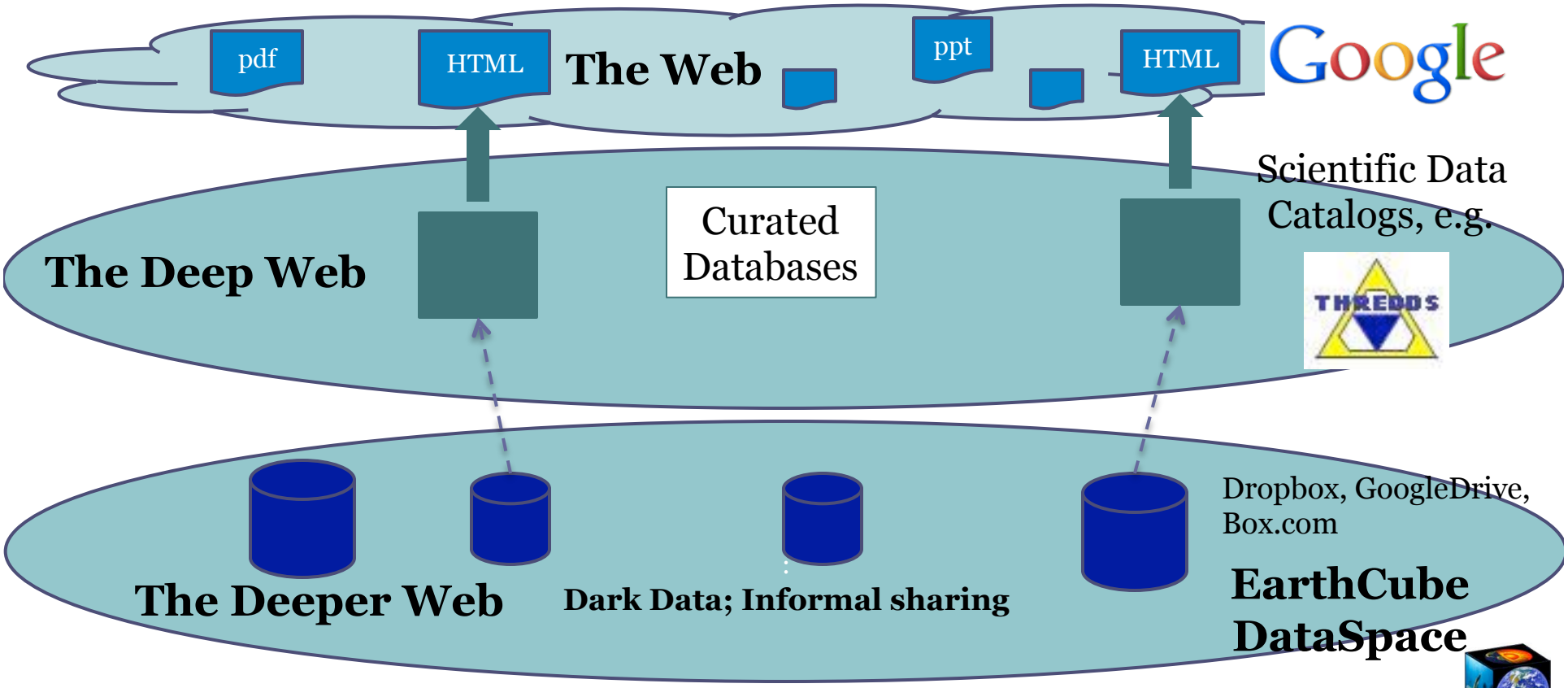
Awarded Amount 2007 [Bryan Heidorn 2008]

EarthCube Data Discovery, Access, Mining Services

EarthCube DataSpace

- Cloud-based service to support data discovery, access, and mining
- Lower barrier to entry for data sharing and reuse
- Use principles like “metadata ASAP” to encourage submission of data
- Enable creation of “Curation Co-ops” among communities, sub-communities
- Serve the NSF Data Management Plan requirement

Reaching into the Long Tail



Requirements for EarthCube Data Services

- Discovery
  - Easily discover *all* available data
  - Especially, all available data that is relevant to me!
  - → Determining "intent" in a computer science challenge
- Access
  - Make it easy to access discovered data
  - Including transforming data as appropriate
- Mining
  - Access data by its contents, not just metadata
  - Analyze/mine large volumes of data for patterns, anomalies, etc., summarize data for me.