# Between Land and Sea: Divergent Data Stewardship Practices in Deep-Sea Biosphere Research # IN53C-1577

Rebekah Cummings[1] and Peter Darch[2]
University of California, Los Angeles, Department of Information Studies
[1]rebekah.cummings@utah.edu, [2]petertdarch@ucla.edu

## ABSTRACT

Deep-sea biosphere researchers collect data in a variety of ways for a variety of reasons, but not all data are managed, stored, and shared equally. While data generated on International Ocean Discovery Program (IODP) expeditions are highly structured, professionally curated, and widely shared, the data practices of deep-sea biosphere laboratories are far more localized and ad hoc, resulting in what is referred to as "dark" data. An in-depth study of the divergent data practices of deep-sea biosphere researchers, supported by our study of data practices in other fields, allows us to:

- Better understand the social and technical forces that shape data stewardship throughout the data lifecycle;
- Develop policy, infrastructure, and best practices to improve data stewardship in small labs;
- Track provenance of datasets from IODP cruises to labs and publications;
- Create linkages between laboratory findings, cruise data, and IODP samples.

We present findings from the first year of a case study of the Center for Dark Energy Biosphere Investigations (C-DEBI), an NSF Science and Technology Center that studies life beneath the seafloor.

## RESEARCH QUESTIONS

1. How do data lifecycles differ between big and little science?
2. What tools and services do researchers in small- and medium-sized laboratories (SMLs) need to manage and share their data effectively?
3. How can we create better linkages between samples and data from IODP cruises with the laboratory data?

## METHODS

Our methods include interviews, participant observation, online ethnography, and document analysis. To date, we have conducted 27 interviews with C-DEBI scientists working on a variety of projects, from different disciplinary backgrounds, and at various stages in their careers.

Interview questions pertinent to the presented research:

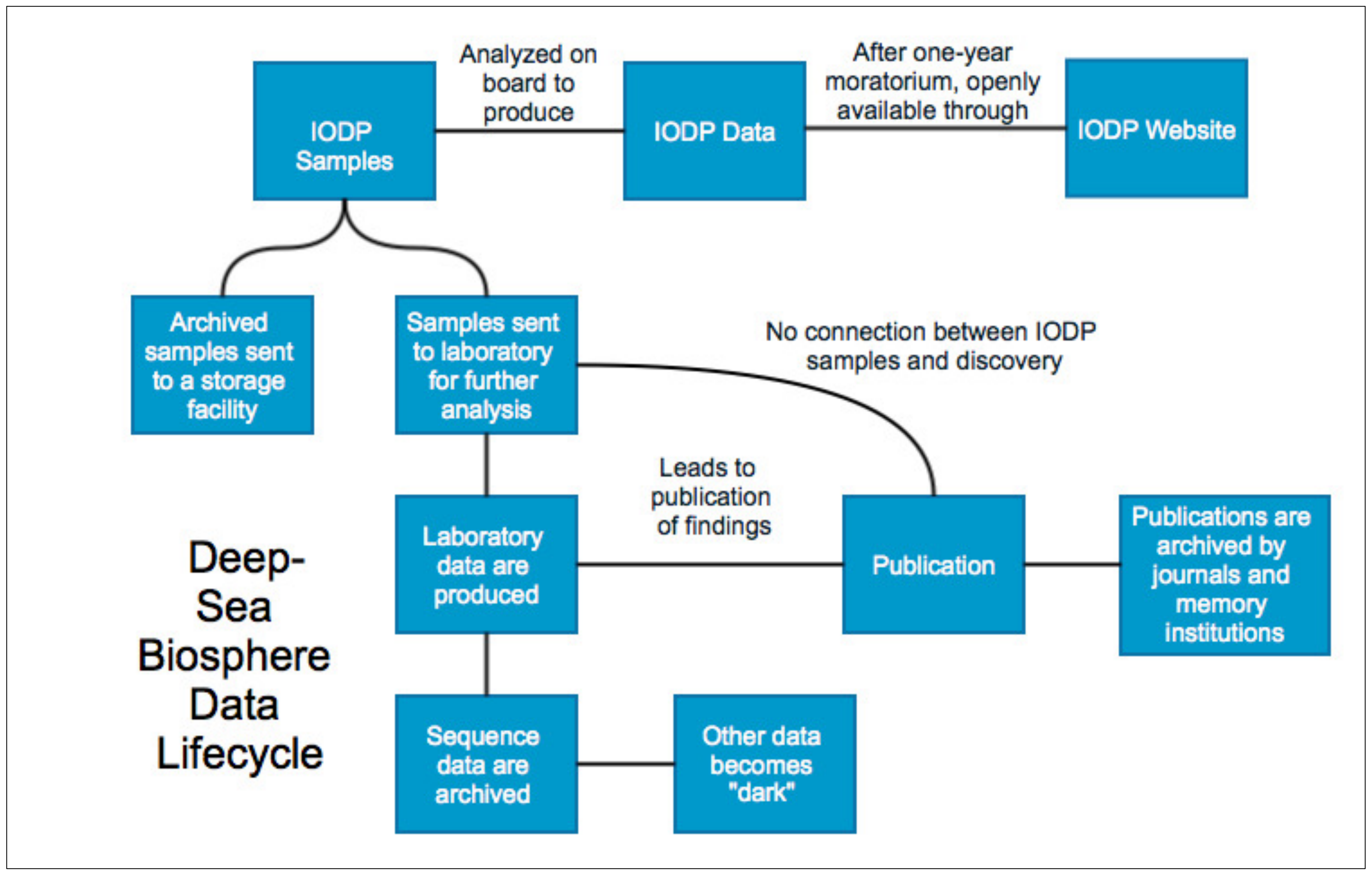
- Within your work, what are typically considered to be "data?"
- What tools are used to collect your data?
- What difficulties do you experience in collecting data?
- Where do you store your data at each stage of the process?
- What tools do you use to share your data if you have to share your data?

## SITE: CENTER FOR DARK ENERGY BIOSPHERE INVESTIGATIONS

The Center for Dark Energy Biosphere Investigations (C-DEBI) is a National Science Foundation (NSF) Science and Technology Center (STC) funded since 2010. C-DEBI aims to explore microbial life beneath the seafloor and to explore its relationship with its environment.[2]

Even though C-DEBI is a large STC, the research is organized around research teams working in small- to medium-sized laboratories. These laboratories are highly-distributed both across multiple scientific disciplines and geographically.

## IODP CRUISE AND LABORATORY DATA PRACTICES



IODP Samples
Analyzed on board to produce
IODP Data
After one-year moratorium, openly available through
IODP Website
Archived samples sent to a storage facility
Samples sent to laboratory for further analysis
No connection between IODP samples and discovery
Leads to publication of findings
Laboratory data are produced
Publication
Publications are archived by journals and memory institutions
Deep-Sea Biosphere Data Lifecycle
Sequence data are archived
Other data becomes "dark"

A broader portrait of deep-sea biosphere data lifecycle, including the physical samples and publications. Figure by Rebekah Cummings

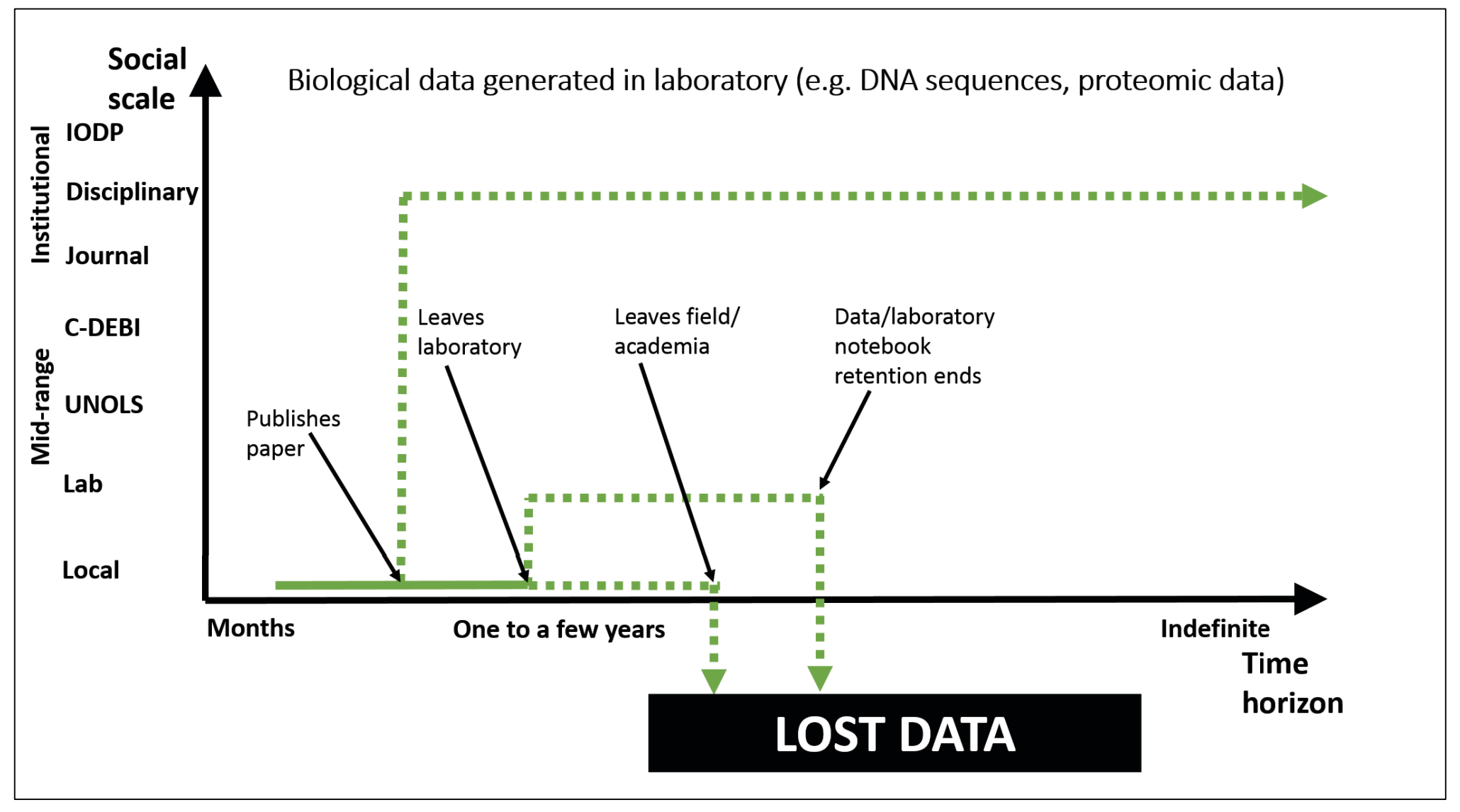
JOIDES Resolution photo by W. Crawford, IODP/TAMU

Photo: Rebekah Cummings

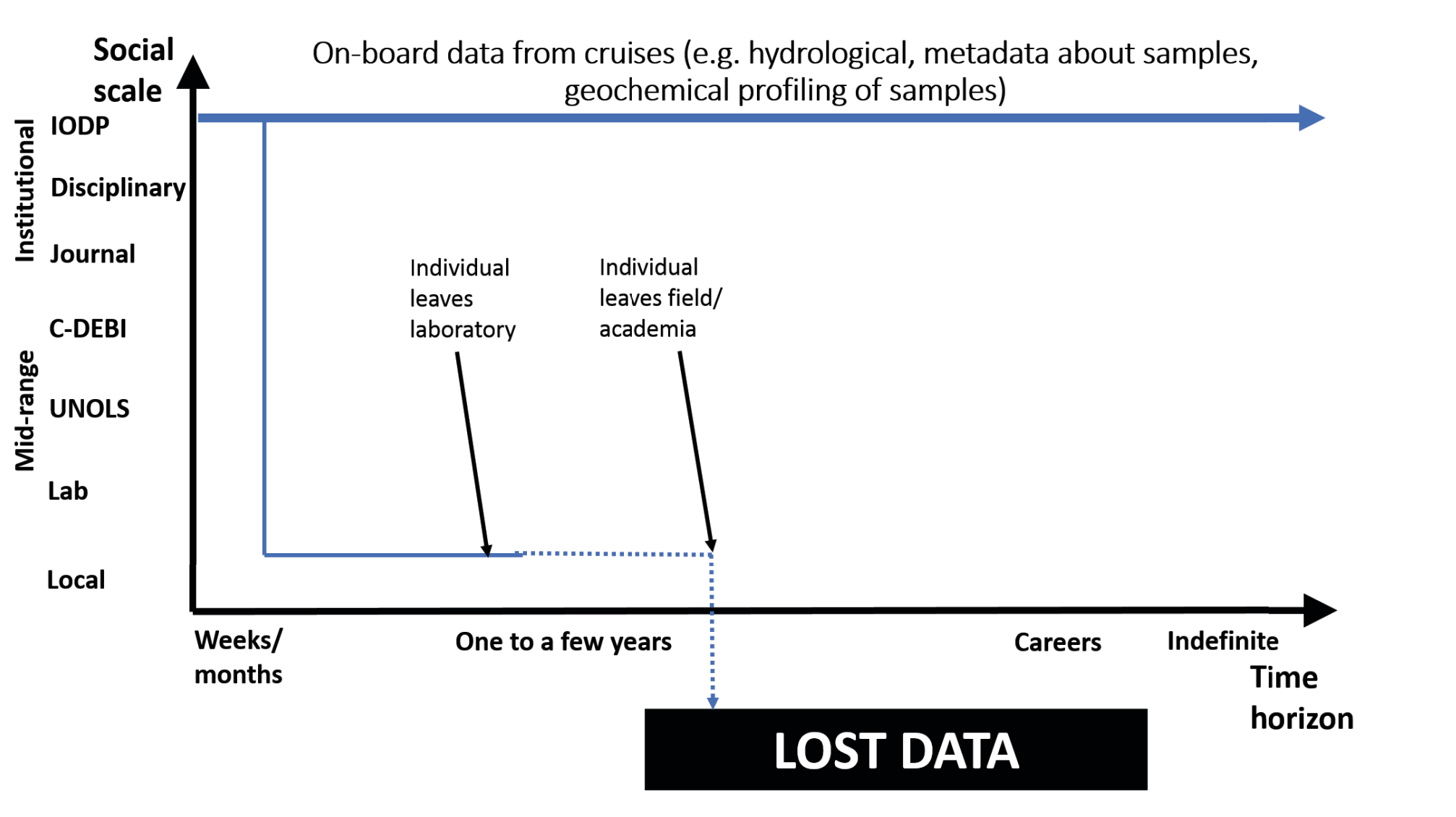
| | IODP Cruises | Laboratory |
|---|---|---|
| Data collected | **Data about samples standardized across cruises**<br>**Types of on-board data generated**<br>• Geochemical profiles of cores<br>• Hydrological data<br>• Some biological data about samples | **Categories of data determined by individual scientist according to their research priorities**<br>**Types of laboratory data generated**<br>• Physical (geological, chemical, hydrological, mineralogical)<br>• Biological (sequence data, proteomics) |
| Who curates? | **Specialist curator on-board** | **Individual scientist/small team** |
| Standards for curation | **Universal across cruises**<br>**Codified in policy** | **Generally particular to individual scientist/small team** |
| What happens to the data and samples once collected? | **Online database for on-board physical data**<br>• One year moratorium<br>- Available only to cruise participants<br>• Then available publicly<br>• Archived cores sent to a storage facility | **Data generated in laboratory stored by scientist**<br>• Laboratory notebooks<br>• Personal computer<br>• Dropbox<br>• University server<br>**Upon publication**<br>• Disciplinary database (where mandated and available)<br>• Shared through a gift exchange culture with trusted colleagues |

Comparing IODP and laboratory-based data practices. Table by Peter Darch

## WHERE DO DATASETS GO?



What happens to laboratory-generated data? Figure by Peter Darch



What happens to IODP data? Figure by Peter Darch

## IMPLICATIONS

Scientists often want to correlate their laboratory-generated data with IODP data. This task is complicated by divergent data practices between the IODP and the laboratory bench.

On the one hand, the more ad hoc practices of scientists in their laboratory contributes to this because:

- These practices, for example in naming or categorizing data can make it difficult to correlate with IODP data
- The many points at which laboratory-generated data may be lost means the required datasets may not be available.

However, the approach of the IODP can also be problematic. C-DEBI research brings together researchers to adapt methods from other domains, or develop methods *de novo*, to address previously-unanswered questions. In such a context, flexibility of data practices can be a virtue. The highly-structured approach of the IODP can promote standardization within the laboratory prematurely.

We recommend that:

- Software development for SMLs should look to big science endeavors such as IODP cruises to find leverage points for data management, but must take into account the differing expertise, resources, and data practices that exist in SMLs.

Work presented here will be inform the Institute for Empowering Long-Tail Research (IELTR) to develop software assisting SMLs.[3]

## REFERENCES

[1] C-DEBI http://www.darkenergybiosphere.org/
[2] IELTR https://sites.google.com/site/ieltrconcept/
[3] UCLA KI Team http://  knowledgeinfrastructures.gseis.ucla.edu/

## ACKNOWLEDGEMENTS

This research is funded by the U.S. National Science Foundation ("Conceptualizing an Institute for Long-Tail Research", #1216872. C.L. Borgman, Co-PI; S. Traweek, Co-PI) and the Alfred P. Sloan Foundation ("The Transformation of Knowledge, Culture, and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective" # 20113194. C.L. Borgman, PI; S. Traweek, Co-PI).