Digital Object Architecture data layer over a network storage system

Yu Luo, Beth Plale, Jeremy Musser, Martin Swany
Indiana University

Ficticious Scenario

Funder: data associated with publications needs to be FAIR (findable, accessible, interoperable, reusable)

Publisher: data associated with publications needs to appear not in supplements but in a research sanctioned trusted repository

Major HPC consortium: we will jointly provide long lived availability for data associated with publications, and use a network storage solution across our institutions to do so

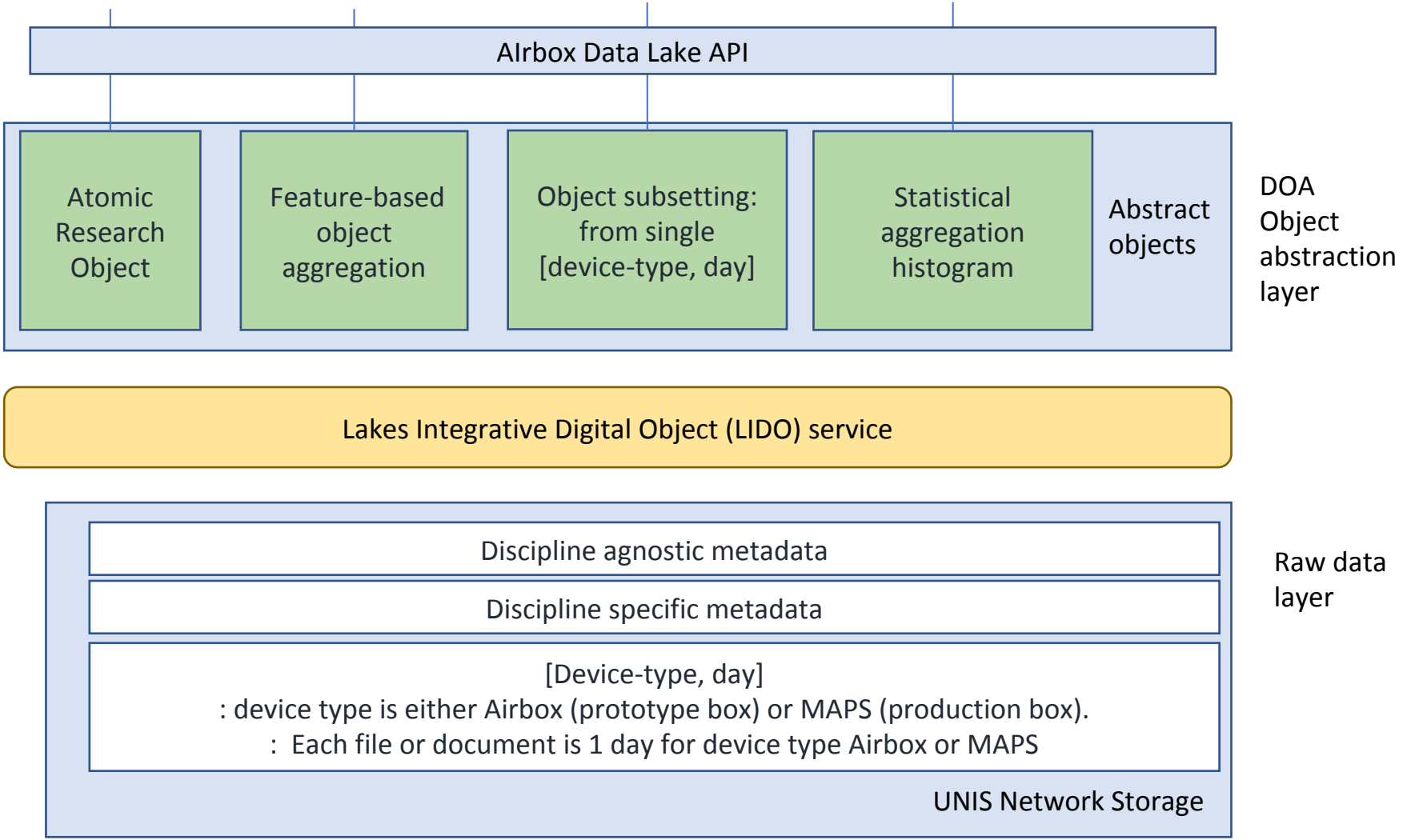
Question: What is overarching architecture for findability for people and machines of data stored in network storage solution and elsewhere?

Research Data Alliance (RDA) is exploring a global Digital Object Architecture (DOA) for FAIR data. FAIR data will reside across databases, in large scale data centers, and in trusted repositories. Through DOA, data can be universally discoverable by people and machines.

We explore how a network storage system (UNIS of Martin Swany) will interoperate with the DOA layer and take advantage of the PID Kernel Information (Beth Plale), and the E-RPID testbed (Rob Quick).

Our research is:

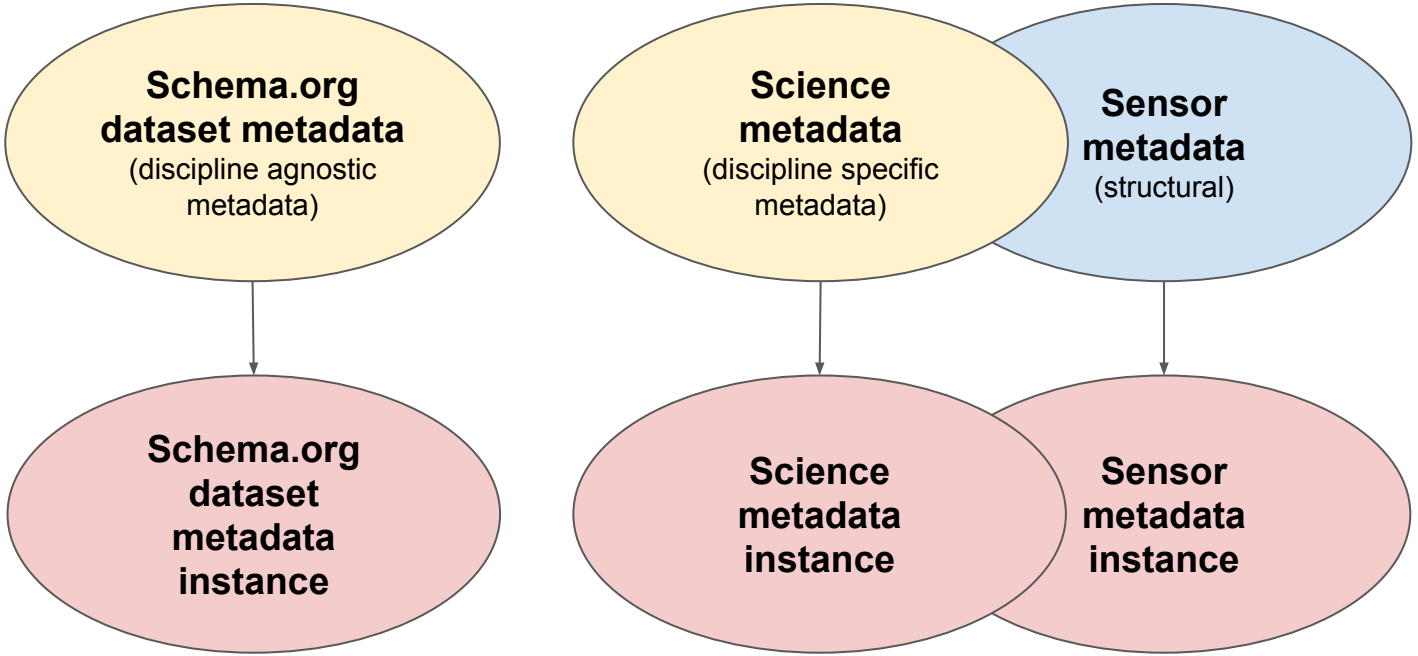
1. developing a ***Lakes Integrative Digital Object*** (***LIDO***) that delivers to an application a set of logical objects in the DOA space, while managing the raw data in Unis, and
2. LIDO's ability to contribute to trust at the local level.


AIrbox Data Lake API
Atomic Research Object
Feature-based object aggregation
Object subsetting: from single [device-type, day]
Statistical aggregation histogram
Abstract objects
DOA Object abstraction layer
Lakes Integrative Digital Object (LIDO) service
Discipline agnostic metadata
Discipline specific metadata
[Device-type, day]
: device type is either Airbox (prototype box) or MAPS (production box).
: Each file or document is 1 day for device type Airbox or MAPS
UNIS Network Storage
Raw data layer

DOA object abstraction layer: Airbox metadata types and instances

Type : resides in DOA Data Type Registry

Instance : resides in DOA LIDO service



Our Exploratory Study

But S&CC draw on not only the local data, such as from local PM 2.5 sensors, but on data from outside the community, and computing may be done both at the edge and on computational resources.

Our work is exploring a ***network storage system*** that is capable of replicating data efficiently anywhere in the network. The network storage system that we use is our Unis [5] network storage system.

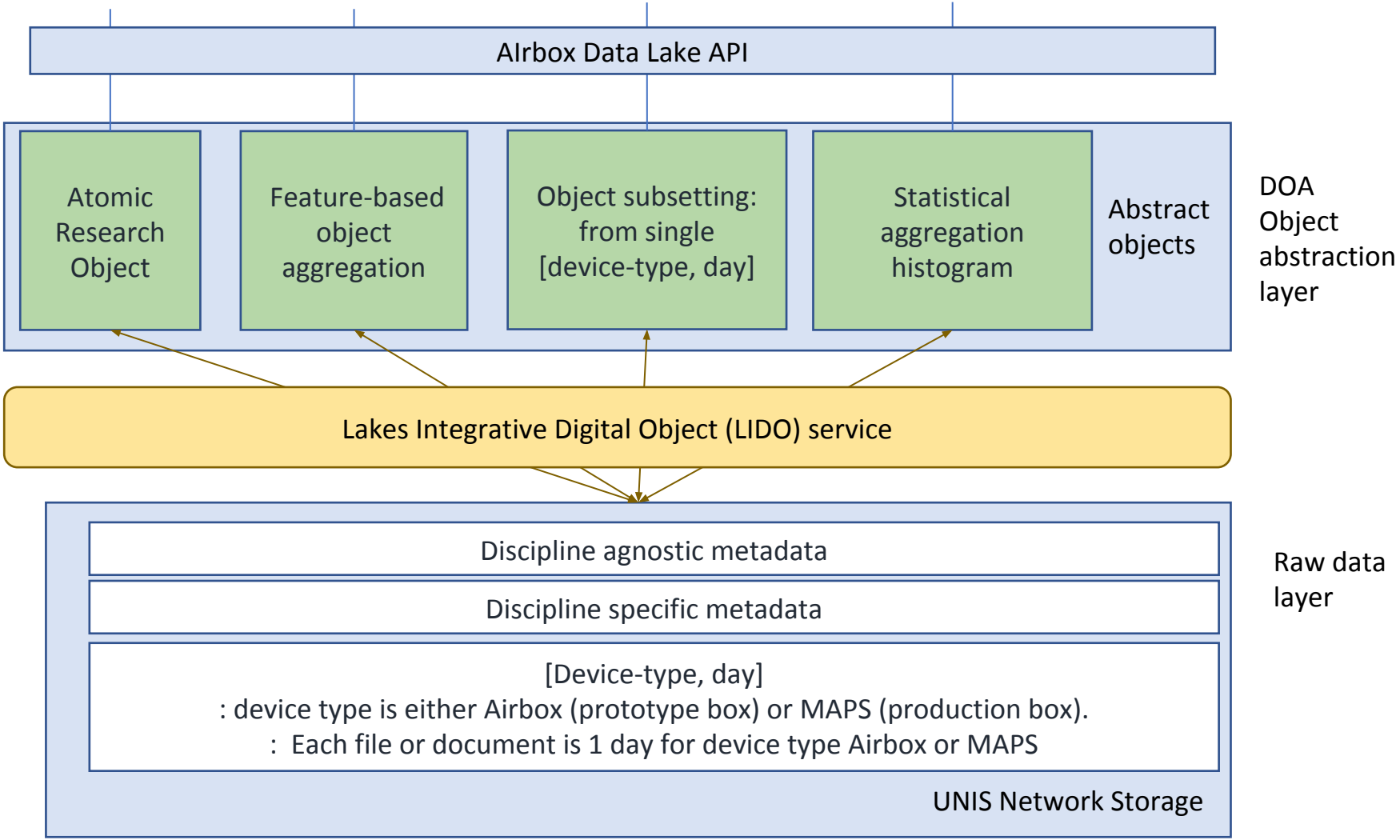
Our research includes

1. developing a ***Lakes Integrative Digital Object*** (***LIDO***) service that delivers to an application a set of logical objects in the DOA space, while managing the raw data in the network through Unis, and
2. studying and identifying the benefits of this layered architecture particularly in its potential to build trust at the local level.

[5] Portnoi, M., & Swany, M. Unified Network Information Services (UNIS) allow users to discover network services and capabilities. *SIGCOMM Computer Communication Review*, *27*, 3-14.

Demo

- Demo
  - Upload a raw data file into Airbox Data Lake
  - Query an abstract object (Atomic Research Object) from Airbox Data Lake
  - Download the abstract object (Atomic Research Object)


AIrbox Data Lake API
Atomic Research Object
Feature-based object aggregation
Object subsetting: from single [device-type, day]
Statistical aggregation histogram
Abstract objects
DOA Object abstraction layer
Lakes Integrative Digital Object (LIDO) service
Discipline agnostic metadata
Discipline specific metadata
[Device-type, day]
: device type is either Airbox (prototype box) or MAPS (production box).
: Each file or document is 1 day for device type Airbox or MAPS
UNIS Network Storage
Raw data layer

Lake Integrative Digital Object (LIDO) service

Relevant functions of LIDO:

1. Building composite DOA abstract objects from raw data/metadata
2. Register PIDs for abstract objects that are returned via queries

Potential Researches:

1. Service/Data Discovery in DOA and LIDO
2. Merge different Data Lakes into One Larger Data Lake

PID Kernel Information

PID Kernel Information (PID KI) is a small amount of metadata stored to the PID (Handle) record; it is stored at the Handle system’s Local Handle Server.

We use PID KI to cache three kinds of information. This information is also managed at LIDO:

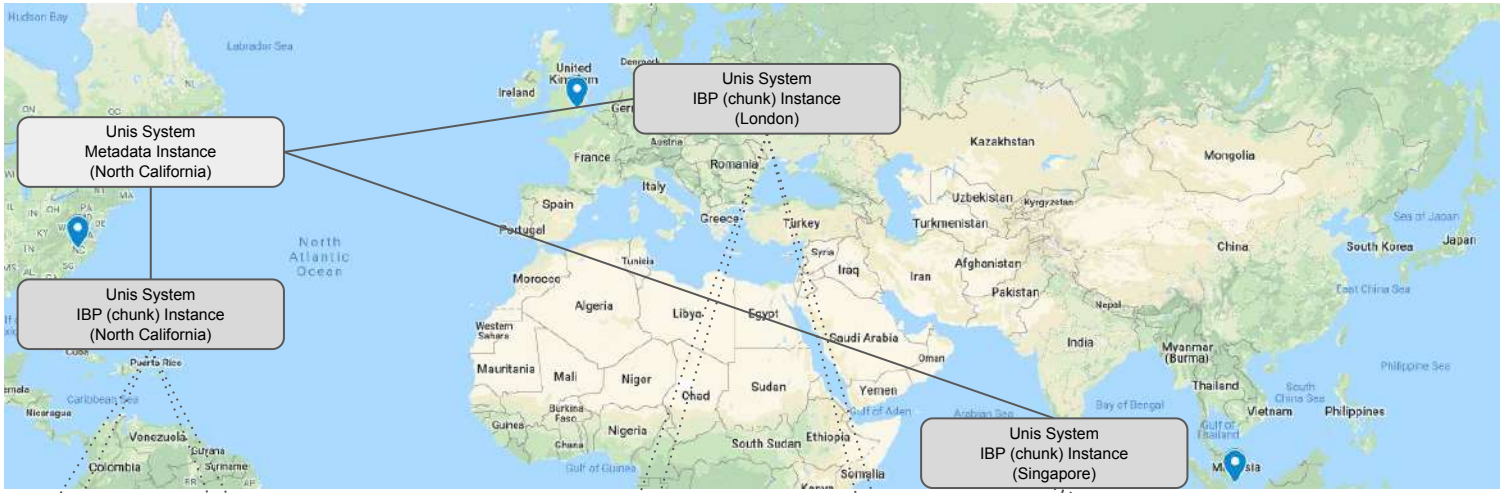
- Unis object location
- Provenance
- Link referencing to schema.org dataset metadata instance

Returned Abstract Object

The abstract object returned from a query is returned as a bundle and contains:

1. PID of the bundle
2. PID content:
    a. PID of the schema.org dataset instance
    b. Location of the bundle zip file
3. Data
4. Metadata instances

Unis Network Storage in AWS



[Device-type, day]
Document 1 replica 1
Chunk 1

[Device-type, day]
Document 1 replica 2
chunk 2

[Device-type, day]
Document 1 replica 2
Chunk 1

[Device-type, day]
Document 3
chunk 2

[Device-type, day]
Document 3
Chunk 1

[Device-type, day]
Document 1 replica 1
chunk 2

[Device-type, day]
Document 3
chunk 3

[Device-type, day]
Document 1 replica 1
chunk 4

[Device-type, day]
Document 1 replica 1
chunk 3

[Device-type, day]
Document 1 replica 2
chunk 4

[Device-type, day]
Document 1 replica 2
chunk 3

[Device-type, day]
Document 2
chunk 4

Three instances: ERPID, Mugo and Smoketree

Instance name	IP address	Service	Database
ERPID	http://149.165.169.46:8080	Local Handle System	None
Mugo	http://129.79.247.9:8082	LIDO	MongoDB
Smoketree	http://129.79.247.8:8082	LIDO	FileSystem

Smoketree	
CPU	48 processors
MEM	131.8GB
Disk	7.6 TB
Database	Linux File System

Mugo	
CPU	48 processors
MEM	131.8 GB
Disk	7.5 TB
Database	MongoDB version 4.0.10

Unis (AWS instance)	
Network Storage	
Metadata Instance Location	North California
Chunk instance Locations	North California, London, Singopro