Development of Earth Science Observational Data Infrastructure of Taiwan

Fang-Pang Lin
National Center for High-Performance Computing, Taiwan

GLIF 13, Singapore, 4 Oct 2013

The Path from Infrastructure to Data

- Sensing for Understanding
  - Sensing: (Networks change the game!)
    - Evolve since 10 years ago: Ecogrid, SARS Grid, … etc
    - Institutional missions based on special vehicles: Satellites, Research Ships & Aircrafts, Met Stations …etc.
    - It is growing even larger and broader, e.g. IOT, social network.
  - Understanding:
    - Modeling from hypothesis to discovery
    - Data dominate

Ecogrid:

Sense the nature in a new way (10 years on)

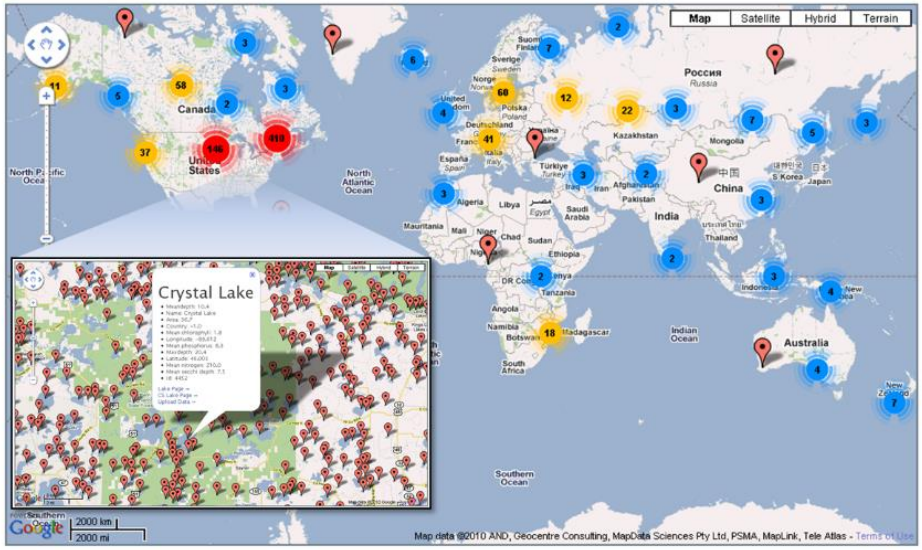
The Path from Infrastructure to Data

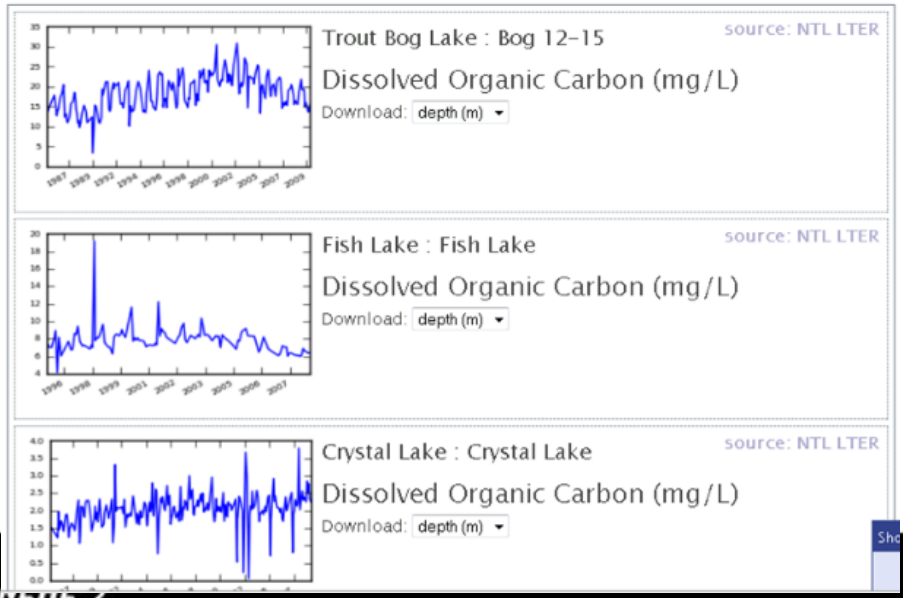
- Sensing for Understanding
  - Sensing: (Networks change the game!)
    - Evolve since 10 years ago: Ecogrid, SARS Grid, … etc
    - Institutional missions based on special vehicles: Satellites, Research Ships & Aircrafts, Met Stations …etc.
    - It is growing even larger and broader, e.g. IOT, social network.
  - Understanding:
    - Modeling from hypothesis to discovery
    - Data dominate

Global Lake Ecological Observational Network (GLEON)

- Lakebase: harvest quality data from internet and collect more than ~25,000 lakes across the world.
- Global compute service through CONDOR
- >10 major real time observational data from selected GLEON sites.

LakeBase

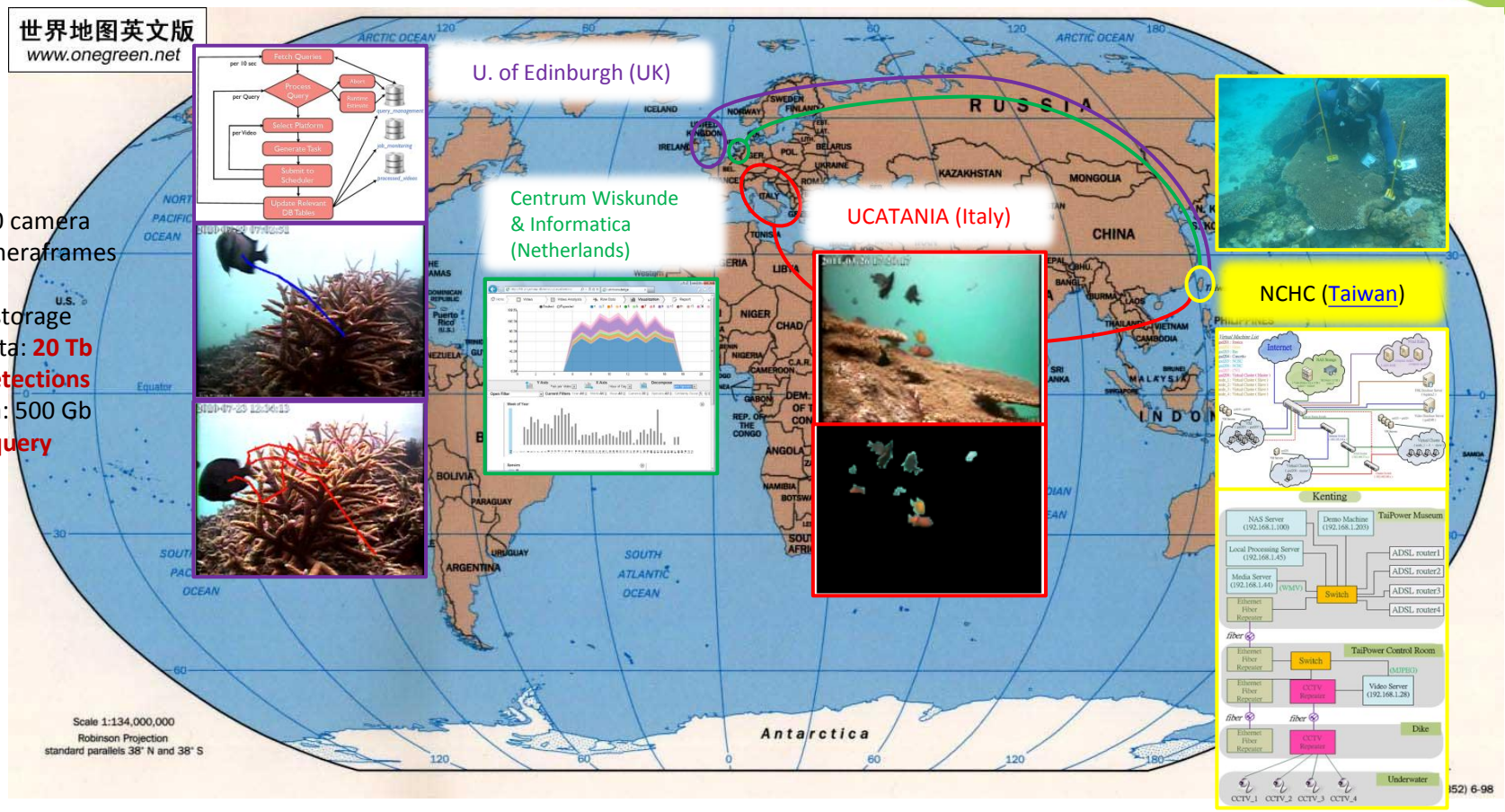




Source: Pau Hanson

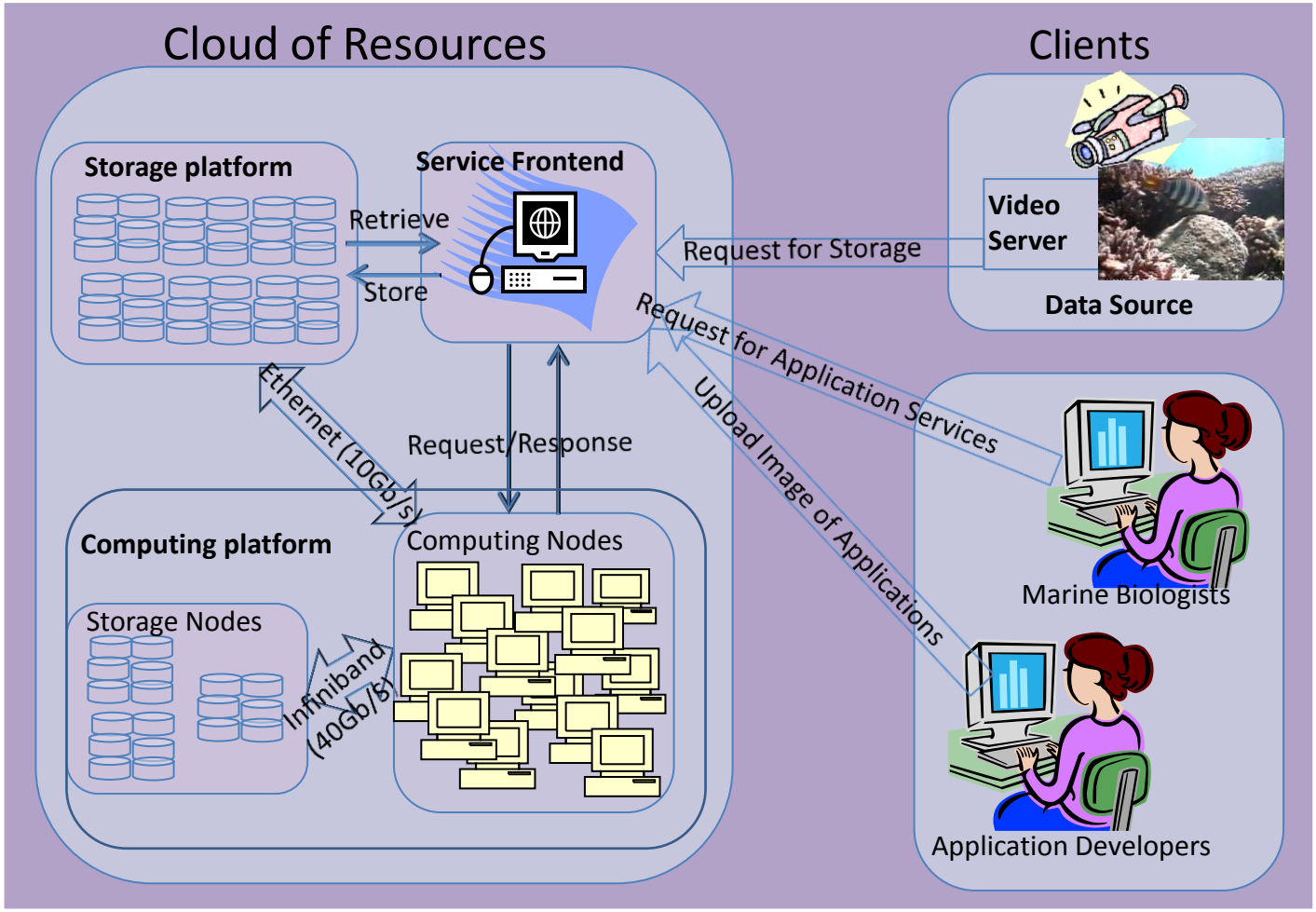
Fish4Knowledge–
human level query for Marine Biology

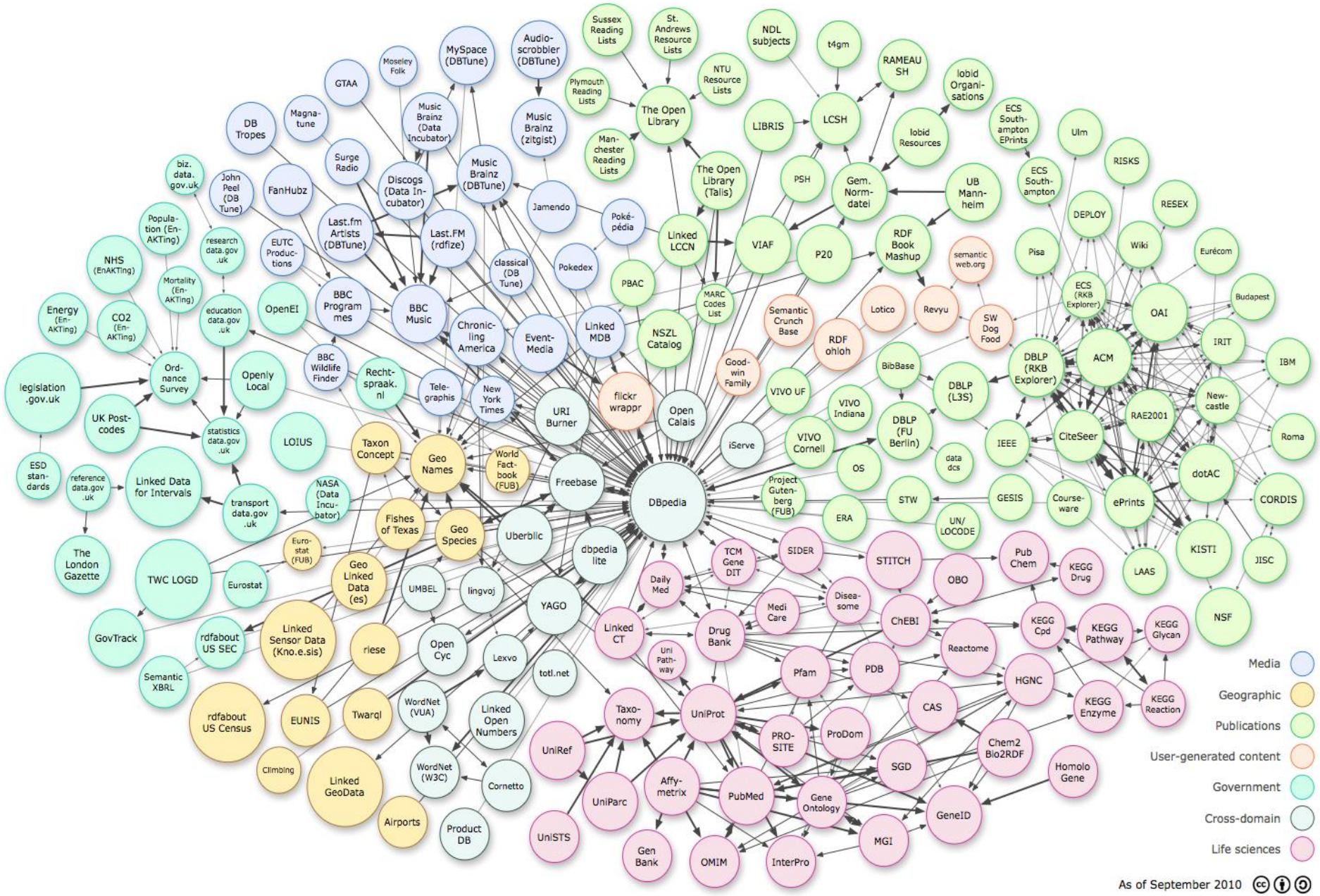
Video Data: 10 camera years = 10 cameraframes = **112 Tb**
massive data storage
Descriptive data: **20 Tb**
Fish: **10^10 detections**
Summary data: 500 Gb
Target: **1 sec query answering**



- NCHC in Taiwan: sustainable system for data acquisition, storage, and computing.
- U. of Catania in Italy: fish detection and tracking.
- U. of Edinburgh in UK: workflow, fish recognition, and fish behavior.
- CWI in Netherlands: user interfaces.

Infrastructure enables Fish4Knowledge

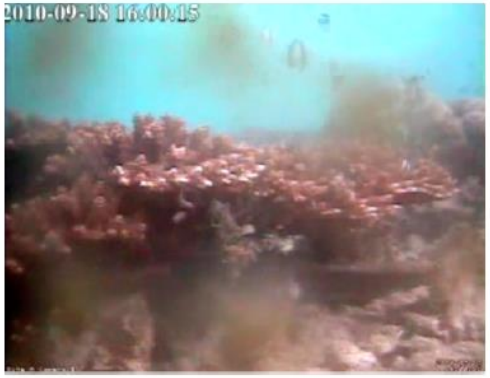




Work on Linking Open Data (e.g. FAO)

Video Classification (~350K videos from 2009 to 2013)

Algae: 9.2%



Blurred: 33.5%

Complex Scenes: 4.3%

Encoding: 23.9%

Highly Blurred: 13.9%

Normal: 12.9%

Unknown: 2.2%

The SEATANK Event & Other

- **SEATANK event**
  - It happened again!
  - Require scientific evidences for Taiwan to win the lawsuit. (TORI, NSPO, NCHC)

2006 Cargo ship 'Tzini' stranded in Yilan water and leaks >100 tons fuel Oil

2013.1 Freighter SEATANK stranded in Penhu water

- **Other**
  - Oceanic temperature data of TW waters only available in JP, if it is before 1990 – K.C. Shao.

The Path from Infrastructure to Data

- Sensing for Understanding
  - Sensing: (Networks change the game!)
    - Evolve since 10 years ago: Ecogrid, SARS Grid, … etc
    - Institutional missions based on special vehicles: Satellites, Research Ships & Aircrafts, Met Stations …etc.
    - It is growing even larger and broader, e.g. IOT, social network.
  - Understanding: (Data change the game!)
    - Modeling from hypothesis to discovery
    - Data dominate

The World of Data Around Us: Data Loss

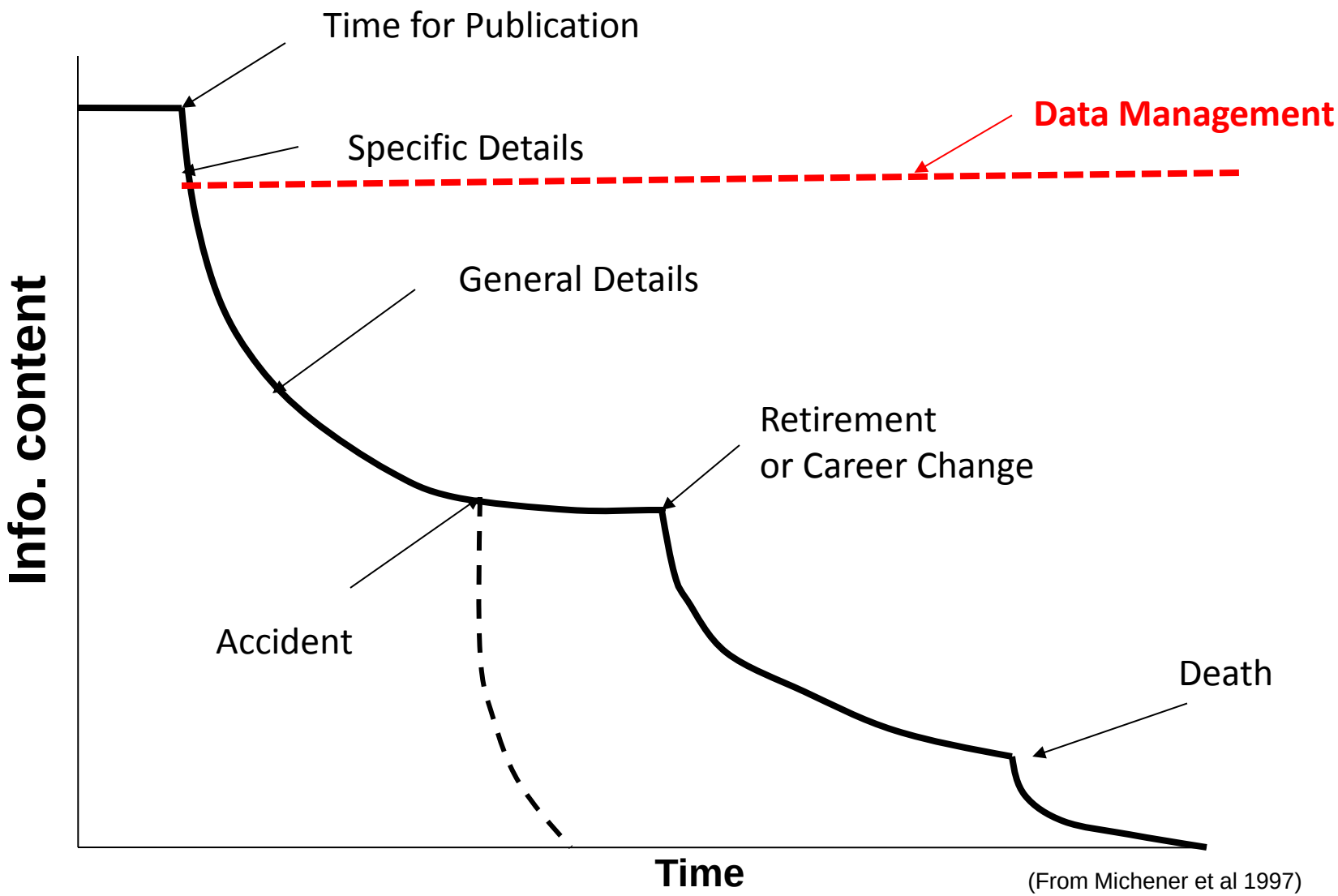
CC image by Sharyn Morrow on Flickr

Source: http://blog.alltop.com.tw/story/archives/3278

- Natural disaster
- Facilities infrastructure failure
- Storage failure
- Server hardware/software failure
- Application software failure
- External dependencies (e.g. PKI failure)
- Format obsolescence
- Legal encumbrance
- Human error
- Malicious attack by human or automated agents
- Loss of staffing competencies
- Loss of institutional commitment
- Loss of financial stability
- Changes in user expectations and requirements

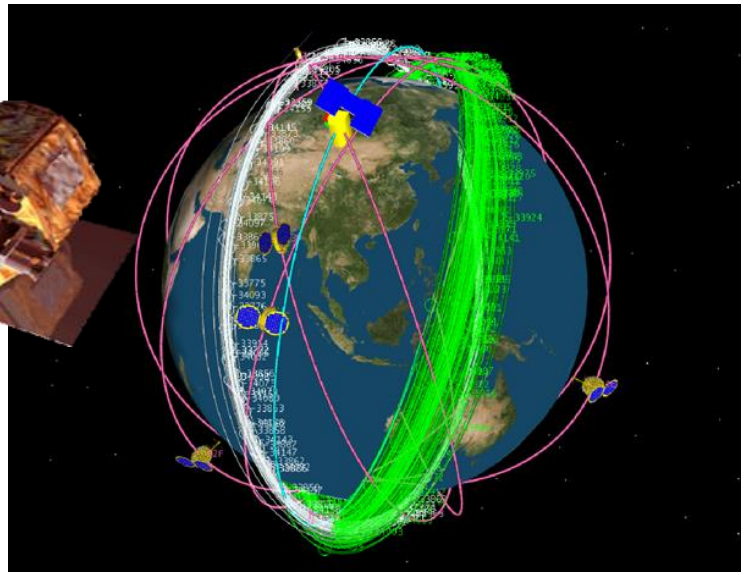
Source: DataONE

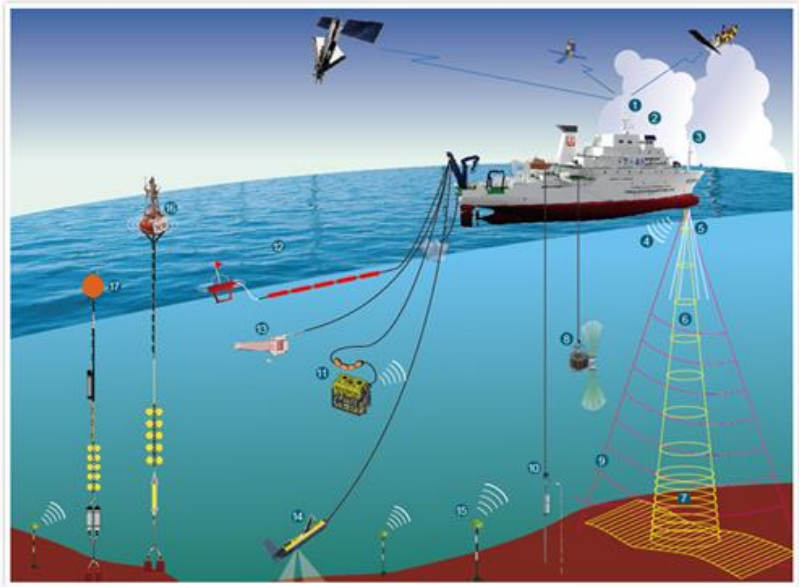
Data Entropy: *Do we need a Data Institute?*

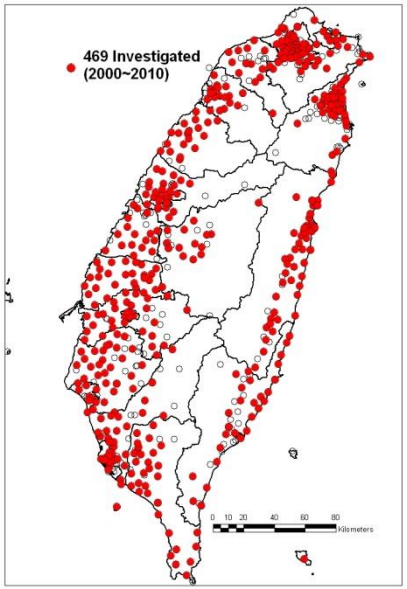


(From Michener et al 1997)

The Treasure of NARLabs:
Earth Science Observational Data



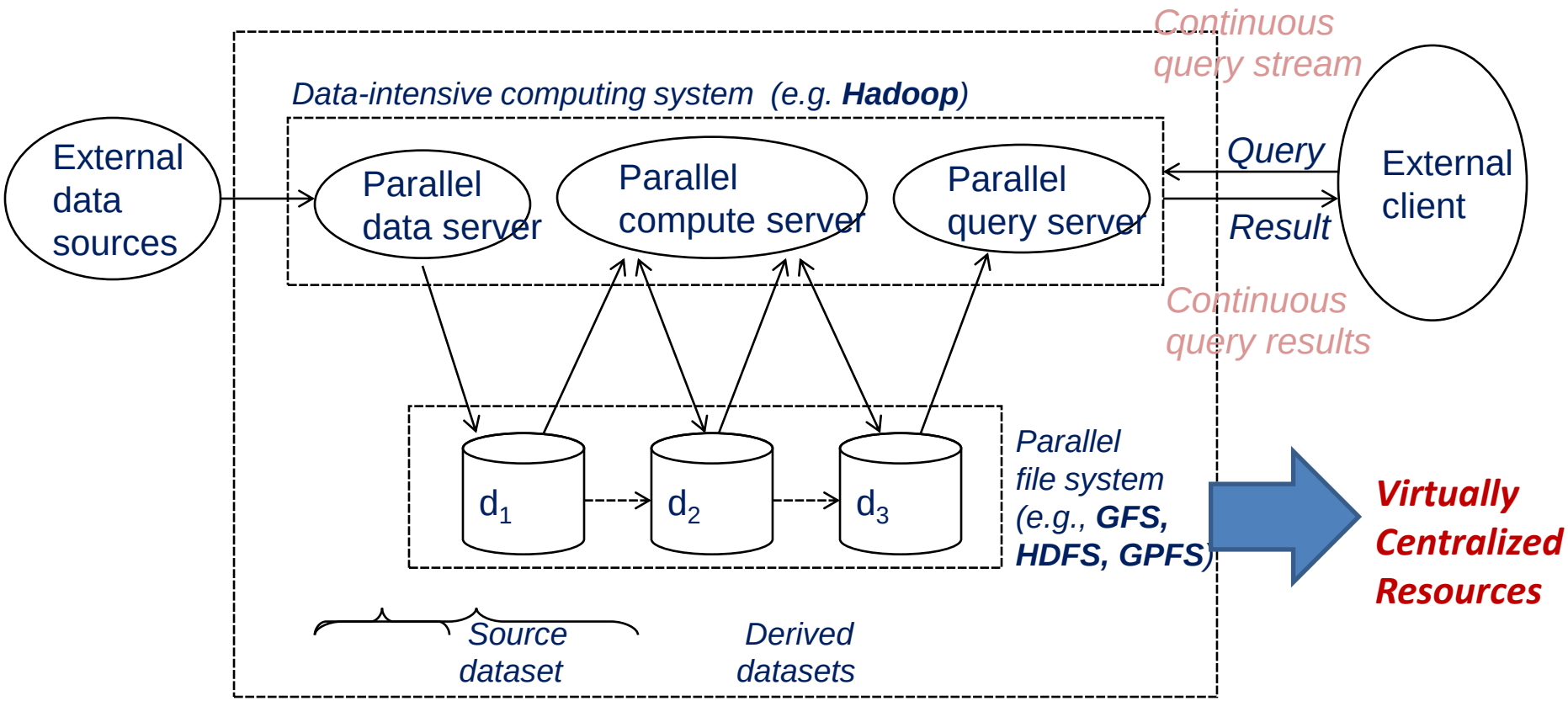




Big Data Infrastructure Challenge in ESOD

- **Data Features:**
  - high complexity, large scale, frequency, real-time and stream → Big Data
- **Integration**
  - of images & data of F2-3 (NSPO), Marine Environmental Databank (TORI), Atmospheric Research Databank (TTFRI), Engineering Geological Database for TSMIP (NCREE) of NARLabs.
- **Data Size:**
  - **155TB/yr.** (Actual data size will 2~3x)
  - Currently, 103 TB for Q1. Simulation data 200TB from TTFRI.
- **Real time, High frequency Data:**
  - **Process > 18,000 records/sec.**
  - Currently 6,000 records/sec.

ESOD Big Data Service ~ 1PB



Characteristics:

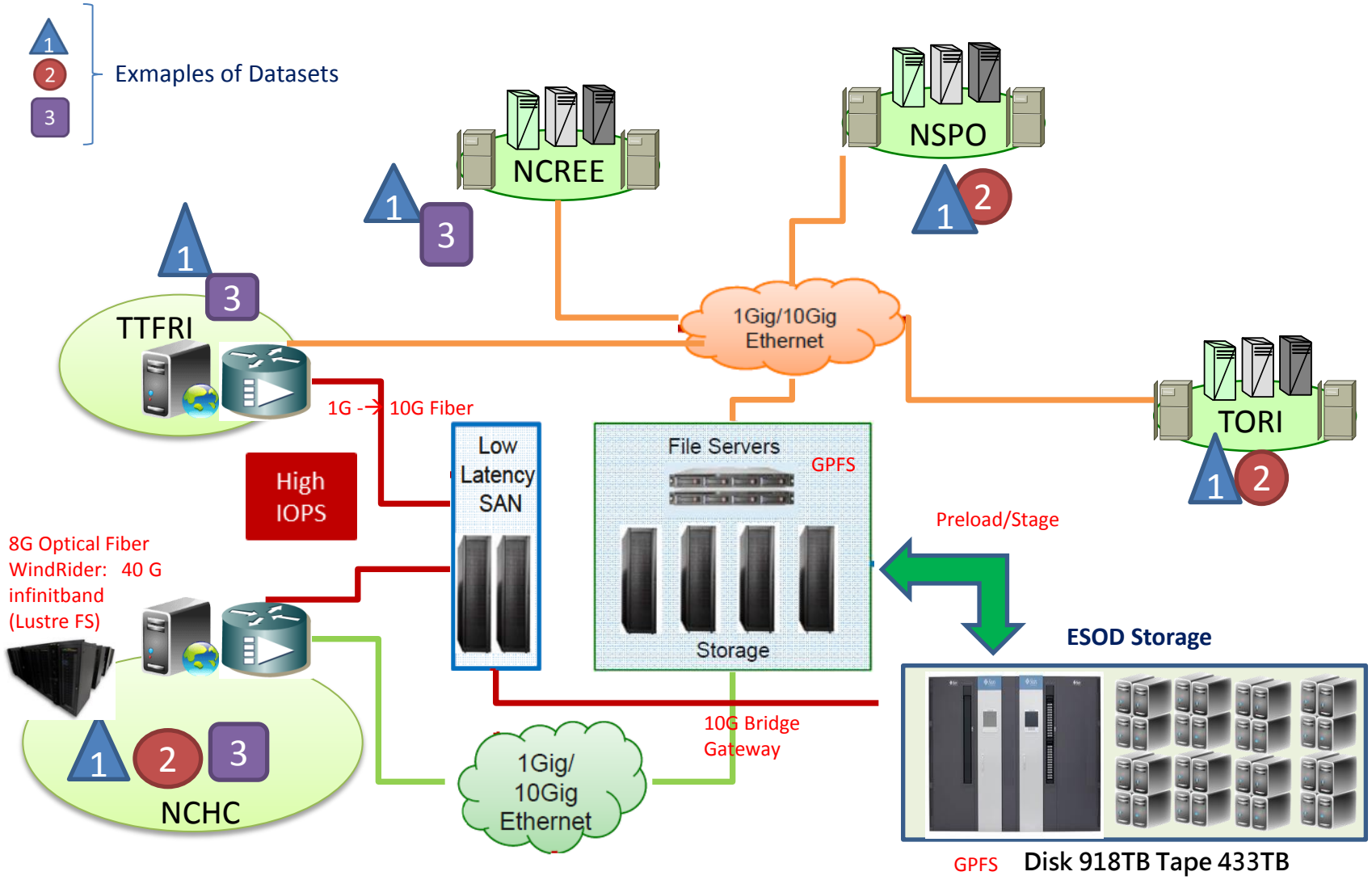
- Small queries and results
- Massive data and computation performed on server

Examples:

- Search
- Photo scene completion
- Log processing
- Science analytics

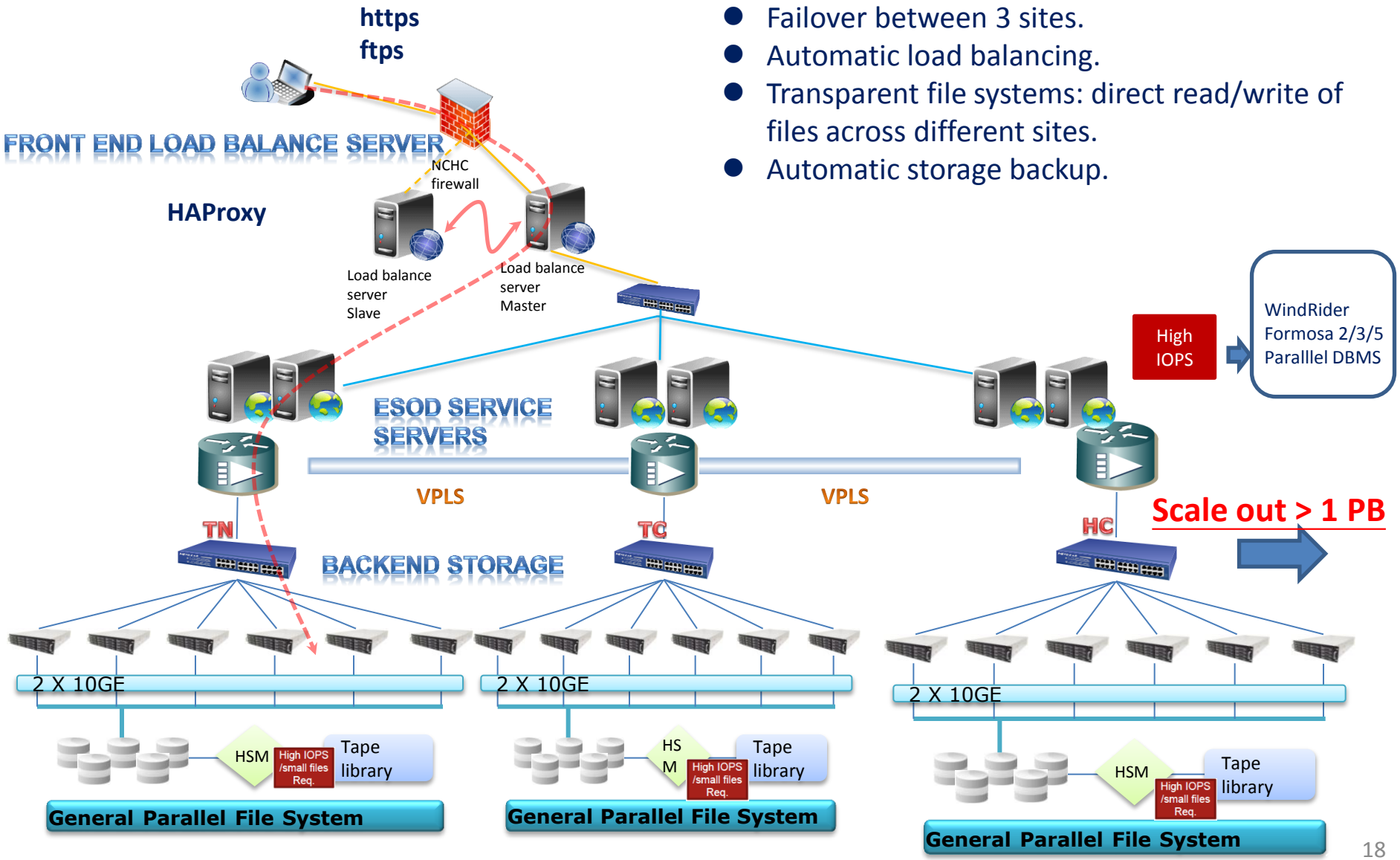
Source: David O'Hallaron

ESOD Hardware Architecture
Challenges of use of distributed resources

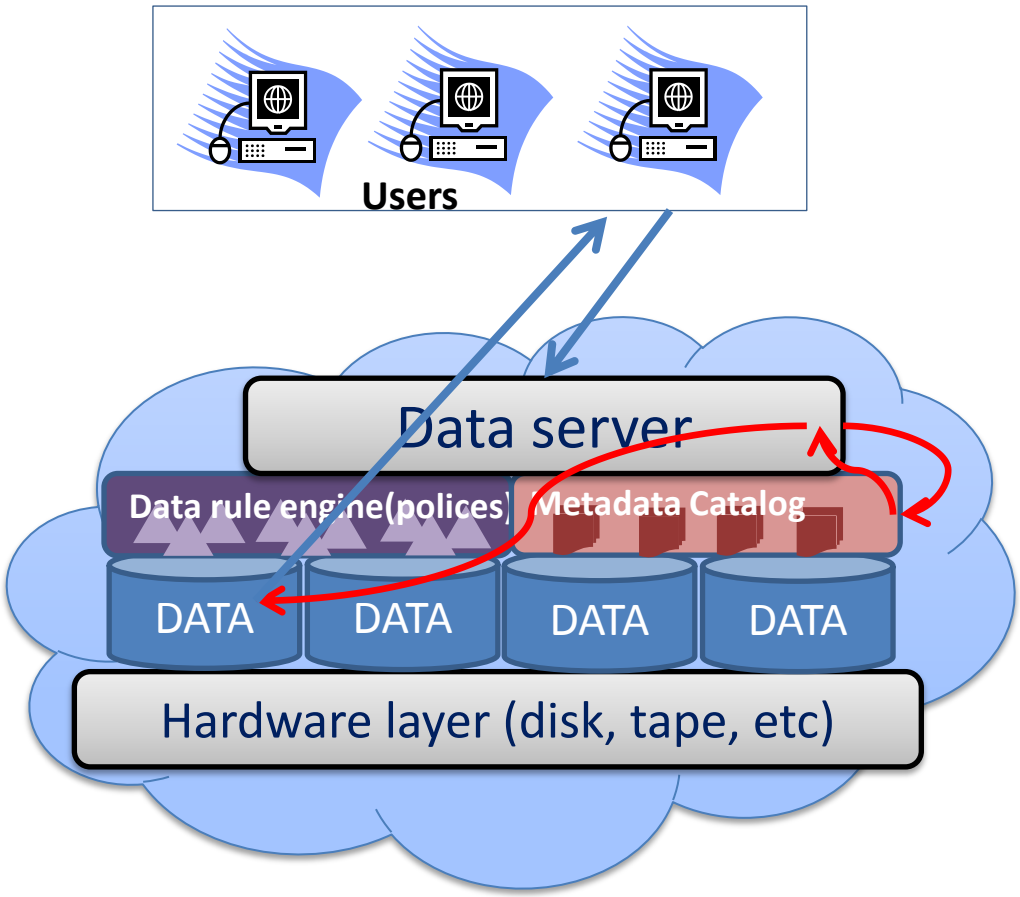


ESOD Data Archiving and System Infrastructure

- Failover between 3 sites.
- Automatic load balancing.
- Transparent file systems: direct read/write of files across different sites.
- Automatic storage backup.



ESOD: Data discovery



- User access to single united data warehouse
- User request for data
- Request goes to data server
- Server looks up information in catalog
- Catalog tells where the data physically located
- Server applied rules and serve data

Data discovery: Metadata Catalog

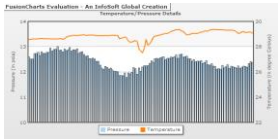
- Use relational database (mysql) with multi-dimensional schema design to speed up searching (migrating to NOSQL solutions now)

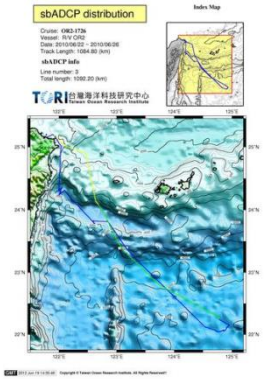
- System Metadata
  - User name space
    - Address / e-mail / telephone number
    - Role (administrator, curator, user)
    - File name space
    - Creation date / size / location / checksum
    - Owner / access controls
  - Storage resource name space
    - Capacity / quotas / Type (archive, disk, fast cache)
- Domain Metadata
  - User-given metadata
    - Key-Value-Unit Triplets, Annotation
    - Relational / XML Metadata
    - Domain-specific Schema
      - Adopt OGC standards

Example: geonetwork

ESOD: Smart query and answer

- Develop a set of **control vocabulary** based on int'l standards, e.g. HDF, NetCDF, OGC … etc.
- Derive **RDF triple dataset** from common query tasks of ESOD.
- Combination of **visual data plus metadata** to support a specific high-level information seeking user task.
- Design of an interface for **graph & visual comparison search.**
- Selected one specific task, that of comparing sets of objects, and designed a prototype interface on top of **linked data sets** used by the experts to support this task explicitly
- Develop methods for data **provenance**.





The Path from Infrastructure to Data

- Sensing for Understanding
  - Sensing: (Networks change the game!)
    - Evolve since 10 years ago: Ecogrid, SARS Grid, … etc
    - Institutional missions based on special vehicles: Satellites, Research Ships & Aircrafts, Met Stations …etc.
    - It is growing even larger and broader, e.g. IOT, social network.
  - Understanding: (Data change the game!)
    - Modeling from hypothesis to discovery
    - Data dominate

Future Key issue: Move Big Data

100G TWAREN/TANET/AS

NARLabs

承諾 · 熱情 · 創新