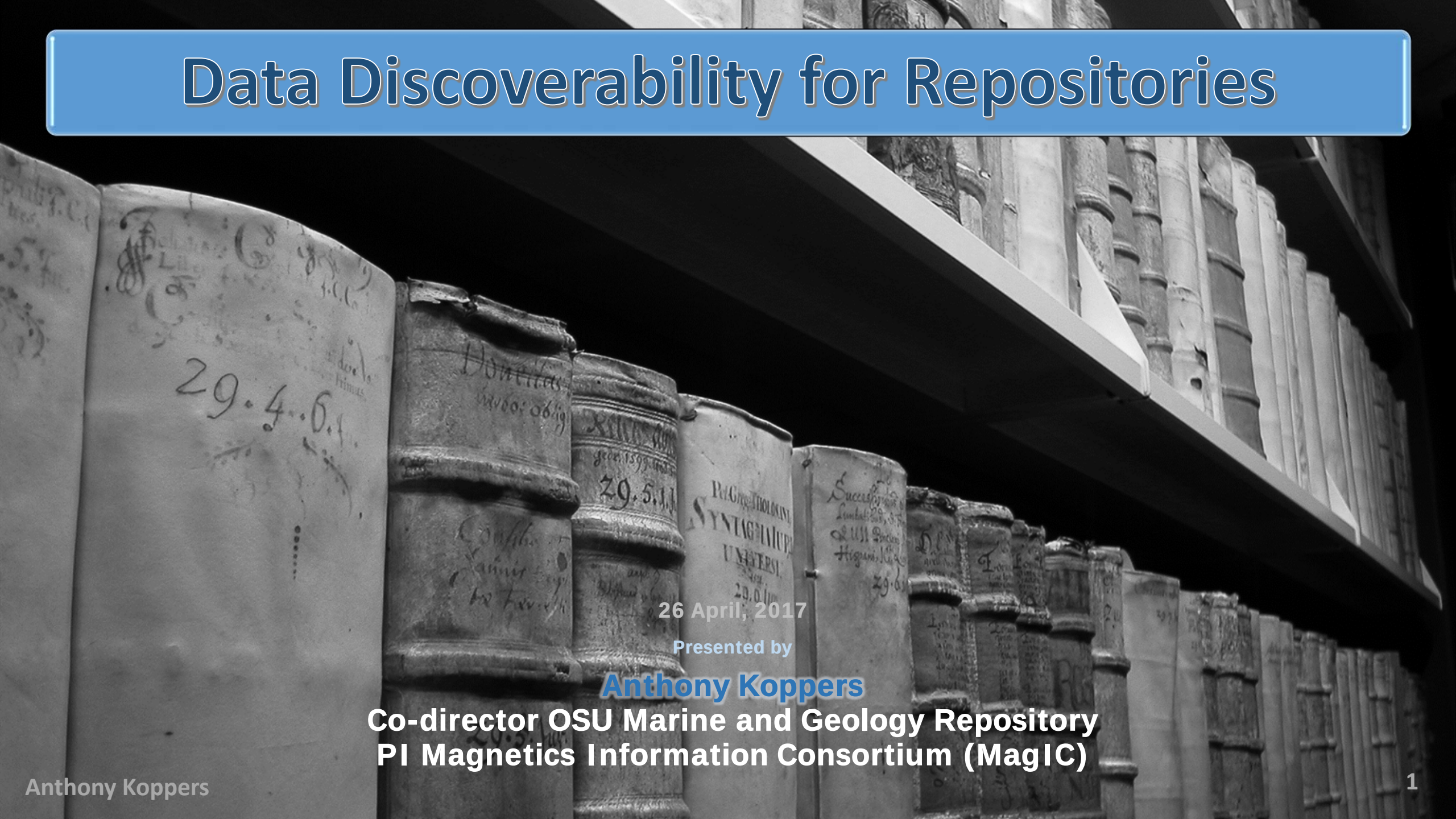
Data Discoverability for Repositories



26 April, 2017

Presented by

Anthony Koppers
Co-director OSU Marine and Geology Repository
PI Magnetics Information Consortium (MagIC)

The Discoverability of Repository Data
The Premise

- **Most Repositories started out in the Pre-historic Digital Ages**
  - Data was collected in analog mode at first
  - Repository mandates have been to (i) curate and (ii) disseminate samples, only recently we started to move into the digital realm
- **It is time to put Data Discoverability on the Front Burner (!)**
  - To help scientists, students and teachers to find their way to all cores and rocks we store in our repositories
  - To make certain analytical data, imagery and core/rock curatorial metadata are intimately linked, as well as the resulting science data appearing in the peer-reviewed journal publications

The Discoverability of Repository Data
Discussion Topics

- Linking in coring/dredging with Rolling Deck to Repository (R2R)
- How to bring Collections and Repositories into the Digital Age
- How to make use of Common Data Portals
  - EarthCUBE in the U.S.
  - NOAA Index to Marine and Lacustrine Geological Samples (IMLGS)
  - The role of this International Curators Consortium (ICC)
- Improving Online Data Discoverability
  - What data do we need to collect and share? What data are required minimal data? Who collects which data? Where does it get stored?

The Discoverability of Repository Data
Two Different Perspectives

OSU-MGR

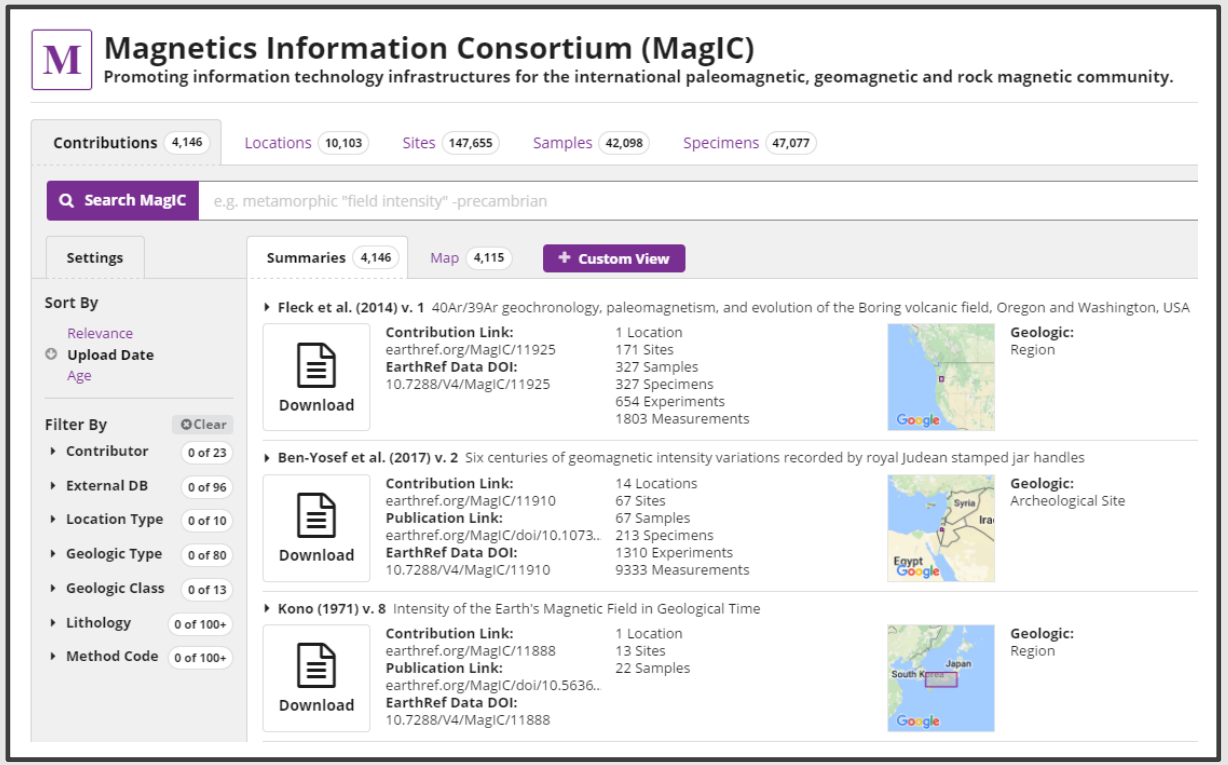
- Co-director for Rock Storage and Data Management since 2010
- Current collection since 1971 (NSF Marine Geology and Geophysics)
- Added NOAA R/V Okeanos Explorer ROV rock collection in 2015
- Adding Antarctic Core Collection in 2018 taking over from Florida State University (NSF Polar Programs)

MagIC Database

- PI for the International Magnetics Information Consortium (MagIC)
- Promoting information technology infrastructures for the international paleomagnetic, geomagnetic and rock magnetic community
- Typical “sample-based” database requiring web linkage to repositories
- NSF-funded since 2003

The Discoverability of Repository Data
Need of Linkage from other Databases

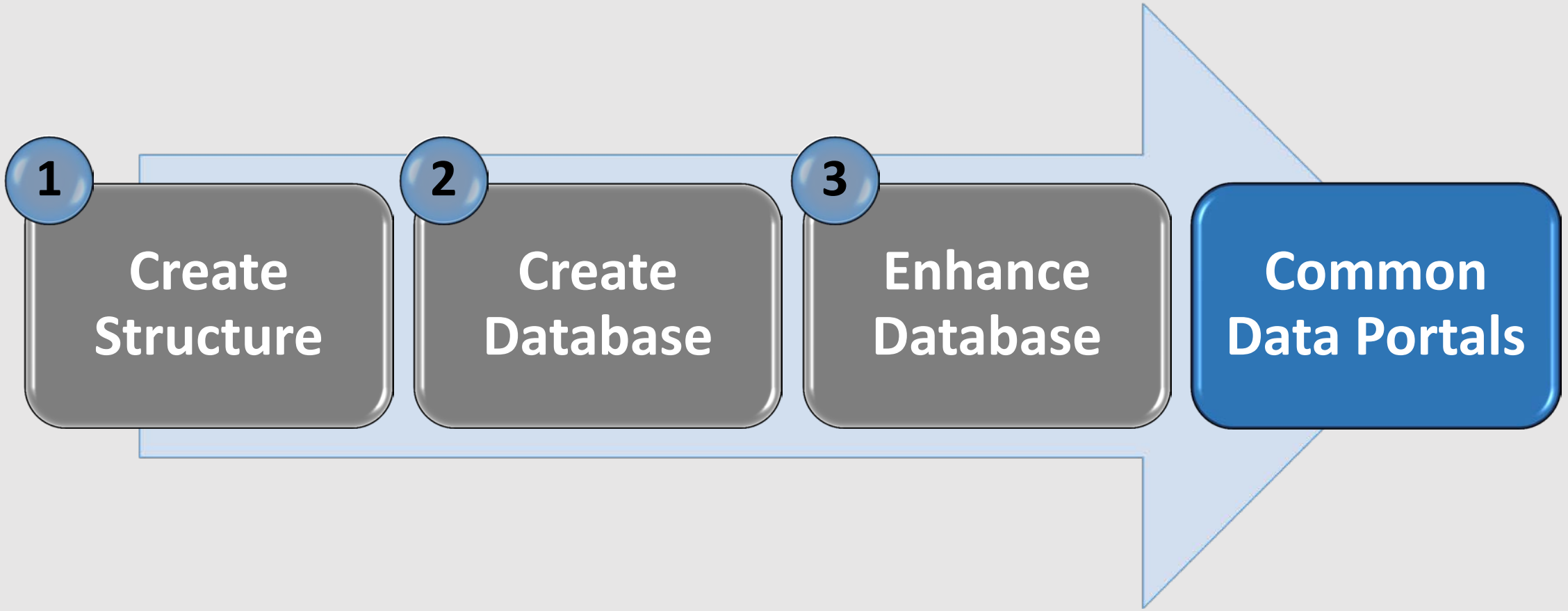
https://www2.earthref.org/MagIC



MagIC Database

- Has peer-reviewed magnetics data from 4,000+ publications
- Data from 10,000+ locations and 147,000+ sampling sites
- More than 4 million measurements
- All data is contributor uploaded within browsers using community-defined data models
- Not linked to other online resources

The Discoverability of Repository Data
OSU-MGR from Analog to Digital



The Discoverability of Repository Data
OSU-MGR from Analog to Digital

STEP 1: CREATE STRUCTURE

1. Collate Cruise Reports and Analog Data
2. Translate into PDF Files and Excel Spreadsheets
3. Assign IGSN Numbers and QR-Codes to All Assets
4. Create File Structure using IGSN Hierarchy
5. Translate into an Online File Structure
6. Use Google Fusion Tables and Google Maps

The Discoverability of Repository Data
OSU-MGR from Analog to Digital

STEP 2: CREATE DATABASE

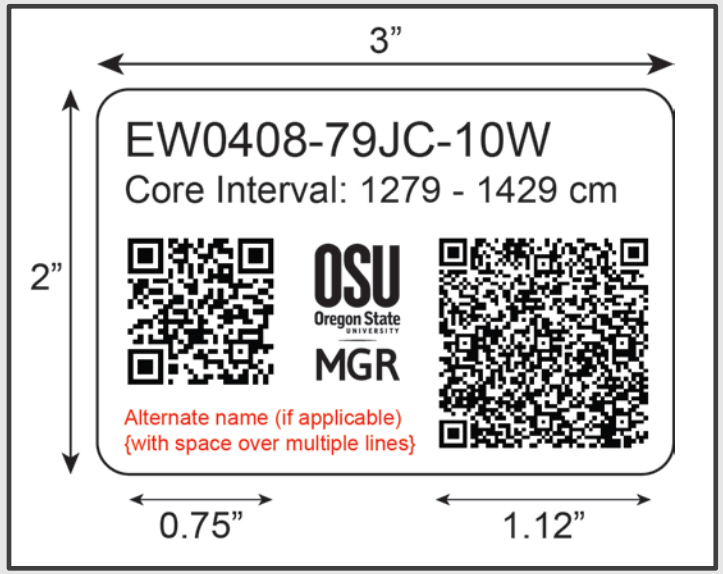
1. Translate Fusion Tables into MongoDB
2. Index All Assets Stored in the Online File Structure
3. Allow Full Filtering based on File Contents
4. Allow Adding New Data without Issues

The Discoverability of Repository Data
OSU-MGR from Analog to Digital

STEP 3: ENHANCE DATABASE

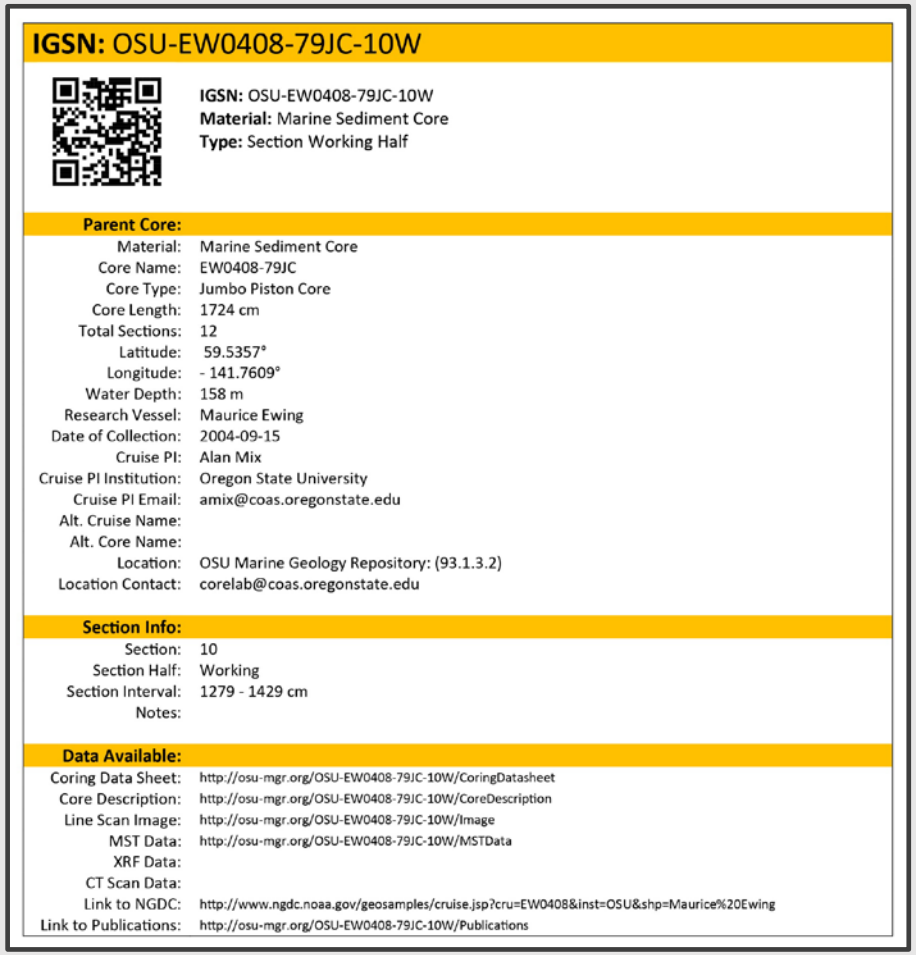
1. Build Web Services and API for Data Sharing
2. Allow Inter-operability between All Repositories
3. Allow Inter-operability with Other Databases
4. Share Tools Built Upon Repository Databases

The Discoverability of Repository Data
OSU-MGR from Analog to Digital



Using Long Syntax IGSN (n<32) with Build-in Hierarchy from Expeditions to Cores to Sections to Subsamples

(or from Expeditions to Dredges to Subsamples)



IGSN: OSU-EW0408-79JC-10W

IGSN: OSU-EW0408-79JC-10W
Material: Marine Sediment Core
Type: Section Working Half

Parent Core:

Material:	Marine Sediment Core
Core Name:	EW0408-79JC
Core Type:	Jumbo Piston Core
Core Length:	1724 cm
Total Sections:	12
Latitude:	59.5357°
Longitude:	- 141.7609°
Water Depth:	158 m
Research Vessel:	Maurice Ewing
Date of Collection:	2004-09-15
Cruise PI:	Alan Mix
Cruise PI Institution:	Oregon State University
Cruise PI Email:	amix@coas.oregonstate.edu
Alt. Cruise Name:	
Alt. Core Name:	
Location:	OSU Marine Geology Repository: (93.1.3.2)
Location Contact:	corelab@coas.oregonstate.edu

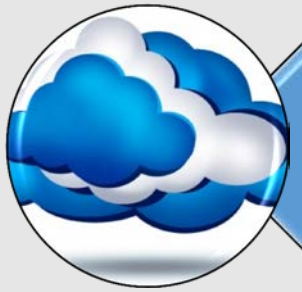
Section Info:

Section:	10
Section Half:	Working
Section Interval:	1279 - 1429 cm
Notes:	

Data Available:

Coring Data Sheet:	http://osu-mgr.org/OSU-EW0408-79JC-10W/CoringDatasheet
Core Description:	http://osu-mgr.org/OSU-EW0408-79JC-10W/CoreDescription
Line Scan Image:	http://osu-mgr.org/OSU-EW0408-79JC-10W/Image
MST Data:	http://osu-mgr.org/OSU-EW0408-79JC-10W/MSTData
XRF Data:	
CT Scan Data:	
Link to NGDC:	http://www.ngdc.noaa.gov/geosamples/cruise.jsp?cru=EW0408&inst=OSU&shp=Maurice%20Ewing
Link to Publications:	http://osu-mgr.org/OSU-EW0408-79JC-10W/Publications

- **Website data is accessible using IGSN in URLs:**
  - http://osu-mgr.org/OSU-EW0408-79JC-10W
  - http://osu-mgr.org/OSU-EW0408-79JC-10W/Image
  - http://osu-mgr.org/OSU-EW0408-79JC-10W/MSTData



The Discoverability of Repository Data
Common Data Portals and Data Discoverability

- Having our repository collections being part of the “cloud” and Common Data Portals requires that each facility has its own “digital house” in order
- Only at that stage will we achieve the needed Online Data Discoverability that enables scientists, students and educators to more effectively use our repositories and our data holdings in their research and teaching

The Discoverability of Repository Data
Discussion Topics

- Linking in coring/dredging with Rolling Deck to Repository (R2R)
- How to bring Collections and Repositories into the Digital Age
- How to make use of Common Data Portals
  - EarthCUBE in the U.S.
  - NOAA Index to Marine and Lacustrine Geological Samples (IMLGS)
  - The role of this International Curators Consortium (ICC)
- Improving Online Data Discoverability
  - What data do we need to collect and share? What data are required minimal data? Who collects which data? Where does it get stored?