Data and Metadata Management at DIAS: Toward More Open Earth Environmental Information Platform

Toshiyuki Shimizu
Graduate School of Informatics, Kyoto University
tshimizu@i.kyoto-u.ac.jp

Dec. 7th, 2017

International Workshop on Sharing, Citation and Publication of Scientific Data across Disciplines

Tachikawa, Tokyo, Japan

Contents

❖ About DIAS

❖ Data and Metadata Management
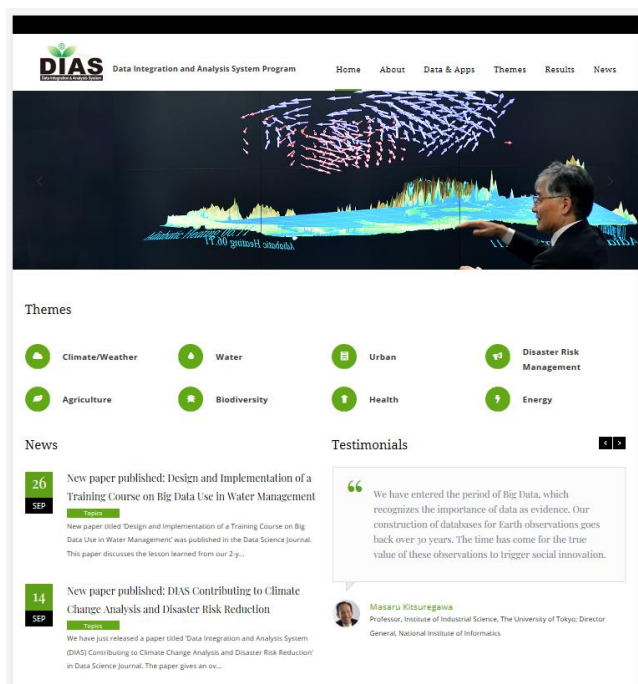 - ❖ Data registration procedure
  - ❖ Metadata management

❖ Open Science Activities

❖ Current and Future Prospects
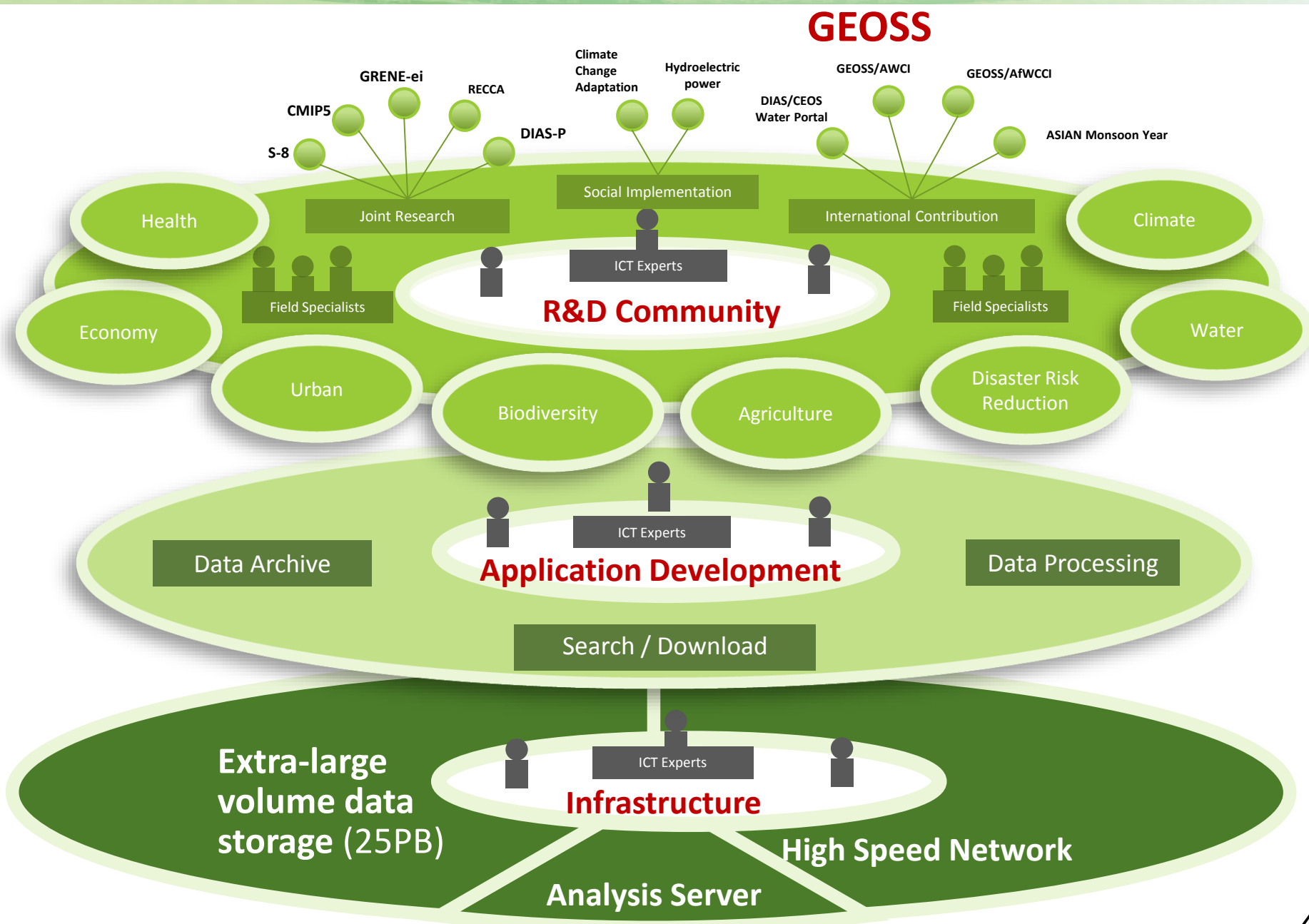 - ❖ DIAS as a national repository
  - ❖ Focusing on metadata quality

DIAS (Data Integration and Analysis System)

❖ DIAS continuously collected and managed earth observation data.

❖ The first phase of DIAS started from 2006, and now we are in the third phase (2016-2020).
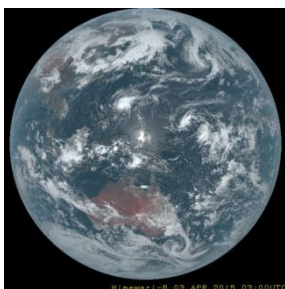

Topics of Datasets available in DIAS

❖ Earth Observation Satellites
❖ Greenhouse Gases Observations
❖ Terrestrial Ecosystems / Carbon Flux Observations
❖ Weather Observations
❖ Watershed Observations
❖ Ocean Observations
❖ Reanalysis
❖ Prediction
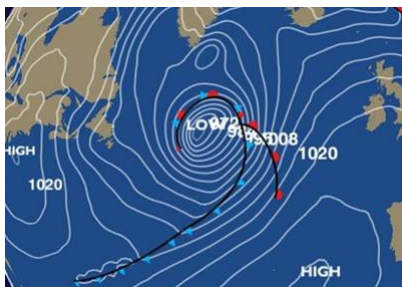❖ Downscaled Data
❖ Natural Disasters
❖ Land Use
❖ Health Hazard

http://www.diasjp.net/en/   http://www.diasjp.net/en/dias-datasetlist/   3

GEOSS

S-8 · CMIP5 · GRENE-ei · RECCA · DIAS-P · Climate Change Adaptation · Hydroelectric power · DIAS/CEOS Water Portal · GEOSS/AWCI · GEOSS/AfWCCI · ASIAN Monsoon Year

Joint Research · Social Implementation · International Contribution

Health · Economy · Urban · Biodiversity · Agriculture · Disaster Risk Reduction · Climate · Water

Field Specialists · ICT Experts · Field Specialists

R&D Community

Data Archive · Data Processing

ICT Experts

Application Development

Search / Download

Extra-large volume data storage (25PB)

ICT Experts

Infrastructure

High Speed Network

Analysis Server

4

Various Applications
http://www.diasjp.net/en/apps_search/

Data Dissemination

Himawari-8 Satellite   Weather Forecast GPV   Radar Data   River Telemeters
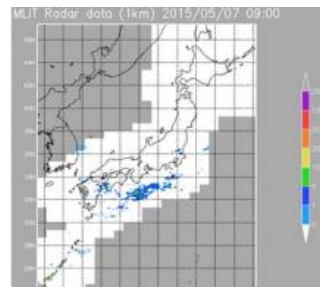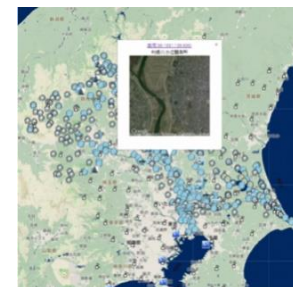
1. Climate
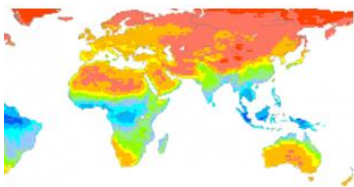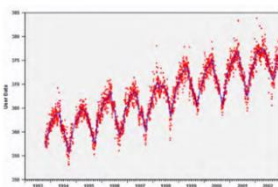
CMIP5 Model   Visualization Tools

2. Water
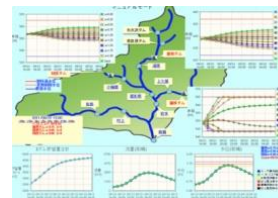
Dam Control   Water Management

3. Agriculture
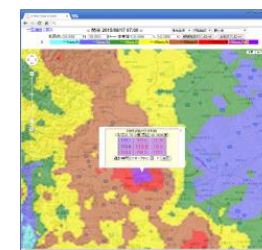

Potential of Rice Crops
after Climate Change

4. Biodiversity

Citizen science-based
observations
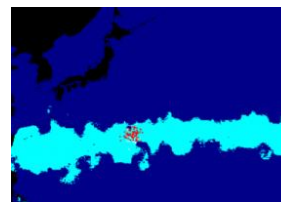
Fish eggs and growth
distribution

Accumulated
radar rainfall

5

Contents

❖ About DIAS

❖ Data and Metadata Management
 ❖ Data registration procedure
 ❖ Metadata management

❖ Open Science Activities

❖ Current and Future Prospects
 ❖ DIAS as a national repository
 ❖ Focusing on metadata quality

Data Deposit Workflow

1. Accept prior consultation	→	2. Submit an application form	→	3. Review and approve

6. Data publicity process	←	5. Data publication process	←	4. Data ingest process
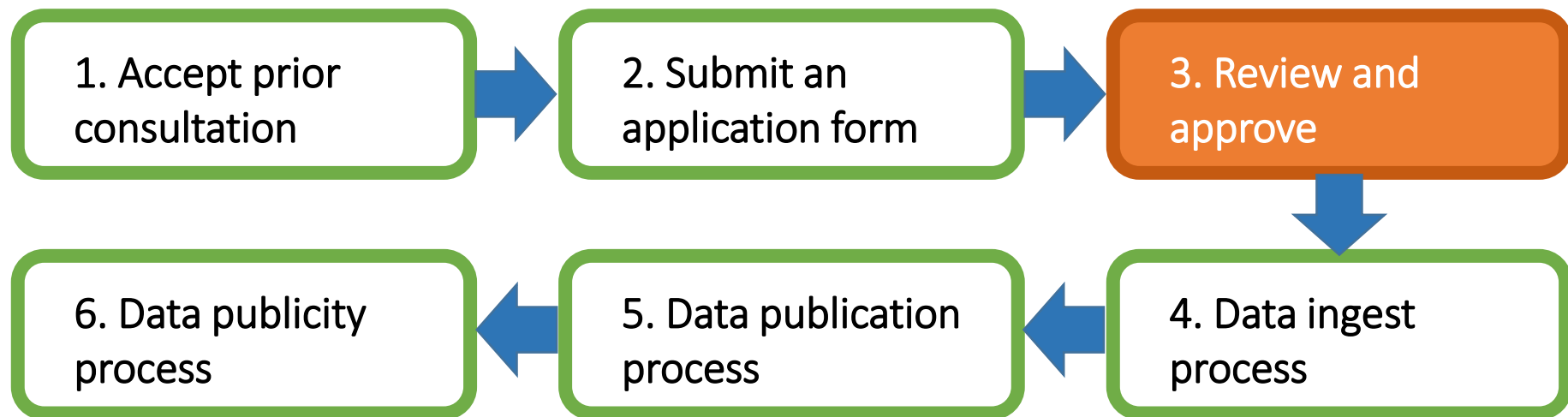
❖ The applications will be reviewed from the viewpoints of value of the data itself, compatibility with DIAS, etc.

❖ You can consult with DIAS Office dias-office@diasjp.net about the data deposit.

7

DIAS Metadata

❖ We are managing various datasets in DIAS

❖ Basic strategy:

Make dataset-level metadata in the common format for all datasets stored in DIAS

❖ The granularity of dataset is decided by the data provider

Examples of datasets

CEOP Satellite Datasets
(TRMM > PR > 3PRECI)
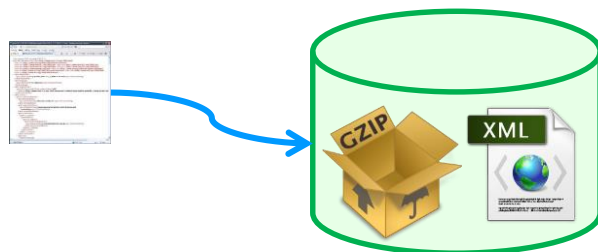
Bombus terrestris and
native bumblebee monitoring

2,694 files (gz, xml, etc.)

5 files (csv)

8
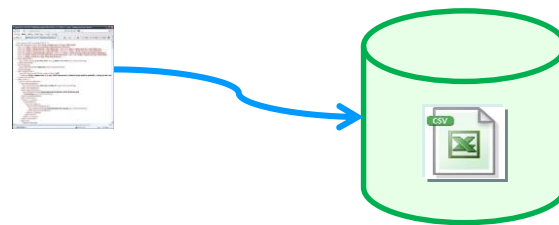
DIAS Metadata (cont.)

❖ Adopt the XML metadata used in geographic information system ISO19115 (ISO19139)

❖ We have developed web-based metadata registration tool

  ❖ Once metadata is created, documents for the dataset is automatically generated in HTML and PDF (document-metadata)

XML metadata
(ISO19115 (ISO19139))

HTML document

PDF document

9

An Example of Metadata

TITLE

Name	MIRAI CTD dataset
Metadata Identifier	MIRAI_CTD20160706144049-DIAS20160706142617-en

CONTACT

CONTACT on DATASET

Name	Data Management Office
Organization	Japan Agency for Marine-Earth Science and Technology
Address	3173-25 Showa-machi, Kanazawa-ku, Yokohama, Kanagawa, 236-0001, Japan
E-mail	dmo at jamstec.go.jp

CONTACT on PROJECT

Data Integration and Analysis System

Name	DIAS Office
Organization	The Earth Observation Data Integration and Fusion Research Initiative (EDITORIA), The University of Tokyo
Address	4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505, Japan
E-mail	dias-office@diasjp.net

DOCUMENT AUTHOR

Name	Data Management Office
Organization	Japan Agency for Marine-Earth Science and Technology
E-mail	dmo at jamstec.go.jp

DATASET CREATOR

Name	Data Management Office
Organization	Japan Agency for Marine-Earth Science and Technology
E-mail	dmo at jamstec.go.jp

DATE OF THIS DOCUMENT

2016-07-06

DATE OF DATASET

- publication : 2010-01-13

DATASET OVERVIEW

Abstract

This data set is vertical profiles of water temperature and salinity observed by CTD on research vessel MIRAI operated by JAMSTEC. This data will be utilized to analyze the characteristics of seawater and current.

Topic Category(ISO19139)

- oceans

Temporal Extent

Begin Date	1998-11-03
End Date	2008-02-10

"MIRAI CTD dataset"
http://search.diasjp.net/en/dataset/MIRAI_CTD

10

An Example of Metadata (cont.)

Geographic Bounding Box

North bound latitude	80
West bound longitude	-180
Eastbound longitude	180
South bound latitude	-60

Keywords

Keywords on Dataset

Keyword Type	Keyword	Keyword thesaurus Name
theme	[Oceans] > [Ocean Temperature], [Oceans] > [Salinity/Density] > [Salinity], [Oceans] > [Ocean Chemistry] > [Oxygen]	GCMD_science
theme	OCEANOGRAPHY PHYSICAL	AGU
theme	Climate	GEOSS
place	INDIAN OCEAN,PACIFIC OCEAN,WESTERN PACIFIC OCEAN	others

Online Resource

- Data Research System for Whole Cruise Information (DARWIN) : http://www.godac.jamstec.go.jp/darwin/e

- file download : http://dias-dss.tkl.iis.u-tokyo.ac.jp/dl/storages/filelist/dataset:5

Distribution Information

name	version	specification
CSV	no information	

DATA REMARKS

Quality levels of data are described on http://www.godac.jamstec.go.jp/darwin/explain/1/e#valueLevel.

USE CONSTRAINTS

Data Policy by Data Provider

Please refer to the data policy at http://www.jamstec.go.jp/e/database/data_policy.html and the terms and conditions at http://www.jamstec.go.jp/cruisedata/e/readme.html

Data Policy for Project

Data Integration and Analysis System

1. With regard to Data policy, if there is any data policy indicated by the data provider, that policy always has priority over this DIAS data policy.
2. DIAS data sets are to be used only for scientific research or educational purposes [*1].
3. Any modification or change of the original DIAS data sets is prohibited.
4. Any Re-export or transfer of the original data sets to a third party is

"MIRAI CTD dataset"
http://search.diasjp.net/en/dataset/MIRAI_CTD

11

DIAS Metadata Management System

❖ A Web Application.
The system manage the registered metadata at the server side.

❖ Metadata input person using this system does not need to be aware of the XML.

❖ There are minimum required fields specified by the metadata schema, and recommended fields by the DIAS.

12

A Search and Discovery System for DIAS Datasets
http://search.diasjp.net/en

❖ Overview of entire DIAS datasets

❖ Search based on keyword/spatial/temporal conditions

❖ Link to the data download system

Login

selection of
external metadata
portals

Axis type selection

Data download

Metadata download

Datasets overview by two axis

File list

Dataset document

13

Management of Data Access Privilege

❖ Access to and search for document-metadata is open to public

❖ Data Access Restrictions:

Login account is required

1. Free access
2. Agreement with data policy is required
3. Approval from data administrator is required
    - ❖ Require manual procedure for approval
    - ❖ Prepare an application form, assist on automatic email and so on.
4. Others / special treatment
    - ❖ Contact with data administrator by email or other media. If an application is approved, the user account is granted permission.
    - ❖ The system provides UI for data administrator to change the access privilege for individual user account.

14

Architecture of DIAS Metadata Systems

DIAS Metadata Management System

Registration of dataset metadata

Metadata

ISO 19139

OAI-PMH

Systems outside of DIAS

Metadata

ISO 19139

DIF   EML
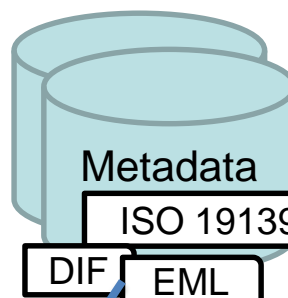
Original metadata page of each system

DIAS Dataset Search and Discovery System

http://search.diasjp.net/en
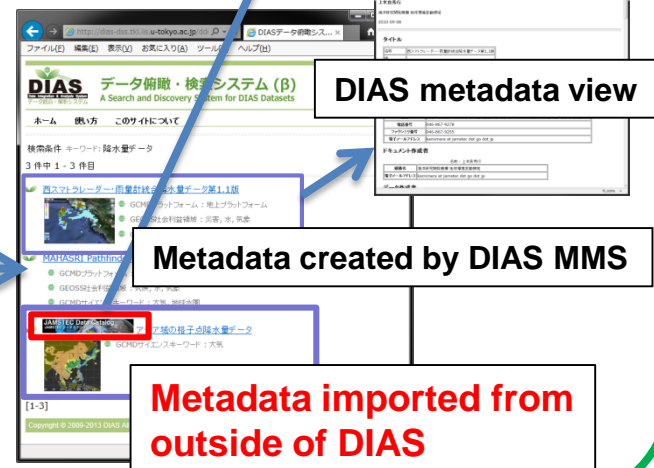
DIAS metadata view

Metadata created by DIAS MMS

Metadata imported from outside of DIAS

15

Metadata Collaboration with Systems outside of DIAS

search

DIAS metadata

Metadata from outside system(s)
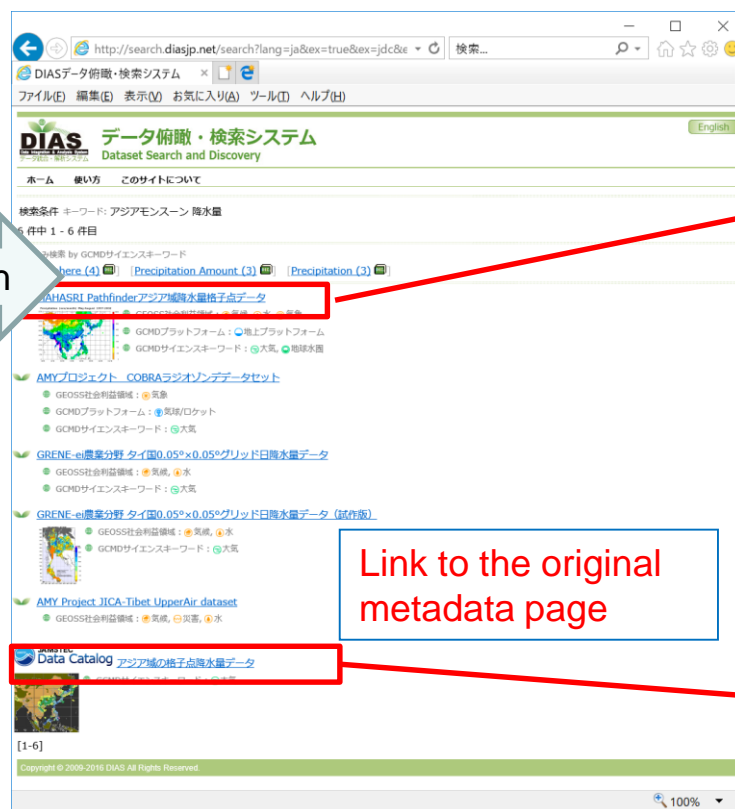
Link to the original metadata page

System	Metadata format	URL
JAMSTEC Data Catalog	DIF	http://www.godac.jamstec.go.jp/catalog/data_catalog/
JaLTER Data Catalog	EML	http://db.cger.nies.go.jp/JaLTER/
NIPR Science Database	DIF	http://scidbase.nipr.ac.jp/
NIPR Arctic Data archive System	ISO19139, DIF	https://ads.nipr.ac.jp/

16

Contents

❖ About DIAS

❖ Data and Metadata Management
   ❖ Data registration procedure
   ❖ Metadata management

❖ Open Science Activities

❖ Current and Future Prospects
   ❖ DIAS as a national repository
   ❖ Focusing on metadata quality

17

DIAS Third Phase and Open Science

1. DIAS Third Phase (2016-2020) : from research phases to the operation phase.
2. Open science : selected as one of strategic keywords in the national-level science and technology policy.
3. DIAS Open Science Special Interest Group (SIG) : planning and implementation to make DIAS ready for open science.
4. More stakeholders: variation of openness.

DOI registration for DIAS data

❖ Digital object identifier (DOI) : architecture of systems and organizations to make resources findable using a global identifier.

❖ DIAS has already started the assignment of DOI since March 2017.

  ❖ We have 26 datasets with DOI assigned in DIAS (Dec. 2017)

❖ DOI registration system from DIAS to JaLC and DataCite

  ❖ Add a new function to DIAS metadata management system to manage DOIs.

  ❖ Add DOI in each DIAS document-metadata (XML, HTML, PDF)

  ❖ Convert DIAS metadata XML to JaLC XML to registrate DOI to DataCite through JaLC

19

First Assignment of DOI on March 2017

doi:10.20783/DIAS.496

http://www.diasjp.net/infomation/
press-release-dias-first-doi-registration/

http://search.diasjp.net/en/dataset/GAME_Tibet

Landing Page with Citation Text (under development)

Domain and National Repository

❖ DIAS is a domain repository in the areas of earth science and environment.

❖ DIAS is a national repository to disseminate research results from Japan.

DIAS can take an important role among the open data policy of Japanese research organizations and funding agencies.

DIAS as a National Repository

❖ Recently, we have accepted some datasets from outside of DIAS.

    ❖ DIAS can be a candidate for storing large data.

❖ DIAS can be used as a repository of evidence data for research articles.

❖ Data deposited in DIAS can be used for submission to a data journal (e.g. ESSD).

    ❖ We are discussing on getting official certificates of trustworthy data repositories so that DIAS can be considered as trustworthy from stakeholders.

https://www.earth-system-science-data.net/

Metadata Quality Issues

❖ Some metadata do not contain enough information

   ❖ due to some reasons, such as metadata specification, usability of systems, motivation of metadata author, etc.

   ❖ Metadata quality affect the findability of datasets.

❖ I am especially focusing on keyword information in metadata.

Keywords in metadata

Keywords in document-metadata

Categorization of datasets using keywords

e.g. http://search.diasjp.net/en/dataset/MIRAI_CTD

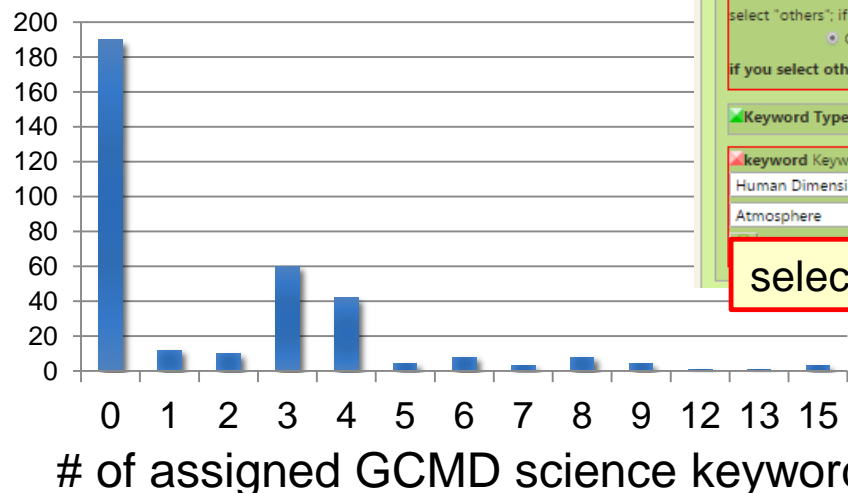**Dataset Search and Discovery**
http://search.diasjp.net/en

- ❖ **We can understand the data through keywords.**
- ❖ **Keywords are also important for search and categorization of datasets.**
  - ❖ DIAS manages various datasets.

Keywords in metadata (cont.)

❖ We don't have enough keywords in metadata
  - ❖ The cost of keyword input is high
  - ❖ It is difficult for novice users to input keywords (lack of knowledge)

❖ We are now developing keyword recommendation function.

# of datasets in DIAS

specification of ontologies (GCMD_science, GCMD_platform, GEOSS, AGU, Country, others)

selection of hierarchical keywords from menu

Current interface for keyword input boxes (in the DIAS Metadata Management System)

# of assigned GCMD science keywords

26

Summary

- ❖ DIAS is not only a data repository, but also an information platform for data science.

- ❖ We are managing various kinds of datasets through the metadata.

- ❖ We will continuously work to make DIAS more open platform.
  - ❖ DOIs to datasets
  - ❖ FAIR Data Principle

Thank you!

DIAS Website
http://www.diasjp.net/en/
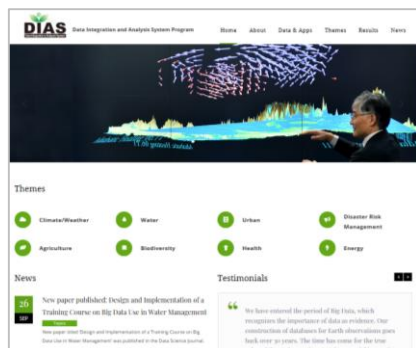
You can search DIAS datasets via

DIAS Dataset Search and Discovery System
http://search.diasjp.net/en