



# Data Integration and Analysis System





# Present DIAS & Future DIAS



# DIAS:power of big data and Future DIAS

Masaru Kitsuregawa

Professor, Institute of Industrial Science, The Univ. of Tokyo  
Director of Earth Observation Data Integration & Fusion Research Initiative(EDITORIA),  
The Univ. of Tokyo  
Director General of National Institute of Informatics(NII)

# DIAS:power of big data and Future DIAS

Masaru Kitsuregawa

Professor, Institute of Industrial Science, The Univ. of Tokyo

Director of Earth Observation Data Integration & Fusion Research Initiative(EDITORIA),

The Univ. of Tokyo

Director General of National Institute of Informatics(NII)

# EDITORIA

(Earth Observation Data Integration & Fusion Research Initiative)

The banner features the EDITORIA logo on the left, which includes a stylized 'E' and a globe. To its right is a vertical poster for '地球のためにアフリカ' (Africa for the Earth) featuring a gorilla. Further right is a 'G-SPACE EXPO 2010' advertisement for September 19-21, 2010. On the far right, there are buttons for '融合させ・アゲル' (Integrate & Elevate) and 'メンバー限定' (Member Only). Below the main banner is a navigation bar with links: 'ニュース・イベント' (News/Events), 'EDITORIAとは' (About EDITORIA), 'プロジェクト' (Projects), 'メンバー' (Members), '会議・シンポジウム' (Conferences/Symposiums), 'リンク' (Links), and '教' (Education).

東京大学  
地球観測データ統融合連携研究機構

Earth observation Data Integration and fusion Research Initiative

地球観測データ統融合連携研究機構  
(EDITORIA) は、学内の地球観測分野、情  
報科学技術分野 ※ 宇宙や農業などの公

## 参加部局

- 生産技術研究所 / ● 空間情報科学研究センター / ● 工学
- 農学生命科学研究科 / ● 大気海洋研究所

What's New (11/1/4更新)



# 4V for Bigdata

Volume

Variety

Velocity

Veracity

Volume

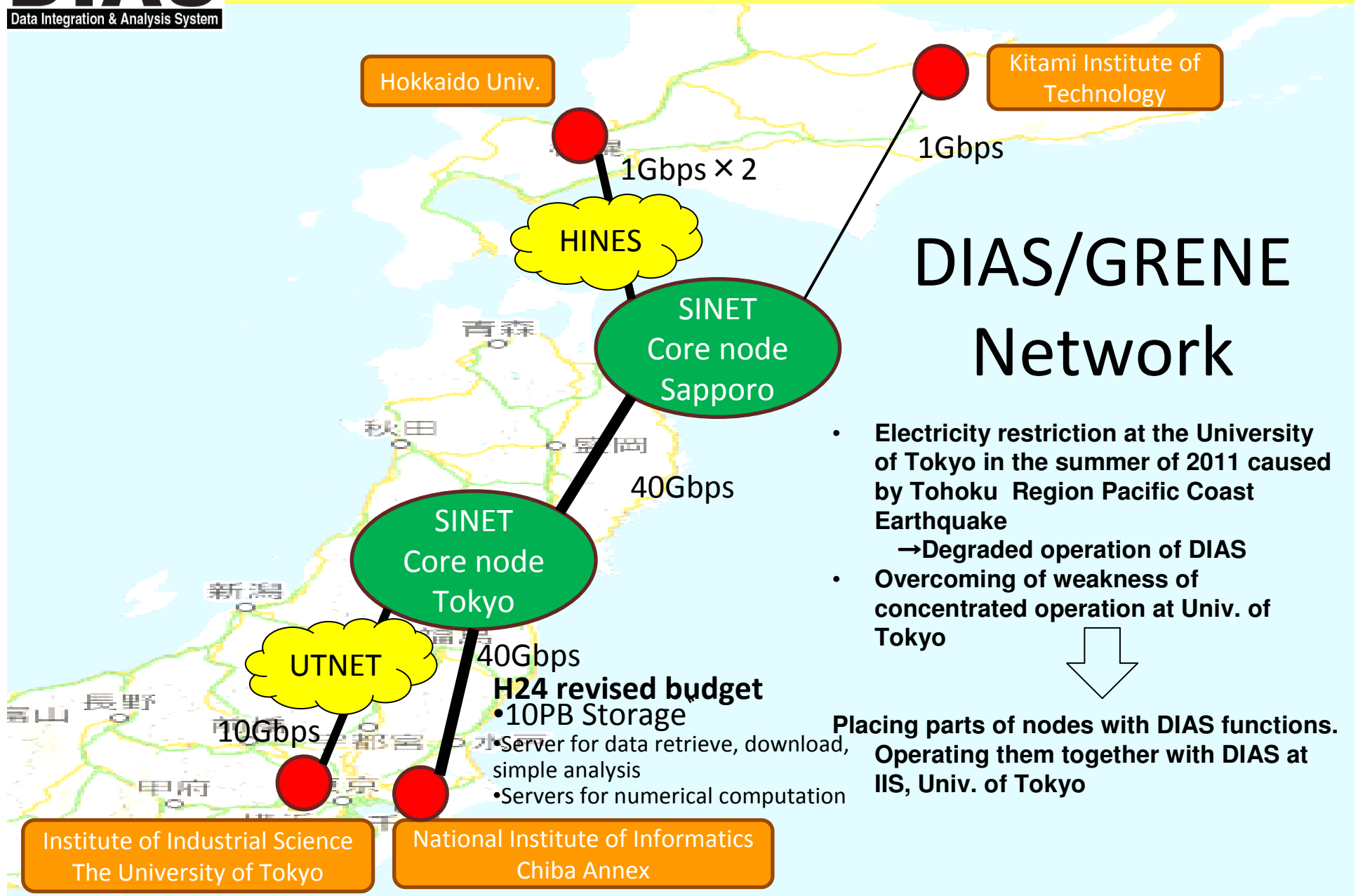




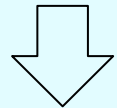
# DIAS System

disk + tape > 20PB





- Electricity restriction at the University of Tokyo in the summer of 2011 caused by Tohoku Region Pacific Coast Earthquake  
→ Degraded operation of DIAS
- Overcoming of weakness of concentrated operation at Univ. of Tokyo



Placing parts of nodes with DIAS functions.  
Operating them together with DIAS at IIS, Univ. of Tokyo

**H24 revised budget**

- 10PB Storage
- Server for data retrieve, download, simple analysis
- Servers for numerical computation

Variety


# So many kinds of Data on DIAS

**DIAS** データ俯瞰・検索システム (β)  
A Search and Discovery System for DIAS Datasets

日本語

Home How to use What's New About

**What?**  
 All:   
 Title:   
 Contact info.:   
 Abstract:

**Where?**  
  
 N   
 W   E  
 S   Global  
 overlaps  encloses

**When?**  
 From     
 Use this condition  
 To     
 Use this condition  
 overlaps  between dates

Search

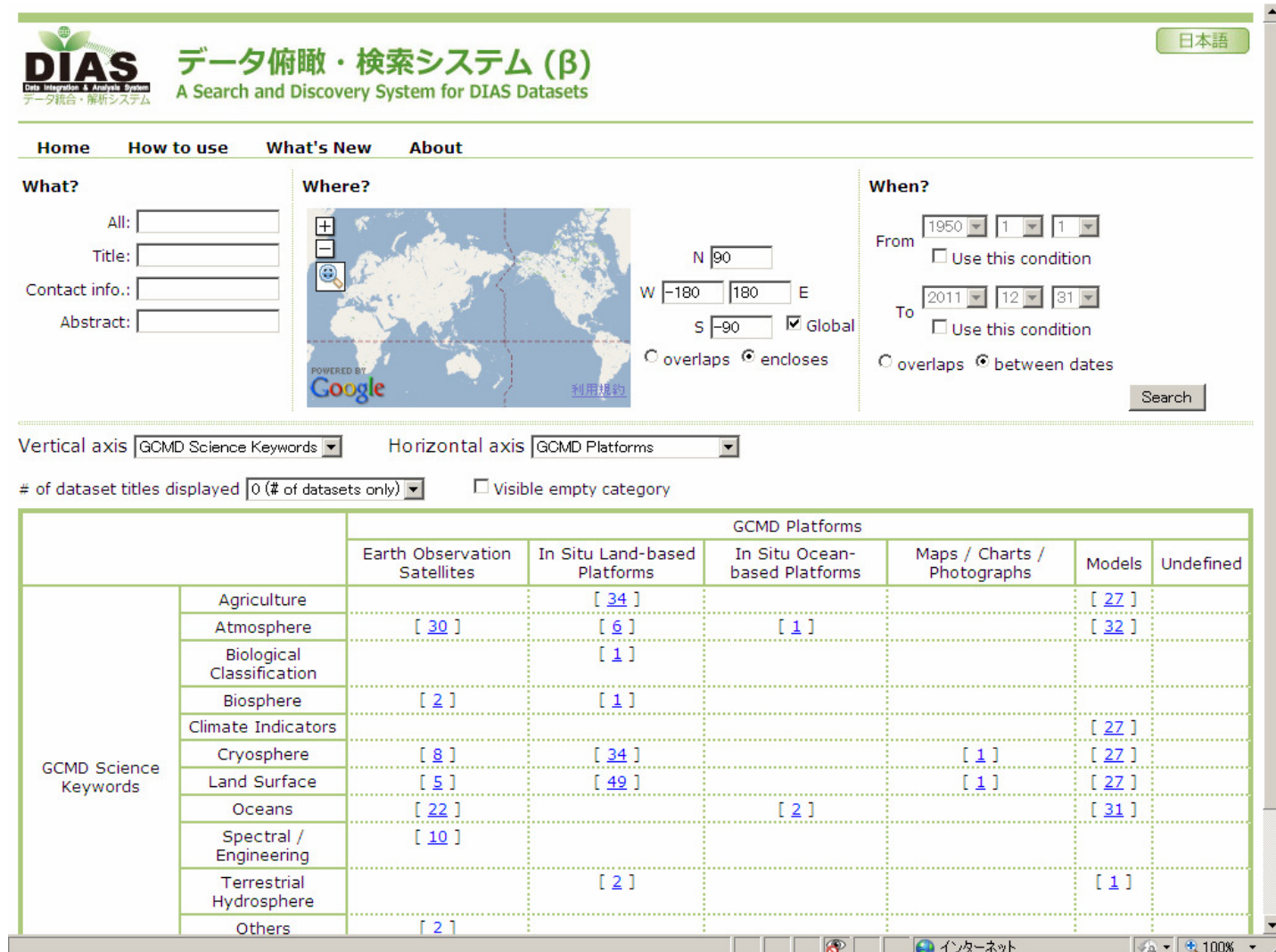
Vertical axis  Horizontal axis

# of dataset titles displayed   Visible empty category

		GCMD Platforms				
		Earth Observation Satellites	In Situ Land-based Platforms	In Situ Ocean-based Platforms	Maps / Charts / Photographs	Models
GCMD Science Keywords	Agriculture		[ 34 ]			[ 27 ]
	Atmosphere	[ 30 ]	[ 6 ]	[ 1 ]		[ 32 ]
	Biological Classification		[ 1 ]			
	Biosphere	[ 2 ]	[ 1 ]			
	Climate Indicators					[ 27 ]
	Cryosphere	[ 8 ]	[ 34 ]		[ 1 ]	[ 27 ]
	Land Surface	[ 5 ]	[ 49 ]		[ 1 ]	[ 27 ]
	Oceans	[ 22 ]		[ 2 ]		[ 31 ]
	Spectral / Engineering	[ 10 ]				
	Terrestrial Hydrosphere		[ 2 ]			[ 1 ]
	Others	[ 2 ]				

インターネット 100%

## enriching data **searching** capability



The screenshot shows the DIAS web interface. At the top, there is a navigation bar with 'Home', 'How to use', 'What's New', and 'About'. Below this are search filters for 'What?' (All, Title, Contact info., Abstract), 'Where?' (a map with latitude/longitude inputs and 'Global' checkbox), and 'When?' (From/To date ranges and 'Use this condition' checkboxes). A 'Search' button is located at the bottom right of the filter section.

Below the filters, there are dropdown menus for 'Vertical axis' (GCMD Science Keywords) and 'Horizontal axis' (GCMD Platforms). There is also a checkbox for 'Visible empty category'.

The main content is a table showing the distribution of datasets across GCMD Science Keywords and GCMD Platforms. The table has the following structure:

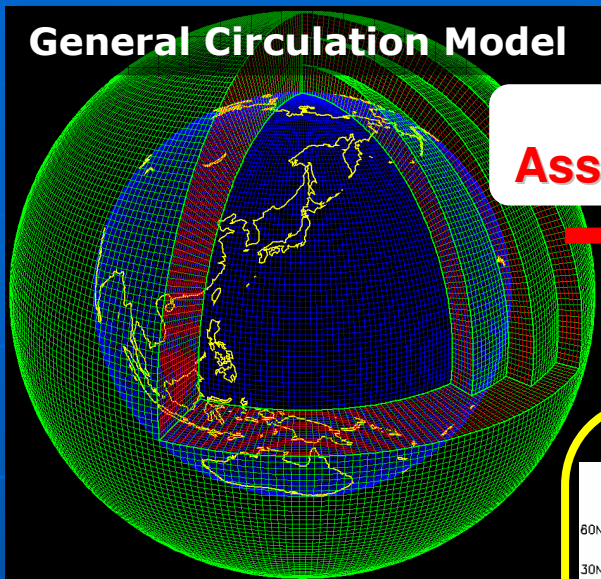
GCMD Science Keywords		GCMD Platforms					
		Earth Observation Satellites	In Situ Land-based Platforms	In Situ Ocean-based Platforms	Maps / Charts / Photographs	Models	Undefined
	Agriculture		[ 34 ]			[ 27 ]	
	Atmosphere	[ 30 ]	[ 6 ]	[ 1 ]		[ 32 ]	
	Biological Classification		[ 1 ]				
	Biosphere	[ 2 ]	[ 1 ]				
	Climate Indicators					[ 27 ]	
	Cryosphere	[ 8 ]	[ 34 ]		[ 1 ]	[ 27 ]	
	Land Surface	[ 5 ]	[ 49 ]		[ 1 ]	[ 27 ]	
	Oceans	[ 22 ]		[ 2 ]		[ 31 ]	
	Spectral / Engineering	[ 10 ]					
	Terrestrial Hydrosphere		[ 2 ]			[ 1 ]	
	Others	[ 2 ]					

# Velocity

(stock and stream)

# Global Data to Local Information

## General Circulation Model

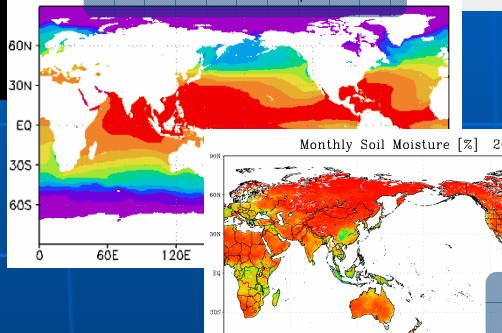


**Data Assimilation**

**Data Assimilation**

**Improved prediction**

**Satellite data**



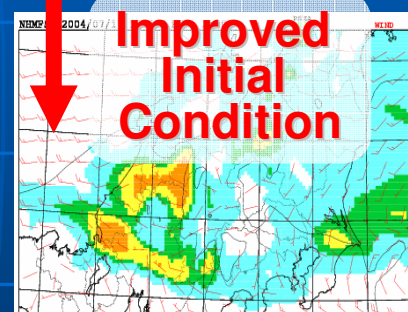
**In-situ data**



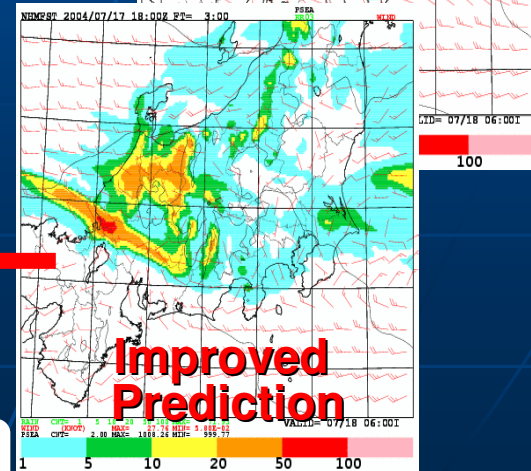
**Centralized Data System**

**Regional/Meso Model**

**Improved Initial Condition**

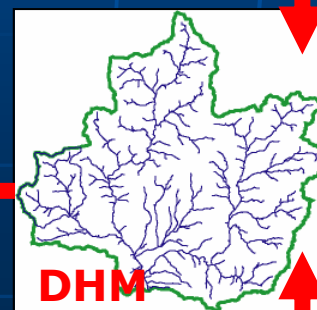
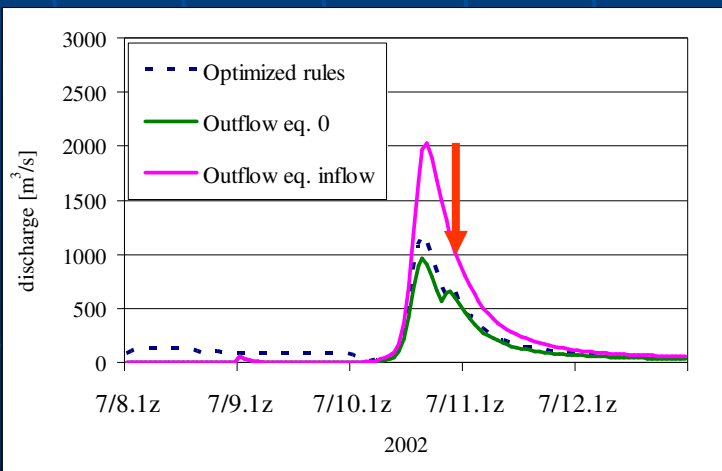


**Improved Prediction**



**Proactive Control of DAM**

**Flood Peak Reduction**

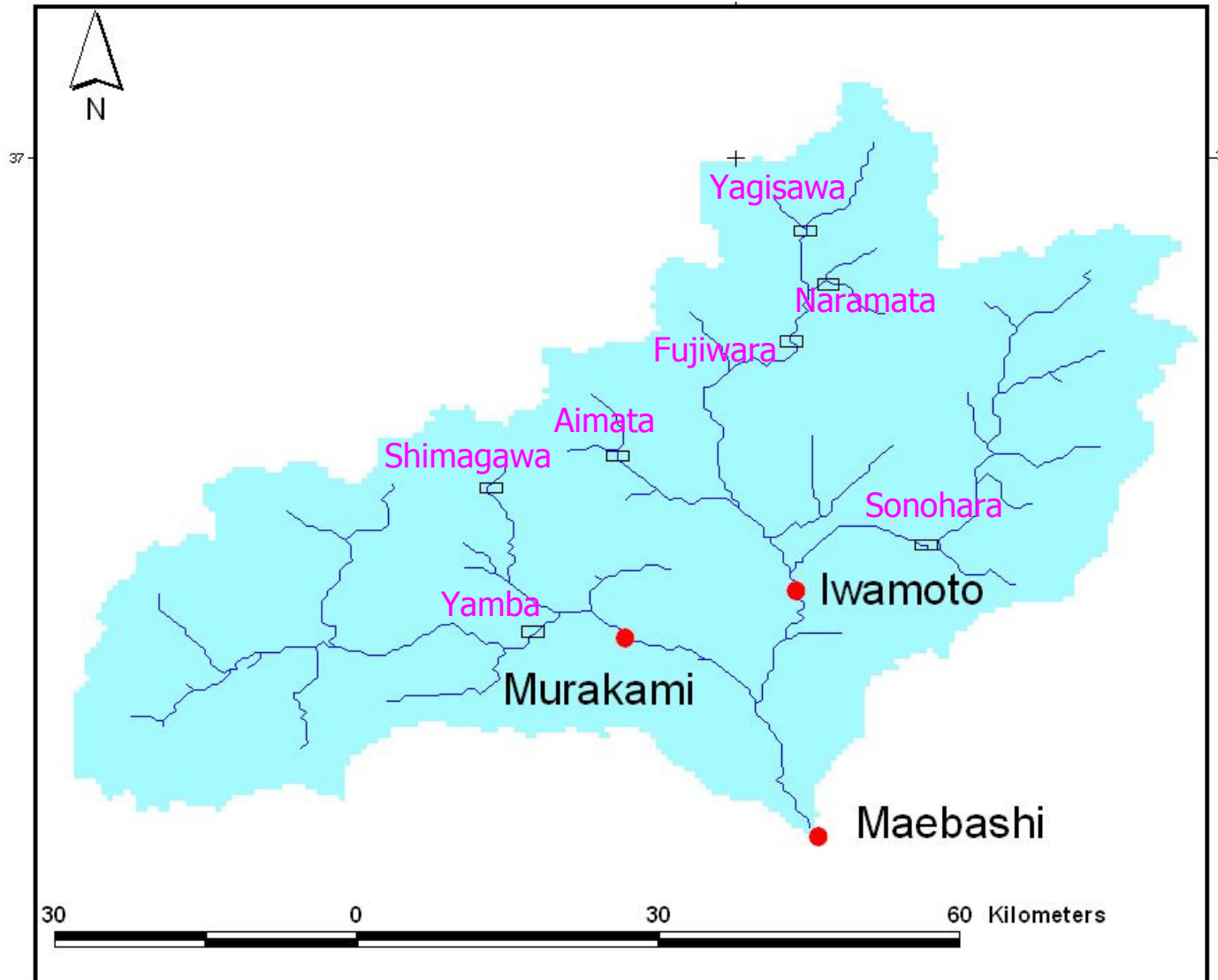


**DHM**

**Socio-Economic Data**



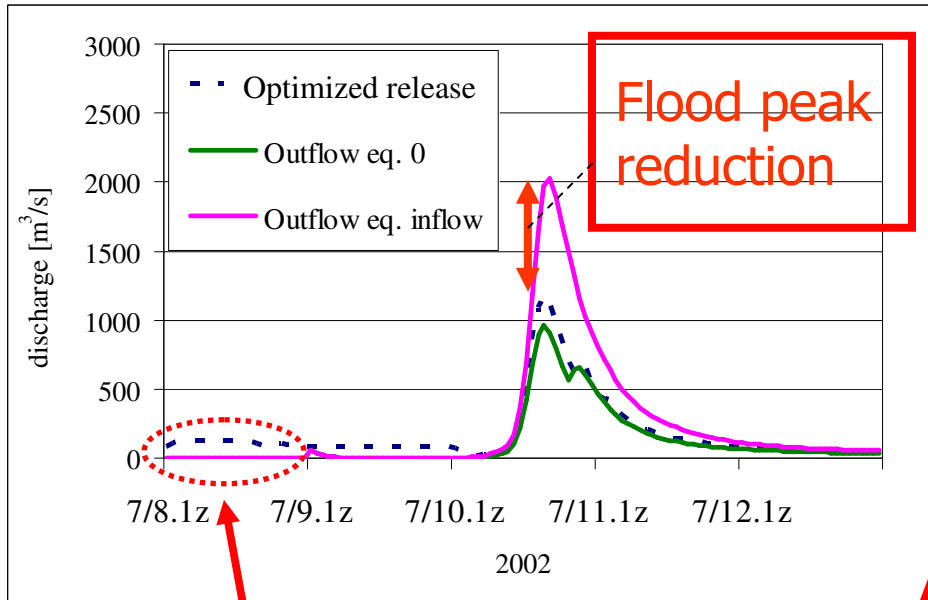
# Upper Tone River Basin



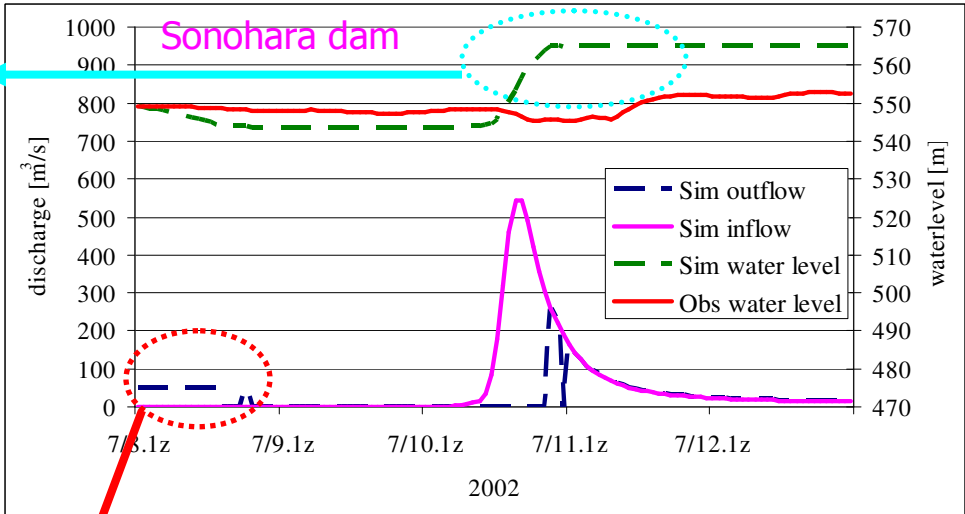
# Flood reduction by *Proactive* Control of Dam Discharge with GPV 13~18

Water is stored until max capacity is reached

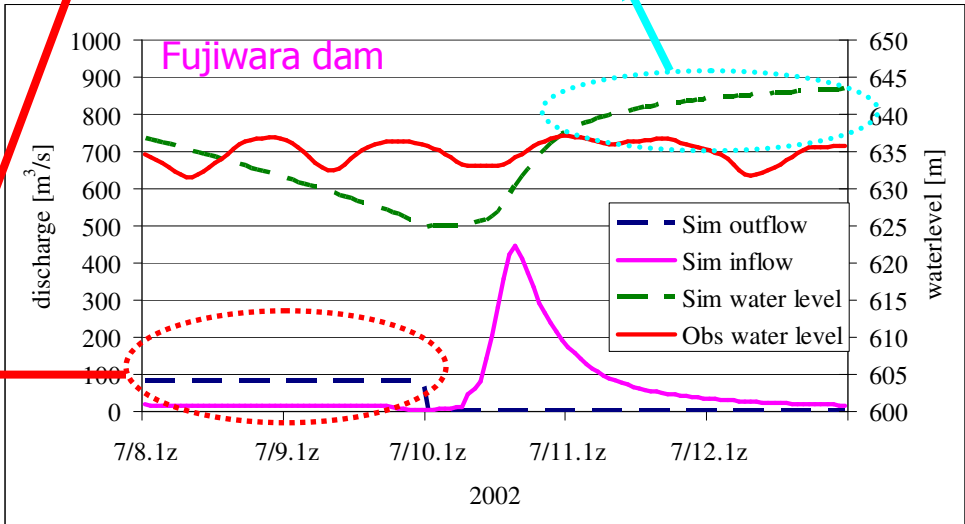
Iwamoto gauge



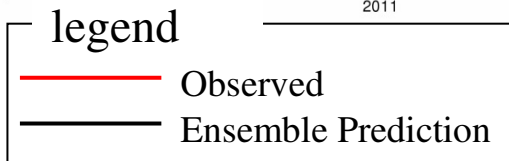
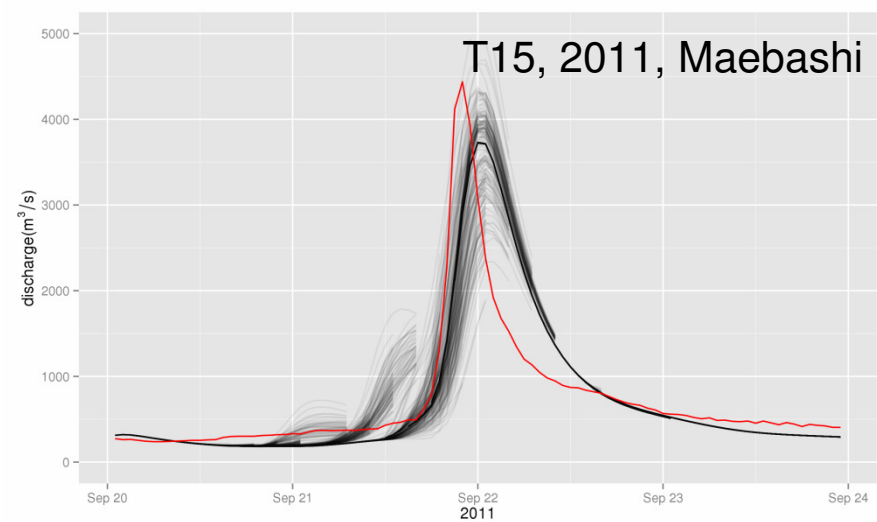
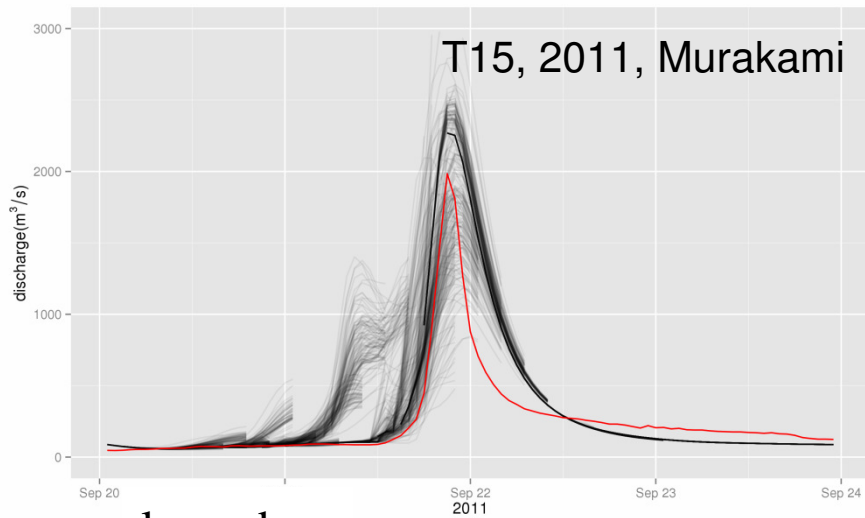
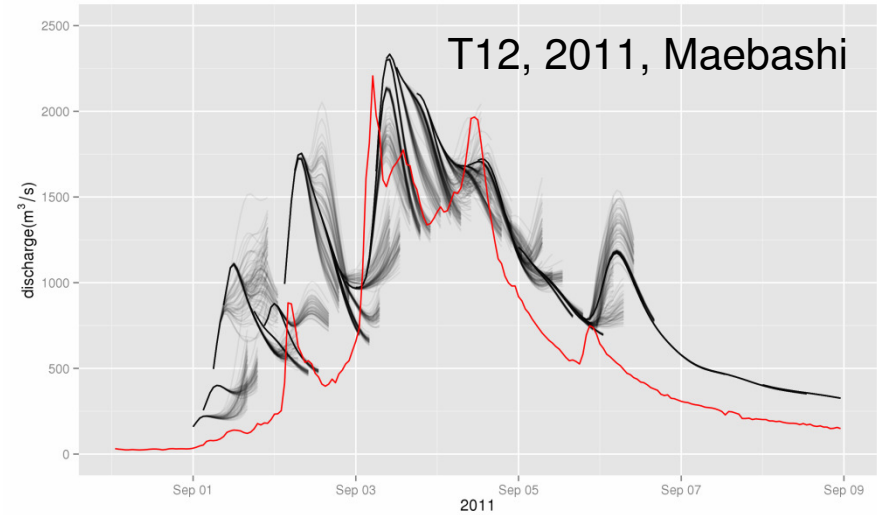
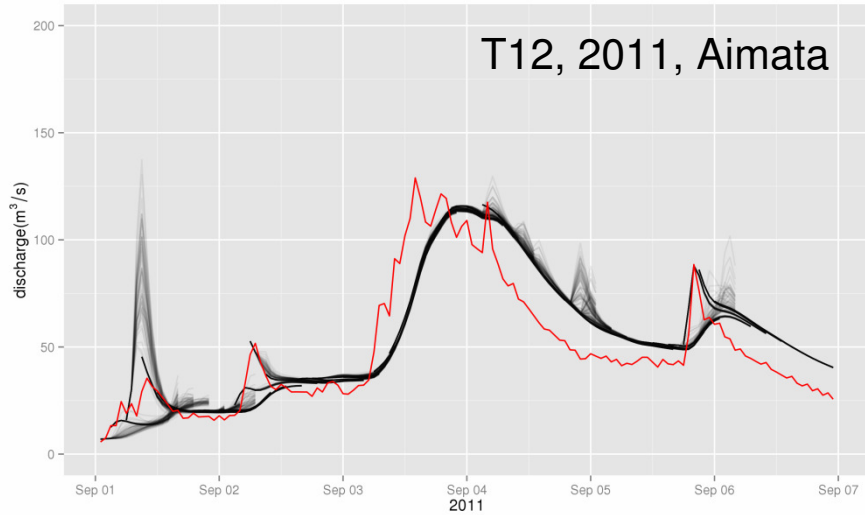
Peak created due to water release from dams



Water level increase due to storage

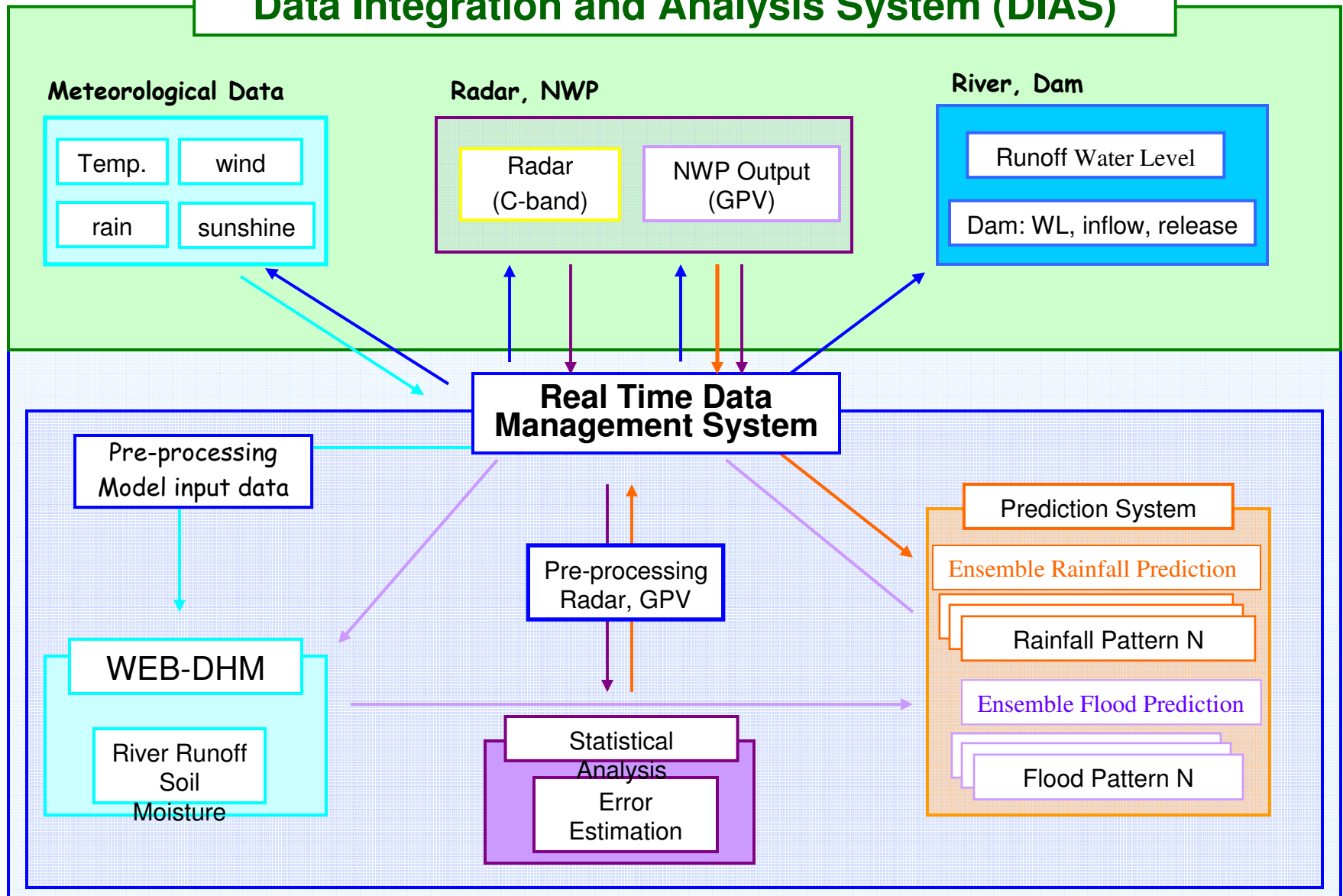


# DIAS Ensemble Flood Prediction



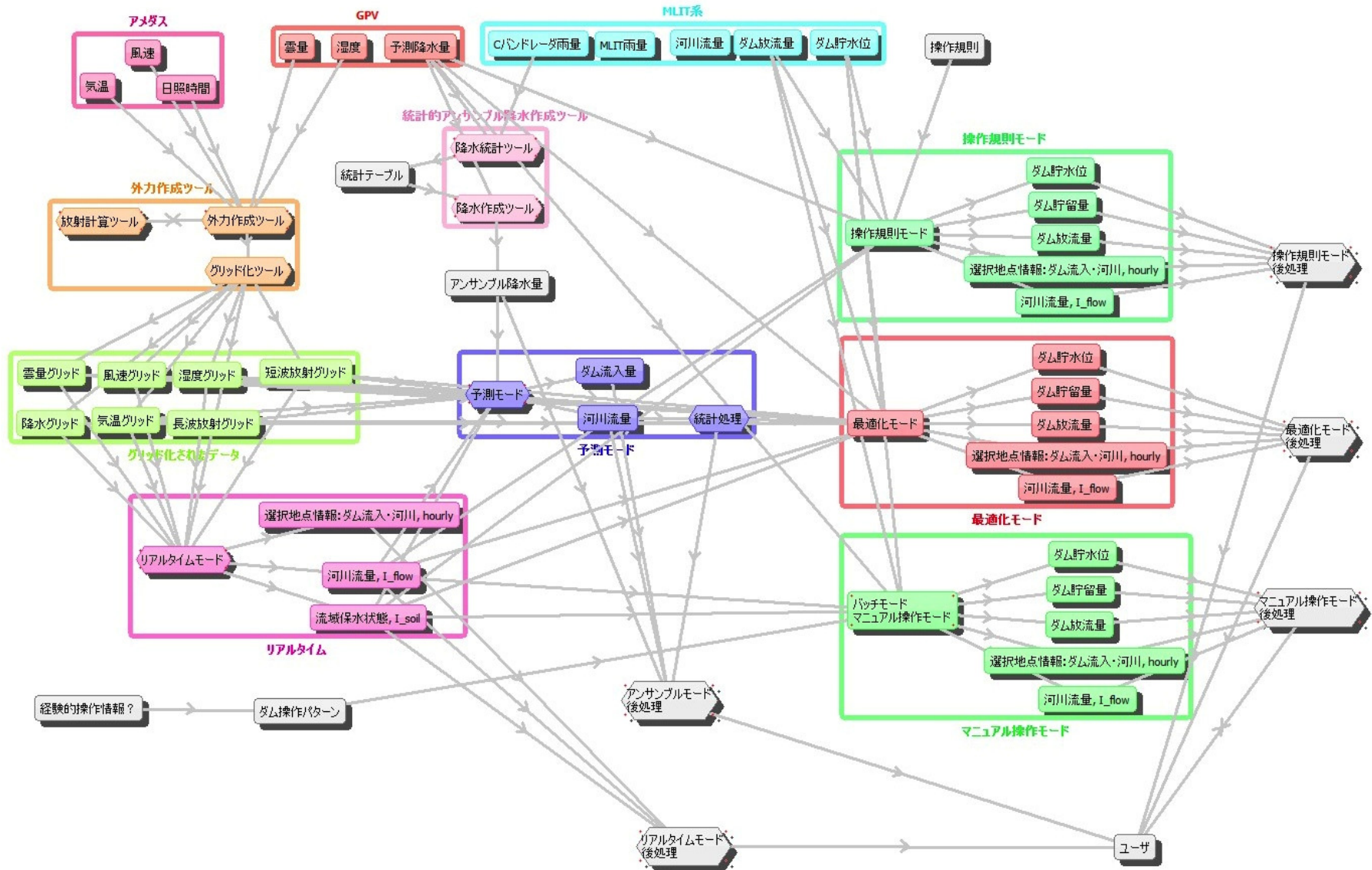
# DIAS Ensemble Flood Prediction

## Data Integration and Analysis System (DIAS)

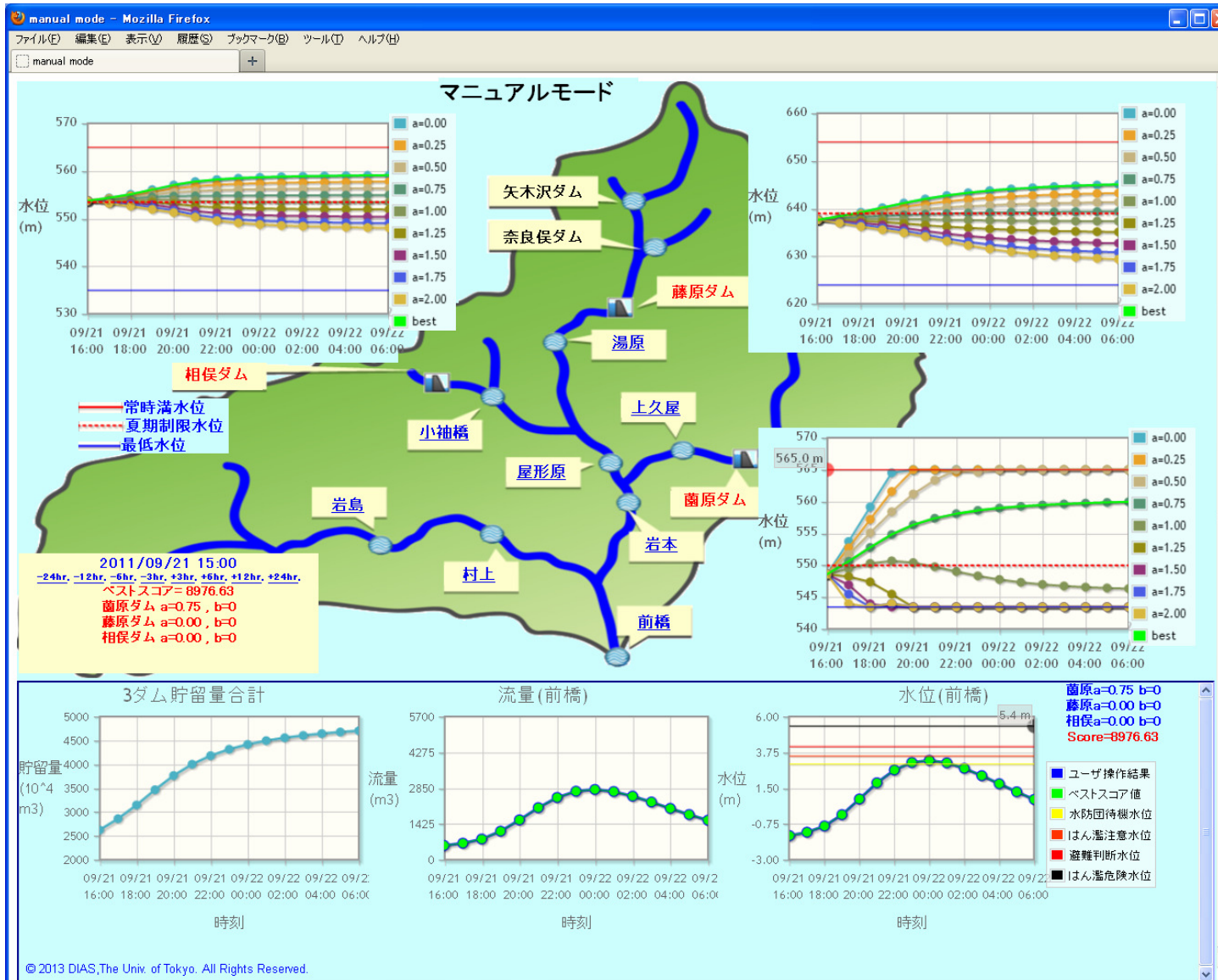


# Software Architecture

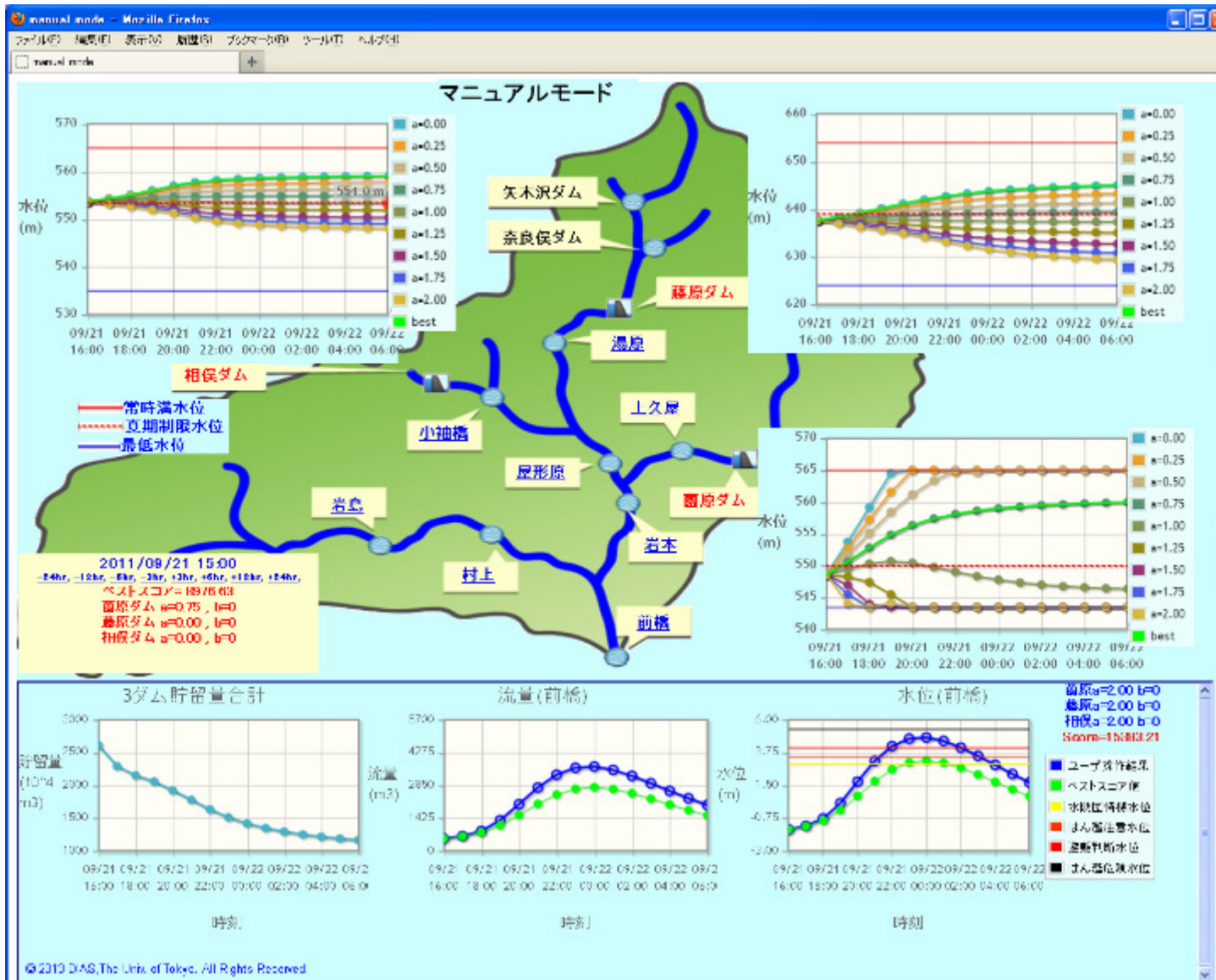
## Real time Control



# Dam Operation Simulator (Optimal)



# Dam Operation Simulator (Human Operation)



Veracity

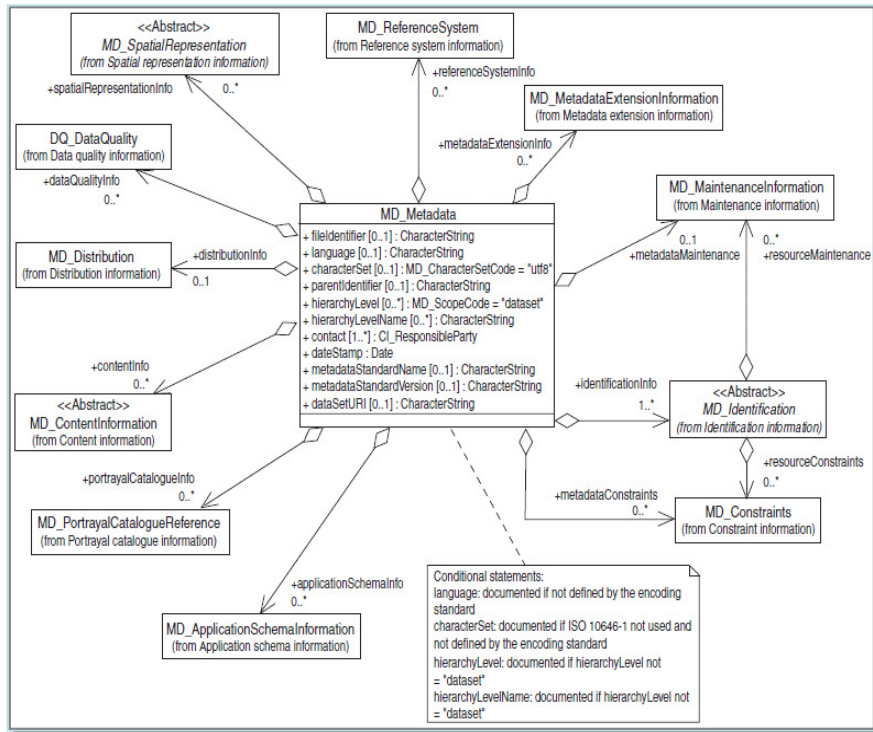


# **DIAS METADATA IMPLEMENTATION**

## ISO 19115: Geographic information - Metadata

- ❖ **ISO 19115** and its parts defines how to describe geographical information and associated services.
- ❖ The objective of this International Standard is to provide a clear procedure for the description of digital geographic datasets.
- ❖ **ISO 19139** provides the XML implementation schema for ISO 19115 specifying the metadata record format and may be used to describe, validate, and exchange geospatial metadata prepared in XML.

# ISO19115, ISO19139



ISO19115 UML

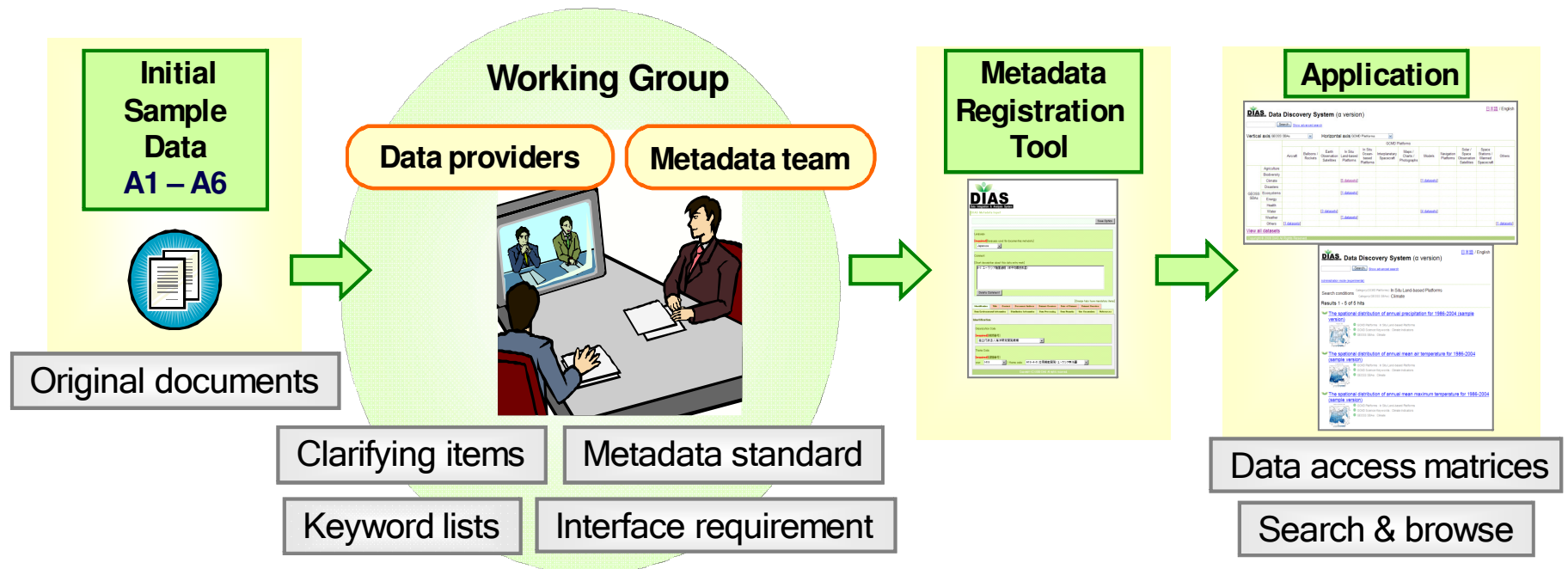
```

<?xml version="1.0" encoding="utf-8" ?>
<xs:schema targetNamespace="http://www.isotc211.org/2005/gmd" elementFormDefault="qualified"
  version="0.1" xmlns:xs="http://www.w3.org/2001/XMLSchema"
  xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:gco="http://www.isotc211.org/2005/gco"
  xmlns:gmd="http://www.isotc211.org/2005/gmd">
  <!-- Annotation -->
  <xs:annotation>
    <xs:documentation>This file was generated from ISO TC/211 UML class diagrams == 01-26-2005
    12:40:00 =====</xs:documentation>
  </xs:annotation>
  <!-- Imports -->
  <xs:import namespace="http://www.isotc211.org/2005/gco" schemaLocation="../../gco/gco.xsd" />
  <xs:include schemaLocation="../../gmd/spatialRepresentation.xsd" />
  <xs:include schemaLocation="../../gmd/metadataExtension.xsd" />
  <xs:include schemaLocation="../../gmd/content.xsd" />
  <xs:include schemaLocation="../../gmd/metadataApplication.xsd" />
  <xs:include schemaLocation="../../gmd/applicationSchema.xsd" />
  <xs:include schemaLocation="../../gmd/portrayalCatalogue.xsd" />
  <xs:include schemaLocation="../../gmd/dataQuality.xsd" />
  <xs:include schemaLocation="../../gmd/freeText.xsd" />
  <!-- Classes -->
  <xs:complexType name="MD_Metadata_Type">
    <xs:annotation>
      <xs:documentation>Information about the metadata</xs:documentation>
    </xs:annotation>
    <xs:complexContent>
      <xs:extension base="gco:AbstractObject_Type">
        <xs:sequence>
          <xs:element name="fileIdentifier" type="gco:CharacterString_PropertyType" minOccurs="0" />
          <xs:element name="language" type="gco:CharacterString_PropertyType" minOccurs="0" />
          <xs:element name="characterSet" type="gmd:MD_CharacterSetCode_PropertyType" />
        </xs:sequence>
      </xs:extension>
    </xs:complexContent>
  </xs:complexType>
  </xs:schema>
  
```

ISO19139 XML

# Collaborative development of DIAS metadata implementation

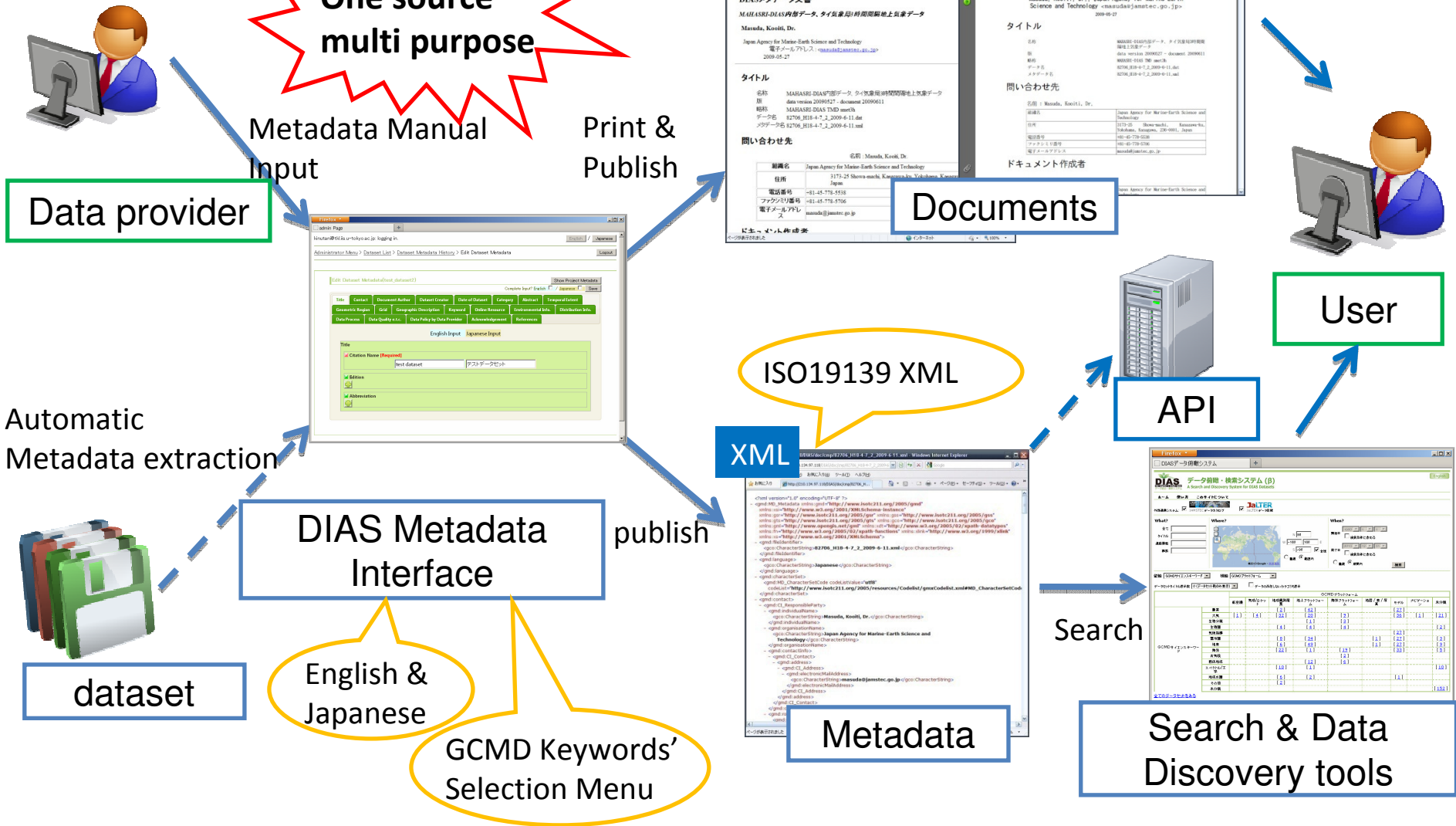
Collaborative development activities with data providers (JAMSTEC) and system developers (Universities)



- ⊕ We analyzed data providers' holding dataset documentation.
- ⊕ We collected metadata in the dataset documentation.
- ⊕ We mapped between ISO 19115 metadata elements and the dataset documentation items.
- ⊕ We selected useful keyword lists in order to search effectively.

# DIAS Metadata & Document Creation System

**One source – multi purpose**



Data provider

Metadata Manual Input

Print & Publish

Automatic Metadata extraction

DIAS Metadata Interface

publish

dataset

English & Japanese

GCMD Keywords' Selection Menu

HTML

PDF

Documents

User

ISO19139 XML

XML

API

Search

Metadata

Search & Data Discovery tools

# DIAS Metadata & Document Creation System

Edit Dataset Metadata(Sample\_in\_Manual)

Show Project Metadata

Complete Input? English  / Japanese  Save

Title	Contact	Document Author	Dataset Creator	Date of Dataset	Category	Abstract	Temporal Extent
Geometric Region	Grid	Geographic Description	Keyword	Online Resource	Environmental Info.	Distribution Info.	
Data Process	Data Quality e.t.c.	Data Policy by Data Provider	Acknowledgement	References			

English Input Japanese Input

**Title**

✖ Citation Name **[Required]**

<input type="text" value="Sample_in_Manual"/>	<input type="text" value="サンプル_マニュアル"/>
---	---

✔ Edition ✖

<input type="text" value="1.0"/>	
----------------------------------	--

✔ Abbreviation ✖

<input type="text" value="SIM"/>	
----------------------------------	--

# DIAS Metadata & Document Creation System

Edit Dataset Metadata(Sample\_in\_Manual) Show Project Metadata

Complete Input? English  / Japanese  Save

Title Contact Document Author Dataset Creator Date of Dataset Category Abstract Temporal Extent  
Geometric Region Grid Geographic Description Keyword Online Resource Environmental Info. Distribution Info.  
Data Process Data Quality e.t.c. Data Policy by Data Provider Acknowledgement References

<ul style="list-style-type: none"> <li>•Title</li> <li>•Contact</li> <li>•Document Author</li> <li>•Dataset Creator</li> <li>•Date of Dataset</li> <li>•Category</li> <li>•Abstract</li> <li>•Temporal Extent</li> <li>•Geometric Region</li> <li>•Grid</li> <li>•Geographic Description</li> </ul>	<ul style="list-style-type: none"> <li>•Keyword               <ul style="list-style-type: none"> <li>•GCMD Science</li> <li>•GCMD Platform</li> <li>•AGU</li> <li>•GEOSS</li> <li>•GEO_COP</li> <li>•Country etc.</li> </ul> </li> <li>•Online Resource</li> <li>•Environmental Info.</li> <li>•Distribution Info.</li> <li>•Data Process</li> </ul>	<ul style="list-style-type: none"> <li>•Data Quality</li> <li>•Data Policy by Data Provider</li> <li>•Acknowledgement</li> <li>•References</li> </ul>
---	--	---

# Dataset Metadata Edit

Input items have been tabbed.  
Please select an item while switching.

You can confirm the related projects' metadata using this button.

Edit Dataset Metadata(Sample\_in\_Manual) Show Project Metadata

Complete Input? English  / Japanese  Save

Title	Contact	Document Author	Dataset Creator	Date of Dataset	Category	Abstract	Temporal Extent
Geometric Region	Grid	Geographic Description	Keyword	Online Resource	Environmental Info.	Distribution Info.	
Data Process	Data Quality e.t.c.	Data Policy by Data Provider	Acknowledgement	References			

English Input

Japanese Input

**Title**

✖ Citation Name [Required]

Sample\_in\_Manual

サンプル\_マニュアル

▶ Edition +

▶ Abbreviation ✖

There are two boxes side by side when you need to input the field in both English and Japanese.



# Imported vocabulary resources

- We have imported the following resources in DIAS Interoperability portal. Properly, we get permissions for our fair use from original resource providers.

No.	Name	Language	Data format
1	WMO Glossary	English	Web pages
2	CEOS Missions, Instruments and Measurements(MIM) Database	English	MS Excel
3	CEOS Systems Engineering Office(SEO)	English	MS Excel
4	GEMET (GEneral Multilingual Environmental Thesaurus)	English	RDF
5	INSPIRE((Infrastructure for Spatial Information in the European Community) Feature Concept Dictionary	English	Web pages
6	NASA SWEET	English	OWL
7	CUAHSI (Consortium of Universities for the Advancement of Hydrologic Science)	English	OWL
8	JAXA glossary	Japanese	Web pages
9	Rremote sensing glossary for Japanese RS academic society	Japanese	Books (paper media)
10	GIS glossary for Japanese GIS academic society	Japanese	MS Word

# Data Index/Search (basic keyword search)

- Users can search DIAS dataset by using basic keyword search function.

① Select "Data Index/Search" tab

② Input keywords

(ex. "flood" or "AWCI")

③ Push "search"

The screenshot shows the DIAS Interoperability Portal search interface. The search bar contains the keyword "flood". The search results list includes several entries related to the Global Earth Observation System of Systems (GEOSS) and the Asian Water Cycle Initiative (AWCI). The results are ranked by similarity scores, with the top three results showing a score of 2.91. The search results are displayed in a table format with columns for Dataset, Document, Title, Contact, Document, Author, and Dataset. The search results are also displayed in a network diagram on the left side of the page, showing the relationship between the search term and related datasets.

④ Visualizing dataset related to the input keyword

⑤ Similarity scores

# Technical Term Search

- Users can check the definition of terms by using 'technical term search' function.

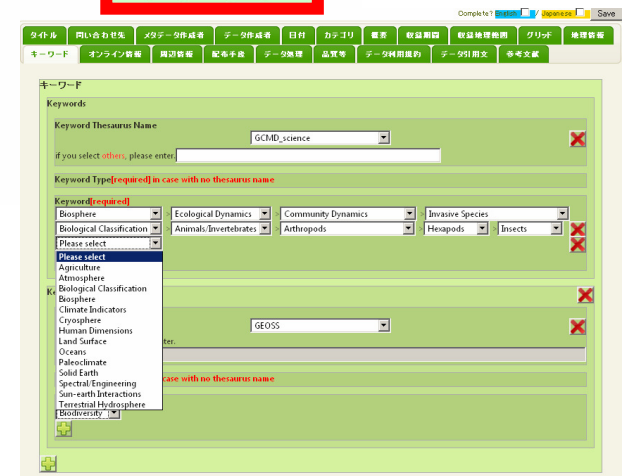
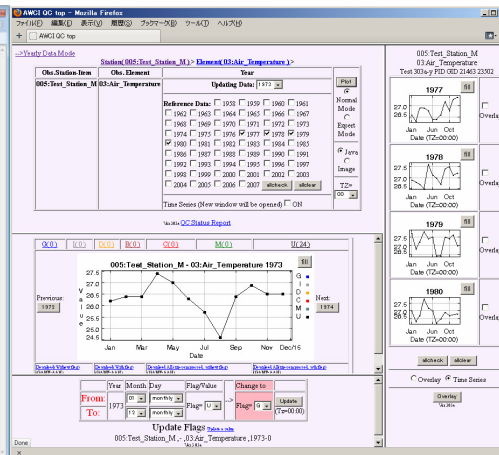
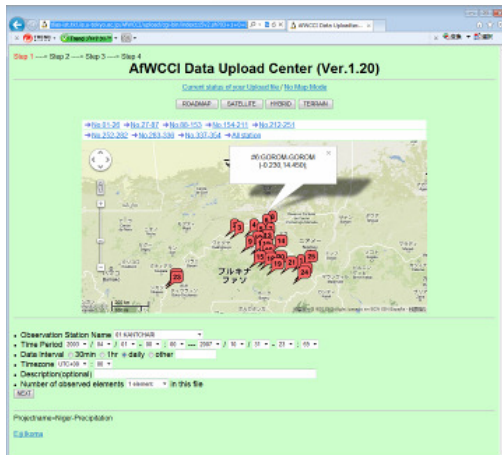
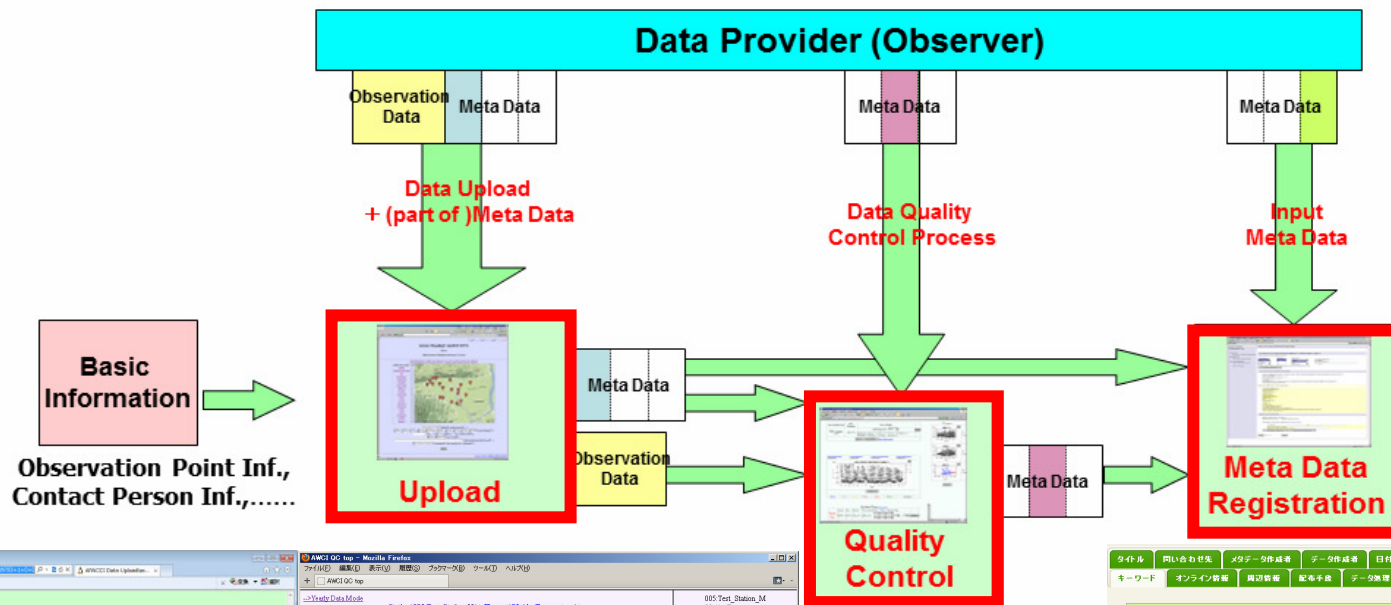
- Select "Technical Term Search" tab
- Select targets in list of ontologies, terminologies, glossaries and dictionaries.
- Input keywords (ex. "basin")
- Push "search"

The screenshot shows the DIAS Interoperability Portal interface. The 'Technical Term Search' tab is selected. The search input field contains the keyword 'basin'. The search button is labeled 'search'. The search results are displayed in a list format, including the following entries:

Term	Score	Relevant
Basin (SWEET)	14.09	relevant term
Hydrographic basin (GEMET)	9.46	relevant term
Storm water basin (GEMET)	8.95	relevant term
Catchment (SWEET)	5.62	relevant term
Area management/restriction/regulation zones and reporting units (INSPIRE FCD)	4.34	relevant term
COSMO-SkyMed 4 (CEOS SEO Database)	4.32	relevant term

- Visualizing terms in each dictionaries related to the input keyword

# 3-stage Pipeline: Data upload, Data Quality Control and Metadata registration



**Noise cannot be Zero  
but minimized!**

We plan to introduce data lineage.

Upon 4V's criteria  
DIAS is truly well-organized  
'Big Data'

How about Big Data in other  
domain?



# What happens on the Cyber space every sixty seconds?



ashahel.wordpress.com/





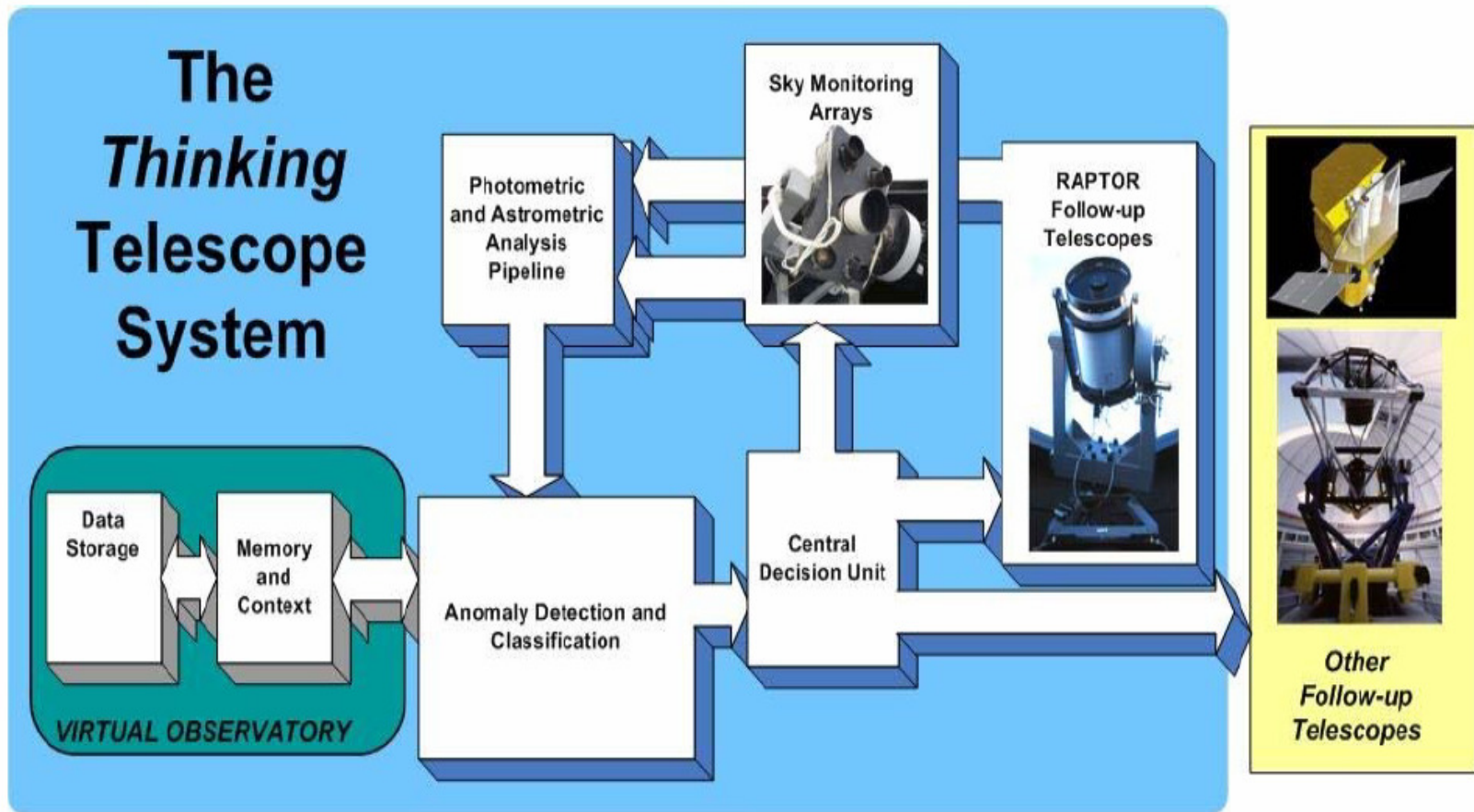
There might be earth monitoring data  
being pushed onto Cyberspace.  
But I have never heard how much!!



We have to orchestrate various activities

# Automatic and Active Monitoring and Injection

# Recent astronomy: Thinking telescope



Style of Data Intensive Science is evolving!

# Conclusion

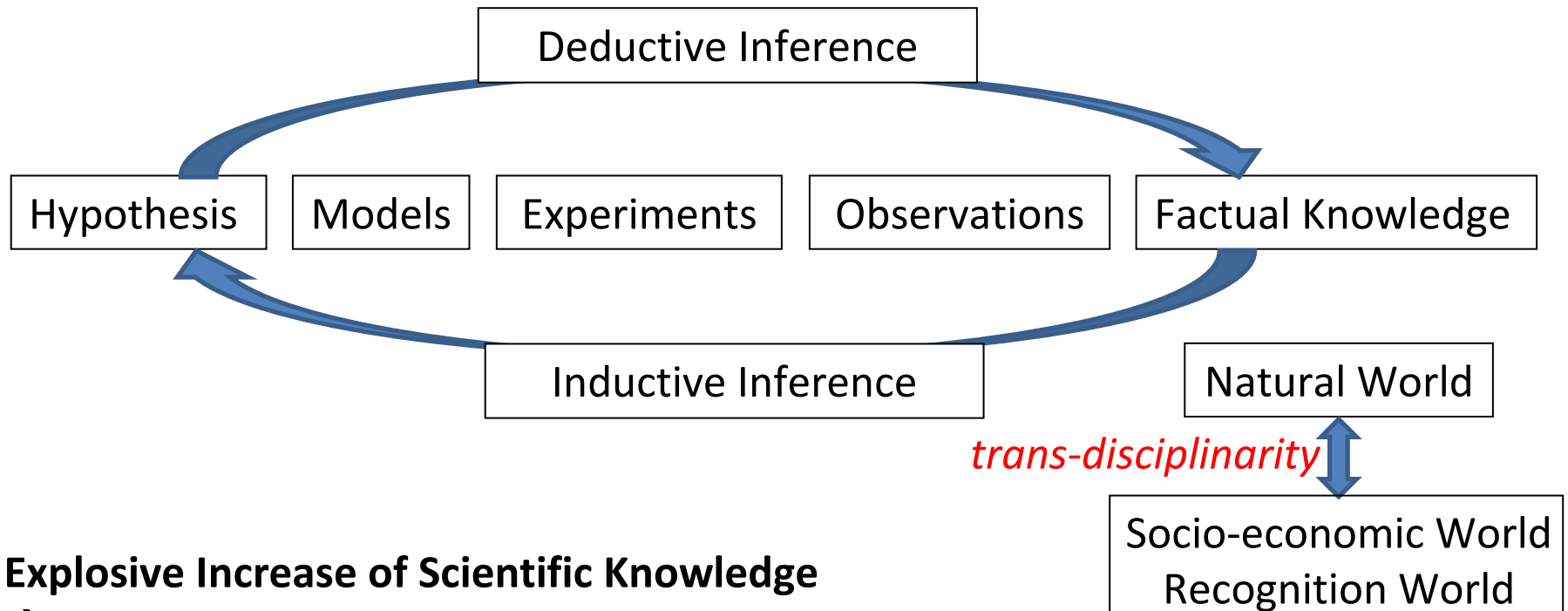
- Extreme IT for Data Automation
- More time to think and design new social solution.
- Once you have an idea on societal innovative service idea, we are more than happy to 'Realize' with our **IT power** for you.

Nothing Is impossible for us.

# *inter-disciplinarity & trans-disciplinarity*

## Scientific Knowledge

Formal Knowledge which can be transferred and shared among vast amount of Factual Knowledge



### Explosive Increase of Scientific Knowledge

→ **Differentiation & Systematization: disciplines**

- Accumulated sub-system knowledge can not be reflected to holistic knowledge.
- Effects of a whole system can not be involved in a targeted sub-system.

→ **Far from fundamental solutions of issues across disciplines** → *inter-disciplinarity*

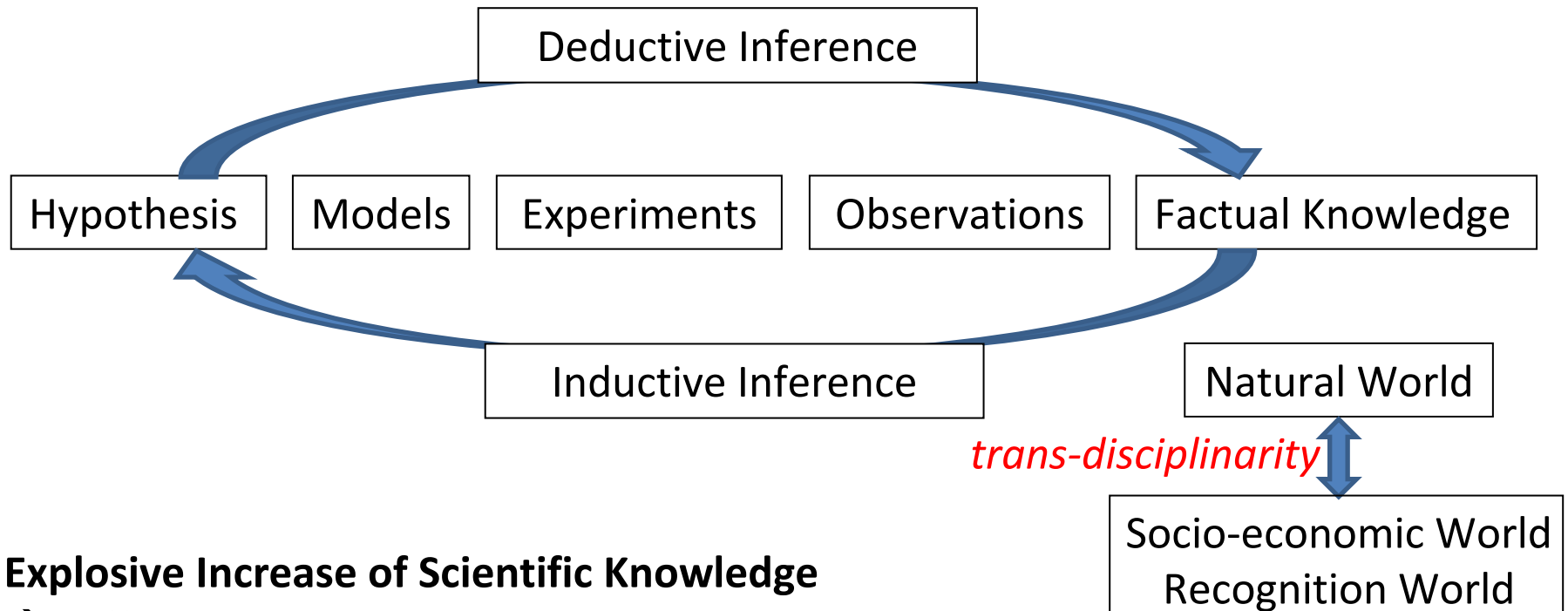
**Data System**

**Integration-Interlinkage System**

# *inter-disciplinarity & trans-disciplinarity*

## Scientific Knowledge

Formal Knowledge which can be transferred and shared among vast amount of Factual Knowledge



### Explosive Increase of Scientific Knowledge

#### → Differentiation & Systematization

- Accumulated sub-system knowledge can not be reflected to a whole system.
- Effects of a whole system can not be involved in a targeted sub-system.

→ Far from fundamental solutions of issues across disciplines → *inter-disciplinarity*

**Data System**

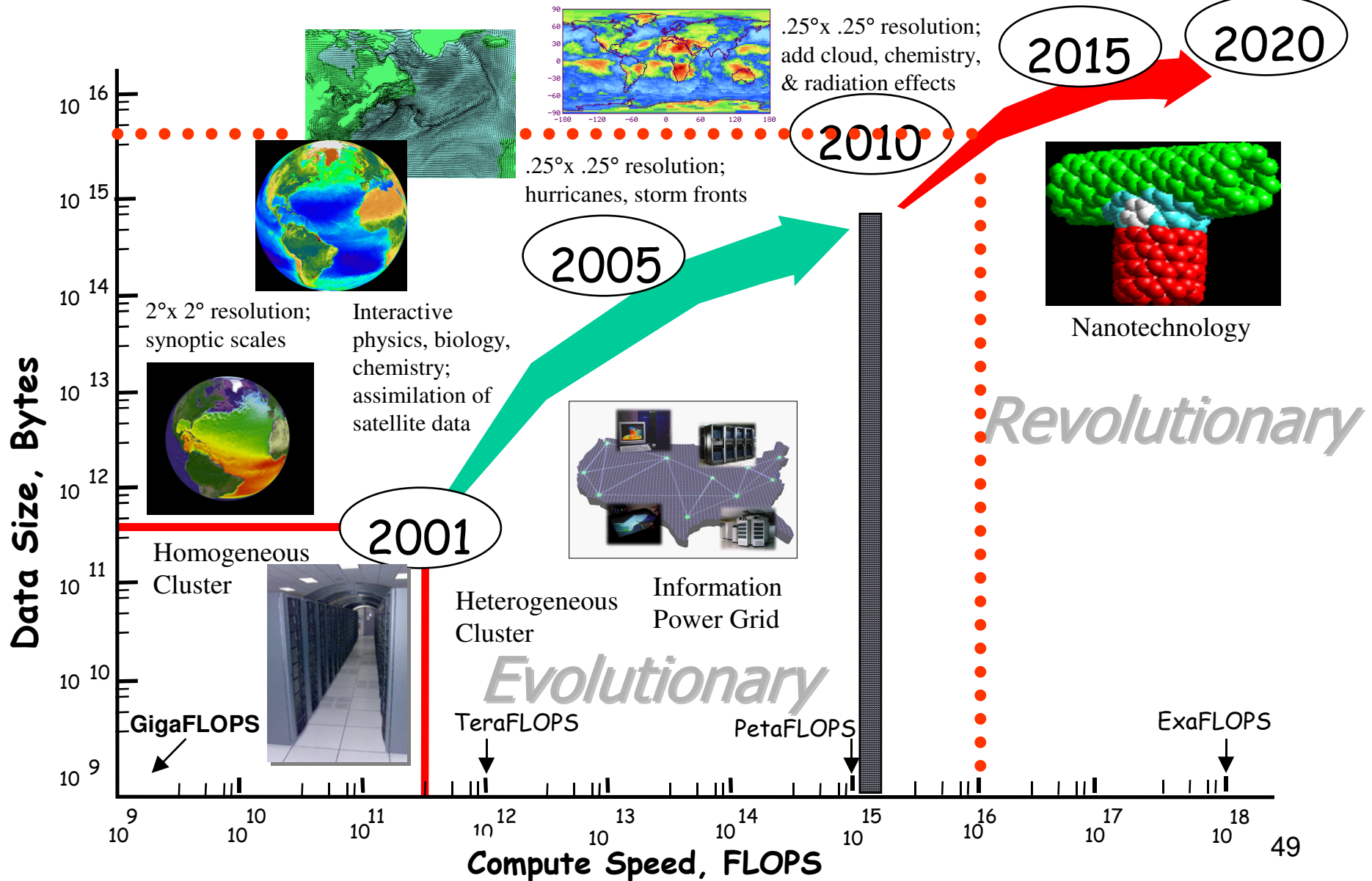
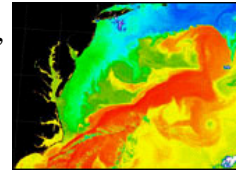
**Integration-Interlinkage System**





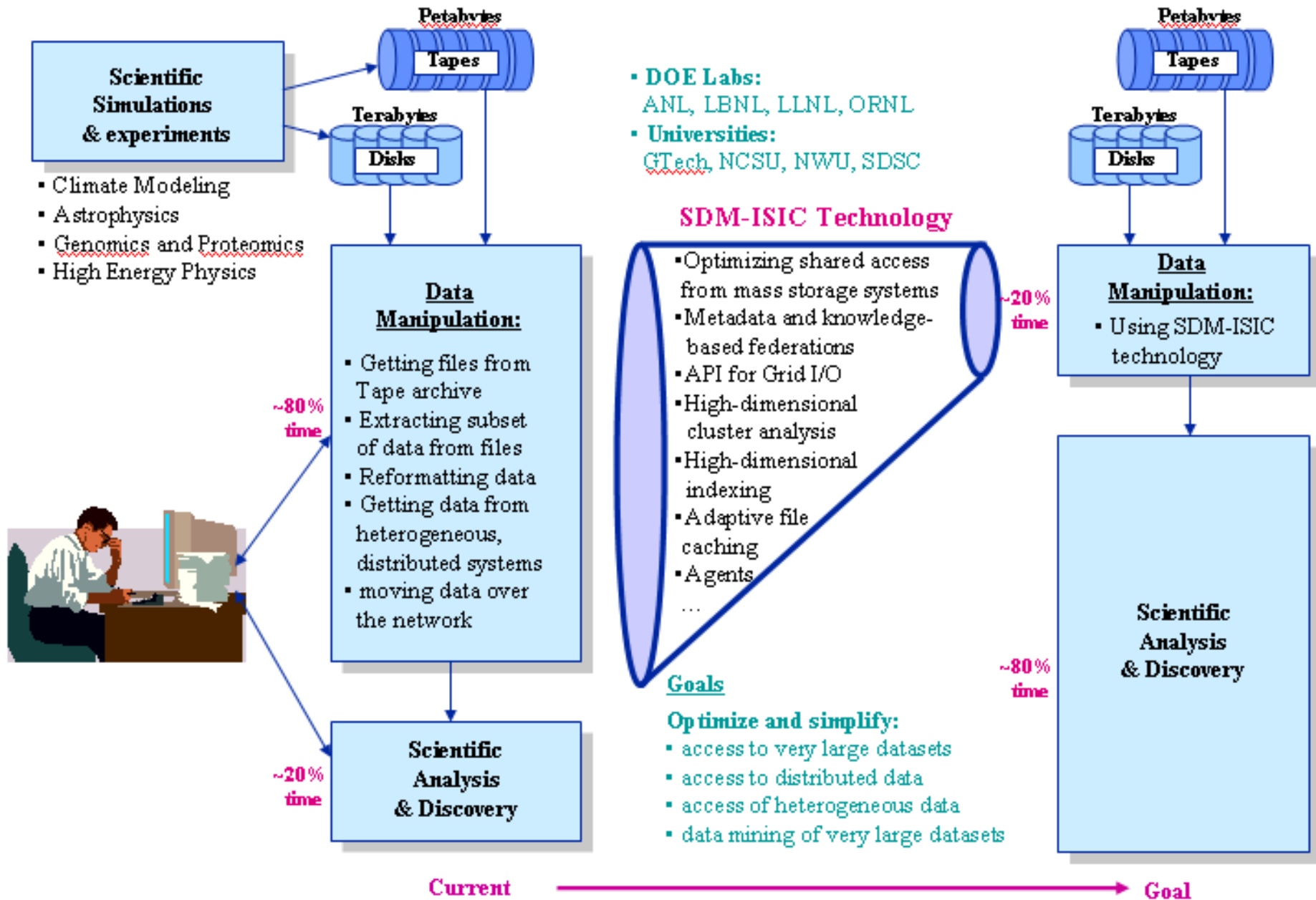
# Computational Modeling in Two Stages; Driving Evolution & Enabling Revolution

Fully interactive (biology, chemistry, physics) ensemble simulations in an operational mode



Real Demonstrated Performance doing useful Science

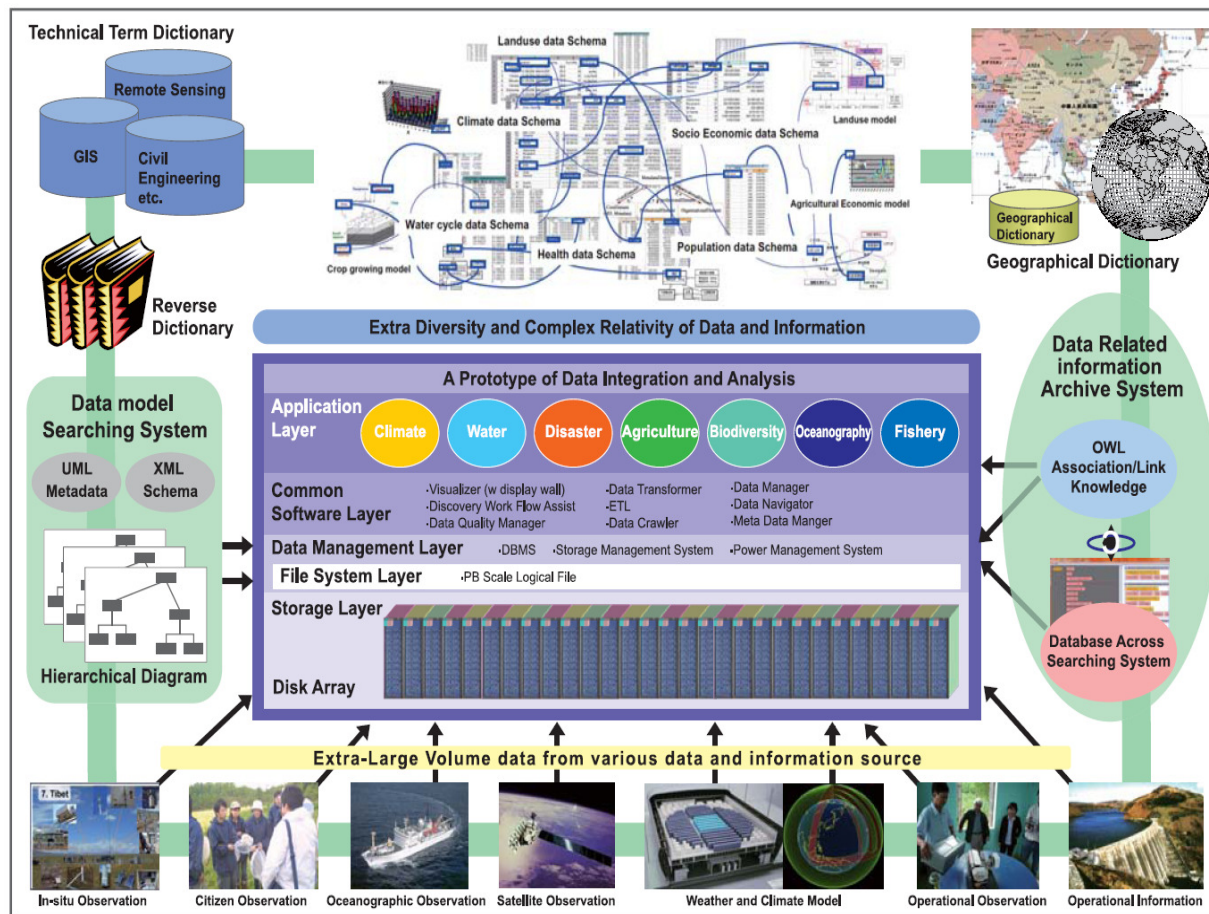
# Scientific Data Management ISIC



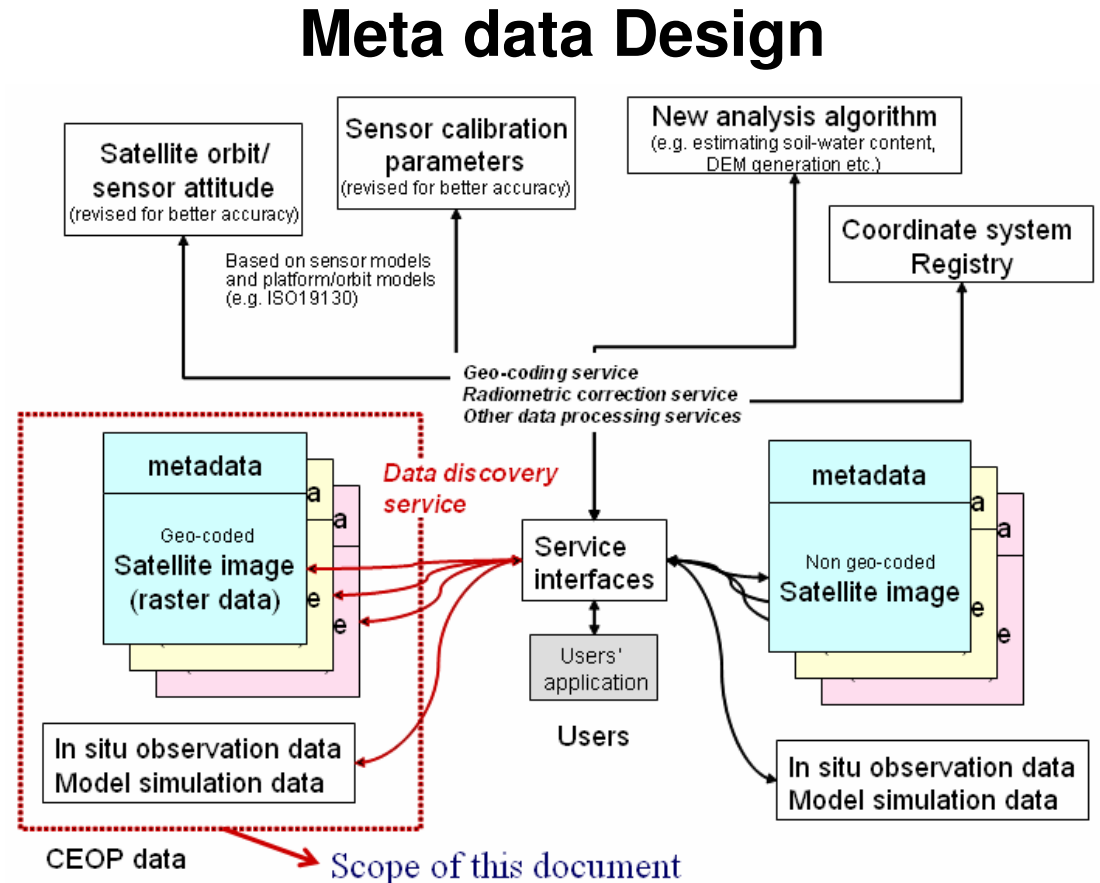
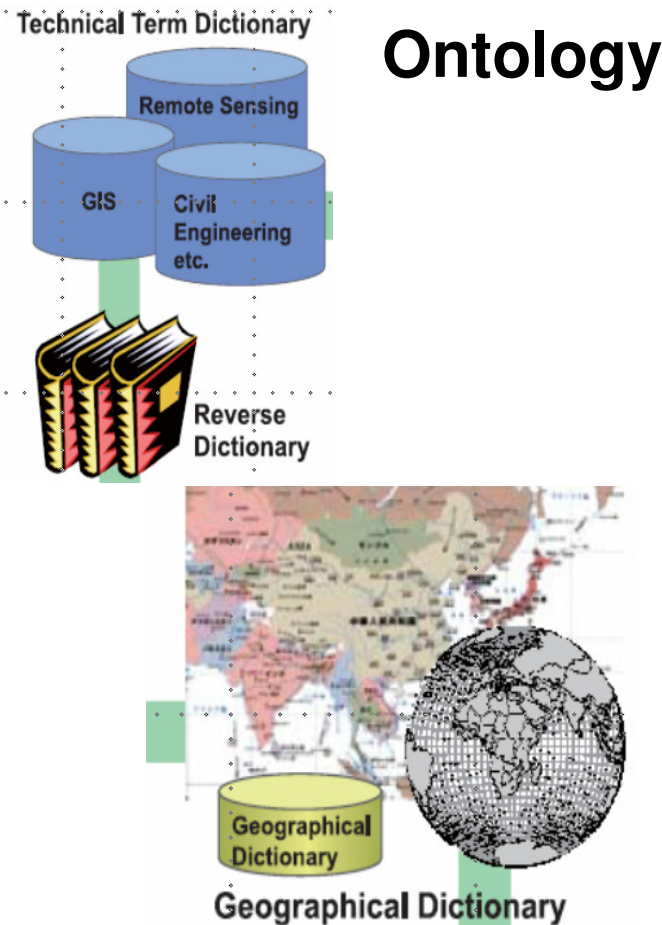
# Data Integration and Analysis System

*a legacy for Japan's contributions to GEOSS*

To create knowledge enabling us to solve the Earth environment problems and to generate socio-economic benefits,

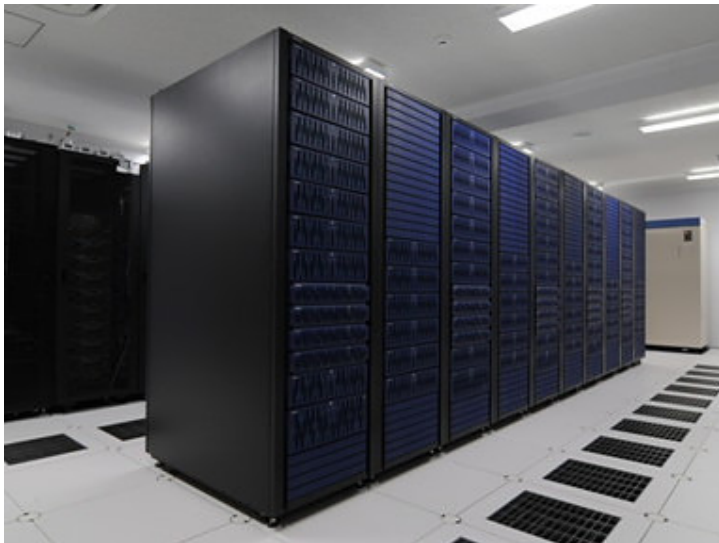


## tackling a large increase in **diversity** of the Earth observation data.

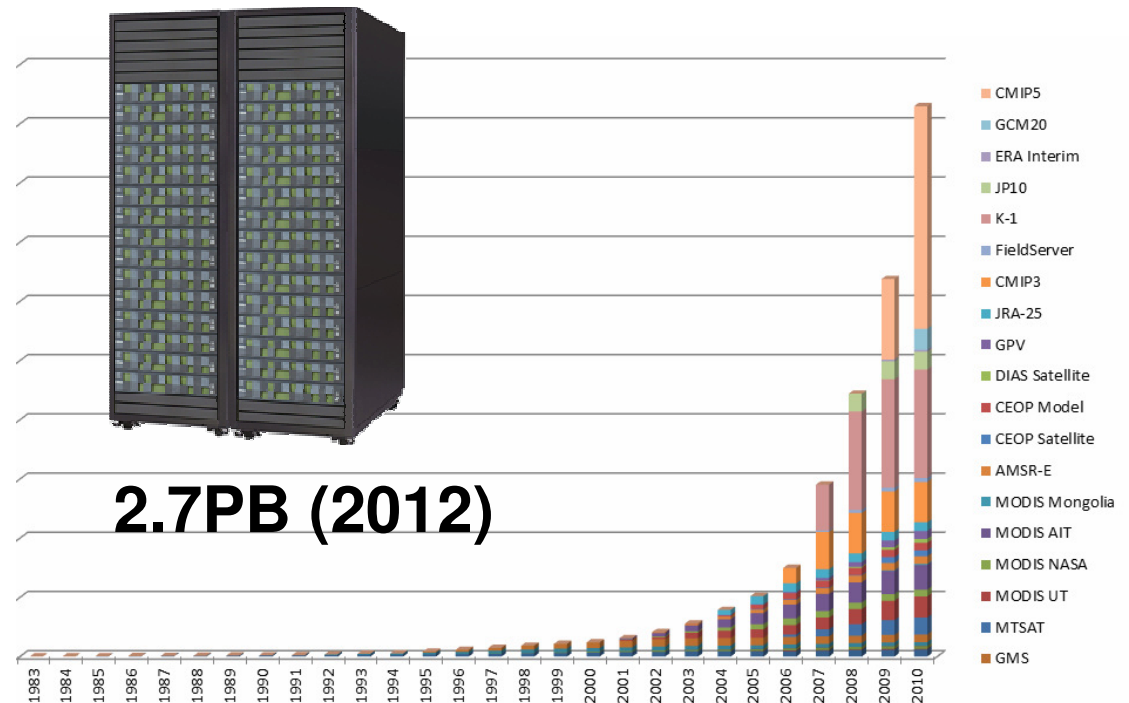


tackling a large increase in **volume** of the Earth observation data.

IPCC AR4 (2007): 40TB → IPCC AR5 (2012): 2.6PB



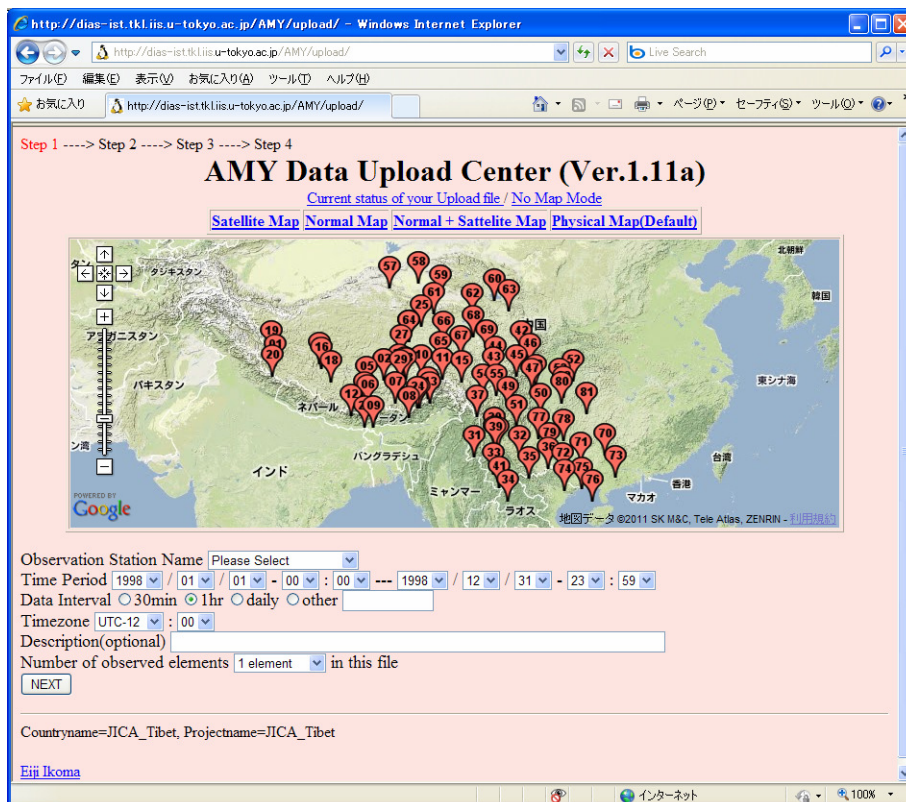
**600TB (2007)**







## accelerating data **archiving**, including data loading, QC and metadata registration



**Asia Monsoon Year**  
**24 project, 277 stations**



**18 River Basin, 280stations**  
**completed**

**Climate Data**  
**16 River Basin, 202 stations**



## enriching data **searching** capability

Data Index/Search > Indexes

Powered by GETA

v2.2.2b    Spring    relocate

DIAS interoperability Portal  
Places  
Keywords  
Persons  
Organisations

hit: 20  
search

### Data Directory (Keywords)

all    unselect     Data Index

temperature    search    synonym    20

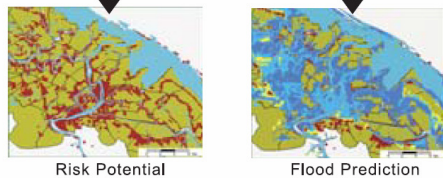
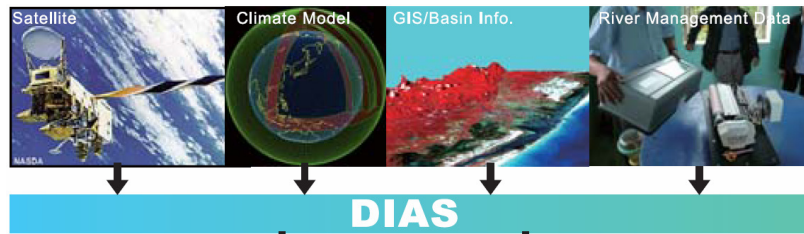
Copyright (C) 2007-2009 The University of Tokyo, Shibasaki Lab. All rights reserved.    graph    table



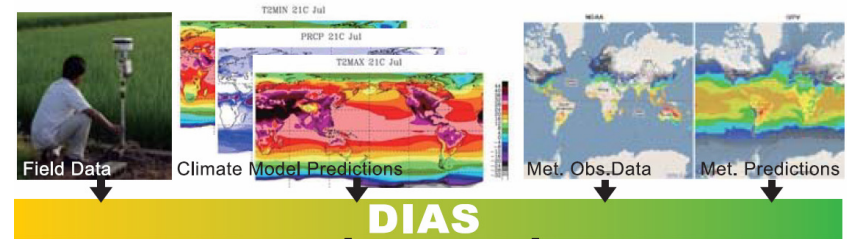
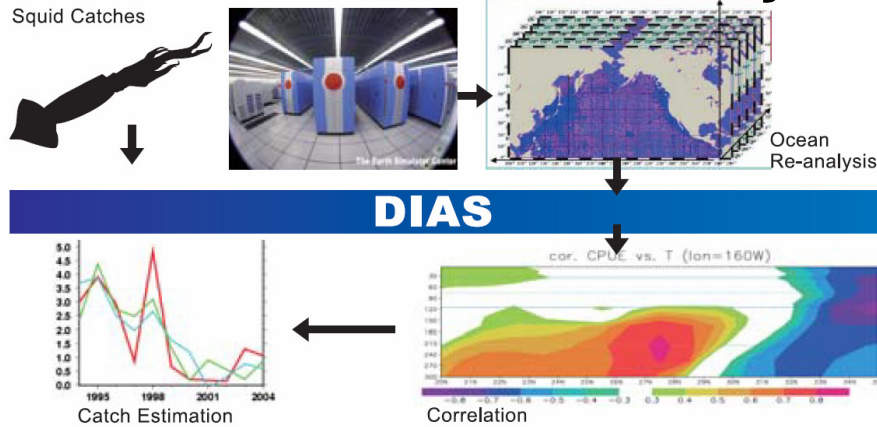
# Data Integration and Analysis System

*a legacy for Japan's contributions to GEOSS*

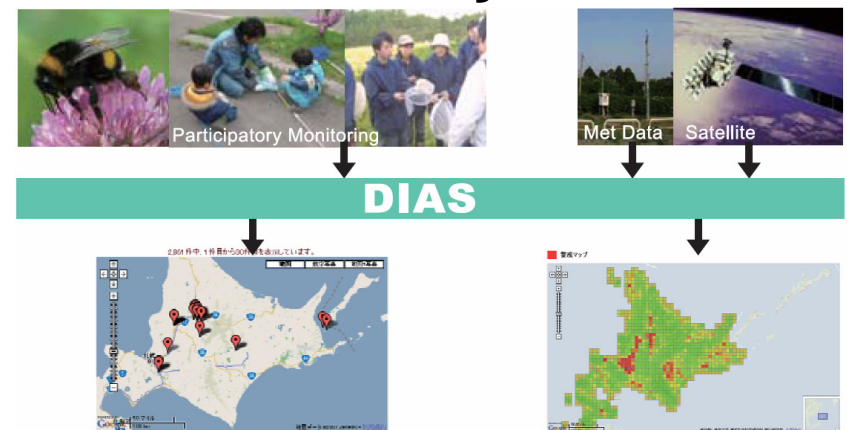
enabling us to do **integrated research** and to realize **inter-disciplinarity**



**Water**  
**Fishery**  
**Climate**



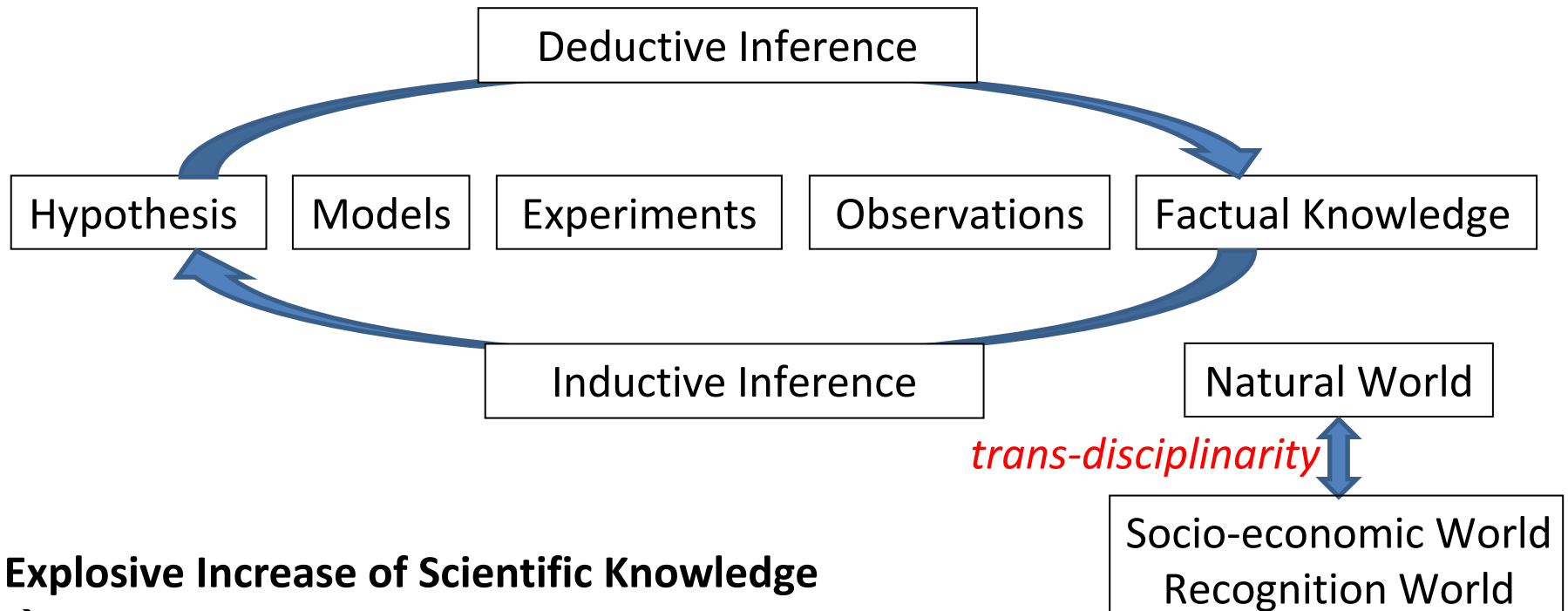
**Food**  
**Biodiversity**



# *inter-disciplinarity & trans-disciplinarity*

## Scientific Knowledge

Formal Knowledge which can be transferred and shared among vast amount of Factual Knowledge



### Explosive Increase of Scientific Knowledge

#### → Differentiation & Systematization

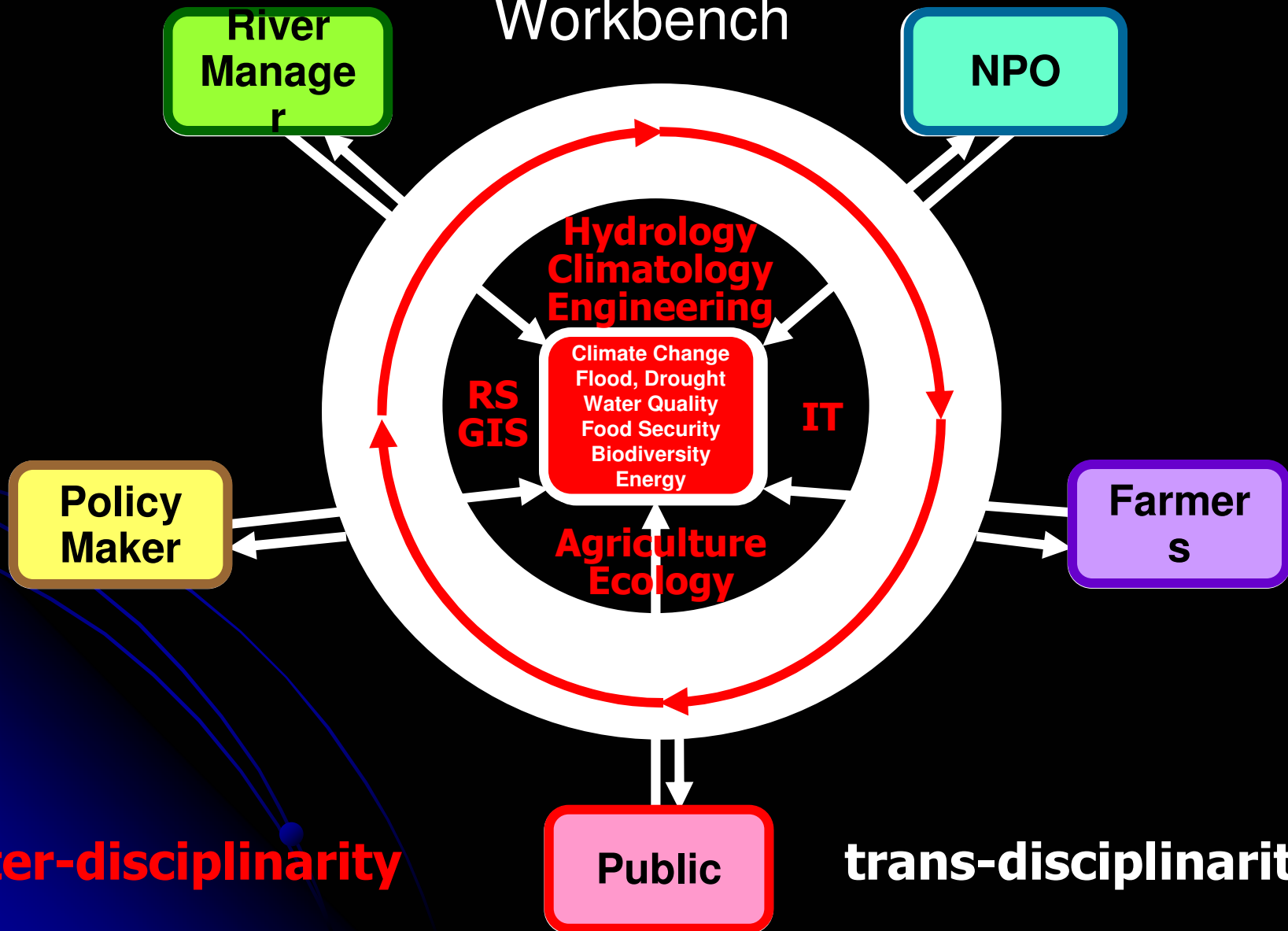
- Accumulated sub-system knowledge can not be reflected to a whole system.
- Effects of a whole system can not be involved in a targeted sub-system.

→ Far from fundamental solutions of issues across disciplines → *inter-disciplinarity*

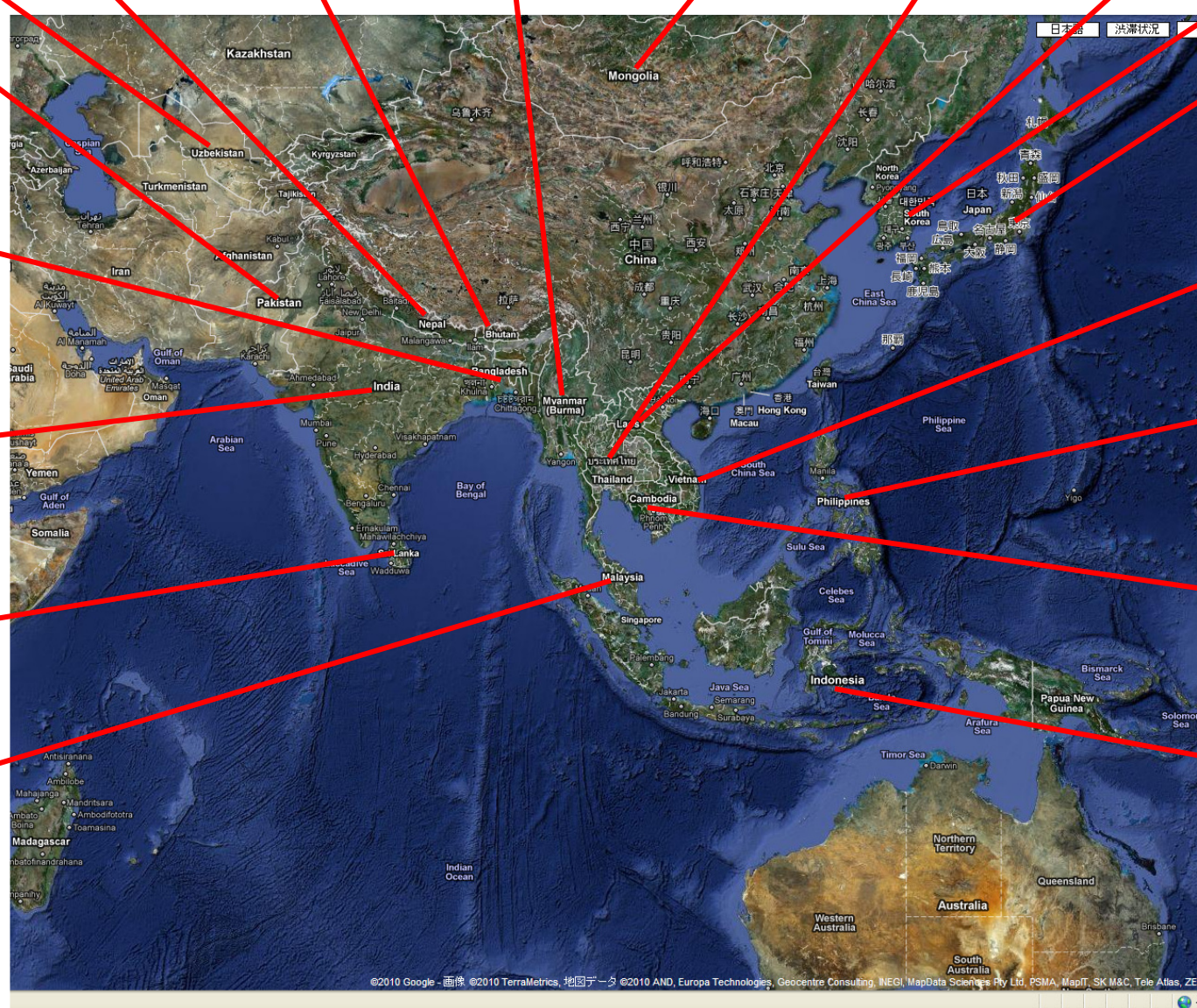
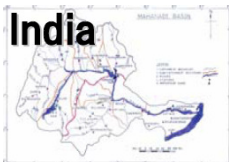
Data System

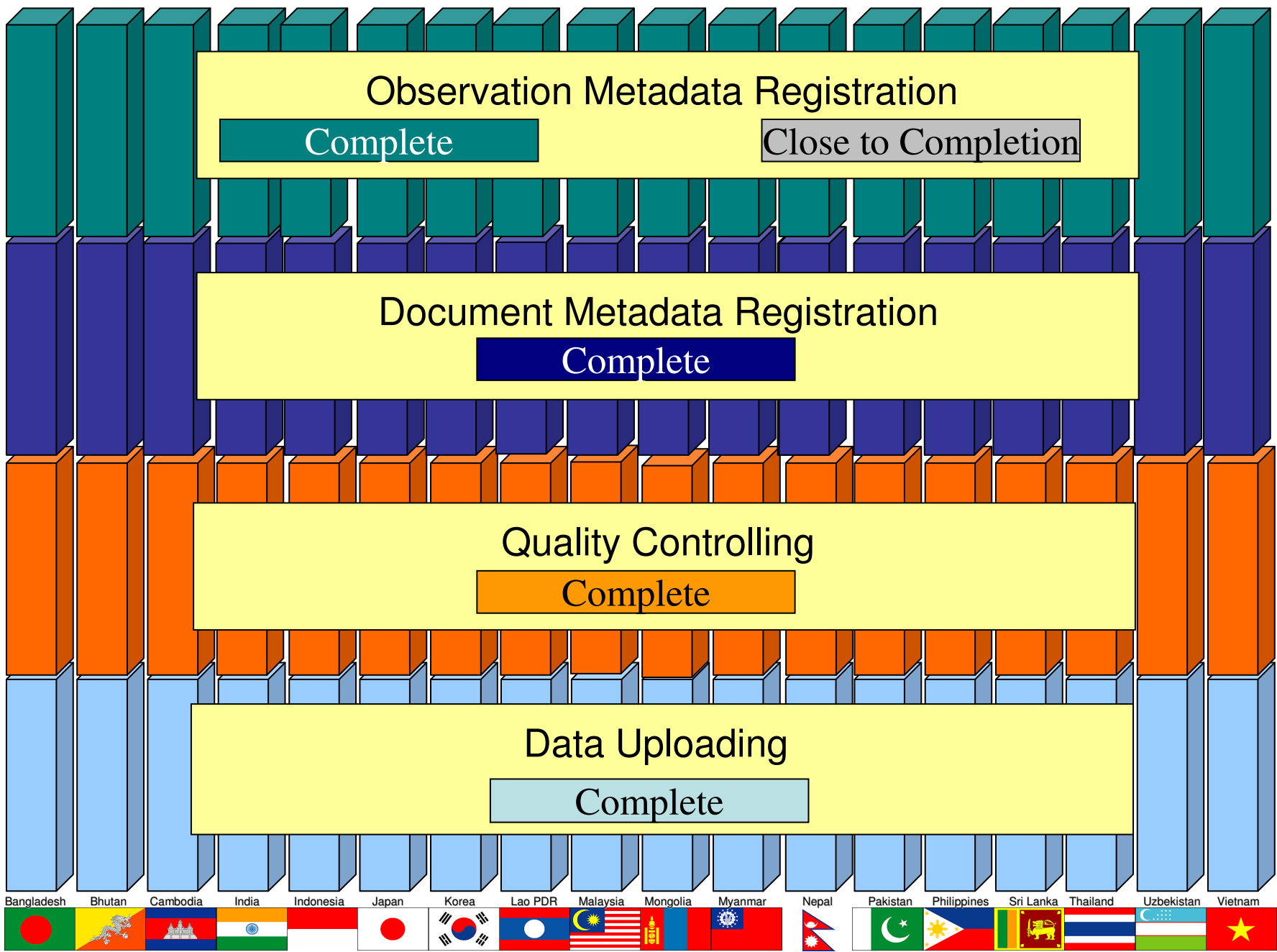
Integration-Interlinkage System

Sharing Data and Information  
Exchanging Knowledge, Experiences and Ideas  
Working Together  
Workbench



# Demonstration River Basins

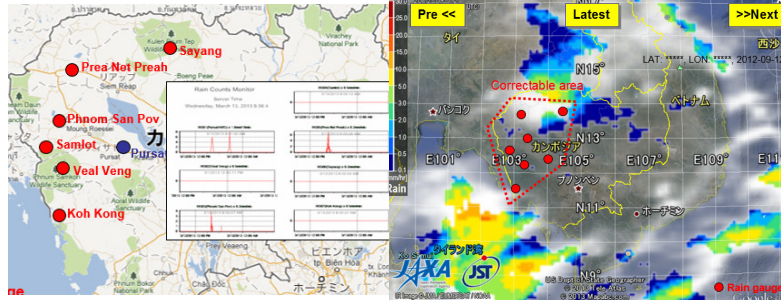




# Water-Climate-Agriculture Workbench in Cambodia



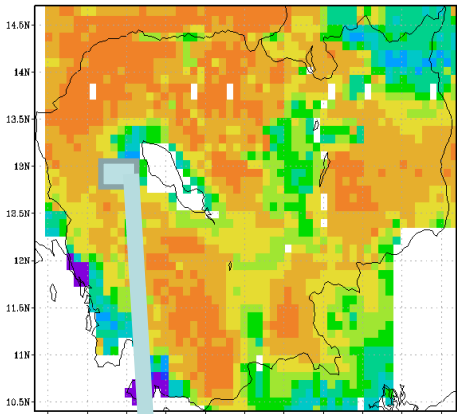
Stakeholder Meeting



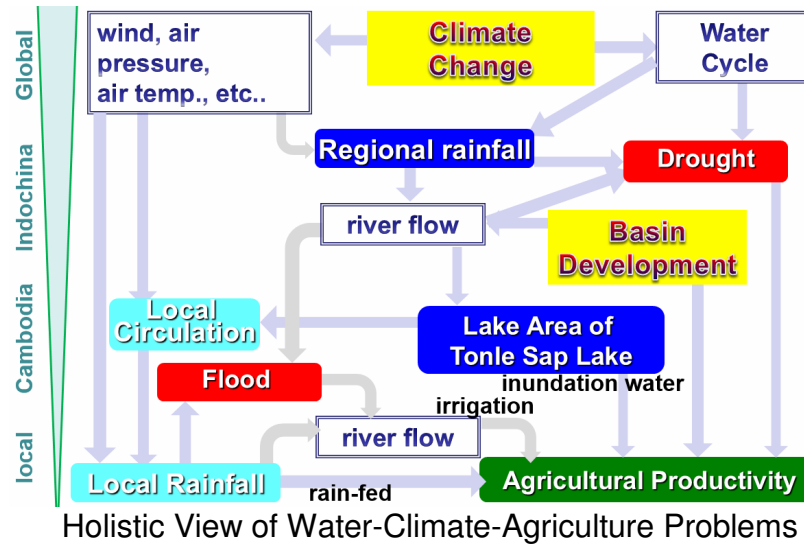
Real-time Rain Gauge → Satellite Data Correction  
→ Wide Data Dissemination



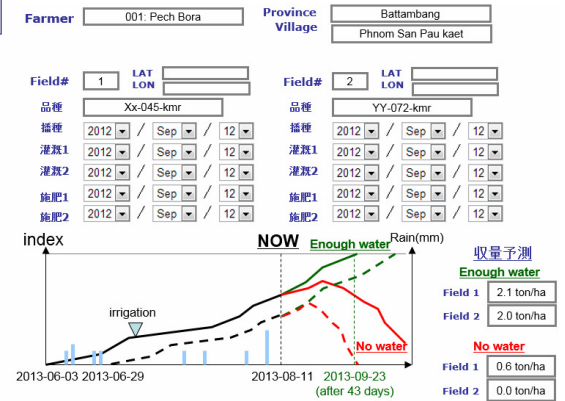
Famers' Needs & Experiences



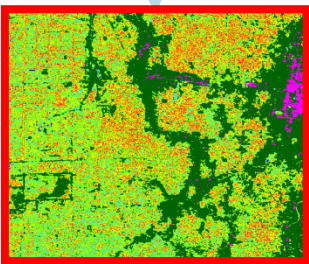
Nation-wide Daily Soil Moisture from Satellite



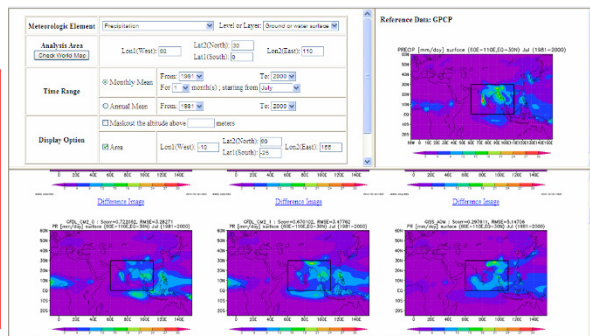
Holistic View of Water-Climate-Agriculture Problems



Water Cycle-Rice Production Coupled Model



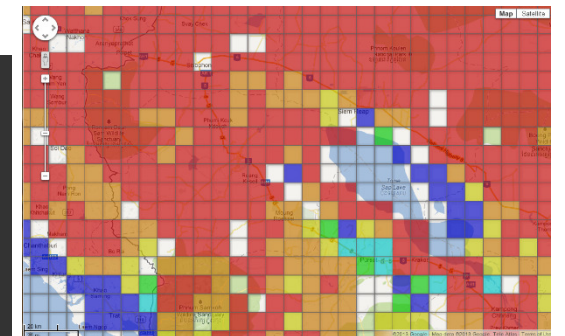
Local Information



Climate Change Analysis Tools



OJT for Local Practitioners



Rice Production Monitoring



# Data Integration and Analysis System



*a legacy for Japan's contributions to GEOSS  
by Promoting Data Sharing  
and Effective Use*

