

# XML Metadata Interoperability and Ingestion of Massive Earth Science Data

*Xinhua Deng<sup>1</sup>, Ruixin Yang, Menas Kafatos, Sean X. Wang,*

Center for Earth Observing and Space Researches  
George Mason University  
Fairfax, VA, USA

(1. Currently with Capital One IT Futures)

*Long B. Pham*

NASA Goddard Space Flight Center DAAC  
Greenbelt, MD, USA



0/01

---

George Mason University



# Presentation Outline

- ◆ XML Metadata Interoperability in SIESIP  
**(The Seasonal-to-Interannual Earth Science Information Partner (SIESIP) – GMU, NASA GDAAC, COLA, and UDel)**
- ◆ XML Metadata Ingestion & Cleaning



# Metadata

- ◆ Diversity of metadata
  - Data catalog/inventory information
  - Data operation information
  - Services information (tools, methods etc) on data
    - ✓ Data access (on-line data acquisition)
    - ✓ Data order (near-line)
  - Scientific knowledge regarding data



## Metadata (cont.)

- Diversity of metadata
- Complexity in metadata contents
  - Text strings
  - HDF(-EOS) headers
  - Owner/scientist annotation
  - DODS (DDS, DAS), data sets URLs
  - GrADS control files
  - Data order (HTML) forms



# Metadata Interoperability

- Exchange their metadata with minimized operations
- Allow metadata searchable by as many users as possible.
- Require Metadata Model – metaModel --  
Be flexible to adapt diverse types of metadata



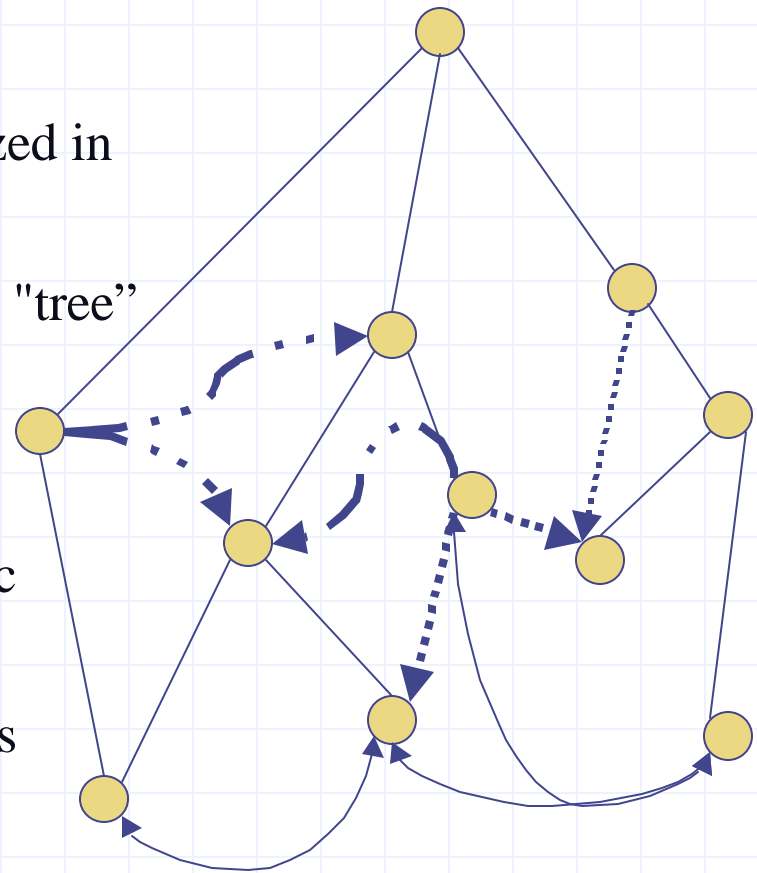
## Metadata Model -- MetaModel

- Indisputable desire for metadata standard, but unrealistic
- XML provides a solution for the need
- Federated metadata management approach in SIESIP:
  - Maintain a metadata repository at each data provider
  - Each repository is described by a metamodel



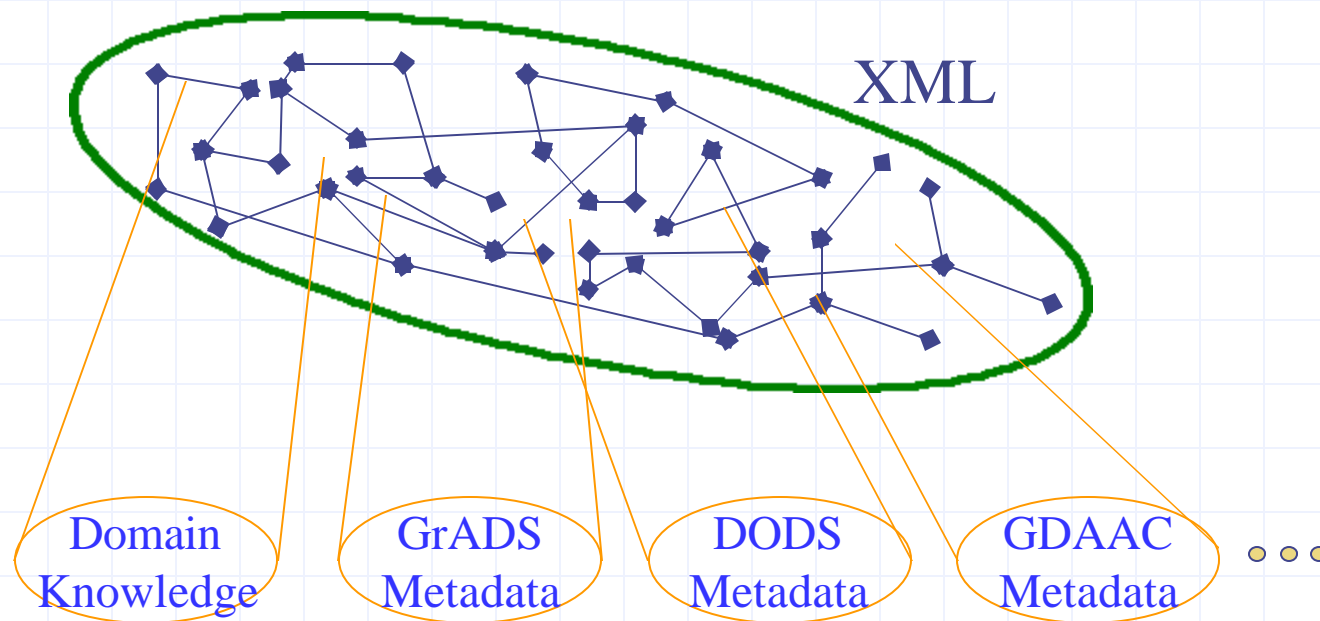
## Metadata Model – MetaModel (cont.)

- Metadata in each repository are organized in a linked XML “tree” or “Graph”
- All metadata entities are “nodes” in the “tree”
- Semantic links bw/ nodes added for the diversity & domain knowledge
- Nodes are grouped in terms of scientific and operational categories
- To add new metadata, simply add nodes and add/update links



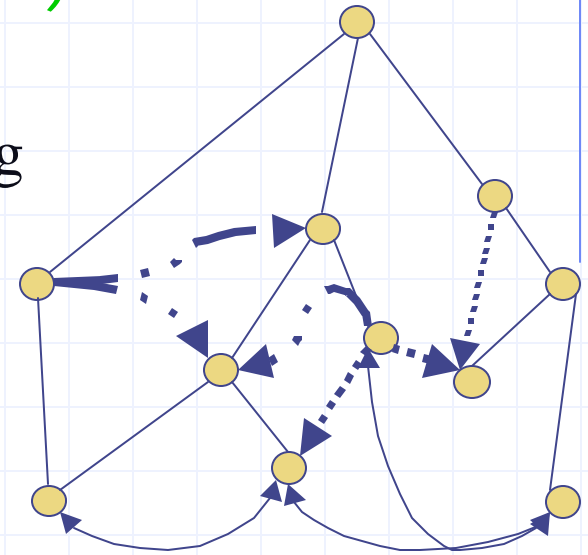
# SIESIP MetaModel -- Features

- If part of metadata conforms to a standard, great! If not, ok.
- Allowing flexible metadata standards or no standards at all
- All metadata become searchable
- ...

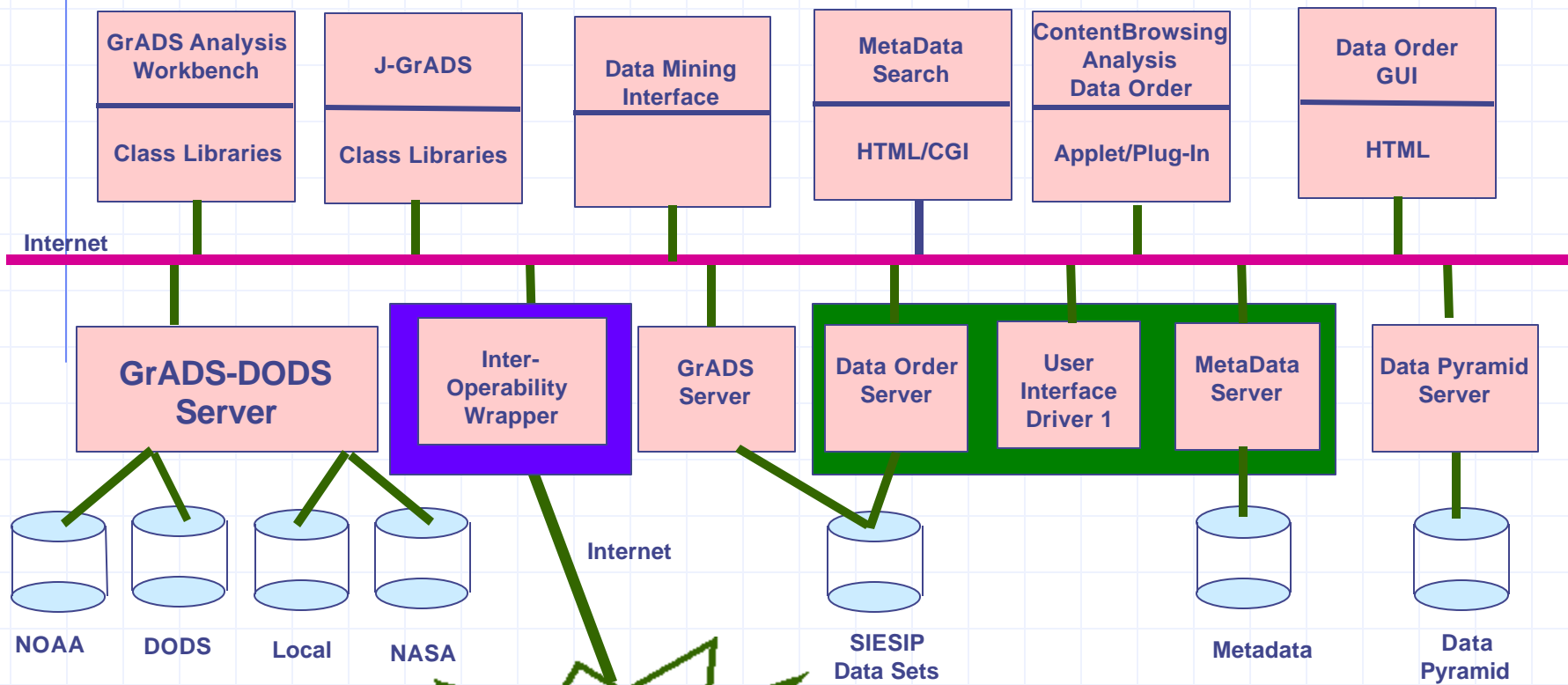


## SIESIP MetaModel – Features (cont.)

- Preserve original data structures, increasing performance
- Support various types of accesses
  - ✓ Browse tree
  - ✓ Closest “Target” search
  - ✓ regular spatial/temporal searches and field/free textual search
- Implement pre-defined queries -- allow queries can be resolved much faster because there is no need to map the XML data tree structure to DBMS tables
- Optimized with “Hot Indexing” – Hush Index



# SIESIP Components



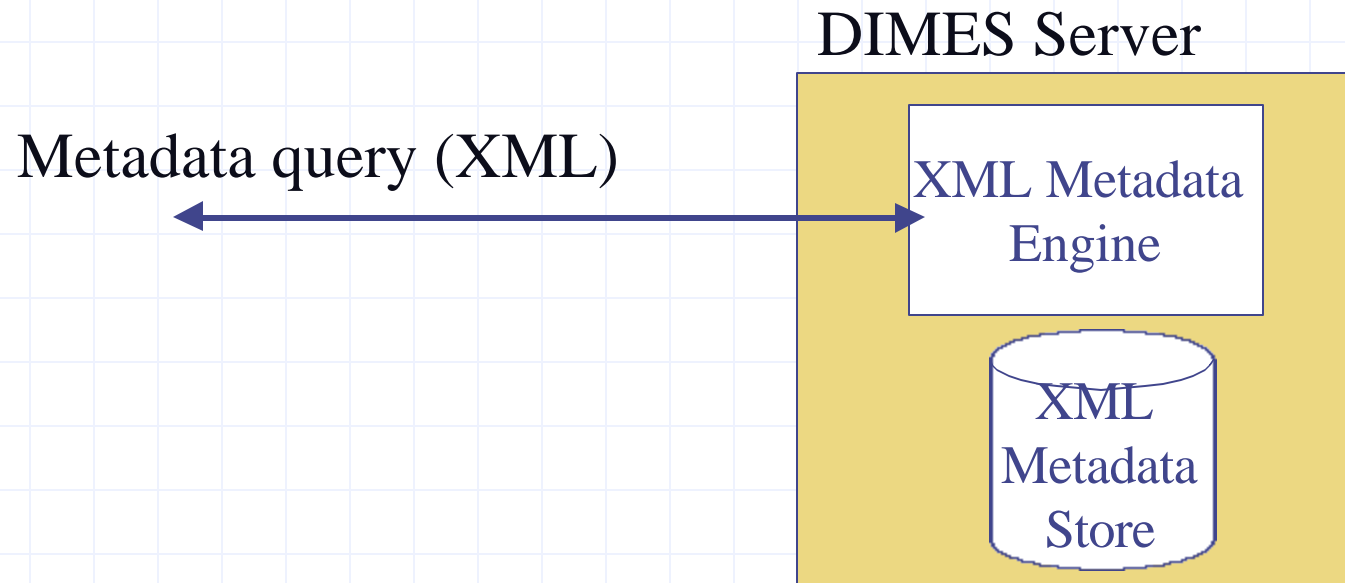
*Data and Metadata Systems on the Internet Outside of SIESIP*

10/30/01

2nd Digital Earth



# Metadata Server

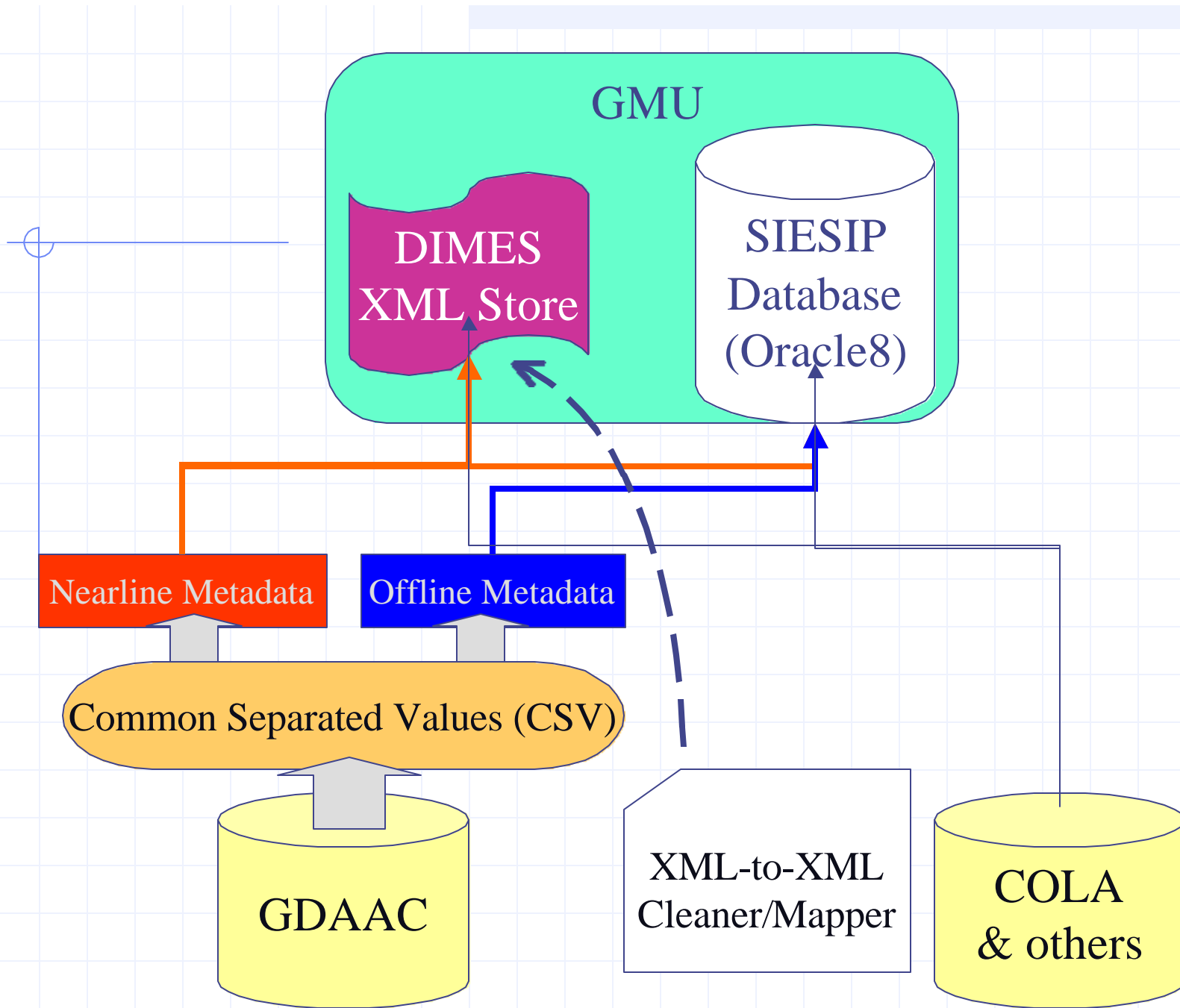


*DIMES: DIstributed MEtadata Server*



- ◆ XML Metadata Ingestion & Cleaning

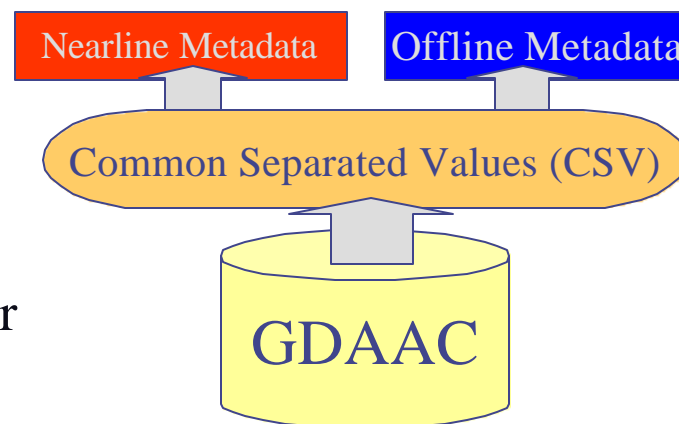




# Metadata Ingestion

## ◆ Approaches

- Automate the transfer of nearline and offline metadata between the GDAAC and GMU
- To make available all archive data from GDAAC using several popular interoperability mechanisms (e.g., metadata publishing, DODS)



## ◆ Completed at GDAAC

- Archive data at the GDAAC is now accessible through GMU via FTP
- Daily nearline and offline metadata are published online daily and made available to SIESIP through the GDAAC

# Metadata Ingestion at GMU

## ◆ Metadata-to-XML Ingestion

- Automate the transfer of nearline and offline metadata between GDAAC and GMU
- A set of tools developed for ingestion
  - ✓ **CSV-to-DBMS Converter** (Perl/DBI module for ingesting offline metadata to Oracle8)
  - ✓ **CSV-to-XML Converter** (Java/XML/DOM for ingesting into the XML Store)
- A cron table is set up for the ingestion automation
- Scientific domain knowledge added manually however



# Metadata Ingestion at GMU (cont.)

## ◆ XML-to-XML cleaner/mapper

- Integrated XML Transformation through XSLT
- Separated the domain knowledge from the programming logic
- Cleaning steps:
  - ✓ Normalized nodes
  - ✓ Checks and reinforces the symmetric property of the links
  - ✓ Remove redundant nodes



## DIMES XML Metadata Server

<http://www.siesip.gmu.edu>

Check out our poster presentation titled as  
**“The Role of XML in the Design and Implementation  
of a Distributed Earth Science Information System”**





Home

Science

Data

Tools

Federation

- DataSet
- Region
- GrADS/data/ncep.1nmago
- Phenomenon
- Time Range
- Observation
- Model
- Parameter
- Project/Experiment
- Format
- Repository
  - COLA
  - DODS
  - GDAAC
  - SIESIP/CEOSR
    - TRMM Combined rain rate**
    - NCEP SSTA
    - UDeI precipitation
    - UDeI air temperature
    - NDVI
    - NCEP SST
    - TRMM Combined rain rate
    - NCEP SSTA
    - UDeI precipitation
    - UDeI air temperature
    - NDVI
    - NCEP SST
- Contact

## Specific Parameter: TRMM Combined rain rate

**Temporal Coverage:**  
 from: GMT Jan 1 00:00:00 1998  
 to GMT Dec 31 00:00:00 1999

**Temporal Resolution:**  
 MONTHLY

**Spatial Resolution:**  
 Longitude: 1.0 degree(s)  
 Latitude: 1.0 degree(s)

**Spatial Coverage**

Name	Start	End
Longitude	-180	180
Latitude	-40	40

**Contact Information:**  
 name: Dong-Bin Shin  
 e-mail: dshin@science.gmu.edu

You may order the data set contains this specific parameter by clicking the Order button.

**Order!**




File Edit View Go Communicator Help

Back Forward Reload Home Search Guide Print Security Shop Stop sgi

Bookmarks Location: <http://esip.gmu.edu:8080/siesip/servlet/SiesipSearchSer>

SGI... Members WebMail Connections BizJournal Mktplace


 [Home](#) [Science](#) [Data](#) [Tools](#) [Federation](#)

**Welcome to Siesip Search Result Page**

Data sets found are listed below. You may click a data set name to find more information about that data set or click the "Order" button to order the data.

1. [TRMM Combined rain rate](#) **Order!**
2. [NDVI](#) **Order!**

100%




File Edit View Go Communicator Help

Back Forward Reload Home Search Guide Print Security Shop Stop

Bookmarks Location: <http://esip.gmu.edu:8080/siesip/servlet/SiesipViewServlet>

SGI... Members WebMail Connections BizJournal Mktplace

 **Home** **Science** **Data** **Tools** **Federation**

**Welcome to Siesip Data View Page**

The major related information about the data set you selected, TRMM Combined rain rate, is given below.

**Temporal Coverage:**  
 from: GMT Jan 1 00:00:00 1998  
 to: GMT Dec 31 00:00:00 1999

**Temporal Resolution:**  
 MONTHLY

**Spatial Resolution:**  
 Longitude: 1.0 degree(s)  
 Latitude: 1.0 degree(s)

Spatial Coverage		
Name	Start	End
Longitude	-180	180
Latitude	-40	40

**Contact Information:**  
 name: Dong-Bin Shin  
 e-mail: [dshin@science.gmu.edu](mailto:dshin@science.gmu.edu)

**You may order this data set by clicking the Order button.**

**Order!**

100% 10/30/01



File Edit View Go Communicator Help

**SIESIP**

### Welcome to Siesip Data Order Page

The data set you selected, TRMM Combined rain rate, is ready for order. The major information is given below.

**Temporal Coverage:**  
from: GMT Jan 1 00:00:00 1998  
to GMT Dec 31 00:00:00 1999

**Temporal Resolution:**  
MONTHLY

**Spatial Resolution:**  
Longitude: 1.0 degree(s)  
Latitude: 1.0 degree(s)

Spatial Coverage		
Name	Start	End
Longitude	-180	180
Latitude	-40	40

**Contact Information:**  
name: Dong-Bin Shin  
e-mail: dshin@science.gmu.edu

Please give your e-mail address which will be used for sending you instructions about how to retrieve the data set. You may choose the time period you are interested in. Please refer to above for available temporal range. Only data files in that range could be ordered.

e-mail:

From    To:

100%



# Metadata Search Interface

*File Edit View Go Communicator Help*

**Spatial Selection Panel [Help](#)**

Spatial Resolution:  degrees

Left  Right  Top  Bottom

**Temporal Selection Panel [Help](#)**

Temporal Resolution:

From:

To:

**Textual Select Panel [Help](#)**

The result must meet  of the following conditions

<input type="text" value="Parameter"/>	<input type="text" value="contains"/>	<input type="text" value="sst"/>
<input type="text" value="DataFormat"/>	<input type="text" value="does not contain"/>	<input type="text" value="grib"/>
<input type="text" value="Select A Category"/>	<input type="text" value="contains"/>	<input type="text"/>

10/30/01 100% znd Digital Earth



# Conclusions

- Integrated XML Transformation through XSLT
- Separated the domain knowledge from the programming logic
- Cleaning steps:
  - ✓ Normalized nodes
  - ✓ Checks and reinforces the symmetric property of the links
  - ✓ Remove redundant nodes

