Edging Toward Interoperability

Mélanie F. Meaux

GO-ESSP: June 6-8, 2005

The Global Change Master Directory:
A Directory of Earth Science Data & Earth Science-Related Services

Edging Toward Interoperability

Moving forward through …

- International Interoperability Standards
- Authoring Tools: web-based & off-line
- Controlled-Vocabulary Query
- Free-Text Query
- Open Application Programmer Interfaces (APIs)
- OAI/HTTP Harvesting Capabilities
  - Metadata Converter
- Local Database Agent (LDA)
- OGC WMS Capabilities
  - GCMD's Prototype Mapserver

I. International Interoperability Standards – crosswalks among metadata standards

2004: ISO-19115 Geographic Information Metadata International Standard formalized

- ISO-19115-2 Extensions for imagery and gridded data

Other formats:

- Dublin Core Metadata Initiative (DCMI) & MARC21
- WMO Core Metadata
- US Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM)
  - -Extensions for Remote Sensing Metadata (2002)
- Directory Interchange Format (DIF)
- European Directory of Marine Environmental Datasets (EDMED)
- ROSCOP/CSR (Cruise Summary Report)
- Ocean Data Acquisition System (ODAS)

GCMD Interfaces
Search / Publish Metadata Content

Web User

1

Search **data set** descriptions through GCMD search interface

Retrieve metadata records

Search **data set** descriptions through **portals** - virtual subsets of the directory

Retrieve metadata records

2

Search **data service** descriptions through GCMD search interface

Retrieve metadata records

GCMD Database / Server

HTML DocBUILDER

DocBUILDER; Metadata records are stored off-line where they can be later submitted to GCMD

Harvester : Retrieves, converts & exchanges metadata through HTTP or Open Archives Initiatives (OAI) servers

Metadata Converter : Converts metadata records (e.g. FGDC) to GCMD's Directory Interchange Format (DIF)

A

Publish metadata through web-based & off-line tools

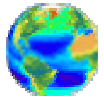
B

Harvest existing records for exchange

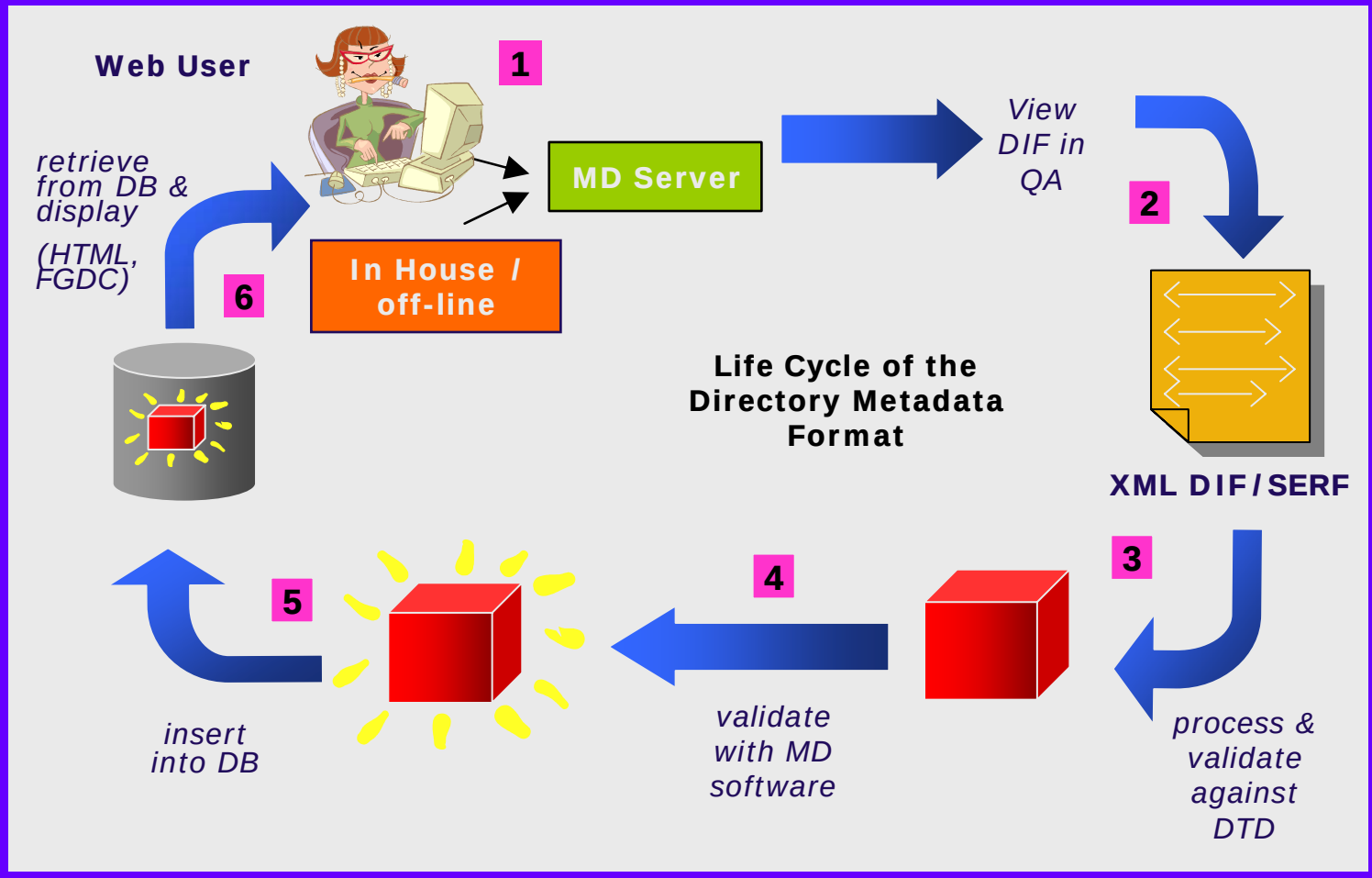
C

Translate existing records using *Extensible Stylesheet Language Transformations (XSLT)*

Metadata Tools



The Life Cycle of the Directory Metadata Format

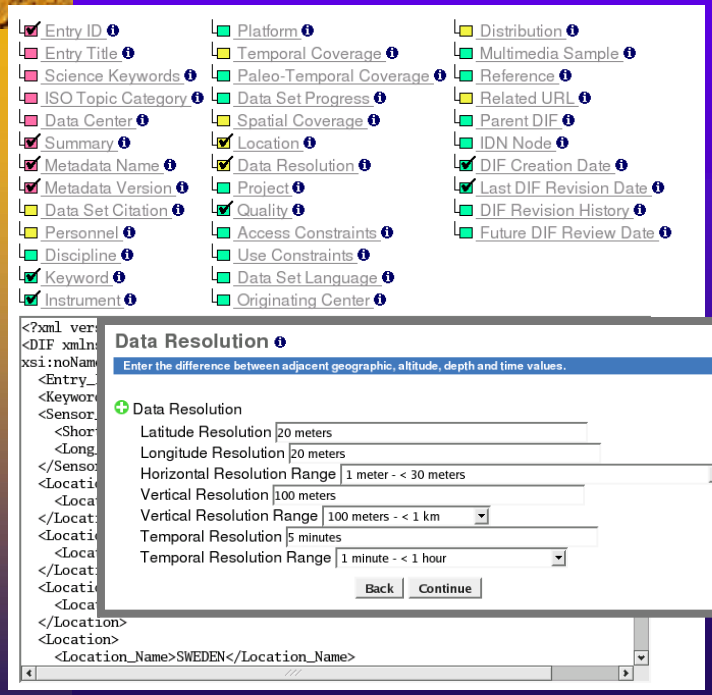


XML DIF Schema can be found at
http://gcmd.nasa.gov/Aboutus/xml/dif/dif.xsd

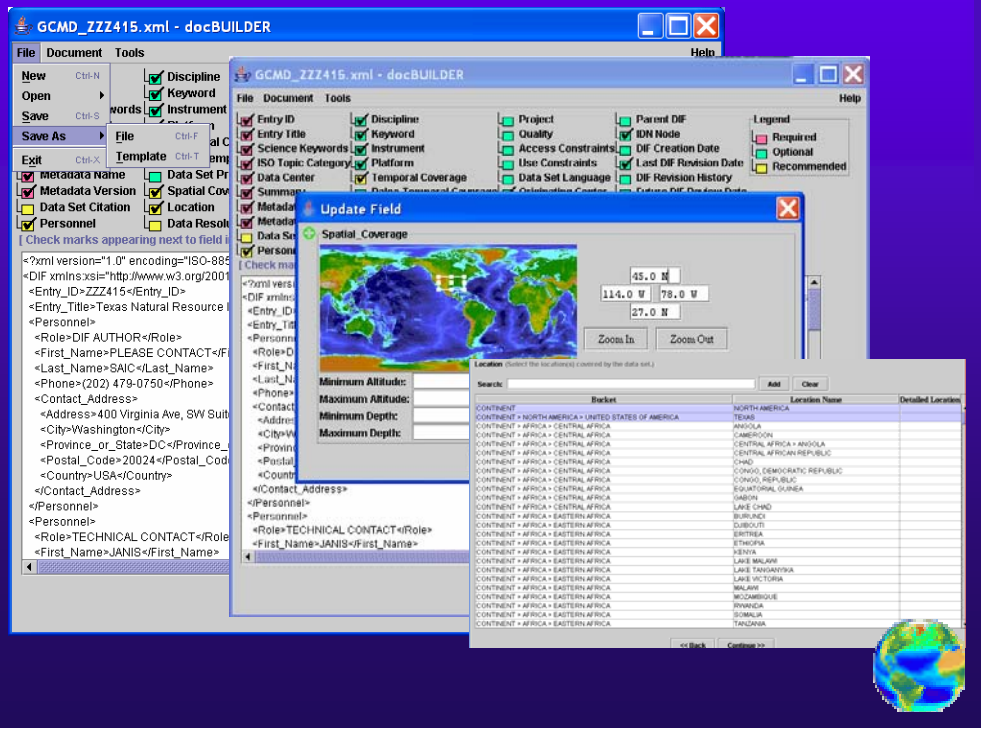
II. Authoring Tools: web-based & off-line

doc BUILDER
metadata authoring tool

HTML docBUILDER



docBUILDER



Data Resolution Proposal “Valids”

- DIF authors can now select from set of Geospatial and Temporal resolution range “valid” values.
- Vertical Resolution refers to both Altitude and Depth resolution.

Horizontal Resolution Range:

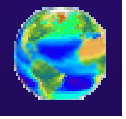
< 1 m
1 m - < 30 m
30 m - < 100 m
100 m - < 250 m
250 km - < 500 m
500 m - < 1 km
1 km - < 10 km
10 km - < 50 km
50 km - < 100 km
100 km - < 250 km
250 km - < 500 km
500 km - < 1000 km
> 1000 km or > 10 degrees

Vertical Resolution Range:

< 1 m
1 m - < 10 m
10 m - < 30 m
30 m - < 100 m
100 m - < 1km
> 1km

Temporal Resolution Range:

< 1 second
1 second - < 1 minute
1 minute - < 1 hour
Hourly - < Daily
Daily - < Weekly
Weekly - < Monthly
Monthly - < Annual
Hourly Climatology
Daily Climatology
Pentad Climatology
Weekly Climatology
Monthly climatology
Annual
Annual climatology
Decadal
Climate Normal (30-year climatology)



Numerical Model Fields Proposed

- Anne Sophie Archambeau from IPSL, France proposed an addition to the DIF to handle numerical output data sets:

Group: Numerical_Experiment
Model_Name:
Model_Version:
Model_URL:
Model_Configuration:
Model_Resolution:
Model_Calendar:
Group: Model_Integration_Period
Start_Date:
Stop_Date:
End_Group
Simulation_Name:
Initial_Conditions:
Perturbation:
Imposed_Boundary_Conditions:
End_Group

- Interop discussions on the proposed new fields are summarized in the CEOS IDN Interop Newsletter for April 2003:

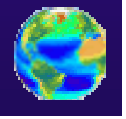
http://gcmd.gsfc.nasa.gov/pipermail/interop/2003-April/000011.html

A new model keyword list for MODELS?

Global Coupled Climate System Models
Global General Circulation Models (GCM)
Earth System Component Models (Global or Regional)
Process Models: Clouds and Precipitation
Phenomenological/Decision Making Models
Geophysical Models (as in Earth Geophysics)
Solar-Atmospheric Space-Weather Models

A higher level of hierarchy?

Diagnostic and Simulation Models
Data Assimilation Models
Forecast, Prediction, Projection (Scenario's) Models
Monitoring and Early Warning Systems



THREDDS

- THREDDS (Thematic Realtime Environmental Data Distributed Services)

  "...for students, educators and researchers to publish, contribute, find, and interact with data relating to the Earth system in a convenient, effective, and integrated fashion."

- Exploring methods to add relative temporal coverage to DIF automatically for THREDDS real-time data.
  - Require that DIF "Keyword:" field be populated with the phrase:
    - **RELATIVE_START_DATE**: -d, where d = any number of days
  - Created script to automatically populate DIF field **Temporal_Coverage** with **Start_Date** and **Stop_Date** based on the current date +# of days specified in **RELATIVE_START_DATE**
  - Cron set up nightly to populate THREDDS DIFs

GCMD Interfaces
Search / Publish Metadata Content

New Look !

Web User

1

Search **data set** descriptions through GCMD search interface

Retrieve metadata records

Search **data set** descriptions through **portals** - virtual subsets of the directory

Retrieve metadata records

2

Search **data service** descriptions through GCMD search interface

Retrieve metadata records

GCMD Database / Server

HTML DocBUILDER

DocBUILDER; Metadata records are stored off-line where they can be later submitted to GCMD

Harvester : Retrieves, converts & exchanges metadata through HTTP or Open Archives Initiatives (OAI) servers

Metadata Converter : Converts metadata records (e.g. FGDC) to GCMD's Directory Interchange Format (DIF)

Metadata Tools

A

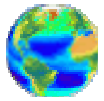
Publish metadata through web-based & off-line tools

B

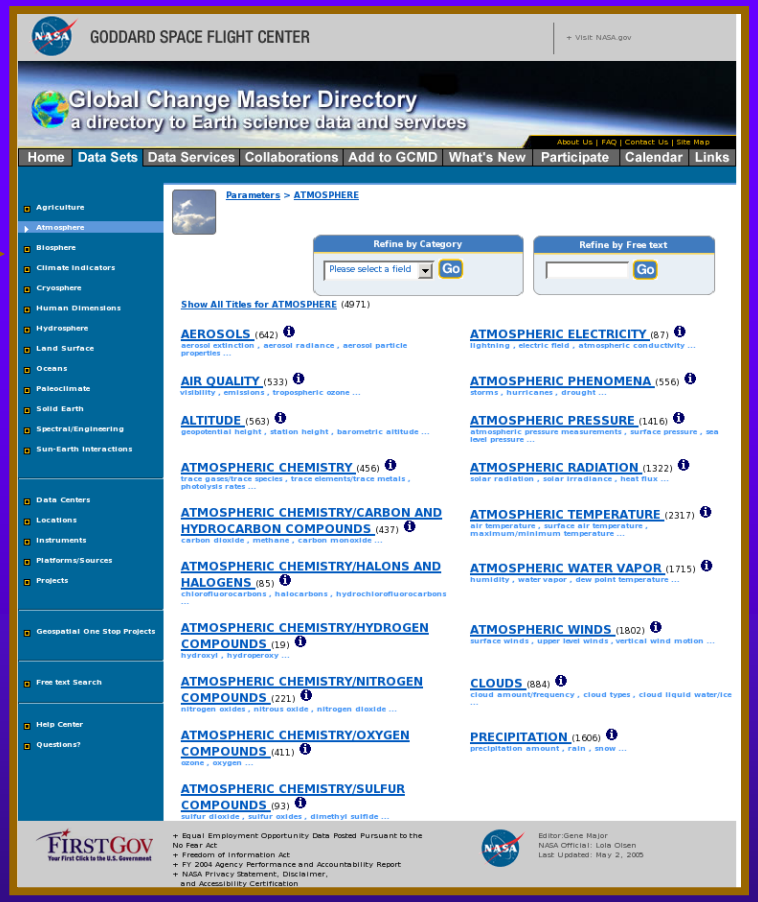
Harvest existing records for exchange

C

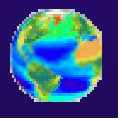
Translate existing records using *Extensible Stylesheet Language Transformations (XSLT)*



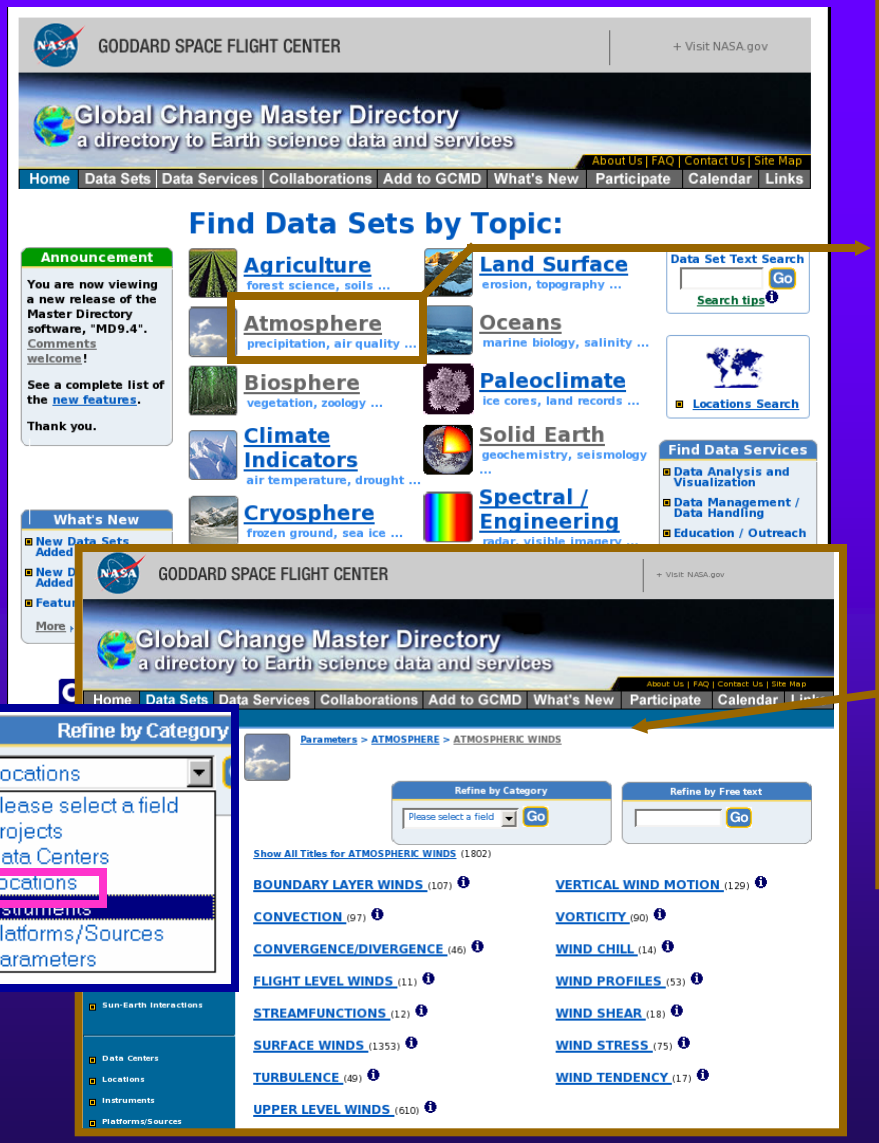
III. Controlled-Vocabulary Query

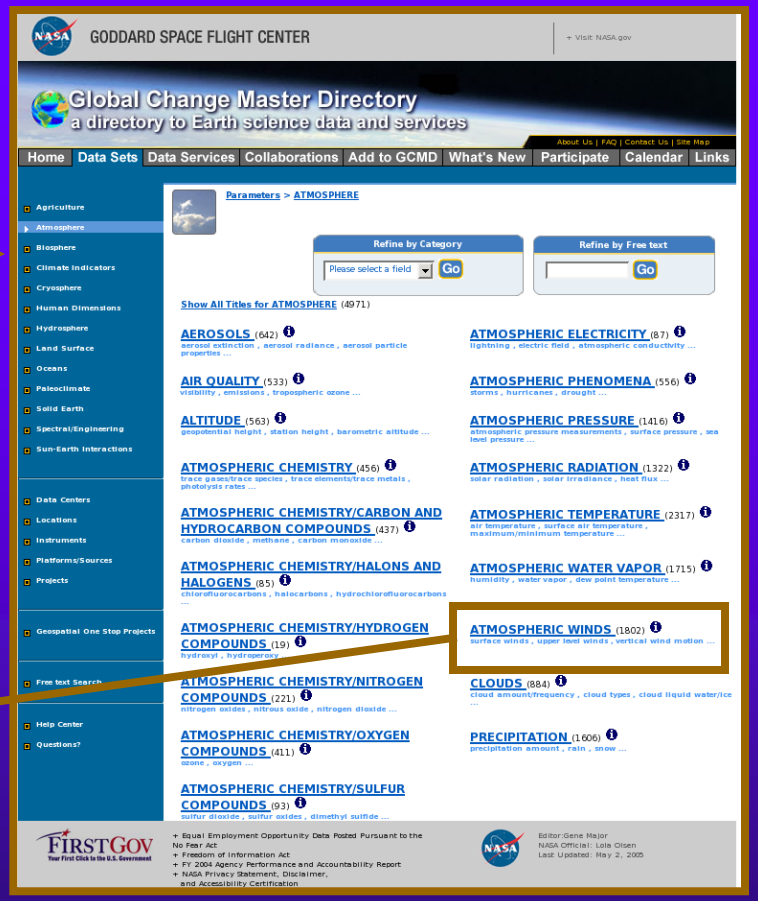


**** For a listing of all GCMD keywords see: http://gcmd.nasa.gov/Resources/valids/keyword_list.html***



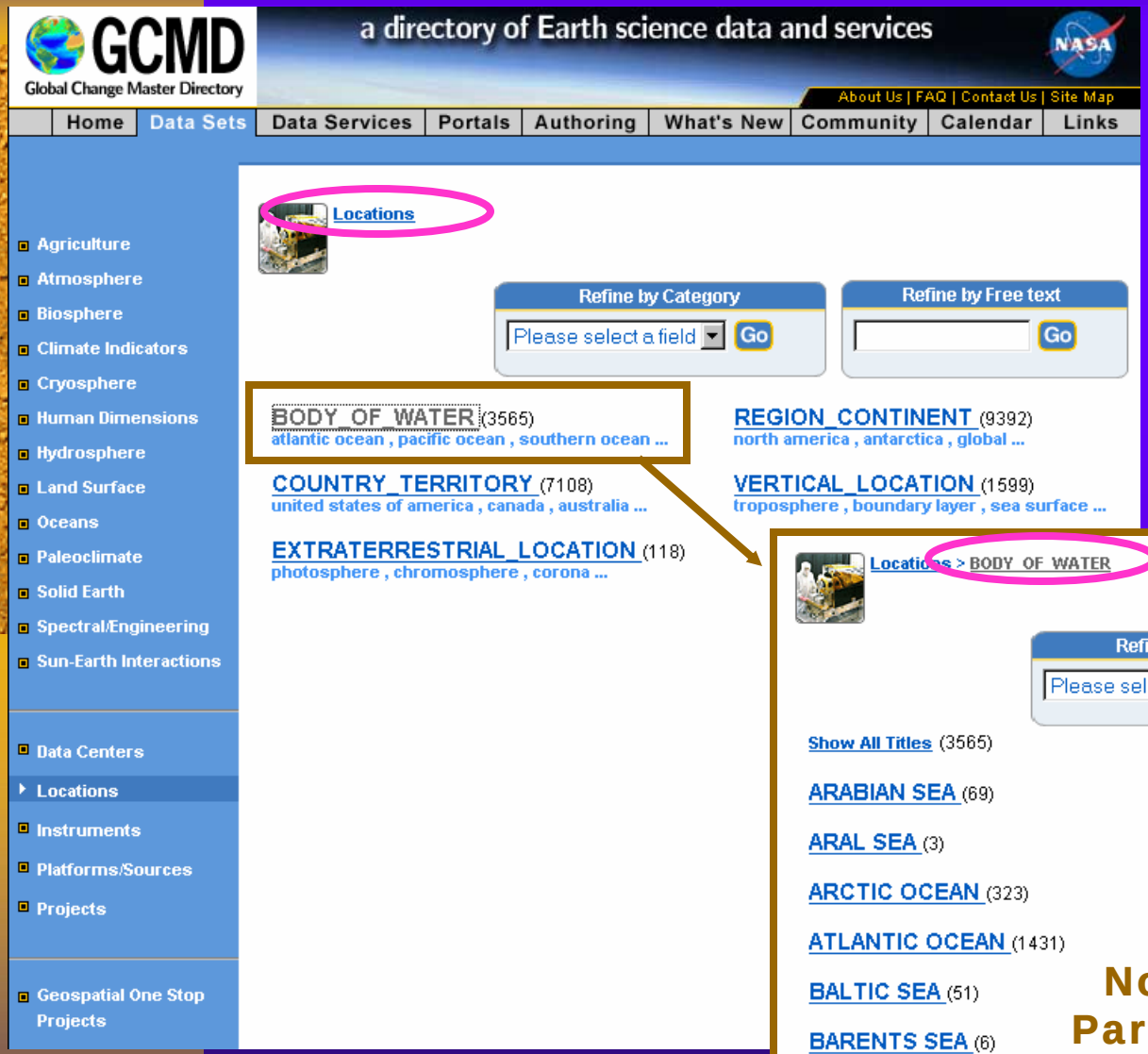
III. Controlled-Vocabulary Query

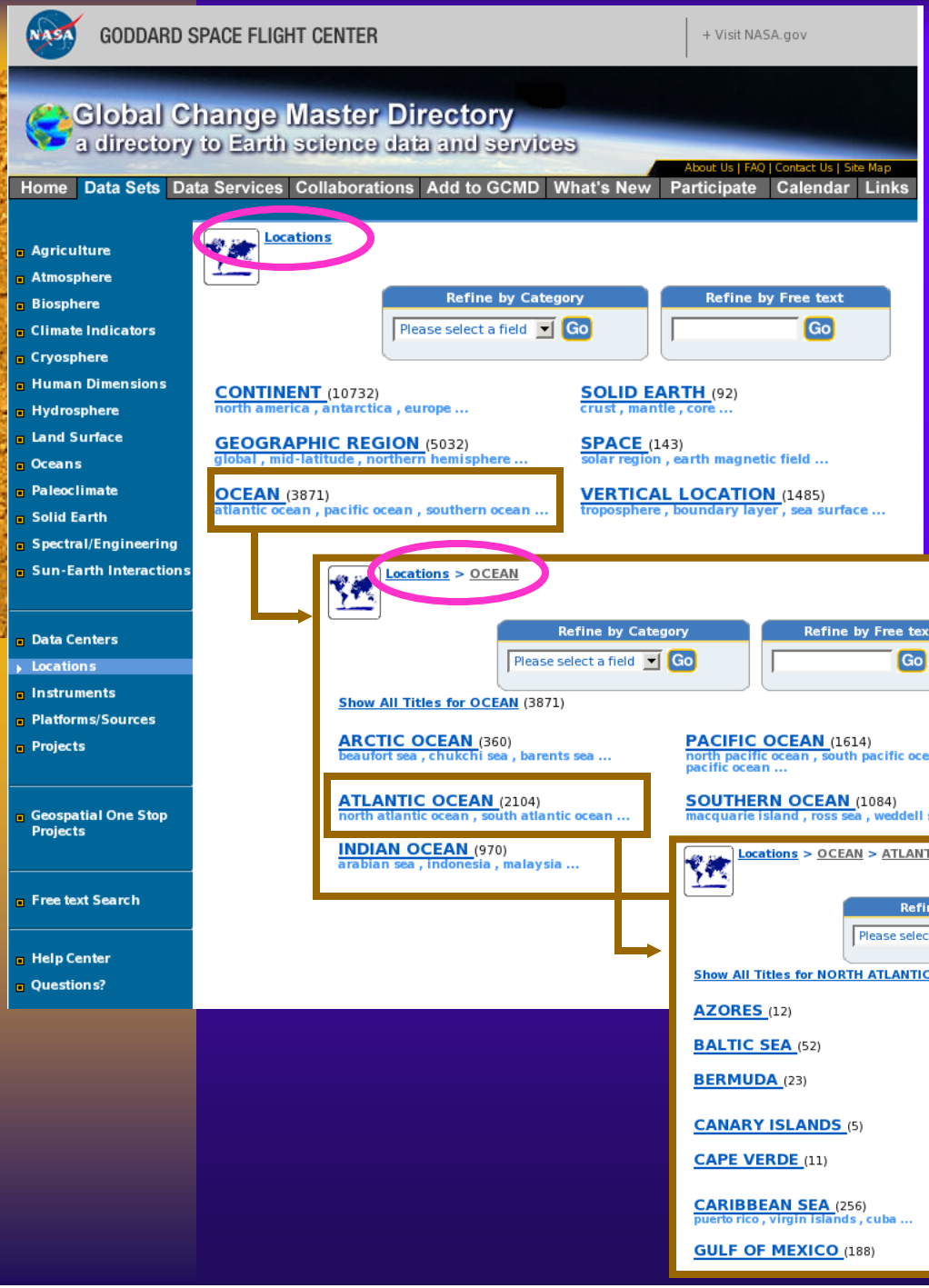




**** For a listing of all GCMD keywords see:***
http://gcmd.nasa.gov/Resources/valids/keyword_list.html

III. Controlled-Vocabulary Query – Old Location Hierarchy





III. Controlled-Vocabulary Query – New Location Hierarchy

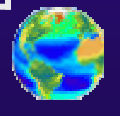
- based on the USGS Coastal and Marine Geology Program InfoBank Atlas and gazetteer

http://walrus.wr.usgs.gov/infobank/gazette/html/regions/gl.html

III. Controlled-Vocabulary

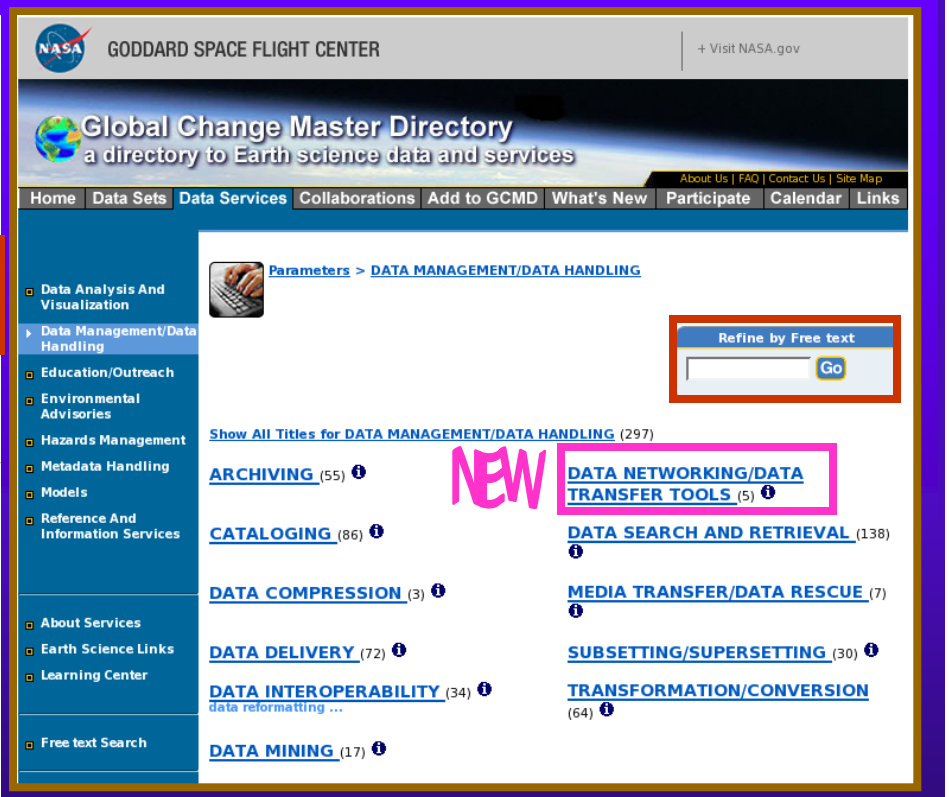
NEW Earth Science Keywords

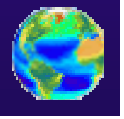
TOPIC	TERM	VARIABLE
PALEOCLIMATE	**Ice Core Records** **Paleoclimate Reconstructions**	➢Electrical Properties ➢Air Temperature Reconstruction ➢Drought/Precipitation Reconstruction ➢Streamflow Reconstruction ➢Ocean Salinity Reconstruction ➢Ground Water Reconstruction ➢Lake Level Reconstruction ➢SST Reconstruction ➢Atmospheric Circulation Reconstruction ➢Sea Level Reconstruction ➢Vegetation Reconstruction ➢Solar Forcing/Insolation Reconstruction
ATMOSPHERE	**Atmospheric Pressure** **Altitude**	➢Planetary/Rossby Waves ➢Topographic Waves ➢Mesopause
OCEANS	**Ocean Circulation** **Ocean Waves**	➢Upwelling/Downwelling ➢Gravity Waves ➢Rossby/Planetary Waves ➢Topographic Waves



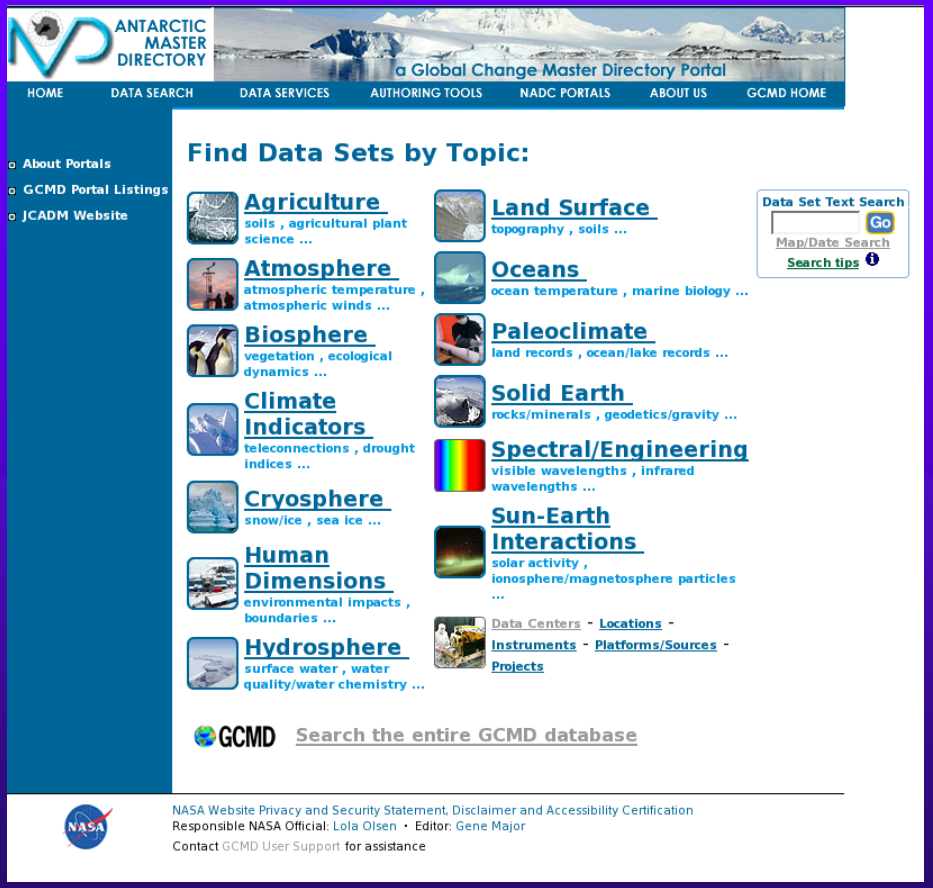
III. Controlled-Vocabulary Query

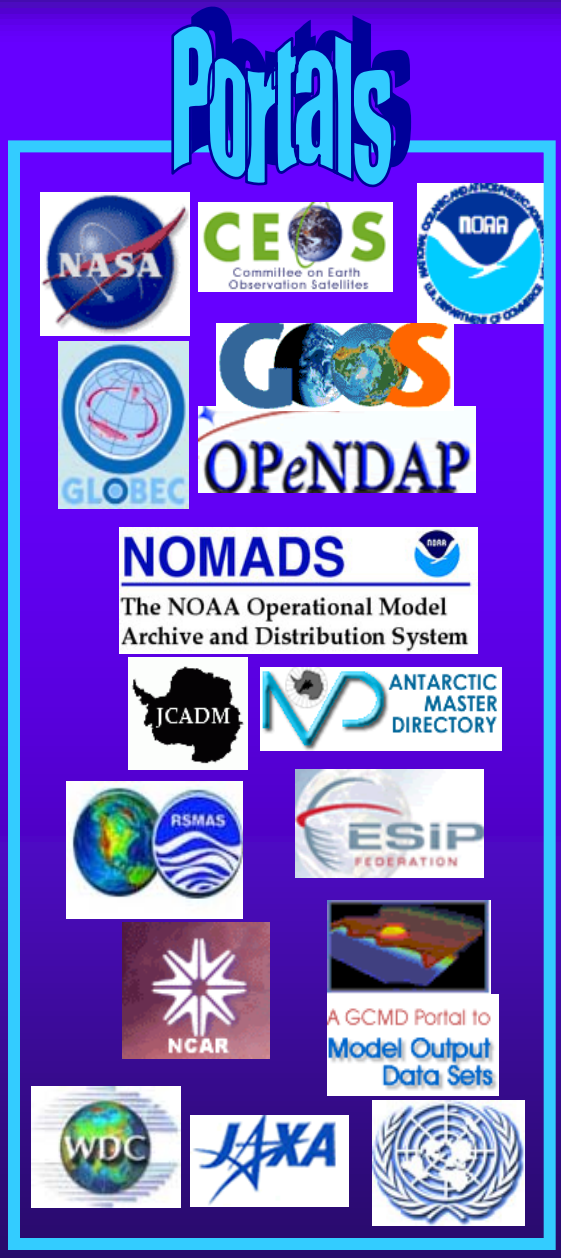
Earth Science Services Keywords





Highly Customized Portals: Virtual Subsets of the Directory



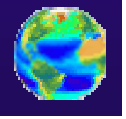


IV. Free-Text Query Language (Lucene)

Lucene

open source

- ✓High performance, full-featured text search engine written in Java
- ✓Cross-platform; supports fielded content; adaptable relevance ranking
- ✓Advanced free-text query language
- ✓Geospatial & temporal search capabilities



Lucene Query Language

Advanced Free-Text Query

- Single Keyword and Phrase Searches: [DODS]
- Boolean Operators: [ESIP AND “sea surface temperature”]
- Wildcard Searches: [NASA*]
- Fielded Searches: [DIF/Project: DODS]
- Fuzzy Search: [rain~0.8]
- Proximity Search [ocean temperature~10]

Query language can be adapted as such:

http://gcmd.gsfc.nasa.gov/KeywordSearch/Freetext.do?KeywordPath=&Portal=GCMD&MetadataType=0&Freetext=**DIF/Project:DODS+AND+DIF/Parameters:"Sea Surface Temperature"&**action=search

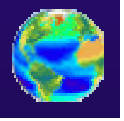
Lucene Query Example

Further Refine By

Query "Breadcrumbs"

Query Results

Metadata Record



V. Open Application Programmer Interfaces (APIs) – Mechanisms for remote clients to interoperate with GCMD

MDServer: Java-based protocol providing Java clients the ability to interoperate using the RMI protocol (communication protocol used by Java)

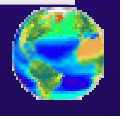
HTTP Protocol: Servlets that respond to GET and POST requests and return XML results.

SOAP: WSDL file implemented that reuses most of the business logic used in the MDServer. (Still in development)

MDServer

- ✓Provides a mechanism for remote clients to interoperate with GCMD using RMI protocol (communication protocol used by Java)
- ✓Assumes a Java client.

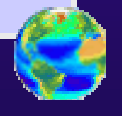
Application Programmers Interface (API)	
Document API	Create, modify, and remove Documents
Query API	Retrieve documents based on entry identifier, object identifier,or query expression.
Valids API	Insert, modify, remove, retrieve valids
Personnel API	Insert, modify, remove, retrieve and merge personnel.
Incoming Queue API	Insert, modify, retrieve incoming items.



HTTP Protocol

- ✓Provides a mechanism for remote clients to interoperate with GCMD using HTTP protocol
- ✓Returns all results in XML

Jython Servlets	
get_entry_ids.py	Retrieve a set of entry identifiers given a specified query.
getdif.py	Retrieve a DIF given its entry identifier or object identifier.
getdifs.py	Retrieve a set of DIFs given a specified query.
get_valids.py	Retrieve a set of valids given its type.
getdifs_by_personnel.py	Retrieve a set of personnel given the first, middle, and last name.



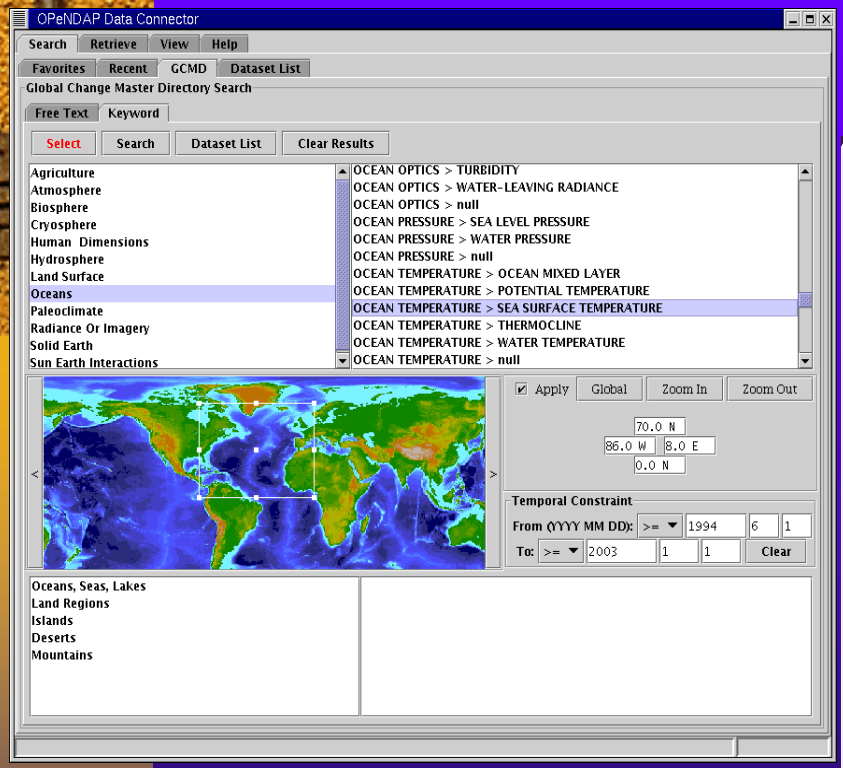
SOAP

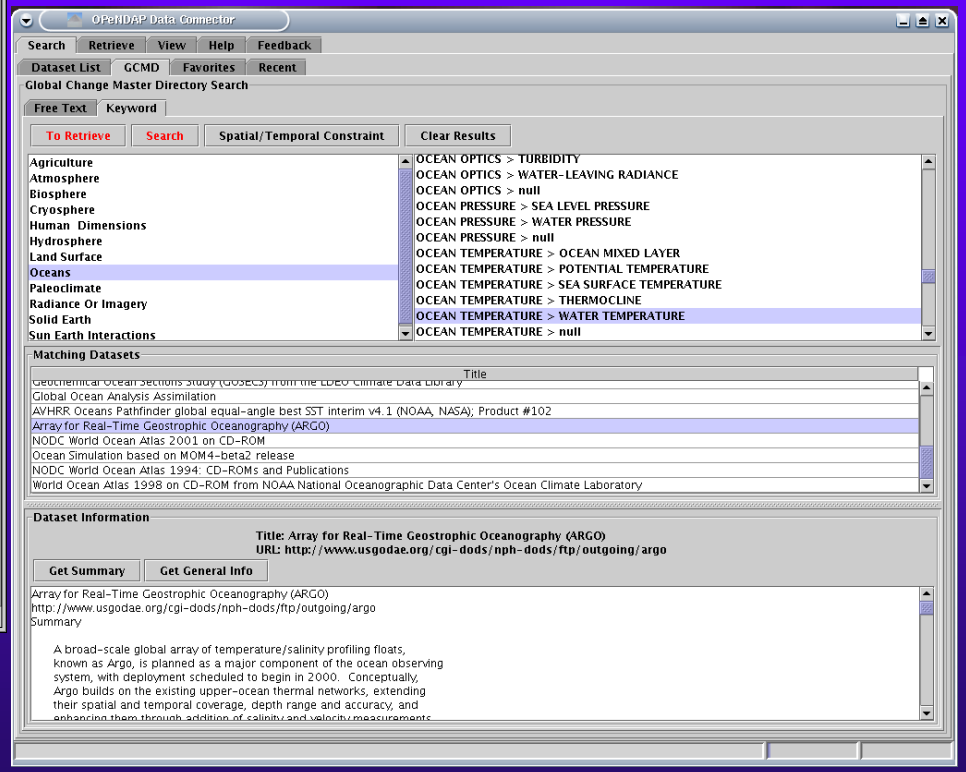
- ✓Provides a simple XML based protocol allowing organizations to inter-operate with the GCMD using a language-independent protocol
- ✓Developed a Web Services Descriptor File (WSDL) that implements most of the MDServer API

Application Programmers Interface (API)	
Document API	Create, modify, and remove Documents
Query API	Retrieve documents based on entry identifier, object identifier,or query expression.
Valids API	Insert, modify, remove, retrieve valids
Personnel API	Insert, modify, remove, retrieve and merge personnel.
Incoming Queue API	Insert, modify, retrieve incoming items for review.

V. Open Application Programmer Interfaces (APIs) – OPeNDAP Data Connector (ODC)

DODS/OPeNDAP uses the GCMD open API and servlets to query the GCMD directly





The "Search" frame provides interfaces to the GCMD's MDServer through APIs HTTP Protocol

GCMD Interfaces
Search / Publish Metadata Content

Web User

1

Search **data set** descriptions through GCMD search interface

Retrieve metadata records

Search **data set** descriptions through ***portals*** - virtual subsets of the directory

Retrieve metadata records

2

Search **data service** descriptions through GCMD search interface

Retrieve metadata records

GCMD Database / Server

HTML DocBUILDER

DocBUILDER; Metadata records are stored off-line where they can be later submitted to GCMD

Harvester : Retrieves, converts & exchanges metadata through HTTP or Open Archives Initiatives (OAI) servers

Metadata Converter : Converts metadata records (e.g. FGDC) to GCMD's Directory Interchange Format (DIF)

A

Publish metadata through web-based & off-line tools

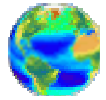
B

Harvest existing records for exchange

C

Translate existing records using *Extensible Stylesheet Language Transformations (XSLT)*

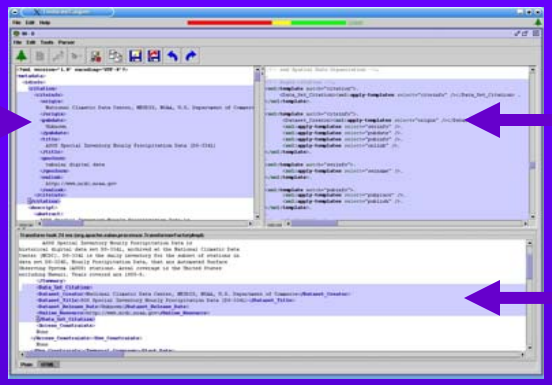
Metadata Tools



VI. OAI/HTTP Harvesting Capabilities – Metadata Converter (NOAA NCDC – DIF)

1. Apply Mapping to an XSLT Stylesheet

Sample CSDGM XML document

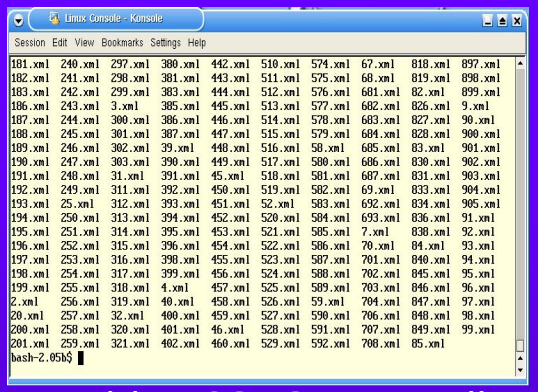


Edit XSLT Stylesheet

DIF XML

Treebeard XSLT Editor

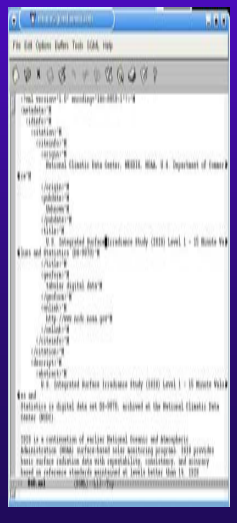
2. Obtain metadata



Resulting CSDGM XML Files

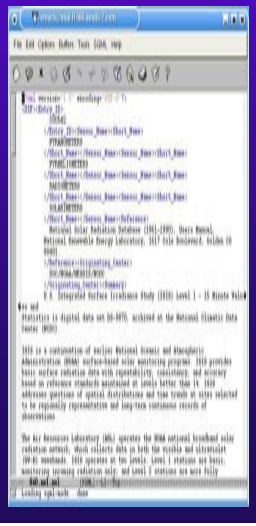
3. Translate, Format, and Submit to the GCMD

CSDGM XML



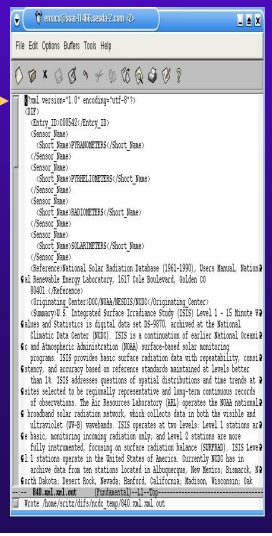
Translate using Xalan

Unformatted DIF XML



Format using HTML Tidy

Formatted DIF XML



VI. OAI/HTTP Harvesting Capabilities – Metadata Converter

- Treebeard
  - XSLT Transformer/Editor; validates XSLT

  http://treebeard.sourceforge.net/

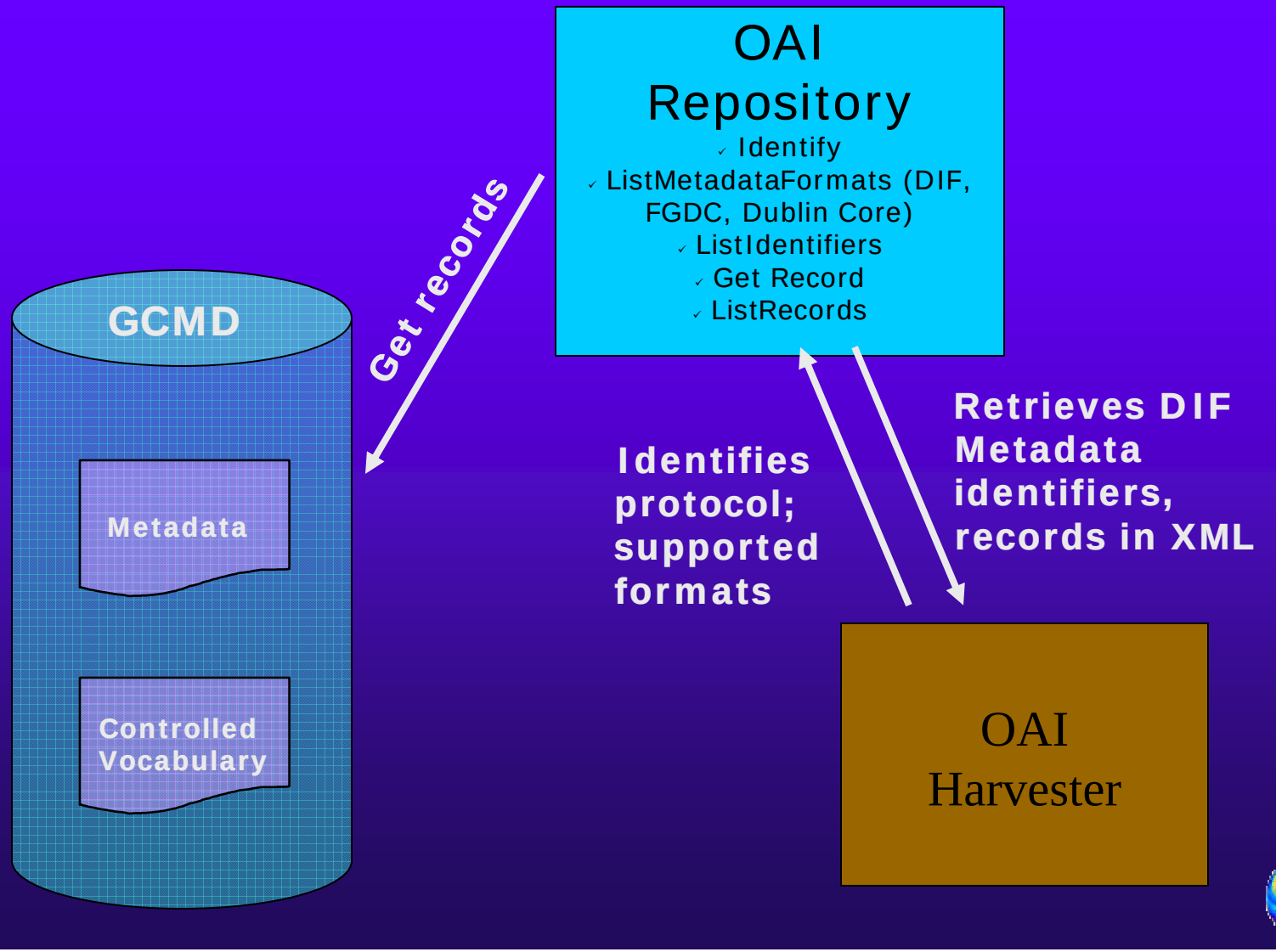
- Xalan-Java
  - XSLT processsor for Transforming XML documents to other XML formats, HTML, or Text

  http://xml.apache.org/xalan-j/

- HTML/XML Tidy
  - Formats and aligns XML/HTML

  http://tidy.sourceforge.net/

VI. OAI/HTTP Harvesting Capabilities – NCAR OAI Harvester Example



NCAR Subset to GCMD Round 1

- Keyword Issues:
  - "Term" Keywords are missing. Minimally the Science Keyword field must extend to the third level "Term". Four levels are recommended "Category > Topic > Term > Variable"
  - Non GCMD-standard keywords are used in some fields (URL content Type, Science Keywords, Personnel "Role", and Location)
- Must minimally have "Last_Name" in Personnel field.
  - Philosophical problem with NCAR
  - "Last Name" does not have to be a person. It just has to be unique. i.e. "DSS General Assistance".
- Free-text fields exceed 80 character line limit.
- Project fields have Data Center information in them?
- Metadata syntax does not validate with the DIF DTD

GCMD Subset to NCAR Round 1

- Show stopper:

  Data has Latin-1 characters embedded in metadata, which is not UTF-8 compatible.

Harvester fails with error:

Invalid byte1 of 1byte UTF-I sequence.

NCAR Subset to GCMD Round 2

- Syntax closer to validation but not quite.
- Most problems are keyword related.
  - Use of non-standard Location keywords.
  - Use of non-standard Science keywords.
  - Use of non-standard Personnel Role keywords.
  - Use of non-standard Related URL Content Type keywords.
- Personnel fields with “Role” only. If choose this field, “Role” and “Last Name” required.
- Two “Summary” fields in several of entries. “Summary” not repeatable.

GCMD Subset to NCAR Round 2

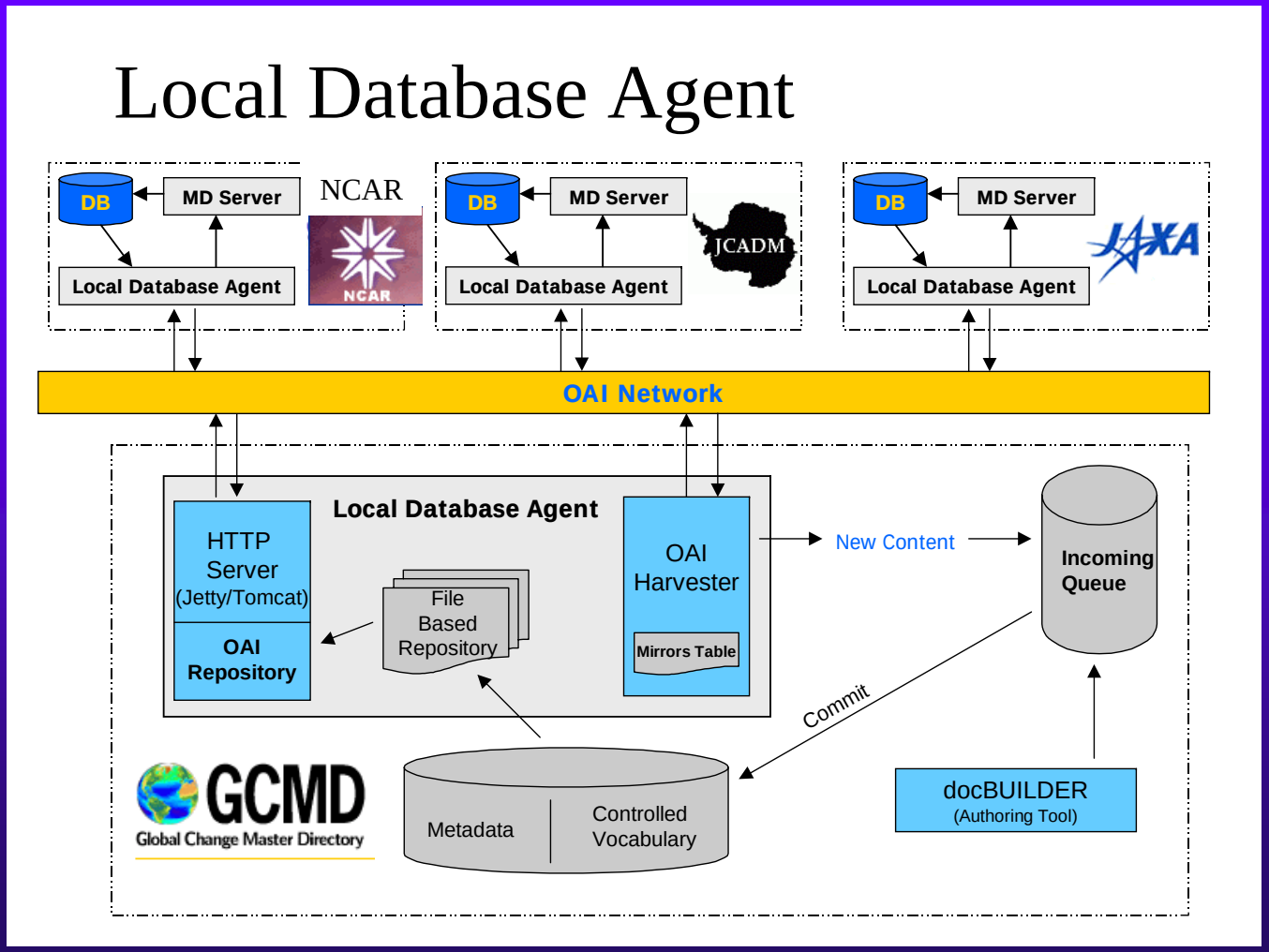
- Jumped character set hurdle.
  Harvesting in progress.
- Trouble with XSLT transform into THREDDS
- Trailing slash problem
- Cryptic validation error against our schema.
  Transform fails.
- Schema issues. (be specific)

VII. Local Database Agent (LDA) – Software for Distributed Content

VII. Local Database Agent (LDA) – Software for Distributed Content

- ✓ Peer-to-peer connectivity to other databases: enables distributed input
- ✓ Captures updates to the local database and shares content
- ✓ Minimal impact to the local DB activities
- ✓ Reduces maintenance related to data exchange
- ✓ Builds a sense of community among partners

VII. Local Database Agent (LDA) – Software for Distributed Content



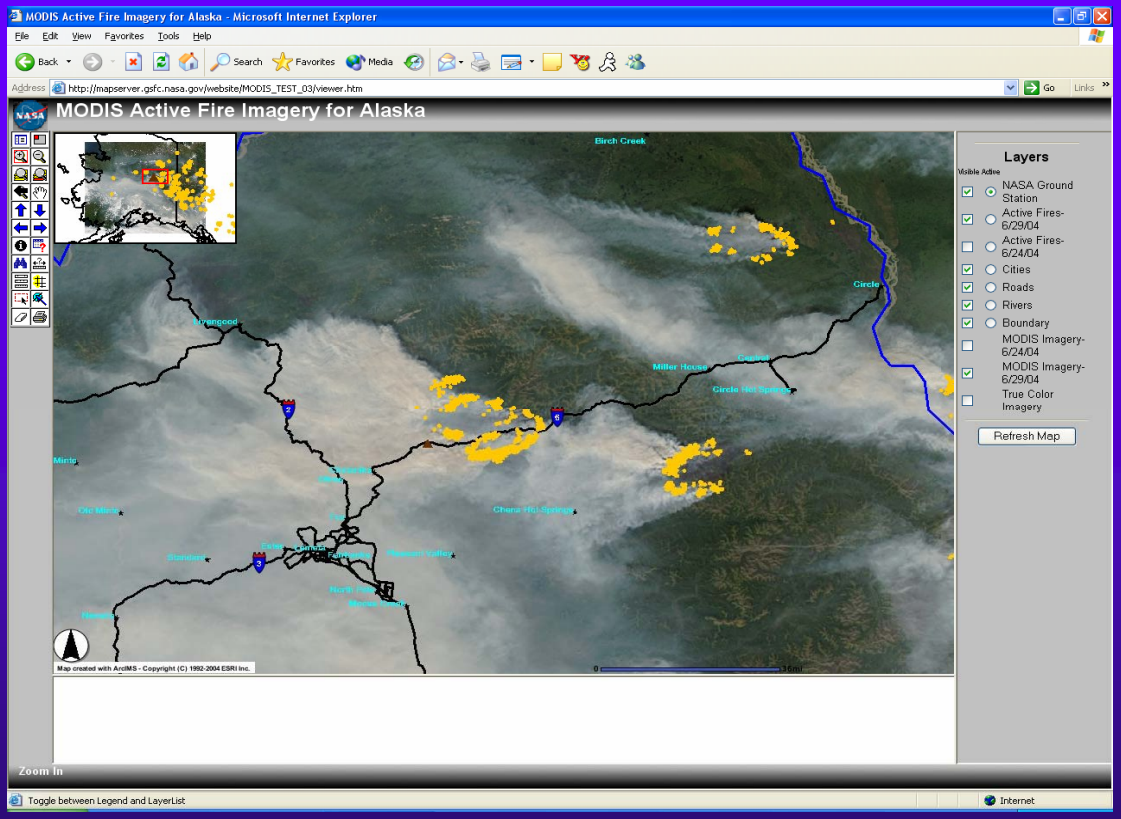
VIII. OGC WMS Capabilities

VIII. OGC WMS Capabilities - GCMD Prototype Mapserver

- Customized Portal for NASA GIS Metadata
- Access to Mapserver displaying vector and raster GIS data using ArcIMS Software
- Access to GIS data that can be downloaded into other GIS system

http://mapserver.gsfc.nasa.gov

VIII. OGC WMS Capabilities - GCMD Prototype Mapserver Provides Real-Time Visualization



Using ESRI OGC connector to provide WMS access. Anyone that has a WMS compliant client can access these WMS.

http://mapserver.gsfc.nasa.gov/website/MODIS_TEST_03/viewer.htm

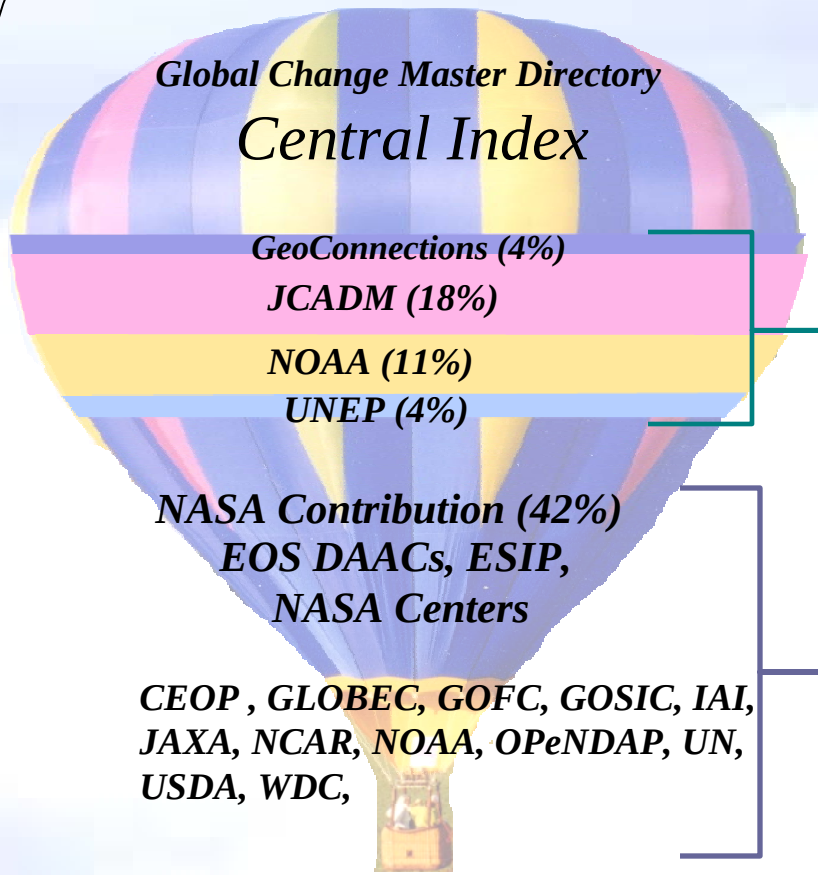
Edging Toward Interoperability

Questions?

Mélanie Meaux
mmeaux@gcmd.nasa.gov

The Global Change Master Directory:
A Directory of Earth Science Data
& Earth Science-Related Services

Middleware Searching Architecture



GCMD Search Engine

User Search

Harvesting Interfaces
(e.g. OAI-PMH)
Data sharing nodes

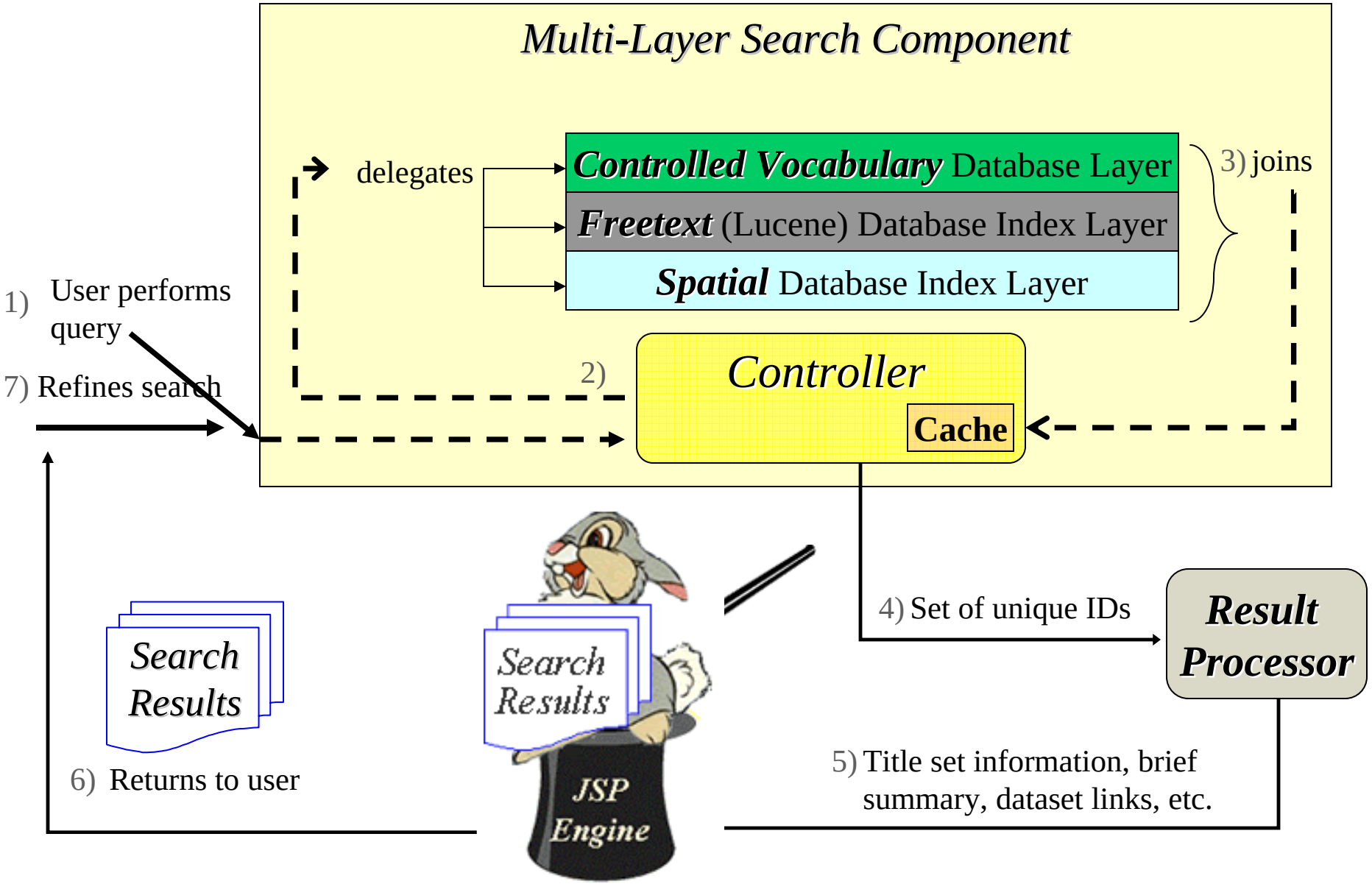
Portals API Interfaces
Data contributed through various portal collaborations

Other Data Contributions

Tools
Data contributed by staff, universities, federal and local agencies, international agencies and organizations, or commercial providers

Middleware Search/Retrieval Component

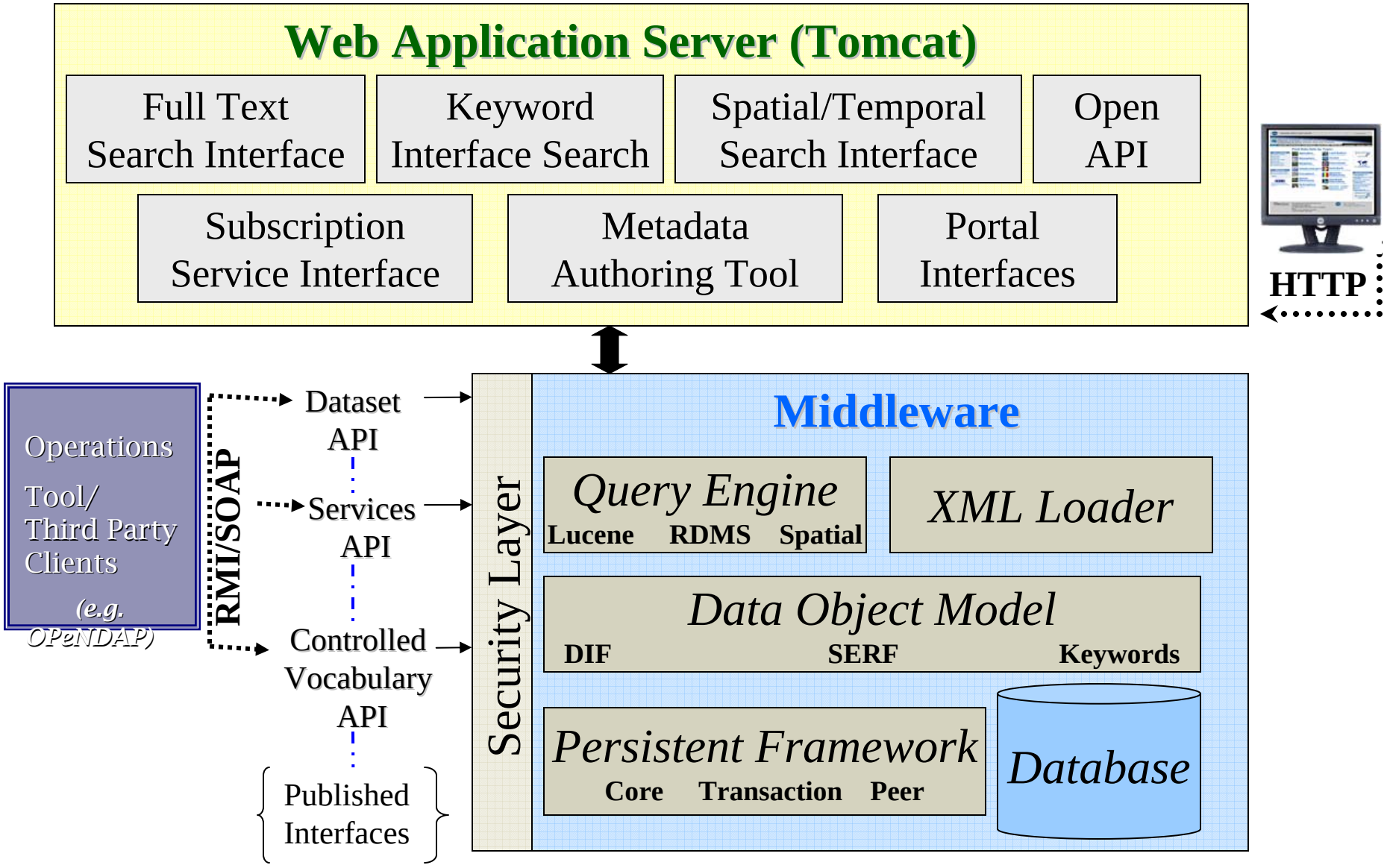
(Integration of spatial, freetext, and controlled queries)



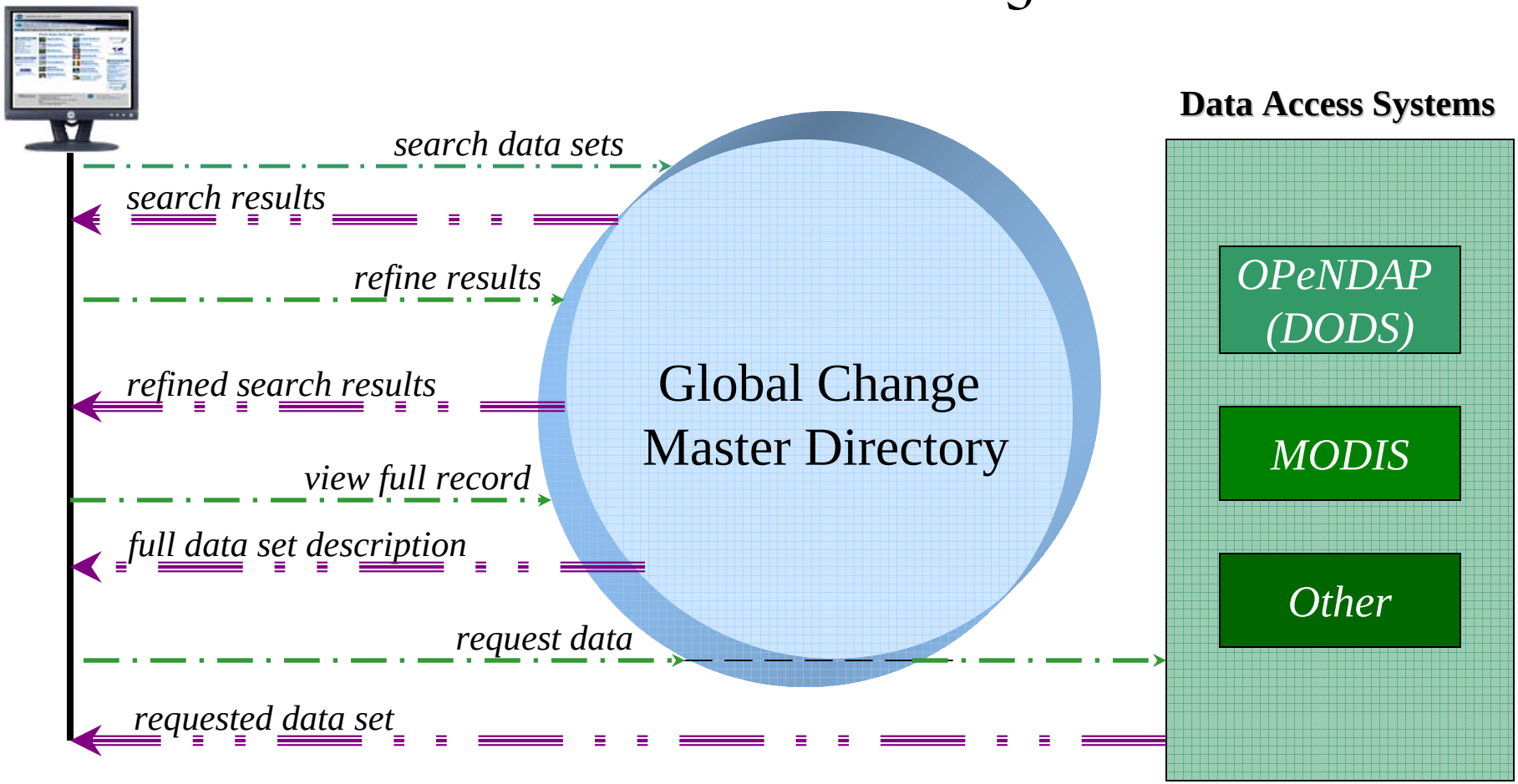
Ingest of Metadata: Sequence diagram

Interaction among User, OPS/docBUILDER and MD Server

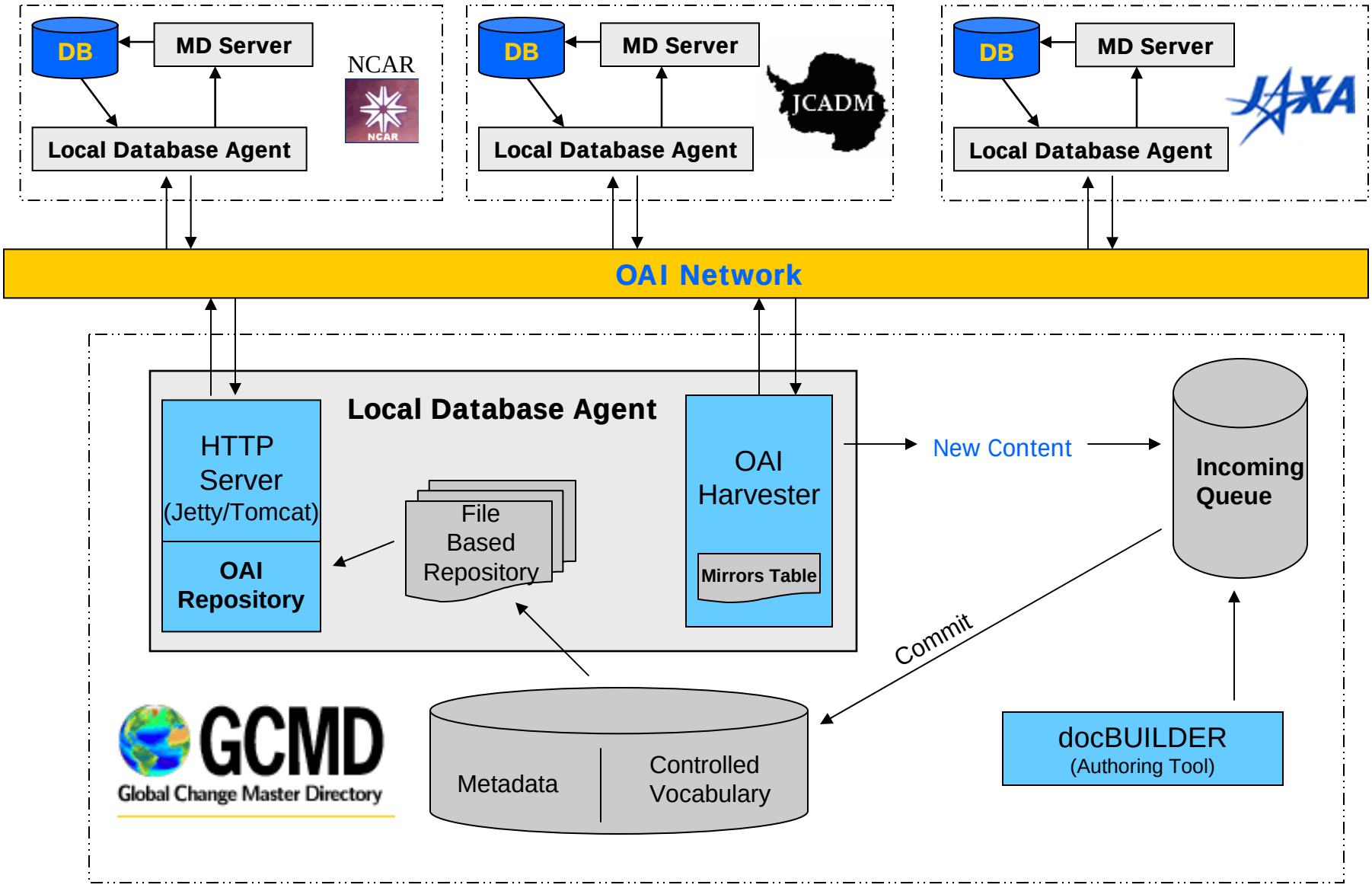
GCMD Architecture



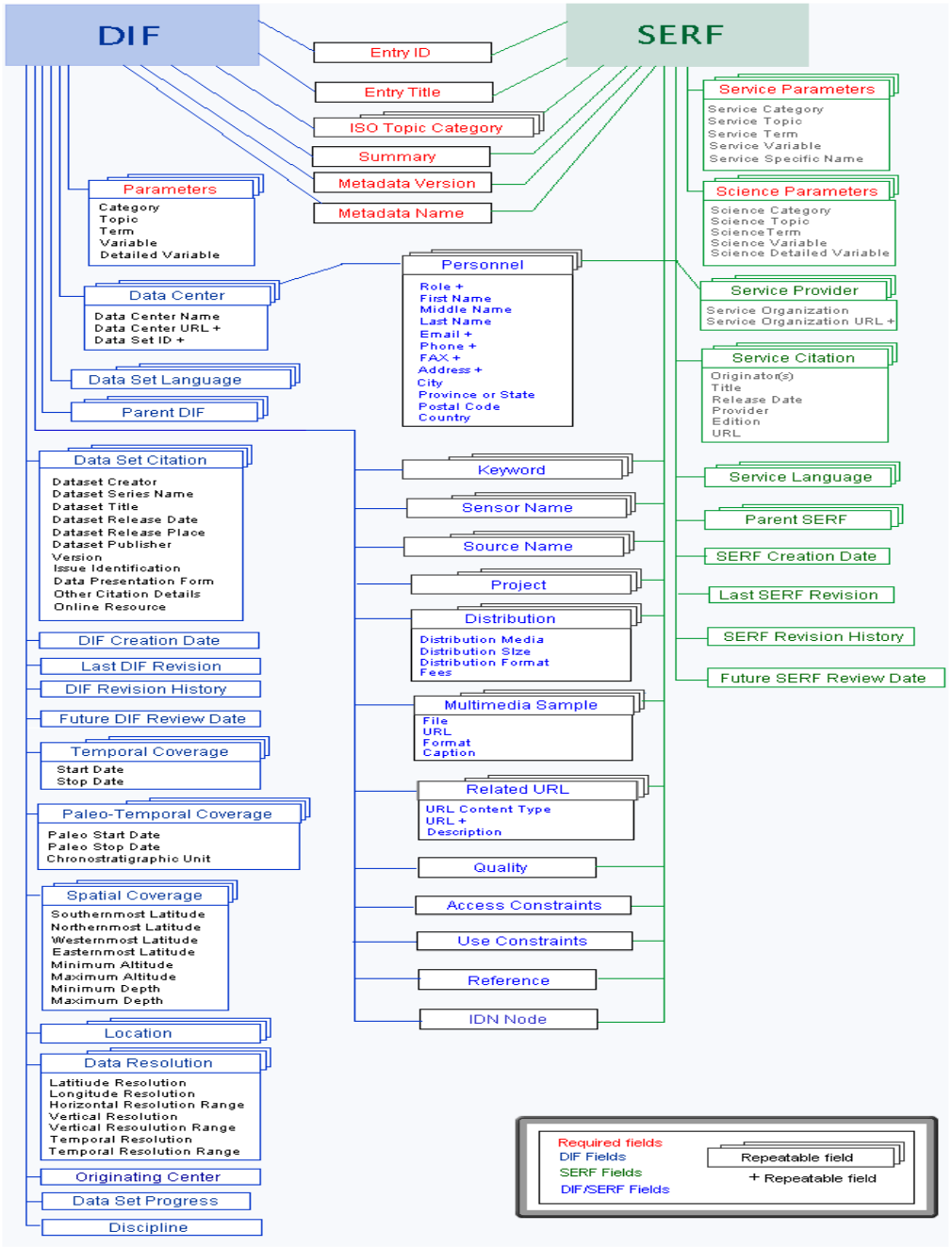
Data Retrieval Flow Diagram



Local Database Agent



GCMD DIF and SERF Diagram



Data Center Bucket Conversion

Original	Modified	Sub-levels	Sub-sub-level
	(Topic)	(Term)	(Variable)
COMMERCIAL	Commercial		
DOC	U.S. Federal Agencies	DOC	
DOD	U.S. Federal Agencies	DOD	
DOE	U.S. Federal Agencies	DOE	
DOI	U.S. Federal Agencies	DOI	
EPA	U.S. Federal Agencies	EPA	
FEDERAL AGENCIES	U.S. Federal Agencies	DHHS	
		DHS	FEMA
		DOC	
		DOD	
		DOE	
		DOI	
		DOT	
		EPA	
		GPO	
		HUD	
		NASA	
		NSF	
		OSTP	
		TVA	
		USDA	
INSTITUTIONS	Consortia/Institutions		
INTERNATIONAL	Government Agencies-Non-US		
INTERNATIONAL AGENCIES	Government Agencies-Non-US		
NASA	U.S. Federal Agencies	NASA	
NOAA	U.S. Federal Agencies	DOC	NOAA
NON-PROFIT ORGANIZATIONS	Non-Government /Non-Profit Organizations		
NSF	U.S. Federal Agencies	NSF	
REGIONAL AGENCIES	U.S. State/Regional/Local Agencies		
UNIVERSITIES	Academic		
USDA	U.S. Federal Agencies	USDA	
USGS	U.S. Federal Agencies	DOI	USGS
WORLD DATA CENTERS	Multinational Organizations		

Data Center Bucket Clean Up

- Each data center name is a hierarchy consisting of a Category (or bucket), Short Name and Long Name
  - Data Center Category > Data Center Short Name > Data Center Long Name
    - Government Agencies-Non-US > JP/JAXA/ISAS > Institute for Space and Astronautical Sciences, Japan Aerospace Exploration Agency, Japan
- Government Agencies – Non-US – broken down by country ISO identifier