Services, Metadata and Interoperability:
Lessons Learned & Use Cases

Brian Wilson
Jet Propulsion Laboratory

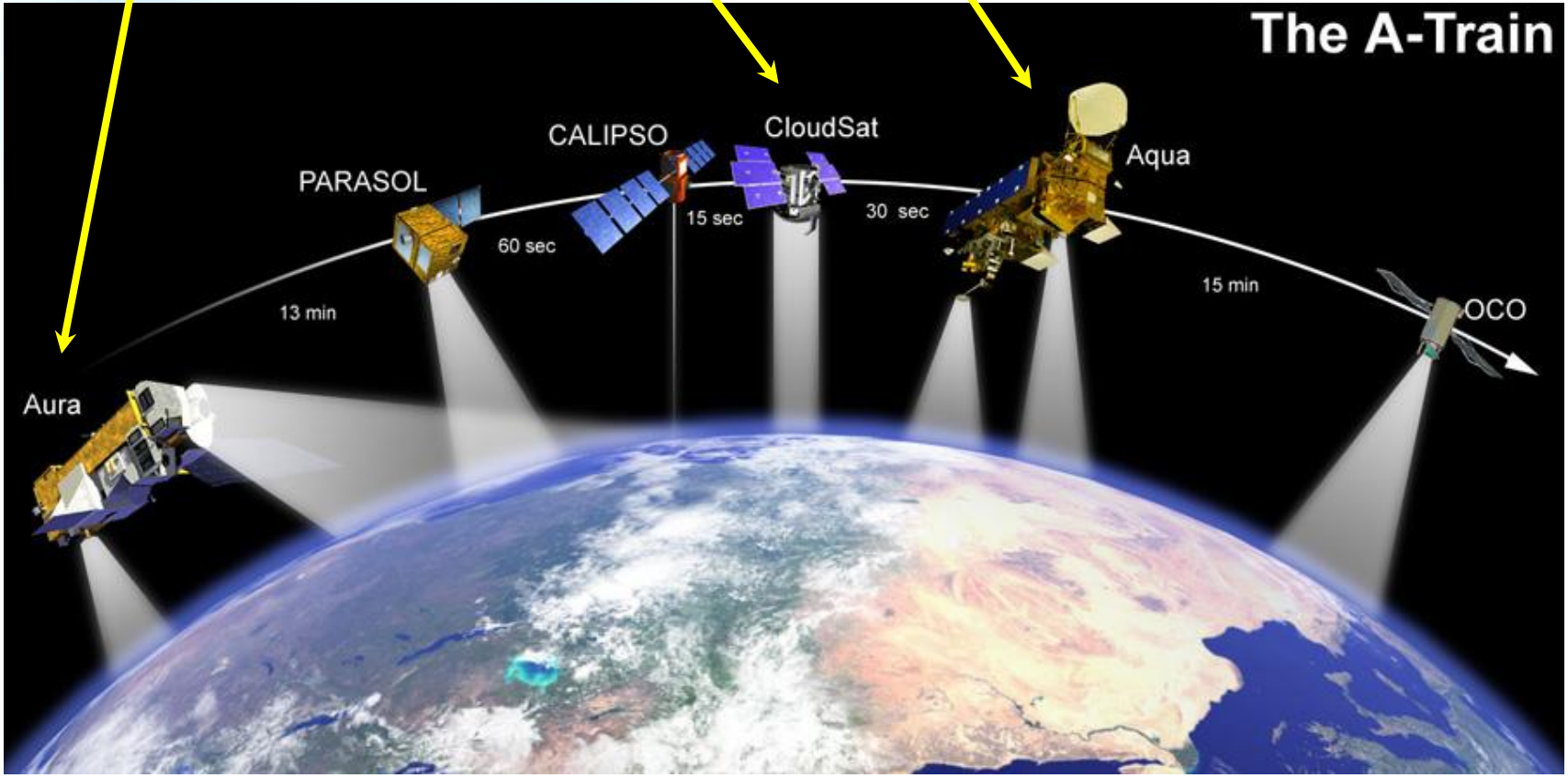
Publish XML Metadata for discovery
Use Loosely-Coupled Web Services
Service Casting (advertise)
Federated OpenSearch

“Take the A-Train”

CloudSat Classes

Water from AIRS, AMSR-E, MODIS

Water from MLS



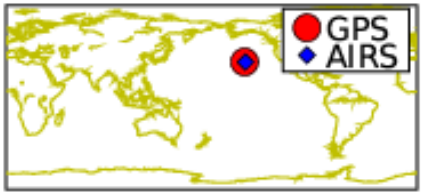
Large-Scale, Distributed Data Fusion

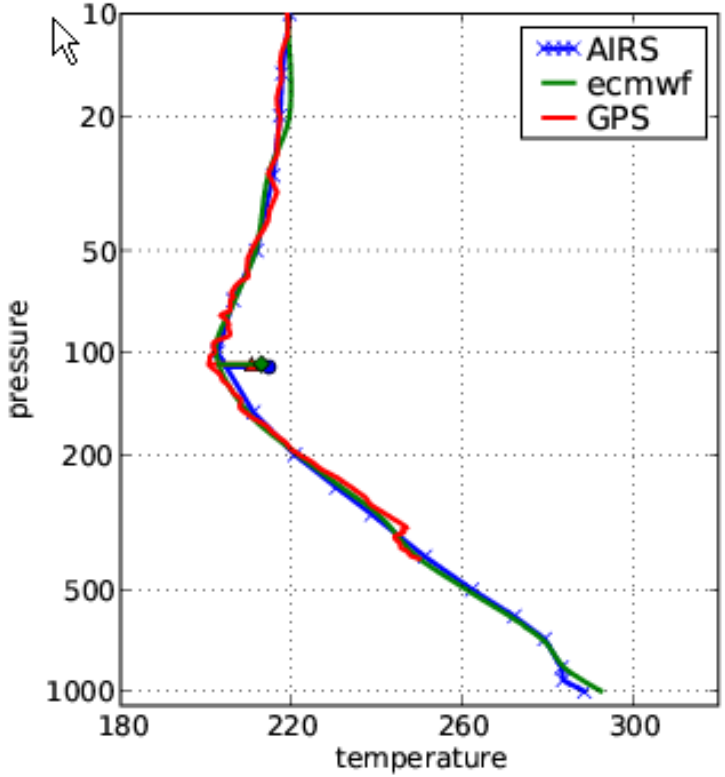
- Find Level-1B/2 datasets across multiple data centers
  - Space/time granule query for multiple EOS ("A-Train") instruments – AIRS, AMSR-E, AMSU, MODIS, Cloudsat, GPS
- Co-locate retrievals using space/time metadata
  - Instantaneous "matchups" in space & time
- Read the data (from HDF)
  - Temperature, water vapor, quality flags, cloud properties
- Understand the data (CF-style attributes help)
  - Units, scale_factor, quality control (non-trivial !!), etc.
- Publish merged products
  - Water vapor climatology, stratified by Cloudsat cloud classes
- Publish multi-sensor "fused" products
  - Determine instrument biases, understand by stratifying
  - Fuse L2 data on a common grid or use joint retrieval algorithm

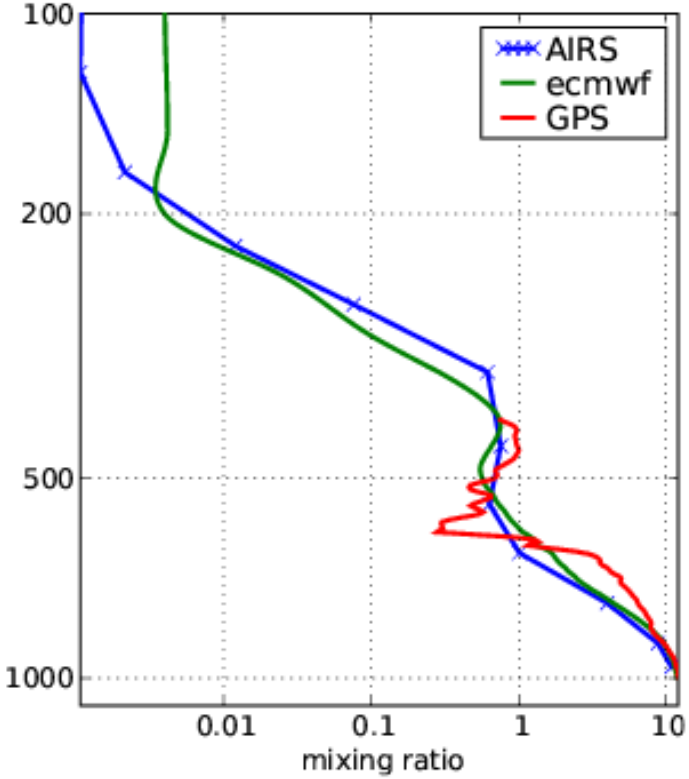
AIRS / GPS Matchups

AIRS/GPS Temperature & Water Vapor Comparison Plots

AIRS: time=Jan 4 20:22:56 2003, loc=(-150.31, 32.57)
GPS: time=Jan 29 23:18:56 2003, loc=(-151.10, 32.68)
diff_time=-2170560 sec, diff_distance=74.7 km

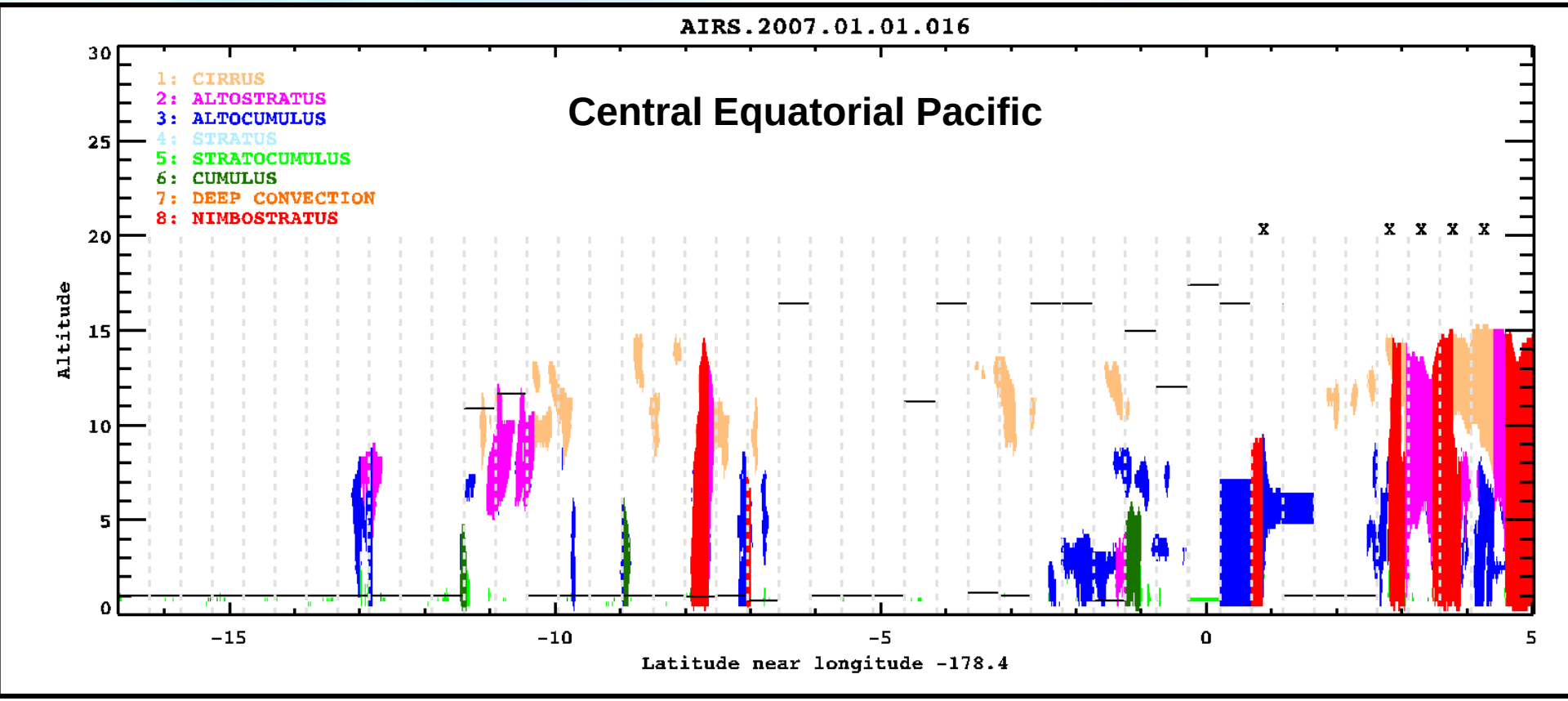






AIRS / CloudSat Matchups

Color fill = CloudSat Class (Sassen and Wang, 2008)



Black lines: AIRS 'best' retrieval altitude
X: no AIRS tropospheric profiling.

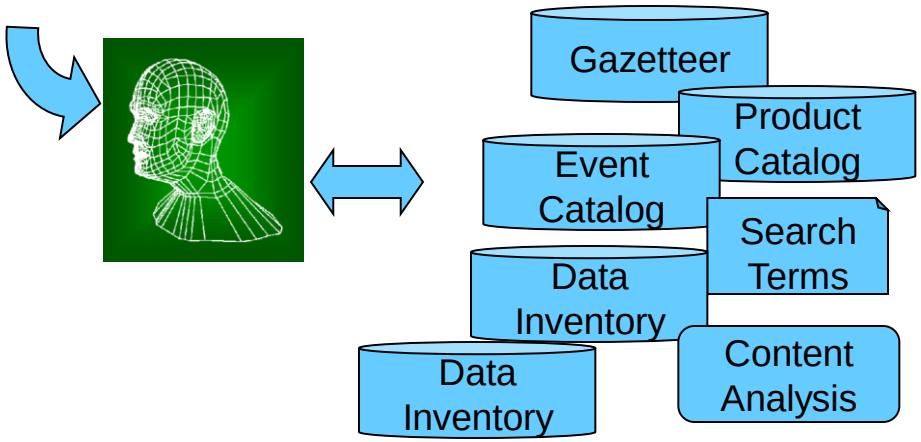
Desired Capabilities / Use Cases

- Dataset Discovery
  - Usually human task
  - Can use keyword search or faceted semantic (“smart”) search
  - Publish rich metadata in XML format (e.g. KML with links)
- Space/Time Granule Query (Level 1B, 2, & 3)
  - Specify geo-region & time range, receive of list of granule URL’s
  - Scaling to ALL datasets presents serious performance problem
- Easy Data Access (to the Numbers)
  - Slicing using OpenDAP URL’s preferred to moving entire files
  - OGC Web Coverage Server (WCS) allows coverage selection
    - But hard for L1B & L2 data (must stitch mosaics)
- Data Understanding
  - CF attributes: short_name, long_name, units, scale_factor, etc.
  - But need more metadata and semantic standards
- Data Provenance
  - Data files should contain links to instrument documentation & retrieval provenance

Assisted Data & Service Discovery: Future Vision

- Researcher uses semantic and content-based search to search for data using proper names, domain-specific jargon, and high-level specifications
- Researcher quickly finds data with the parameters, resolution, and coverage needed for the heat flux analysis
- Researcher finds sensor web resources with coverage of the storm region

Select from
Semantic Web of Earth Data where
parameter = esipfed:precipitation
instrument = gcmd:TRMM
date = between 6 Sept and 16 Sept 1996
region = ogc:South Atlantic
phenomena = esipfed:hurricane
function = rainfall(region=ogc:Bermuda) >3



Outline

- Loosely-Coupled Web Services (SOA)
  - Expose fine-grained functionality as reusable web services (Publish algorithms!)
  - Use REST or SOAP, or both
  - Use workflow engine to orchestrate service chains: BPEL, SciFlo
  - Formal workflow yields data provenance for free
- Best Practices for Services
  - Use standard protocols: OGC WMS/WCS, OpenDAP, OpenSearch
  - Modeless REST (HTTP one-liner) services are easiest to use
- Use Cases / Lessons Learned
  - Binary file format (presume HDF5/netCDF4 with CF attributes)
  - Publish rich dataset metadata for search (XML standards)
  - Space/time query for data granules
  - Subsetting by parameter, space, and time
  - Matchups: two instruments, or retrieval to model comparisons
  - [Your Mission Use Case here.]

“The Grid” Concept

“On-Demand”
Virtualization

Portal Data Storage CPU Algorithms Networks

Web Services "Software Stack"

- Metadata: XML, XML schema, and RDF
- Messaging: Simple Object Access Protocol (SOAP) or Representational State Transfer (REST)
- Service interfaces: Web Service Description Language (WDSL) and OWL-S for SOAP, WADL for REST
- Service registries: Universal Description, Discovery and Integration (UDDI)
- Semantic Web: Ontologies or "concept maps" in Resource Description Framework (RDF) and Ontology Web Language (OWL)
- Grid Workflow: Many dataflow engines
  - SciFlo – GENESIS-II ESIP
  - Business Process Execution Language (BPEL) – GMU LAITS BPELPower engine

REST/WADL and SOAP/WSDL

- REST growing in popularity due to simplicity and browser “mashups”.
- Best to implement REST or both.
- SOAP (with flat interface):
  GeoRegionQuery('AIRS.RetStd', 'L2', 'latest', '2003-01-03 00:00:00', '2003-01-03 04:59:59', -90., 90., -180., 180., 'Medium')
- REST equivalent in SciFlo:
  endpoint?_service=EOSServices&_method=GeoRegionQuery &datasetId=AIRS.RetStd&level=L2&version=latest &startTime=2003-01-03 00:00:00&endTime=2003-01-03 04:59:59 &latMin=-90&latMax=90&lonMin=-180&latMax=180 &ResponseGroups=Medium
- Security Issues:
  - SOAP: WS-Security, SAML
  - REST: OpenId, OAuth

Data Query/Access Services

- Many possible services related to data query/access:
  - Time & Geolocation Query → yielding data inventory
  - Query on Quality Flags
  - General Metadata Query → catalog satisfying conditions
  - Data Access - by object or granule ID and variable name
  - Data Slicing - as in OpenDAP
  - Data Subsetting – by lat, lon, alt, & time ranges
  - Data Reprojection – coordinate transformations
  - Parameter Subsetting – select only desired physical variables
  - Data Reformatting – choose output format, as in WMS/WCS
  - On-Demand Grid Computations – grid diff in GraDS/DODS
  - Variable Bundles – grouping specified by scientist
  - Return Composite Data Objects
    - 4DGeoGrid, 4DGeoSwath
    - Variable grid with lat/lon/alt/time coordinates & attributes
  - User Composites – add custom object to data model
  - Semantic Support
    - Use generic variable names
    - Reasoning about physical variables & metadata

SciFlo Workflow Engine

- Automate large-scale, multi-instrument science processing by *authoring* a dataflow document that specifies a *tree* of *executable* operators.
  - Visual Authoring Tool (VizFlow)
  - Distributed Dataflow Execution Engine
  - Data Grid: Move data "granules" to the operators using FTP, HTTP, or OpenDAP URLs.
  - Compute Grid: Move operators (executables) to the data.
  - Built-in reusable operators provided for many tasks such as subsetting, co-registration, regridding, data fusion, etc.
  - Custom operators easily plugged in by scientists.
  - Leverage convergence of Web Services (SOAP) with Grid Services (Globus toolkit v4).
- Network of SciFlo Nodes
  - Space/time query & access for many Earth Science datasets
  - Move computations "inside" large Data Centers (e.g. DAACs)

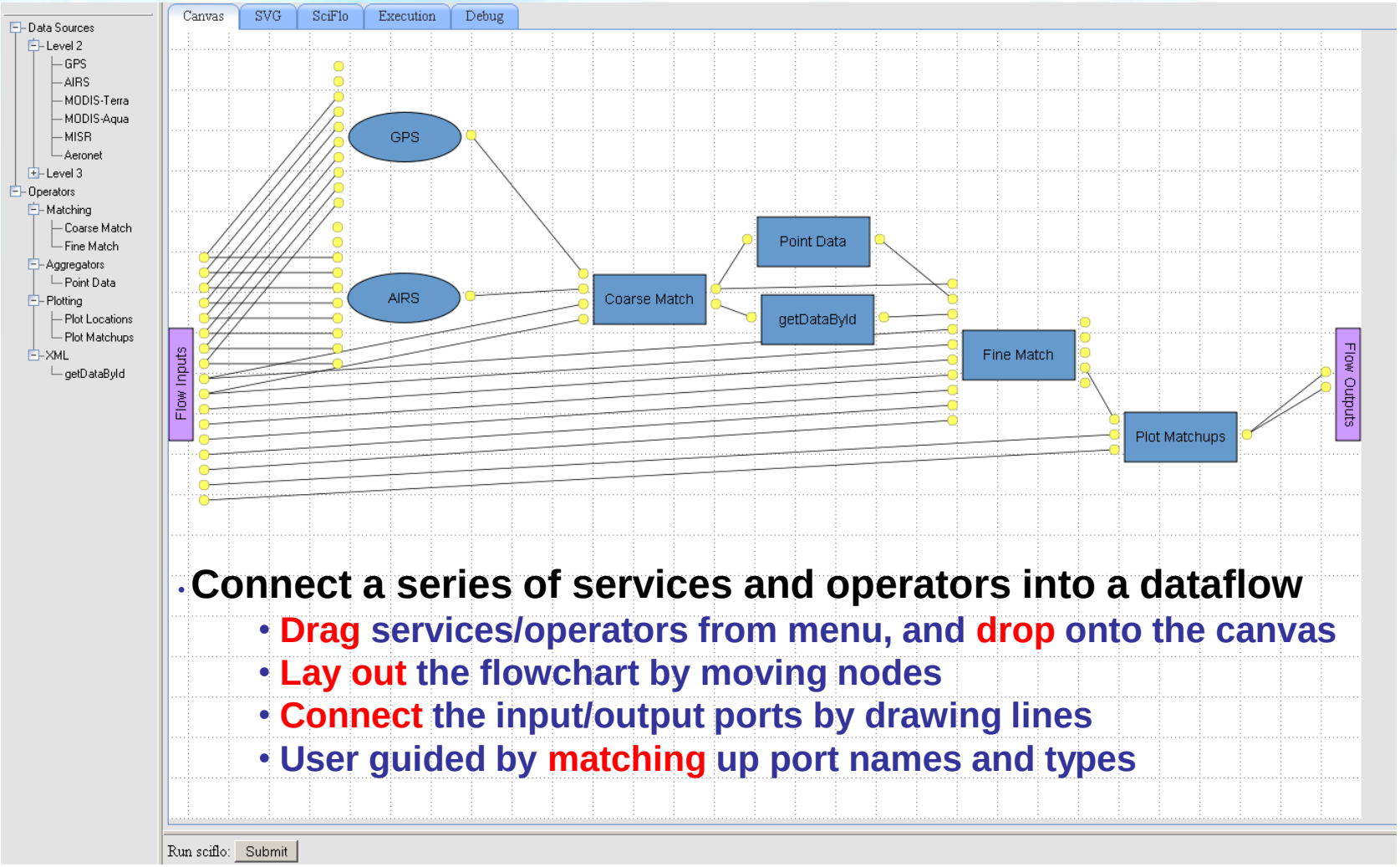
SciFlo's 4DGeoGrid MicroFormat

- Generic XML standard exploiting netCDF (CF 1.0) and HDFEOS conventions
  - In-line small vectors of numbers
  - Use OpenDAP URLs to "point to" larger data arrays
  - Read/write grid from/to HDFEOS or netCDF files.
  - Metadata are crucial to reuse. Arrays of numeric data are a detail.
- 4DGeoGrid or 4DGeoSwath

```
<4DGeoGrid using="@http://gsfc.nasa.gov/dods/AIRS.XX.hdf" >
   <Latitude><Series>-90, 90, 2</Series></Latitude>
   <Longitude>@http://gsfc.nasa.gov/dods/AIRS.hdf?Longitude
   <Altitude>1, 5, 10, 50, 100, 500, 1000</Altitude>
   <Time>@?Time</Time>
   <Variable>
      <Name>atmosphericTemperature</Name>
      <Unit>degrees Kelvin</Unit>
      <MissingValue>-9999.</MissingValue>
      <Data>@http://gsfc.nasa.gov/dods/AIRS.XX.hdf?TAirStd
   </Variable>
   <Variable>
      <Name>waterMixingRatio</Name>
      <Data>@http://gsfc.nasa.gov/dods/AIRS.XX.hdf?MMRH2OStd
   </Variable>
</4DGeoGrid>
```

Visual Dataflow Programming

GPS & AIRS Level-2 Space/Time Matchup



- **Connect a series of services and operators into a dataflow**
  - **Drag services/operators from menu, and drop onto the canvas**
  - **Lay out the flowchart by moving nodes**
  - **Connect the input/output ports by drawing lines**
  - **User guided by matching up port names and types**

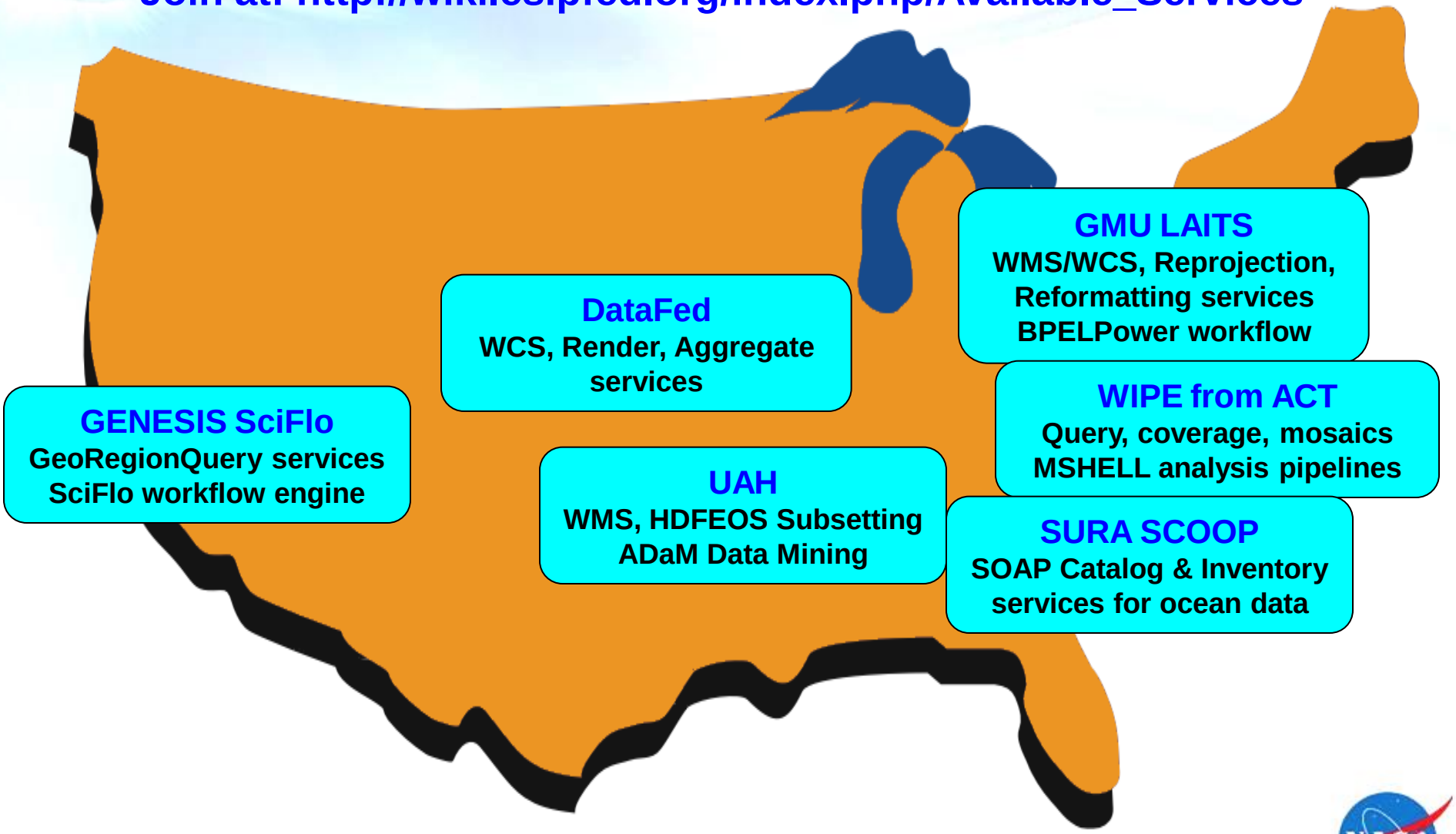
(AJAX Javascript and SVG Programming by Gerald Manipon & Wilson)

Federation Web Service Chaining Demo

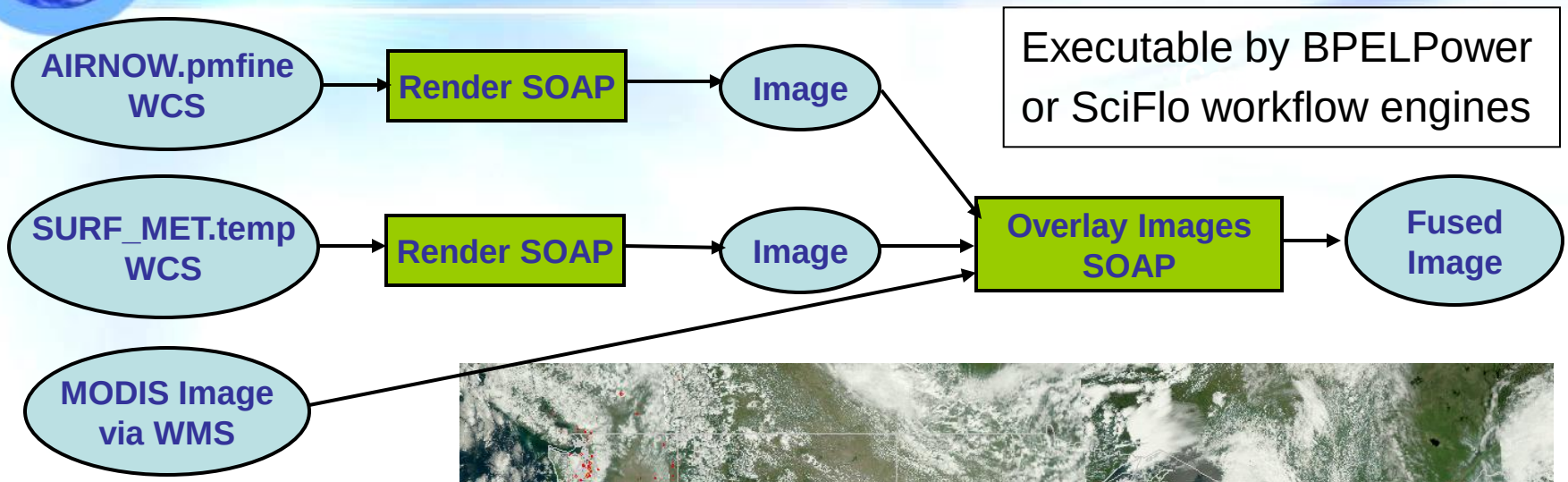
- Success Story (2006)
  - Developed as a collaboration between TIWG and ESIP Federation Web Services Cluster
    - Activity organized using the Federation Wiki, WS Cluster telecons, and TIWG telecons (participants are geographically dispersed)
    - Nine groups published callable services on the Wiki
    - Six groups participated in service choreography demos
  - *Demonstrated* service chains for data query & access, visualization, image overlay, and simple data fusion
  - *Demonstrated* interoperability between different types of services: OGC WMS/WCS, SOAP, OpenDAP
    - A service chain can combine all three
  - Identical Air Quality analysis chains executed by *both* BPELPower & SciFlo workflow engines
    - Two different environments for authoring XML workflow documents]

Geographically Distributed Participants & Services

Join at: http://wiki.esipfed.org/index.php/Available_Services



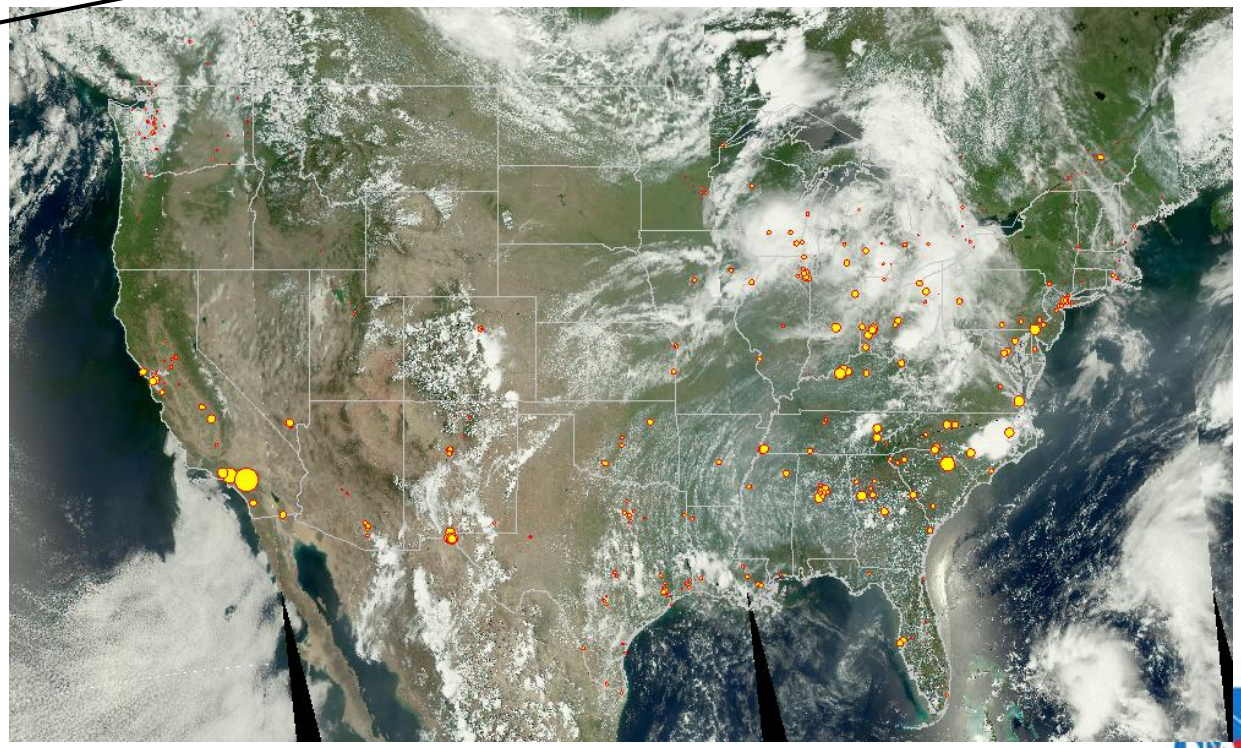
Example: Air Quality Scenario



Executable by BPELPower or SciFlo workflow engines

Dataflow Combines 5 remote services:

- OGC WCS calls (2)
- OGC WMS call
- SOAP services to render & overlay images



Lessons Learned

- Syntactic Interoperability is relatively easy
  - Can connect WCS, SOAP, & XML message- and OpenDAP-based services
- Semantic Interoperability is harder
  - Service interfaces vary in semantics, not just structure
  - Some convergence on interfaces for space/time query and access: WCS & SOAP/XML services
  - Often have to handcraft “glue” code for semantic mediation
- Simple, flat interfaces are more reusable
  - Complex, hierarchical XML docs. as inputs are a barrier to reuse
  - Most inputs should have smart defaults to hide complexity
  - If interface is flat, then REST and SOAP are interchangeable
- Standards, standards, standards
  - Adopt standard protocols like: OGC/WMS, WCS, WFS
  - Need many more standard XML microformats (standard payloads)
  - Need standardized search protocol (i.e. OpenSearch)

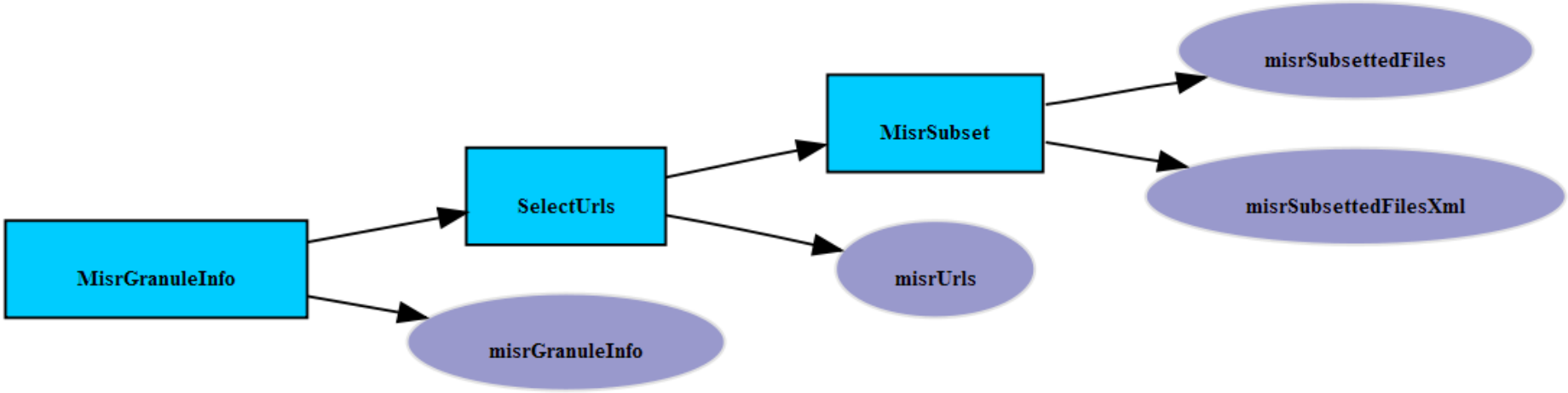
Provenance in Production Systems

- Two Approaches:
- Instrument the Production Scripts
  - Call out to ‘provenance capture’ API to record metadata
  - Could be web service calls
  - Intrusive
  - Only retain what you explicitly save
- Use Formal Workflow for Production
  - Annotated workflow document contains provenance
  - Versions of operators
  - Web service endpoints
  - Intermediate & final results, or pointers to them
  - Use links to limit combinatorial explosion
  - Workflow points to entire provenance, if URI’s are permanent

Example: Subsetting Workflow

MISR L2 Aerosol Granule Subsetter

Execution monitoring:



Work Unit Monitoring

index	procId	type	dependencies	status	execution time	results
0	MisrGranuleInfo	soap		done	0.970	granuleInfo: xml
1	SelectUrls	python function		done	0.060	misrGranules: list
2	MisrSubset	python function		done	17.270	misrSubsettedFiles: list [download]

Work Unit Status/Color Legend: waiting, ready, staging, working, finalizing, done, cached, exception, cancelled, paused

Execution log: log

Auto-generated SciFlo Input Form

MisrSubsetter

Subset a list of MISR granules by lat/lon region and variable list, yielding a list of netCDF files containing the space and parameter subset.

Sciflo Inputs

Input widgets

datasetName: MISR

level: AS_AEROSOL

startTime: 2001-01-01 00:00:00 --> 2001-01-01 00:00:00

endTime: 2001-01-05 00:00:00 --> 2001-01-05 00:00:00

latMin: -30.

latMax: 10.

lonMin: 0.

lonMax: 30.

☑ round?

misrVariables: Groups:

Dimensions:

Variables:

“Practical” Provenance

- Commit to Workflow (many other benefits)
  - Declarative production streams
  - Auto-parallel execution
  - Workflow can be distributed using web services (multi-sensor, multi-data center science)
- Get Basic Provenance For Free
  - No need to instrument production systems
  - Provenance graph implicit in SciFlo document & REST URLs
  - Web graph of versioned, permanent resources
  - By policy, can archive resource replicas (with URL redirection)
- Web Services Era
  - Distributed provenance is important
  - SciFlo or Open Provenance Model (OPM) docs. link to each other
  - Trace full graph by following REST URL’s
  - Reproducible workflows

Some Best Practices

- Easy data access: drill into data files
  - OpenDAP URLs
  - Parameter subsetting and array slicing for free
- Auto-data understanding
  - CF-1.0 compliant attributes in data file
  - Or similar attributes in HDFEOS file
- Advertise your web services & apps. for discovery
  - Service Casting (published Atom feed)
- Standardized data discovery & granule space/time query
  - Federated OpenSearch
- Importance of permanent names for data objects
  - When granule produced, provider assigns permanent URL
- Social collaboration on distributed analysis workflows
  - Informally on wiki
  - Talkoot: workflows shared & modified within collaboration env.

Open Data Access Protocol

- www.opendap.org
  - Drill down into a “deep web” of science data.
  - Use a one-line query URL to retrieve a slice of a variable grid from a netCDF or HDF file anywhere in the world
  - Binary wire protocol for fine-grained data transfer
- OpenDAP URL
  - http://gen-dev.jpl.nasa.gov/genesis/cgi-bin/dods/nph-dods/genesis/data/airs/L2/20030113/airx2ret/AIRS.2003.01.13.171.L2.RetStd.hdf?TAirStd(1:3, 3:6, 4:17)
- OpenDAP Servers
  - netCDF, HDF, GRIB, FreeForm, JGOFS, other file formats
  - Easy to implement another server
- OpenDAP clients
  - Matlab and IDL, any web browser
  - Python (pydap or SciFlo)

Some Best Practices

- Easy data access: drill into data files
  - OpenDAP URLs
  - Parameter subsetting and array slicing for free
- Auto-data understanding
  - CF-1.0 compliant attributes in data file
  - Or similar attributes in HDFEOS file
- Advertise your web services & apps. for discovery
  - Service Casting (published Atom feed)
- Standardized data discovery & granule space/time query
  - Federated OpenSearch
- Importance of permanent names for data objects
  - When granule produced, provider assigns permanent URL
- Social collaboration on distributed analysis workflows
  - Informally on wiki
  - Talkoot: workflows shared & modified within collaboration env.

Climate & Forecast (CF) Conventions

- Standard Attributes for each variable
  - standard_name, long_name, units, scale_factor, add_offset, description, valid_range, _FillValue, flag_values, flag_masks, flag_meanings, etc.
  - Check compliance!
- Standard Coordinates for variables
  - latitude in degrees_north, longitude in degrees_east
  - Vertical levels: height in meters, or pressure in millibars
  - time in “days since 1990-01-01 00:00:00”
- Standard Names for variables
  - Concatenation of multiple facets
  - sea_surface_height_above_sea_level
  - tropopause_air_temperature
  - tropopause_downwelling_longwave_flux
  - Consult definitions at http://cf-pcmdi.llnl.gov/.

Some Best Practices

- Easy data access: drill into data files
  - OpenDAP URLs
  - Parameter subsetting and array slicing for free
- Auto-data understanding
  - CF-1.0 compliant attributes in data file
  - Or similar attributes in HDFEOS file
- Advertise your web services & apps. for discovery
  - Service Casting (published Atom feed)
- Standardized data discovery & granule space/time query
  - Federated OpenSearch
- Importance of permanent names for data objects
  - When granule produced, provider assigns permanent URL
- Social collaboration on distributed analysis workflows
  - Informally on wiki
  - Talkoot: workflows shared & modified within collaboration env.

Service Casting

- Advertise bundle of Web Services via Atom Feed
  - Service Casts (scast) from many providers
  - With links pointing to callable interface (WSDL) & docs.
  - Service provider pushes new ads when services change, or periodically
- Simple Aggregation & Display
  - All feed readers aggregate multiple feeds & import OPML
  - Publish bundled ads on Federation web site as OPML file
  - Display feeds in HTML or custom feed reader
- Discovery
  - Can search for feeds in Google FeedReader
  - Provide ‘opensearch’ interface that returns scasts
- Machine Readable => Auto Service Invocation
  - SOAP/REST service interfaces described by WSDL
  - Or REST interfaces described by WADL
  - Or known interfaces for OGC WMS/WCS/WFS

Service Casting Features

- Reduce advertising services to publishing an “scast” (Atom) feed
  - Feed readers already aggregate & filter feeds
  - Search also available
- Decentralized repository with auto-aggregation
  - Service providers control publishing of advertisements
  - Scast discoverable in the cloud
  - Publish scast to Federation list by creating wiki link
  - Script auto-aggregates to list of scasts (OPML file)
- Scast contains machine-readable metadata to categorize services and enable auto-invocation
  - Links to interface, service endpoint, human documentation
- So simple it will be adopted (unlike previous attempts).
  - Make free tools available
  - Evangelize, see http://sciflo.jpl.nasa.gov/scast

Feed Reader View

Firefox's Feed Reader:

Subscribe to this feed using Live Bookmarks

☐ Always use Live Bookmarks to subscribe to feeds

Subscribe Now

SciFlo Data Query/Access Services

On SciFlo node sciflo.jpl.nasa.gov

GeoRegionQuery

Space/time query and granule URL lookup services for multiple EOS satellite datasets at Levels L2/L3: AIRS, MODIS, MISR, GPS, and AERONET (ground network).

FindDataById

Translate a list of SciFlo data granule IDs to a list of on-line URLs.

For SOAP services, default links point to WSDL interface.

Browsing Service Casts

In Specialized Feed Reader / Aggregator

Signed in as bdwilson | logout

Feeds

Add Feed | Remove

My Feeds

- **SciFlo Data Query/Access Services (df3)**
- RGIS Base Imagery Service
- ADaM Services: Data mining and image process
- QuikSCAT Hurricane Ultra High-Res Winds
- GHRSST AMSR-E SST
- ADaM Mining Web Services: Optimization Algorit
- ADaM Mining Web Services: Image Processing
- ADaM Services: Data mining and image process
- HD - NASA's Jet Propulsion Laboratory
- NSIDC Atlas of the Cryosphere
- RGIS Base Imagery Service
- NSIDC Snow and Ice Cover
- ADaM Mining Web Services: Utility Services
- ADaM Mining Web Services: Texture Algorithms
- ADaM Mining Web Services: Pattern Recognitior
- ADaM Mining Web Services: Domain Algorithms
- OOSTethys/MMI (Marine Metadata Interoperabilit
- Freeze/Thaw Earth System Data Record Data Vi

SciFlo Data Query/Acc

Open All | Reading Pane | Summary

Title	Date
GeoRegionQuery Brian Wilson Space/time query and granule URL lookup services for multiple EOS satellite datasets at Levels L2/L3: AIRS, MODIS, MISR, GPS, and AERONET (ground network).	3/13 1:32 am
FindDataById Brian Wilson Translate a list of SciFlo data granule IDs to a list of on-line URLs.	3/13 1:32 am

View in New Tab | Go to Post

GeoRegionQuery
by Brian Wilson

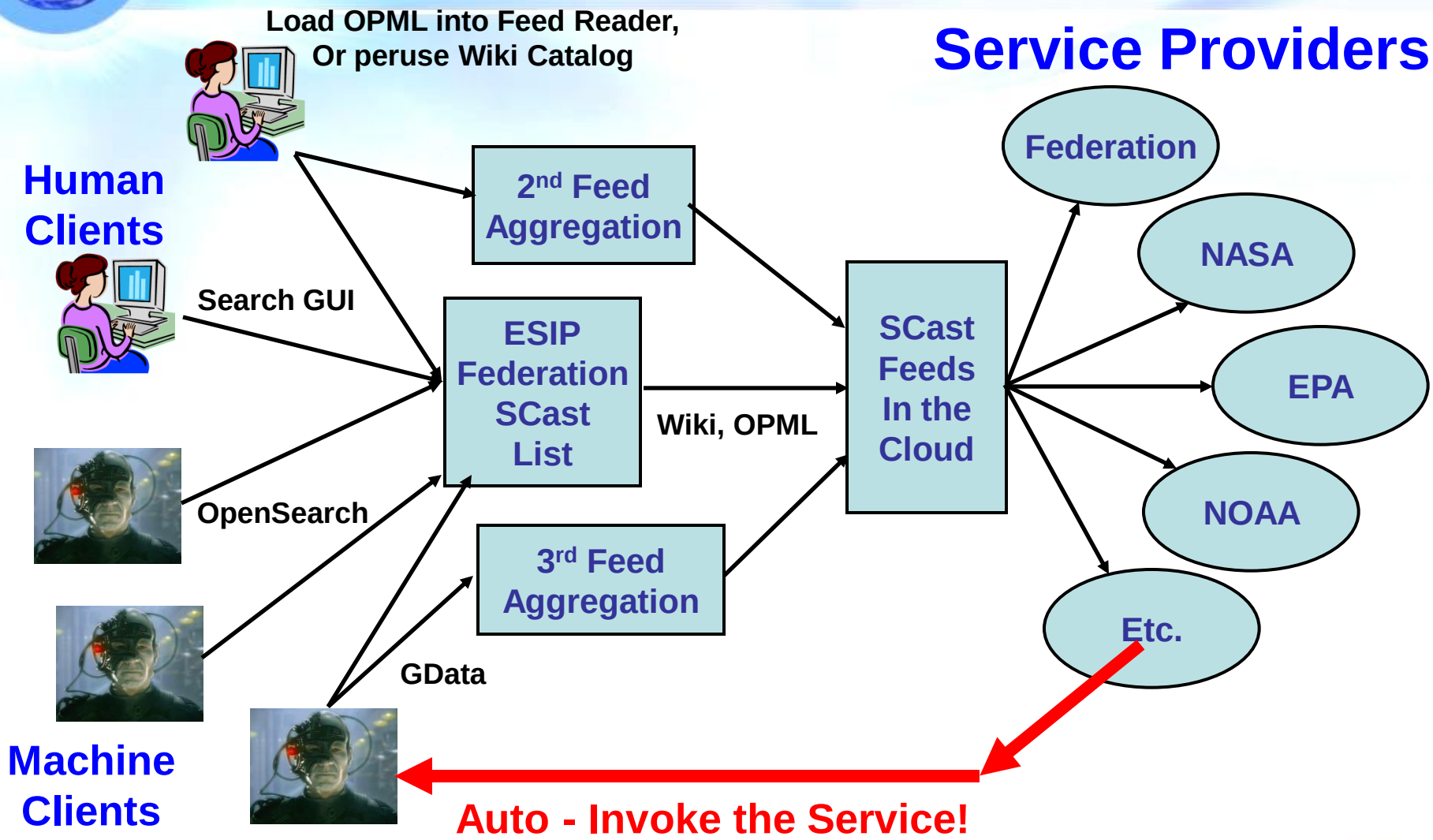
Mar 13, 2008, 1:32 am

Example Call

```
GeoRegionQuery(dataSetId='AIRS', level='L2', version=None, startTime='2006/01/01T00:00:00', endTime='2006/02
/01T00:00:00', latMin=-90., latMax=90., lonMin=-180., lonMax=180., responseGroups='Large' )
```

Service interface description | Server endpoint | Service documentation | SCast Source

SCast Distributed Architecture



Categorizing Service Semantics

- Service Semantics
  - OGC.WXS: WMS, WCS, WFS, etc.
  - OpenSearch, OpenDAP
  - AdHoc, Simple (for SOAP or REST services)
  - Human (e.g. Web apps & AJAX GUI's for humans)
  - [Your service type here]
- Service Protocols (syntax)
  - Tells machine how to interpret links
  - SOAP / WSDL - parse WSDL interface and call service
  - REST / WADL
  - For OGC, do GetCapabilities call
  - HTTP, HTTPPOST
  - AJAX, JavaWebStart – for Human GUIs
- How do we further type SOAP/REST semantics?
  - Use <category> to further specify semantics of call
  - What services are well known besides OGC.WXS?

Some Best Practices

- Easy data access: drill into data files
  - OpenDAP URLs
  - Parameter subsetting and array slicing for free
- Auto-data understanding
  - CF-1.0 compliant attributes in data file
  - Or similar attributes in HDFEOS file
- Advertise your web services & apps. for discovery
  - Service Casting (published Atom feed)
- Standardized data discovery & granule space/time query
  - Federated OpenSearch
- Importance of permanent names for data objects
  - When granule produced, provider assigns permanent URL
- Social collaboration on distributed analysis workflows
  - Informally on wiki
  - Talkoot: workflows shared & modified within collaboration env.

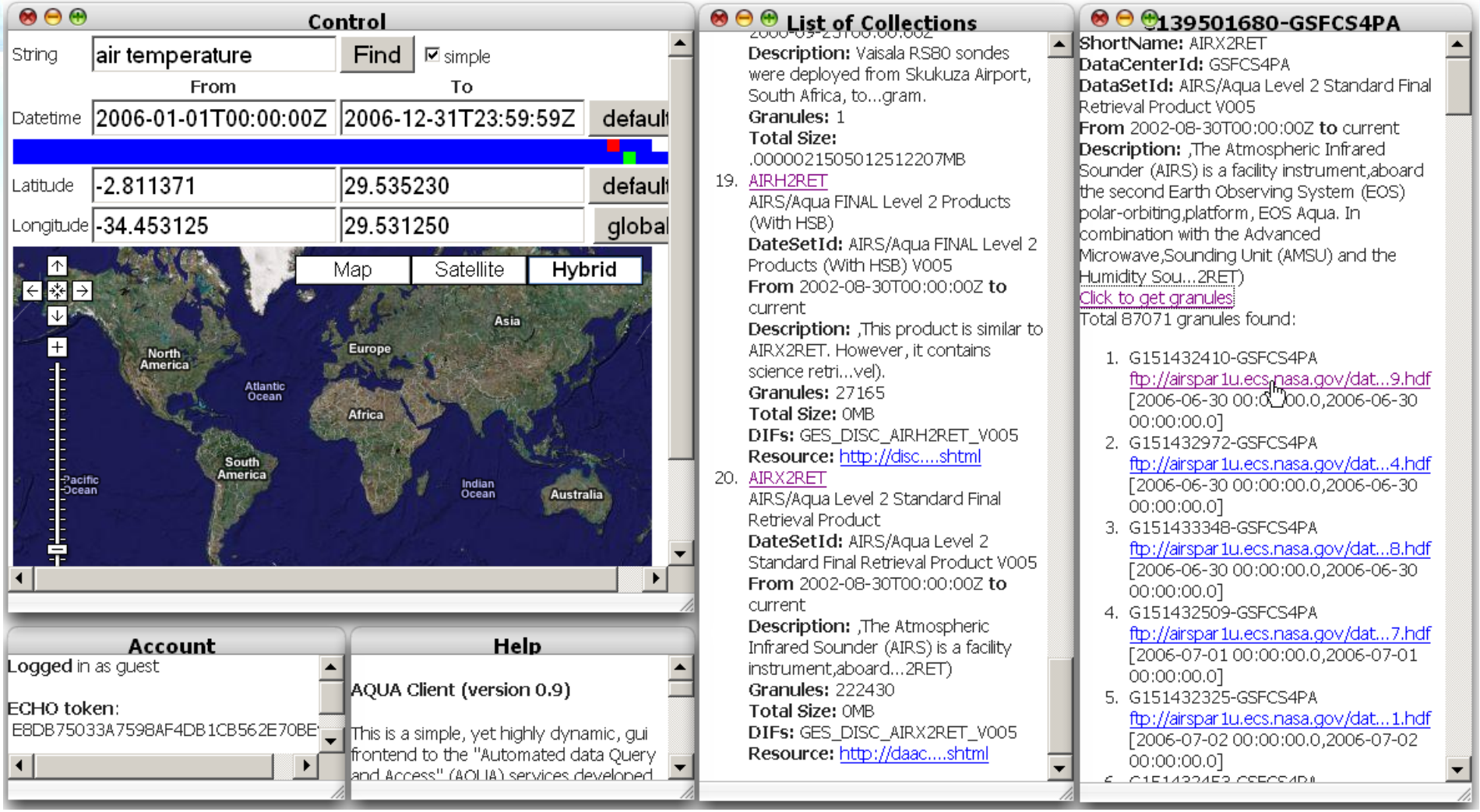
Approaches to Space/Time Query

- SciFlo: Crawl & Extract bounding boxes (polygons)
  - Either crawl metadata files or extract bbox from data
  - GeoRegionQuery services for AIRS, MODIS, MISR, etc.
  - Allows custom metadata for event detection & science analysis
- Query via ECHO metadata repository (AQUA Client)
  - Keyword search to find relevant datasets (collections)
  - Then space/time granule query
- Federated OpenSearch
  - Data Centers provide space/time granule query in-house
  - Everyone exposes services via opensearch protocol
  - Granule query for a specific dataset is re-directed to the proper service via distributed (federated) search
- Massively Scale Using Google "Crawl" Approach
  - Providers publish space/time metadata in uniform format (KML)
  - Indexers crawl & aggregate to support search
  - Who indexes? Google or NASA.

Specifying Datasets in SciFlo

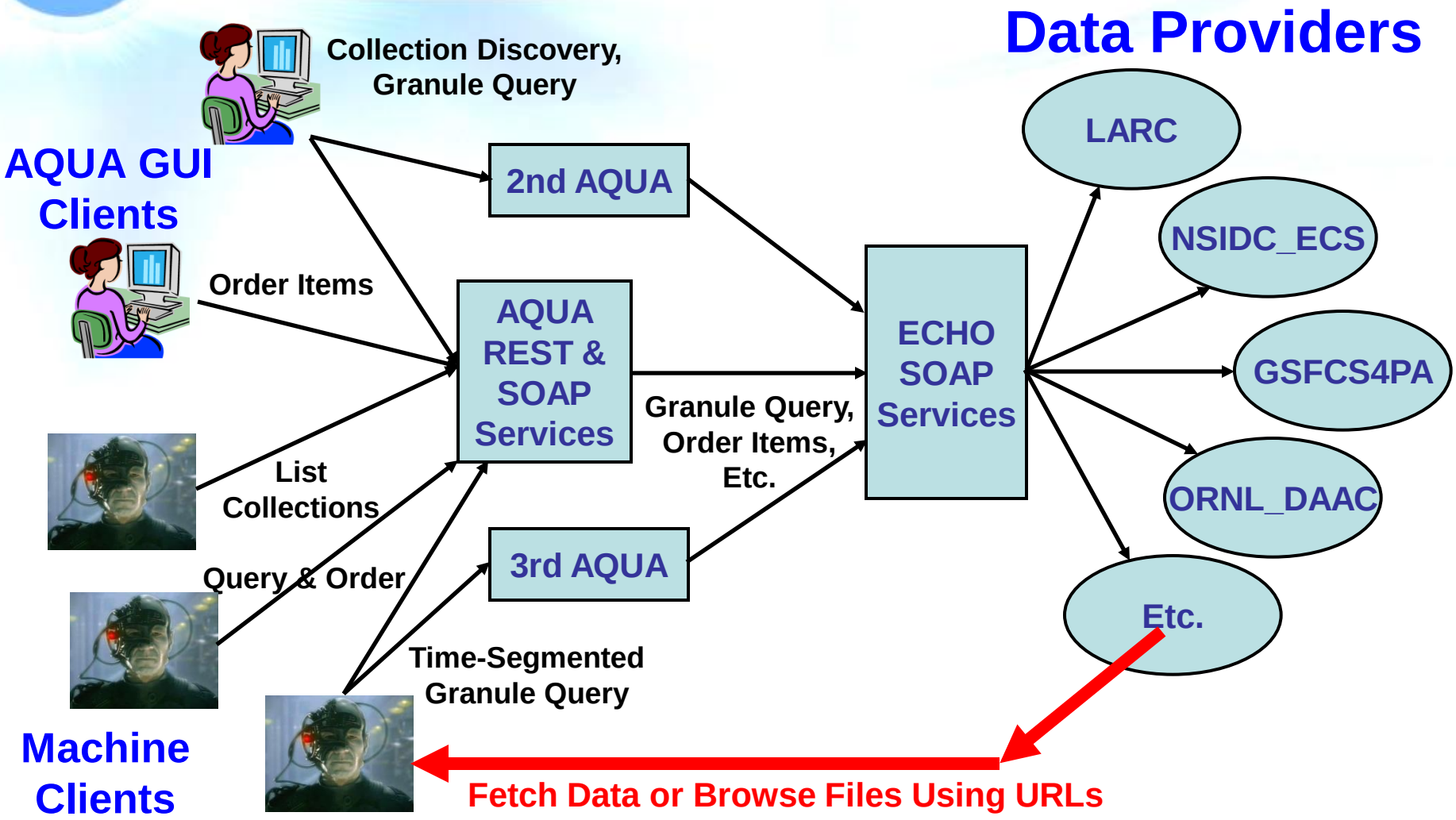
- A SciFlo Dataset is:
  - *Specified* as a space/time query over collections of data products (or retrieved physical variables)
    - **GeoRegionQuery**(DataProduct, TimeRange, LatLonRegion)
    - **GeoRegionQuery**(PhysicalVariable, TimeRange, LatLonRegion)
  - *Realized* as a list of object ID's or URI's (permanent names)
    - GeoRegionQuery returns unique objectIds along with geolocation metadata
  - *Accessed* using a list of URL's pointing to on-line replicas of the data objects (files).
    - FindDataById(objectIds) → URLs (ftp, http, or OpenDAP)
    - Translate unique object ID's into list of on-line locations in DataPools or any SciFlo node
    - DataPools & SciFlo P2P network are "crawled" to update distributed translation tables
    - No need to publish presence of data, continuously discovered
    - SciFlo network is a distributed cache for scientific datasets

AQUA Client GUI (example)



Sequence: Keyword search on "air temperature", Select AIRS L2 dataset AIRX2RET (#20) from list of 278 Collections, Granule search over Africa for year 2006 yields 87,000 hits with URL's.

AQUA Client Architecture



OpenSearch Protocol

- Everything is search, all results are feeds (or KML).
- REST search interface
  - http://your.music.com/-/albums?q=lady+madonna &startIndex=1&count=200&format=atom
  - Returns Atom feed listing first 200 albums that satisfy query keywords
  - Metadata embedded in XML feed entries
- Protocol
  - Feed header contains links to description doc. for server
  - Also contains links to previous, next, & last batch of results
  - Opensearch aggregators can display results in HTML (using stylesheet) and auto-traverse result set
  - Google Data extends opensearch with more conventions, used pervasively in GApps and GFeedReader

AQUA Open Search Interface

- Discover a Collection
  - http:/server/aqua//collections?q=water+vapor&startIndex=1-&count=200&format=atom
  - Returns Atom feed listing collections satisfying query keywords
  - Metadata included in XML feed
- Space/Time Query for Granules
  - http://server/aqua/-/granules/providerId/datasetShortName?start=2006-01-01T00:00:00&end=2006-02-01T00:00:00&bbox=south,west,north,east&responseGroups=Large&startIndex=1&count=200&format=kml
  - Returns Atom feed (or KML document!)
  - Granule metadata (time, georss:box) and URL's included
- OpenSearch (GData) Features
  - GoogleData standard uses and extends OpenSearch
  - Search aggregators auto-handle Atom feed, traverse result sets
  - Ingest data feed into Google Earth or Google Apps

AQUA Space/Time Query Result


```
<feed xmlns="http://www.w3.org/2005/Atom"
      xmlns:georss="http://www.w3.org/2003/01/geo/wgs84_pos#"
      xmlns:opensearch="http://a9.com/-/spec/opensearch/1.1/">
  <title>OpenSearch space/time granule query results for provider GSFCS4
  <updated>2008-07-08T18:16:30.508269</updated>
  <author><name>AQUA ECHO Client</name></author>
  <id>uri:http://sciflo.jpl.nasa.gov/aqua/-/granules/GSFCS4PA/AIRX2RET?t
  <opensearch:totalResults>510</opensearch:totalResults>
  <opensearch:startIndex>1</opensearch:startIndex>
  <opensearch:itemsPerPage>200</opensearch:itemsPerPage>
  <opensearch:Query role="request" searchTerms="??" startIndex="1"/>
  <fubar:time>2006-01-01T00:00:00,2006-02-01T00:00:00</fubar:time>
  <georss:box>0.,0.,60.,179.</georss:box>
  <link href="http://sciflo.jpl.nasa.gov/aqua/-/granules/GSFCS4PA/AIRX2R
  <link rel="self" type="application/atom+xml" href="http://sciflo.jpl.n
  <link rel="alternate" type="application/vnd.google-earth.kml+xml" href
  <link rel="alternate" type="text/html" href="http://sciflo.jpl.nasa.go
  <link rel="previous" type="application/atom+xml" href="http://sciflo.j
  <link rel="next" type="application/atom+xml" href="http://sciflo.jpl.n
  <link rel="first" type="application/atom+xml" href="http://sciflo.jpl.
  <link rel="last" type="application/atom+xml" href="http://sciflo.jpl.n
  <link rel="search" type="application/opensearchdescription+xml" href="
  <entry>
```

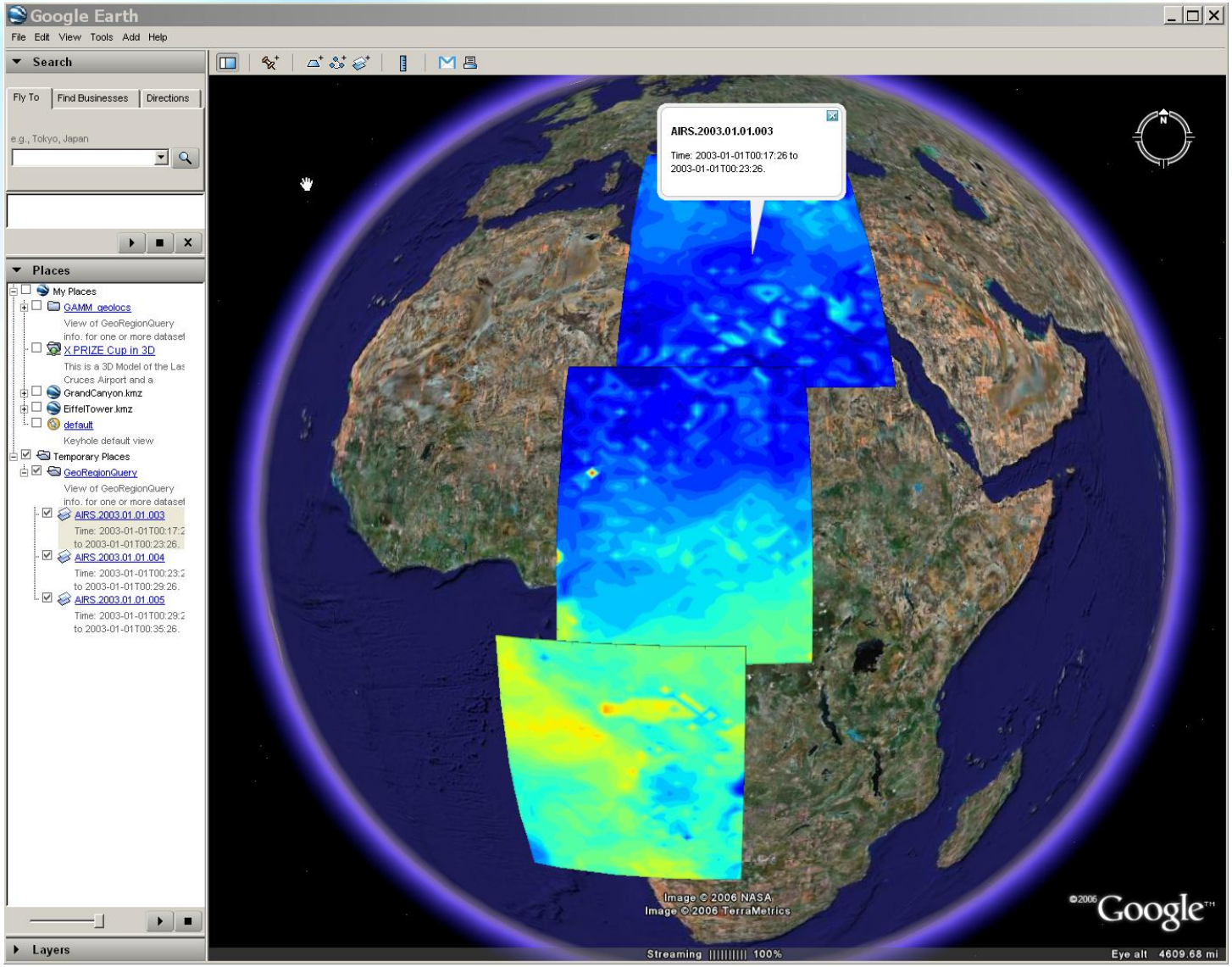
AQUA Space/Time Query Result (2)

```
<entry>
  <title>SC:MYD10_L2.004:9936826</title>
  <link href="urn:SC:MYD10_L2.004:9936826"/>
  <link rel="data" href="urn:SC:MYD10_L2.004:9936826"/>
  <link rel="dap" href="http://ecs.nasa.gov/dap/modis/-"/>
  <link rel="browse" href="http://browse.echo.nasa.gov/-
      NSIDC_ECS/2006/01/04/:BR:Browse.001:9936830:1.BINARY"/>
  <id>urn:SC:MYD10_L2.004:9936826</id>
  <updated>2006-03-02T13:04:48.0000</updated>
  <fubar:time>2006-01-01T10:30:00,2006-01-01T10:35:00</fubar:time>
  <georss:box>south west north east</georss:box>
  <content type="xml">
</entry>
</feed>
```

- Granule metadata includes:
  - Time range
  - Bounding box
  - Link to browse image
  - DAP URL pointing to data file (if possible), otherwise ftp

Displaying data on Google Earth

AIRS Total H_2O from three 6-min. granules



The FROST Solution

Federated	Inventories supply search results
Recursive Open-Search	Open-Search dataset results point to open-search description docs for inventory level search
Tools	Off-the-shelf tools are provided to inventories to index and supply search results

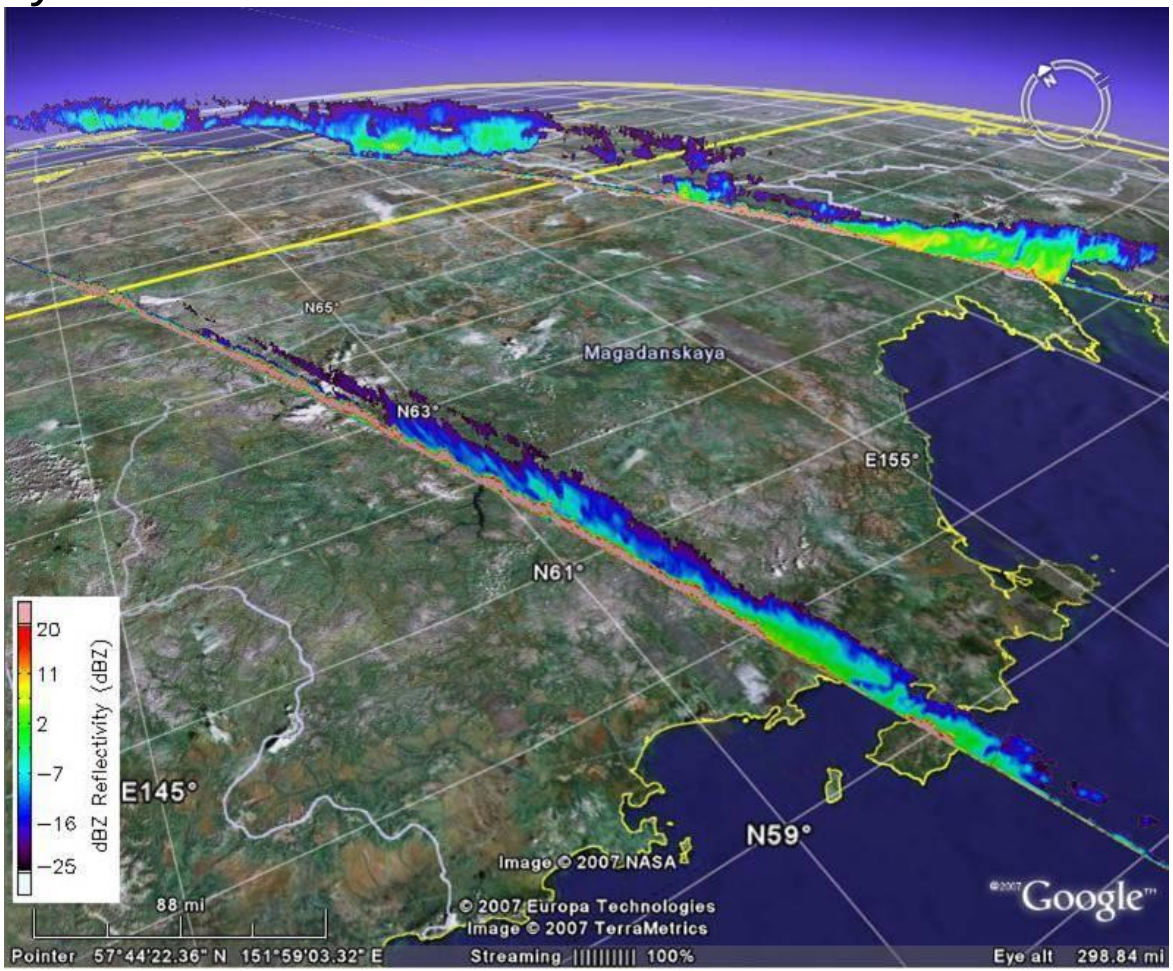
Chris Lynnes, Wed. afternoon

Approaches to Space/Time Query

- SciFlo: Crawl & Extract bounding boxes (polygons)
  - Either crawl metadata files or extract bbox from data
  - GeoRegionQuery services for AIRS, MODIS, MISR, etc.
  - Allows custom metadata for event detection & science analysis
- Query via ECHO (AQUA Client)
  - Keyword search to find relevant datasets (collections)
  - Then space/time granule query
- Federated OpenSearch
  - Data Centers provide space/time granule query in-house
  - Everyone exposes services via opensearch protocol
  - Granule query for a specific dataset is re-directed to the proper service via distributed (federated) search
- Massively Scale Using Google "Crawl" Approach
  - Providers publish space/time metadata in uniform format (KML)
  - Indexers crawl & aggregate to support search
  - Who indexes? Google or NASA.

Massively Scaling Data Discovery

- **Publish → Index → Query → Perm. Names → URL's → Data Slice → Visualization/Analysis**
  - Transfer of data from tape to disk could be hidden inside an asynchronous Name Resolution service



Global-Scale Data Discovery

- Publish metadata in KML or standard (ECS) format
  - Metadata file published alongside data file
  - Granule metadata includes space/time box, domain & dataset keywords or ontology terms, variable list, etc.
  - Permanent name for data object
  - New metadata advertised via datacast (RSS feed)
- Crawl and Index
  - Google or someone crawls and indexes metadata
  - Search by time, space, metadata fields, keywords
  - Also ontology-enhanced search: term broadening, etc.
  - Results are lists of permanent names (URI, XRI)
- Delegated Name Resolution
  - Resolution delegated to proper naming authority
  - Translate each URI/XRI to one or more URLs (preferably DAP).
- Data Access
  - Browse & viz. using KML
  - Slice data using DAP URLs

Publish KML Browse Image

- Dual Purpose
  - Browse image of key variable representing data granule
  - Specifies space/time bounding box for granule index & search

```
<kml xmlns="http://www.opengis.net/kml/2.2">
  <GroundOverlay>
    <name>AIRS.2006.01.12.001.L2.RetStd.hdf</name>
    <description>Total water vapor browse image for AIRS Level-2
      Standard Retrieval granule . . .</description>
    <Icon>
      <href>http://aurapar2u.ecs.nasa.gov/Aqua_AIRS_Level2/-
AIRX2RET.005/2006/012/AIRS.2006.01.12.001.L2.RetStd.v5.0.14.0.hdf.png</href>
    </Icon>
    <LatLonBox>
      <north>37.9</north>
      <south>30.4</south>
      <east>18.3</east>
      <west>12.6</west>
    </LatLonBox>
    <TimeSpan>
      <begin>2006-01-12T00:00:00</begin>
      <end>2006-01-12T06:00:00</end>
    </Timespan>
    <link rel="metadata"></link>
  </GroundOverlay>
</kml>
```

KML / GEarth Platform

- Scalable publishing & commodity visualization
  - The innovation is this powerful combination.
  - Tell a science story with an accompanying visualization
  - Google crawls and indexes KML files
  - GEarth provides spatial Placemark search, and visual animations over time
- Embed (or link to) metadata in KML (or Atom) files
  - Georss tags based on GML, different from KML spatial
  - Micro XML formats: georss:bbox (spatial bounding box)
- GEarth provides Placemark (KML) spatial search
  - Viewport defines implicit lat/lon rectangle
  - Then do keyword search into indexed KML files
  - Google plans to add “time span” search
- Opportunity
  - Exploit scalable publishing & search
  - Link to more XML or RDF metadata

Some Best Practices

- Easy data access: drill into data files
  - OpenDAP URLs
  - Parameter subsetting and array slicing for free
- Auto-data understanding
  - CF-1.0 compliant attributes in data file
  - Or similar attributes in HDFEOS file
- Advertise your web services & apps. for discovery
  - Service Casting (published Atom feed)
- Standardized data discovery & granule space/time query
  - Federated OpenSearch
- Importance of permanent names for data objects
  - When granule produced, provider assigns permanent URL
- Social collaboration on distributed analysis workflows
  - Informally on wiki
  - Talkoot: workflows shared & modified within collaboration env.

Alogrithmic Naming

- Transparent, Content-Full Names
  - Not opaque identifiers, like DOI, LSID, UUID
- Construct name using deterministic algorithm
  - URN = us:gov:nasa:eos:airs:AIRS.2003.01.02.004.L2.RetStd
  - Reversible: can extract metadata from the name
- Using URLs as Permanent Names
  - http://airs.jpl.nasa.gov/data/L2/2003/01/AIRS.2003.01.02.004.L2.RetStd.hdf
  - URL never changes: support using multi-homing hosts and symbolic links
  - Each project designs filenames and **directory** structure
- Crawler Not Necessary
  - To locate a data file, construct its permanent URL using the algorithm and see if it exists

Using URLs as Permanent Names

- Delegated Naming Authority
  - The Real Estate grab has already occurred. Use it.
  - http://airs.eos.nasa.gov/data/L2/RetStd/2003/01/AIRS.2003.01.02.004.L2.RetStd.hdf
  - URL never changes: support by using multi-homing hosts, Unix symbolic links, etc.
- Cultural Problem
  - URL's aren't expected to be permanent, but they can be.
  - Data providers need to think permanently!!
  - Each project designs "algorithmic" filenames, dataset directory structure, permanent URLs
- Crawler & Name Resolver Not Necessary
  - To locate a data file, construct its permanent URL using the algorithm and see if it exists.

Some Best Practices

- Easy data access: drill into data files
  - OpenDAP URLs
  - Parameter subsetting and array slicing for free
- Auto-data understanding
  - CF-1.0 compliant attributes in data file
  - Or similar attributes in HDFEOS file
- Advertise your web services & apps. for discovery
  - Service Casting (published Atom feed)
- Standardized data discovery & granule space/time query
  - Federated OpenSearch
- Social collaboration on distributed analysis workflows
  - Informally on wiki
  - Talkoot: workflows shared & modified within collaboration env.

Talkoot*: Discover, Tag, Share, and Reuse Collaborative Science Workflows

*barn raising in Finnish

Rahul Ramachandran, Sunil Movva, Helen Conover

Univ. of Alabama Huntsville

Chris Lynnes, *NASA/GSFC*

Brian Wilson, *NASA/JPL*

rramachandran@itsc.uah.edu

2009 ESIP Federation Summer Meeting

Santa Barbara, CA

Enter Talkoot!

- We are currently developing “Talkoot”
- Talkoot is a customizable “software appliance” to build collaborative portals for Earth Science services and analysis workflows.
- Talkoot will allow researchers (not just information technologists) be able to build collaborative sites around service workflows within a few hours
- Talkoot is leveraging ***Drupal, an open architecture platform*** to provide the core Content Management System capabilities required by an online collaborative portal
  - Drupal also has a vast array of specialized modules that have been developed by its user community to provide additional features
- Talkoot is adding Earth science specific modules to provide data searching, processing and analysis capabilities.

UAHuntsville
THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

Talkoot Portal

Chat Rooms | Contact | Forums | Project Information | Talkoot Developers

Talkoot: Drupal extensions to create Earth science portals

Search

Workflow Tags

Image Processing

Supervised Unsupervised

more tags

User login

Username: *

Password: *

Log in

- Request new password

Navigation

- Data Search

To leverage and expand Drupal's existing CMS capabilities to allow scientists create Science Portals specifically for processing and analyzing data.

Talkoot

Tue, 04/14/2009 - 21:22 — RxR

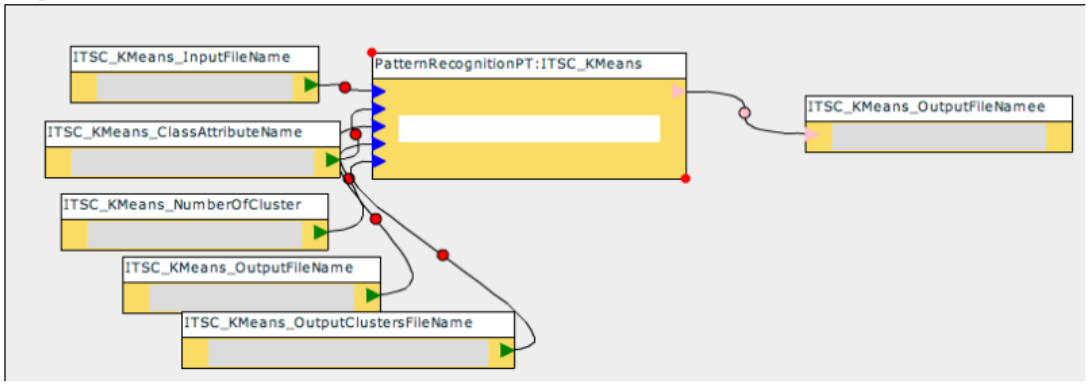
A small but growing number of scientists and researchers are beginning to harness Web 2.0 as a transformative way of doing science. Since communication is at the heart of these technologies provide researchers easy mechanisms to share ideas, data and These technologies complement formal means of sharing knowledge via conferences published papers, where it is impossible to share all the research details, and where negative are rarely included.

Test simple clustering

Mon, 10/12/2009 - 18:46 — RxR

Simple kmeans

Image :



Workflow: *http://miningsolutionsdev.itsc.uah.edu/talkoot/files/userfiles/Workflow_20091012_134632.xwf*

Unsupervised

Who's new

- kkeiser
- demo
- sudeshna
- dbudzak
- rrankin

Who's online

There are currently *1 user* and *0 guests* online.

Online users

- smovva

Active forum topics

- Demo notes
- Talkoot feature requests
- Talkoot forum type added
- RSS feed for upper right
- Talkoot Technologies?

more

Recommendations for Data Providers

- Assign a permanent name (URI) to every data granule
  - Use your owned web real estate (http://daac.nasa.gov/airs/v5)
- Publish space/time granule metadata alongside data
  - Could be KML browse file, or ECHO or ISO19115 format
  - Ensure that metadata contains useful bounding polygons
  - Make high-res. geometry easily accessible (MODIS L2 issue)
- Provide dataset discovery & space/time granule query using the opensearch (REST) protocol
  - Return DAP URL's if possible
  - Performance for large, automated queries is important
  - Join the proposed Federated OpenSearch group
- Serve all L1B, L2, & L3 data via an OpenDAP server
  - Parameter subsetting & space/time slicing for free
- Alternatively, run Web Coverage Server
  - Harder for L1B & L2 data (mosaic stitching)
  - Data Discovery consists of "finding" the WCS servers

Summary

- Loosely-Coupled Web Services (SOA)
  - Expose fine-grained functionality as reusable web services (Publish algorithms!)
  - Use REST or SOAP, or both
  - Use workflow engine to orchestrate service chains: BPEL, SciFlo
  - Formal workflow yields provenance graph for free
- Best Practices for Services
  - Use standard protocols: OGC WMS/WCS, OpenDAP, OpenSearch
  - Modeless REST (HTTP one-liner) services are easiest to use
- More Use Cases / Technologies
  - Compatibility with Metadata Standards: ECS, ECHO, ISO 19115
  - Virtualization: compute load balancing, failover
  - Cloud Computing: re-processing on demand, virtual datasets
  - Services in the Cloud: data mining over years of data
  - Multi-instrument joint retrieval algorithms, instrument simulation
  - Large-scale model validation (e.g. JPL's CDX project)
  - [More Mission Use Cases here.]