Knowledge Provenance

CMSP Vocabulary Workshop

Dec 1-3, 2010 (Woods Hole, MA)

Peter Fox (RPI)
pfox@cs.rpi.edu
Tetherless World Constellation – http://tw.rpi.edu

Provenance

- Origin or source from which something comes, intention for use, who/what generated for, manner of manufacture, history of subsequent owners, sense of place and time of manufacture, production or discovery, documented in detail sufficient to allow reproducibility
- Or … provenance are the types and instances of metadata in a particular multi-faceted context
- *Knowledge provenance*; enriched with semantics (especially the relations between concepts previously isolated, and retaining context) and semantically-aware tools

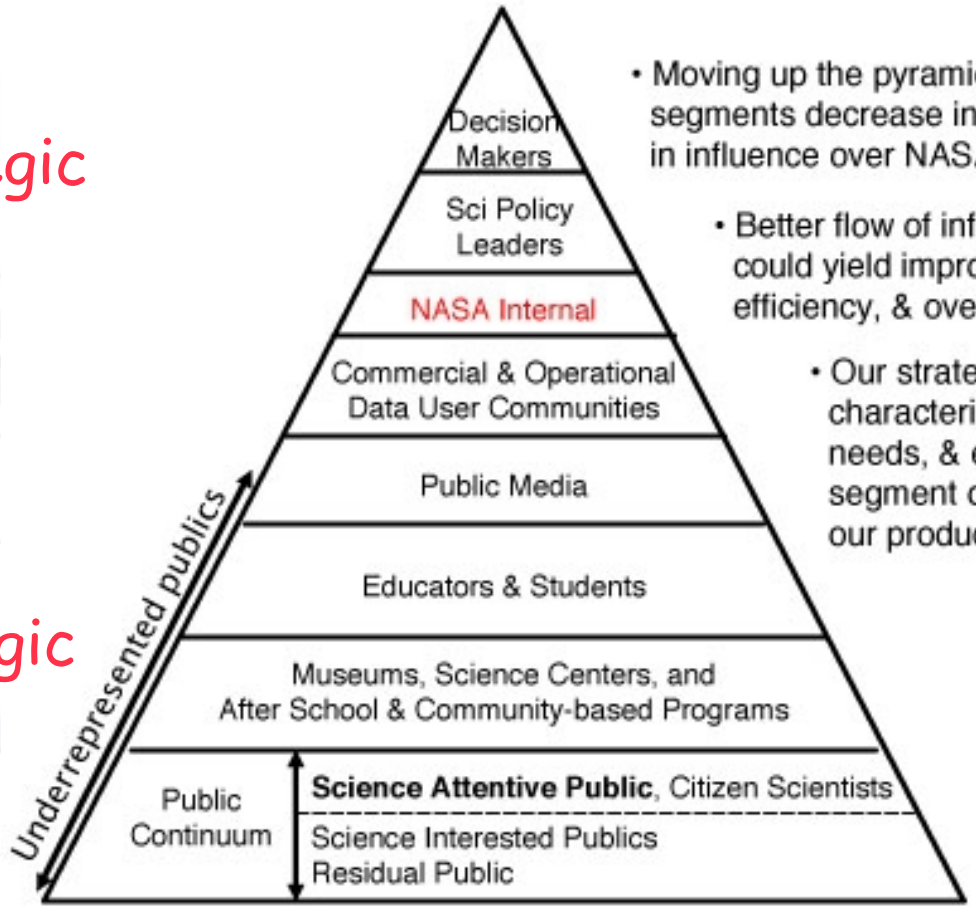
Problem definition

- *Data is coming in faster, in greater volumes and outstripping our ability to perform adequate quality control*
- *Data is being used in new ways and we frequently do not have sufficient information on what happened to the data along the processing stages to determine if it is suitable for a use we did not envision*
- *We often fail to capture, represent and propagate manually generated information that need to go with the data flows*
- *Each time we develop a new instrument, we develop a new data ingest procedure and collect different metadata and organize it differently. It is then hard to use with previous projects*
- *The task of event determination and feature classification is onerous and we don't do it until after we get the data*

Data has Lots of Audiences

More Strategic

Less Strategic



- Moving up the pyramid, these audience segments decrease in size while increasing in influence over NASA's budget & agenda.
- Better flow of information within NASA could yield improved synergy, efficiency, & overall effectiveness.
- Our strategy should be to characterize the information wants, needs, & expectations of each segment of the public, & then tailor our products/programs accordingly.
- The 'science attentive' public considers itself knowledgeable & willing to participate in policy-relevant discussions. Thus, this is a particularly beneficial audience to target.

SCIENTISTS!

From "Why EPO?", a NASA internal report on science education, 2005

Background

Fact: Scientific data services are increasing in usage and scope, and with these increases comes growing need for access to provenance information.

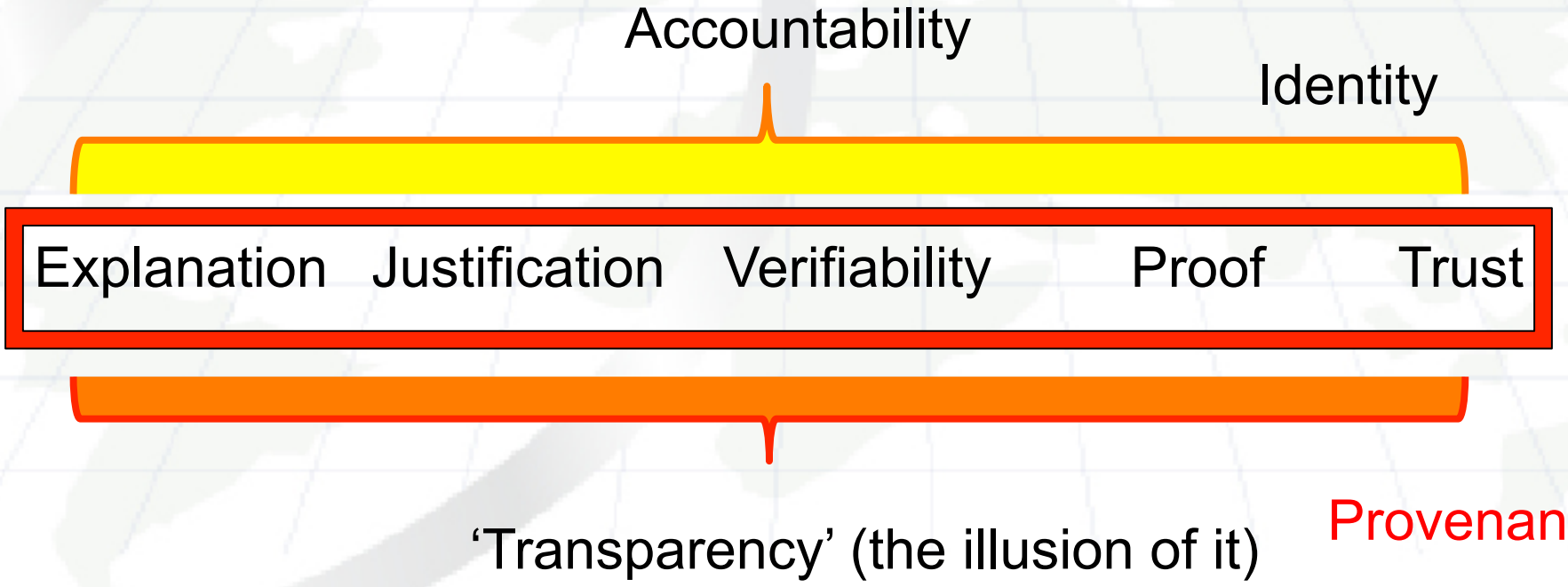
Our Research Goal: design and implement an extensible provenance solution that is deployed at the science data ingest/ product generation time.

Provenance Infrastructure Goal: to design a reusable, interoperable provenance infrastructure.

Outcome: implemented provenance solution in one science setting AND operational specification for other scientific data applications – both achieved.

What is the ecosystem?

- Many elements, and they are scattered
- But these are what enable scientists to explore/ confirm/ deny their ‘hunches’



But back to reality

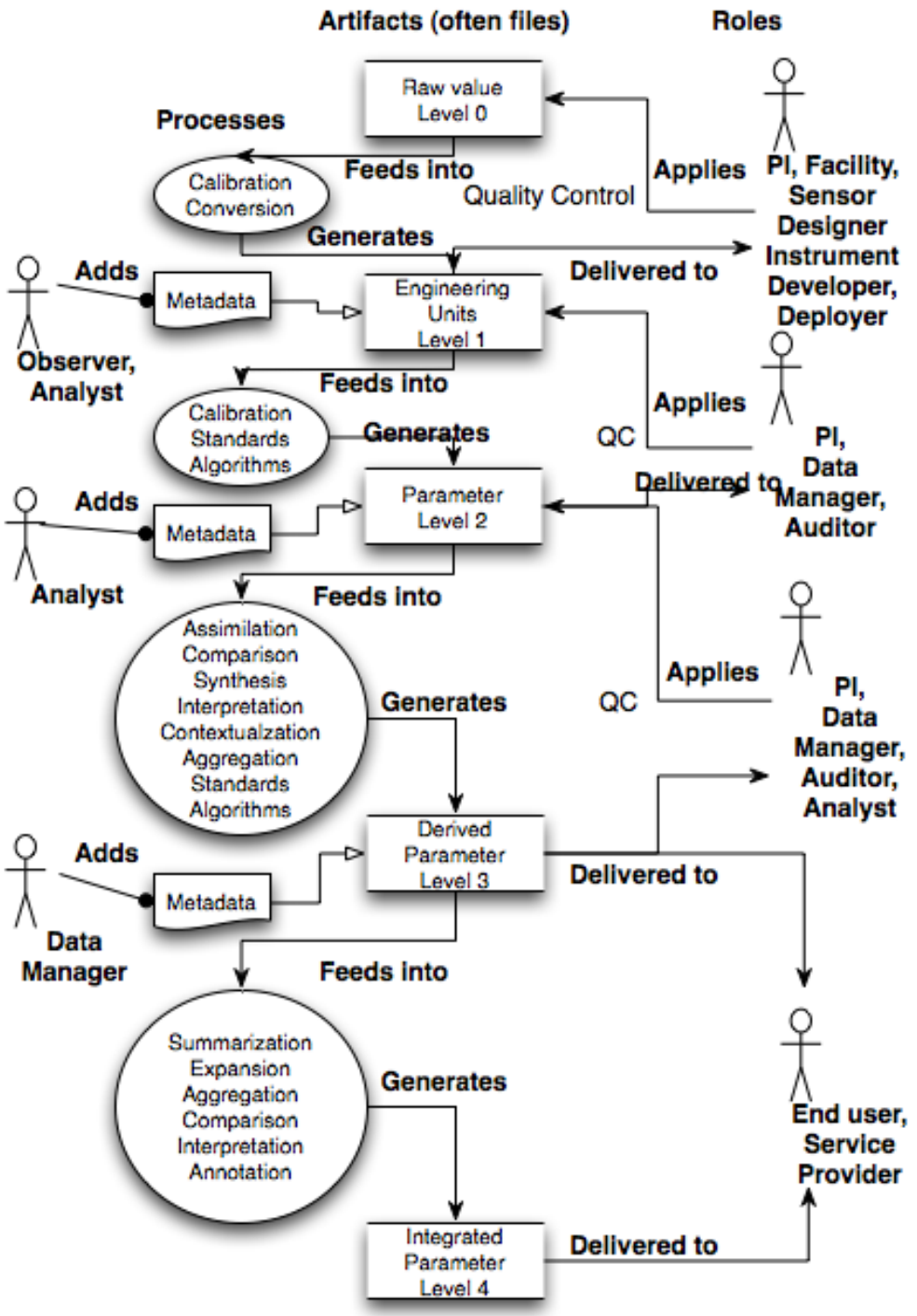
Fragmentation

Disconnection

- Internal/Ext.

Encapsulation

… all are bad for … transparency



MLSO Analysis Center Homepage

Mauna Loa Solar Observatory

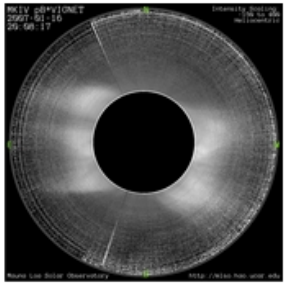
HAO

Welcome to the Mauna Loa Solar Observatory (MLSO) Website. The MLSO, operated by the High Altitude Observatory in Boulder Colorado, houses several instruments designed to observe the sun at many different wavelengths. The MLSO instruments provide observations needed to understand the sun's continuous release of plasma and energy into interplanetary space.

ACOS	Advanced Coronal Observing System. A suite of instruments designed to observe the solar atmosphere at a variety of heights. Includes Chromospheric Helium Imaging Photometer (CHIP, 1083.0nm), H–alpha prominence and solar disk monitor (PICS, 656.2nm), and the Mk4 K–coronameter, which observes the white light K–corona from 1.12–2.79 solar radii.
ECHO	Experiment for Coordinated Helioseismic Observations. A network of two instruments which observe solar oscillations as seen in the radial velocity of the solar surface.
PSPT	Precision Solar Photometric Telescope. Observes the solar disk in three bandpasses: 605–610 nm (red), 408–412 nm (blue), and 393 nm (CaIIK).

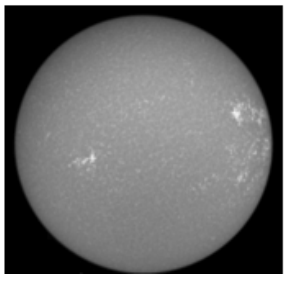
Latest MLSO Images

ACOS Mark-IV

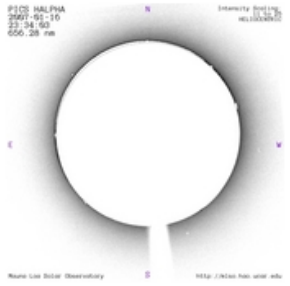


K-Corona
700-950 nm
16-Jan-2007
Watch Movie

PSPT CaIIk

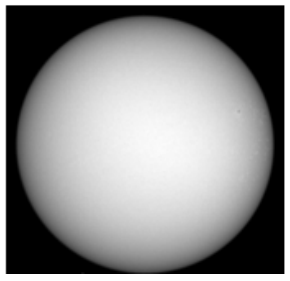


ACOS PICS Limb

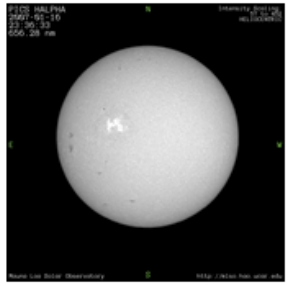


H-Alpha Limb
656.3 nm
16-Jan-2007
Watch Movie

PSPT Blue

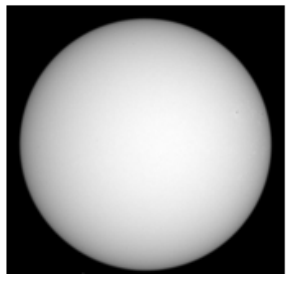


ACOS PICS Disc

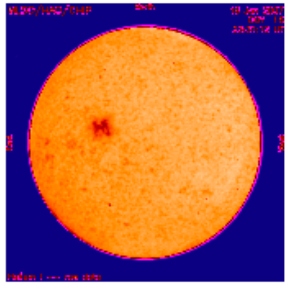


H-Alpha Disk
656.3 nm
16-Jan-2007
Watch Movie

PSPT Red

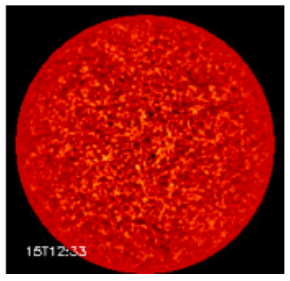


ACOS CHIP

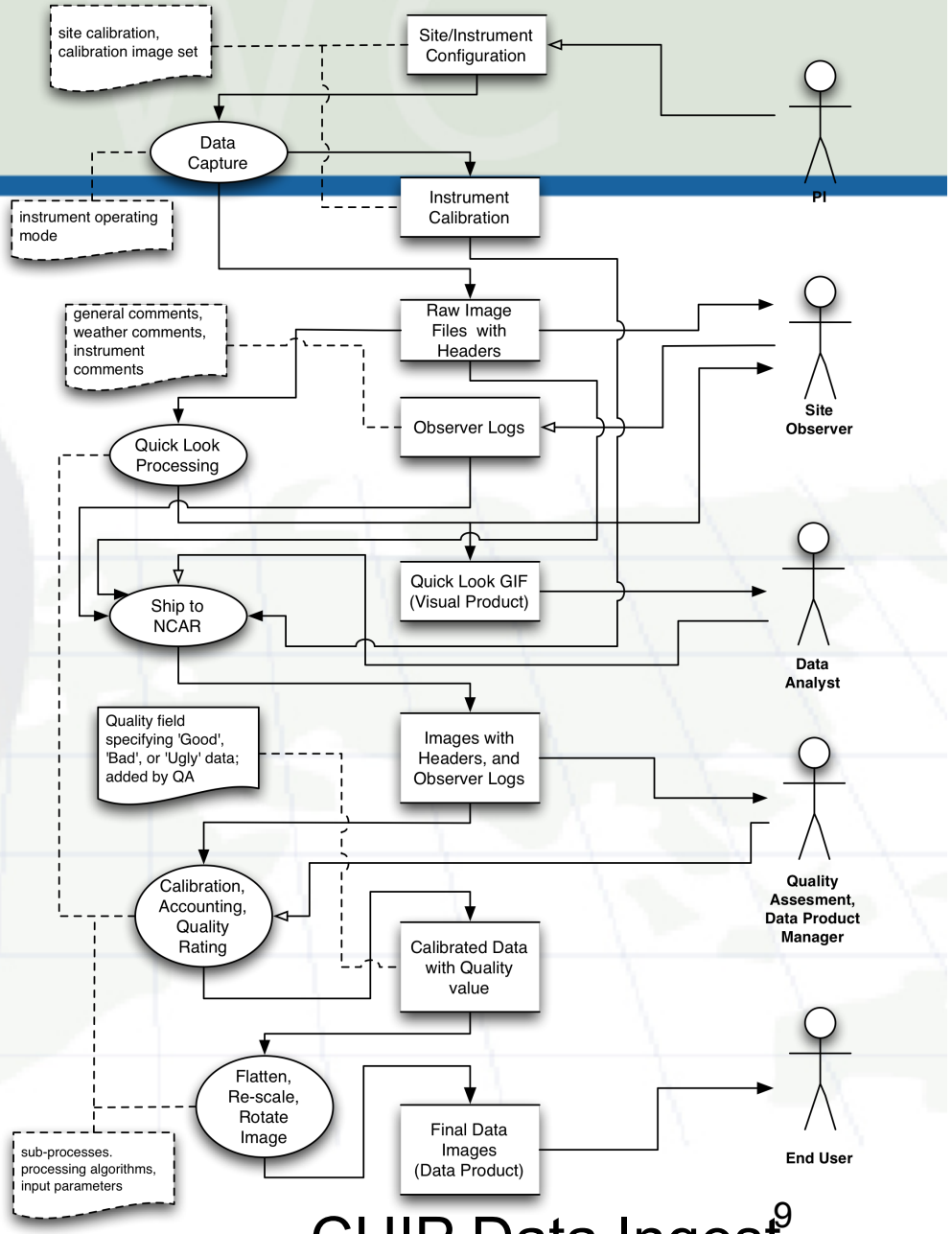


Helium-I
1083 nm
16-Jan-2007
Movie [merged-GIF]

ECHO Sample Velocity Image



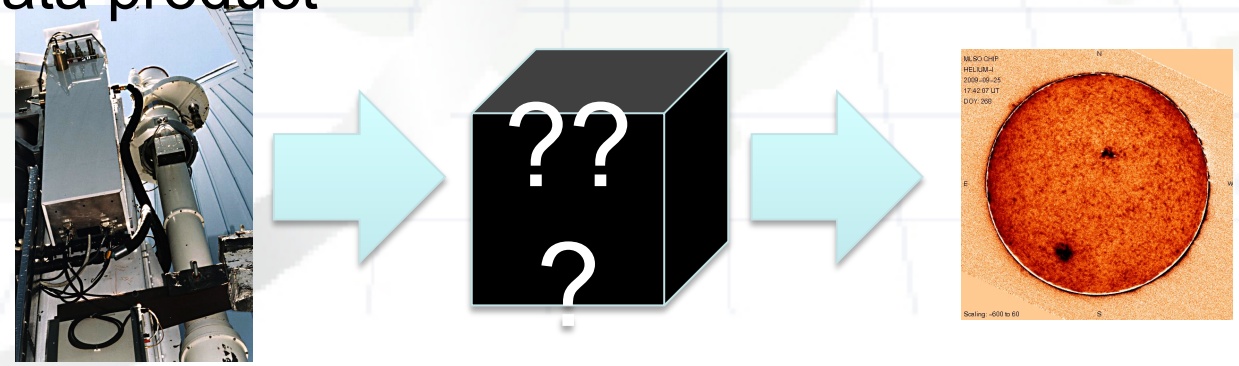
- Typical science data processing pipelines
- Distributed
- Some metadata in silos
- Much metadata lost
- Many human-in-loop decisions, events
- No metadata infrastructure for any user



CHIP Data Ingest

The ACOS case for Provenance

- Provenance metadata currently not propagated with or linked to the data products
  - Processing metadata
  - Origin (observation) metadata
- Data products are the result of “black box” systems
  - Most users do not know what calibrations, transformations, and QA processing have been applied to the data product

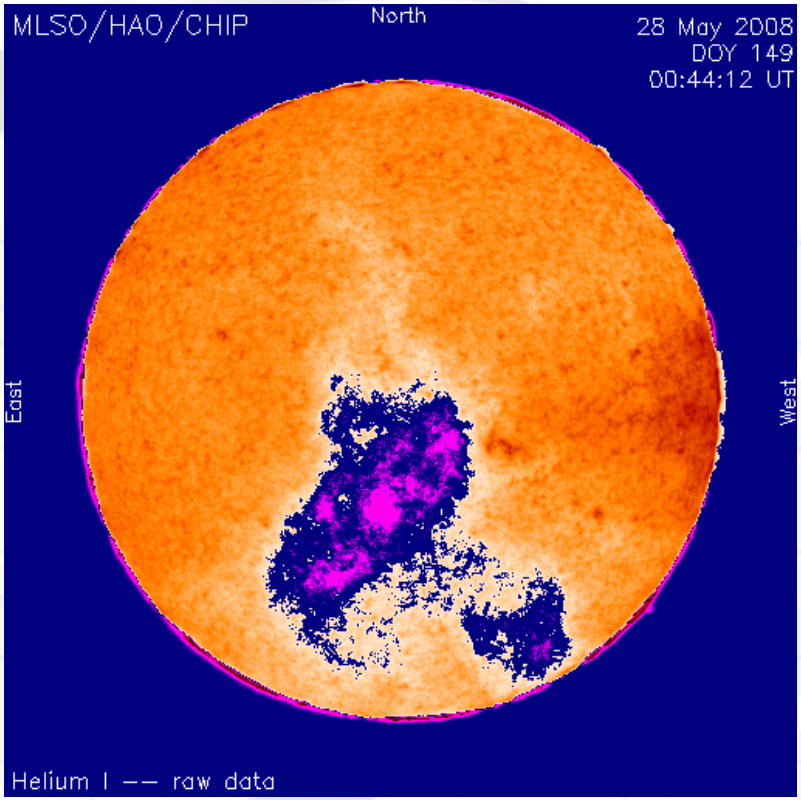


Source Processing Product

ACOS Provenance Use Cases

- What were the cloud cover and seeing conditions during the observation period of this image?
- What calibrations have been applied to this image?
- Why does this image look bad?

Why *does* this image look *bad*?



Observer Log
5/28/2008 00:42:29UT
Observer: Daryl Koon
Weather Comment:
"Low clouds are passing over."

Provenance and Domain concepts in the use cases

solar science

solar science

- What were the *cloud cover* and *seeing conditions*

provenance and solar science

during the observation period of this image?

data processing

provenance and data processing

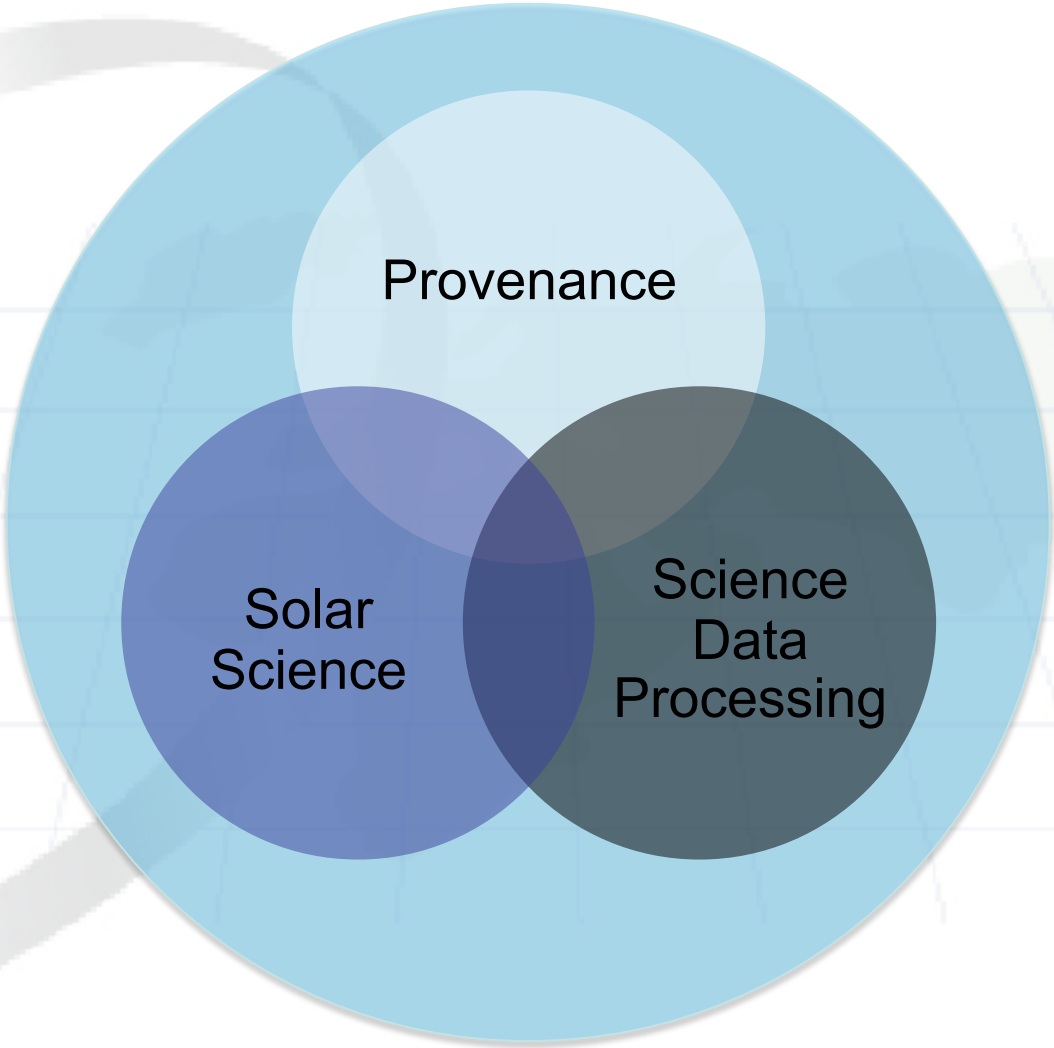
- What *calibrations* have *been applied to this image*?

Provenance, data processing (QA), and solar science

data processing

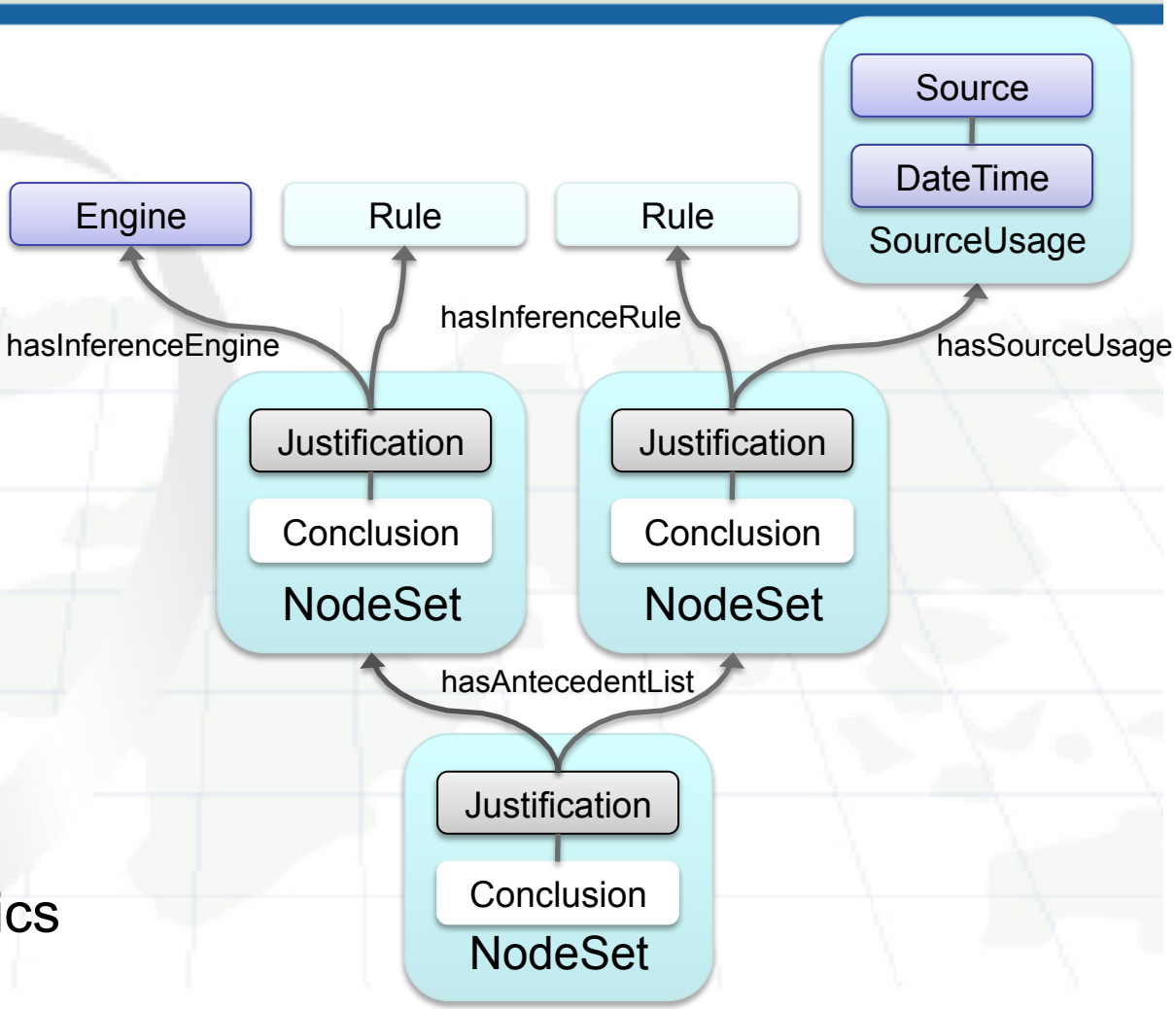
- *Why does this image* look *bad*?

Multi-domain Knowledge Base

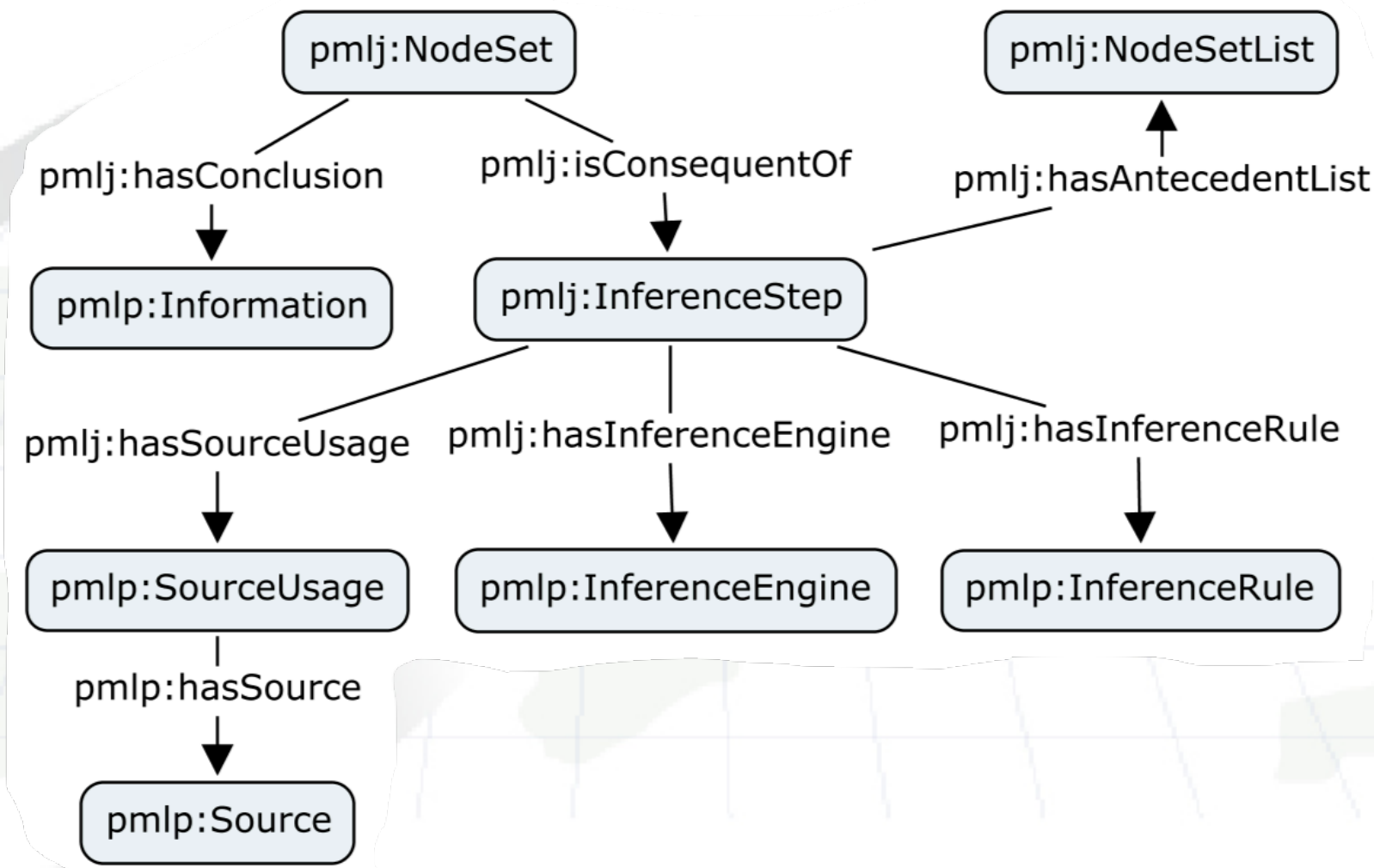


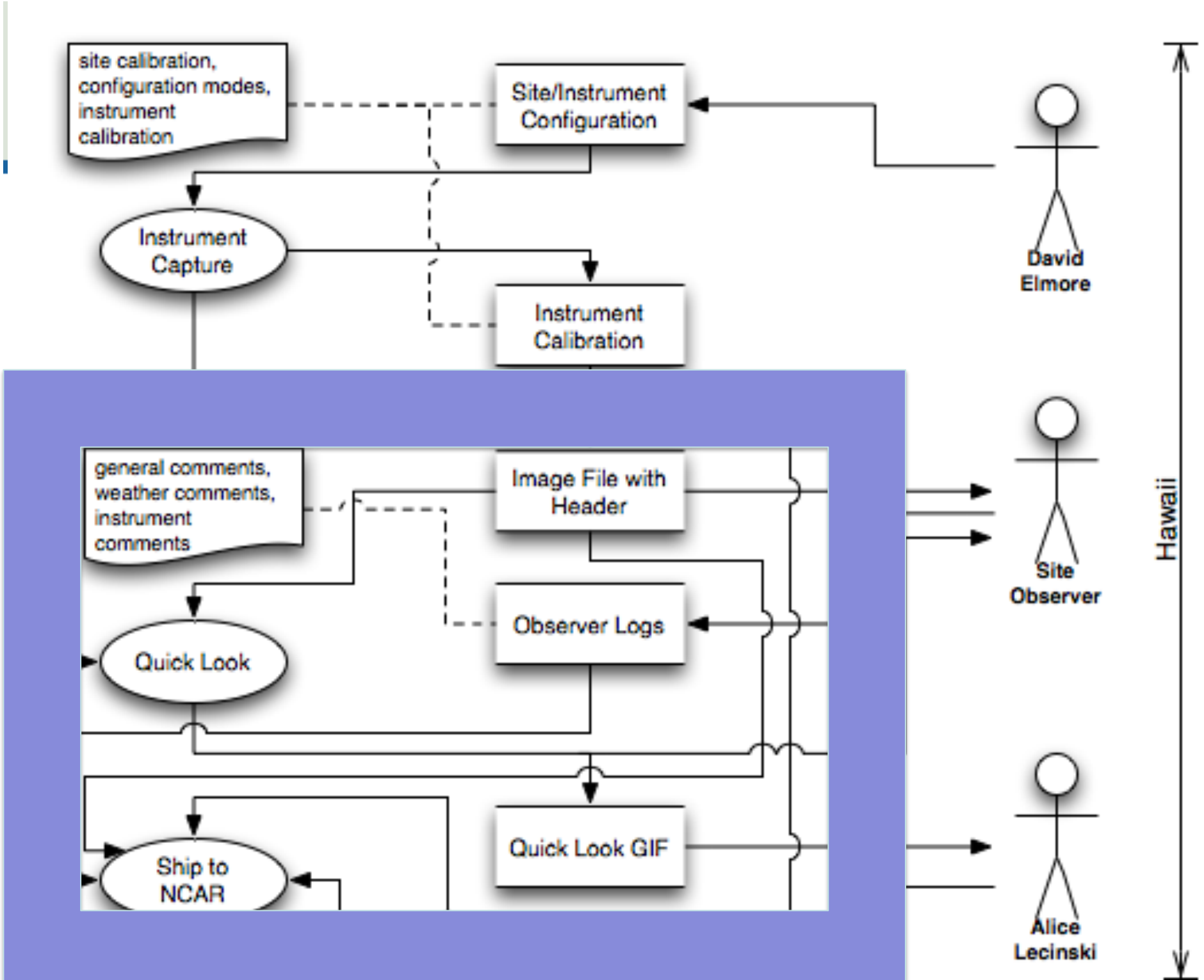
Proof Markup Language (PML)

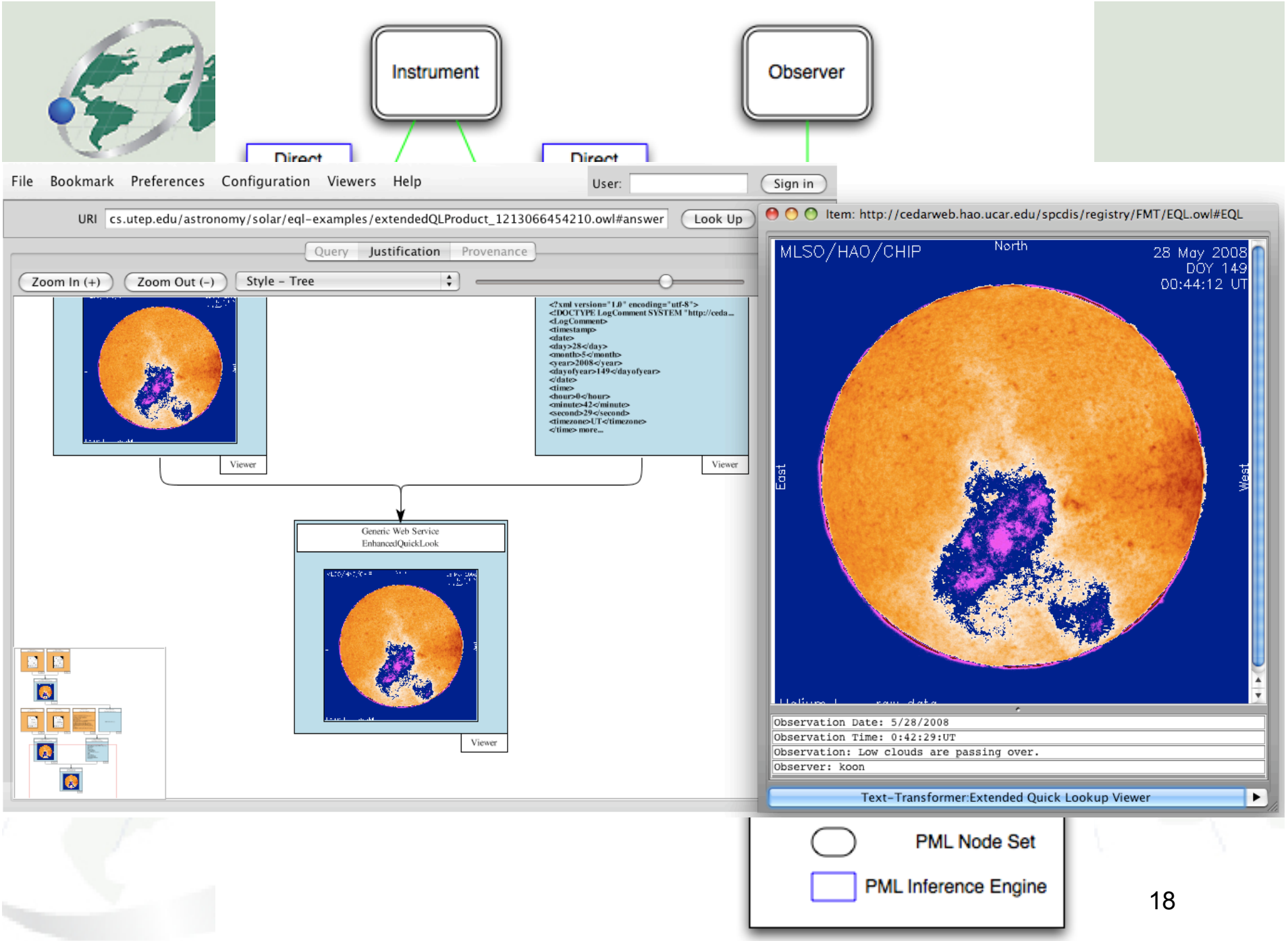
- Justification
  - Explanation
  - Causality graph
- Provenance
  - Conclusion
  - Source
  - Engine
  - Rule
- Trust
  - Trust/Belief metrics



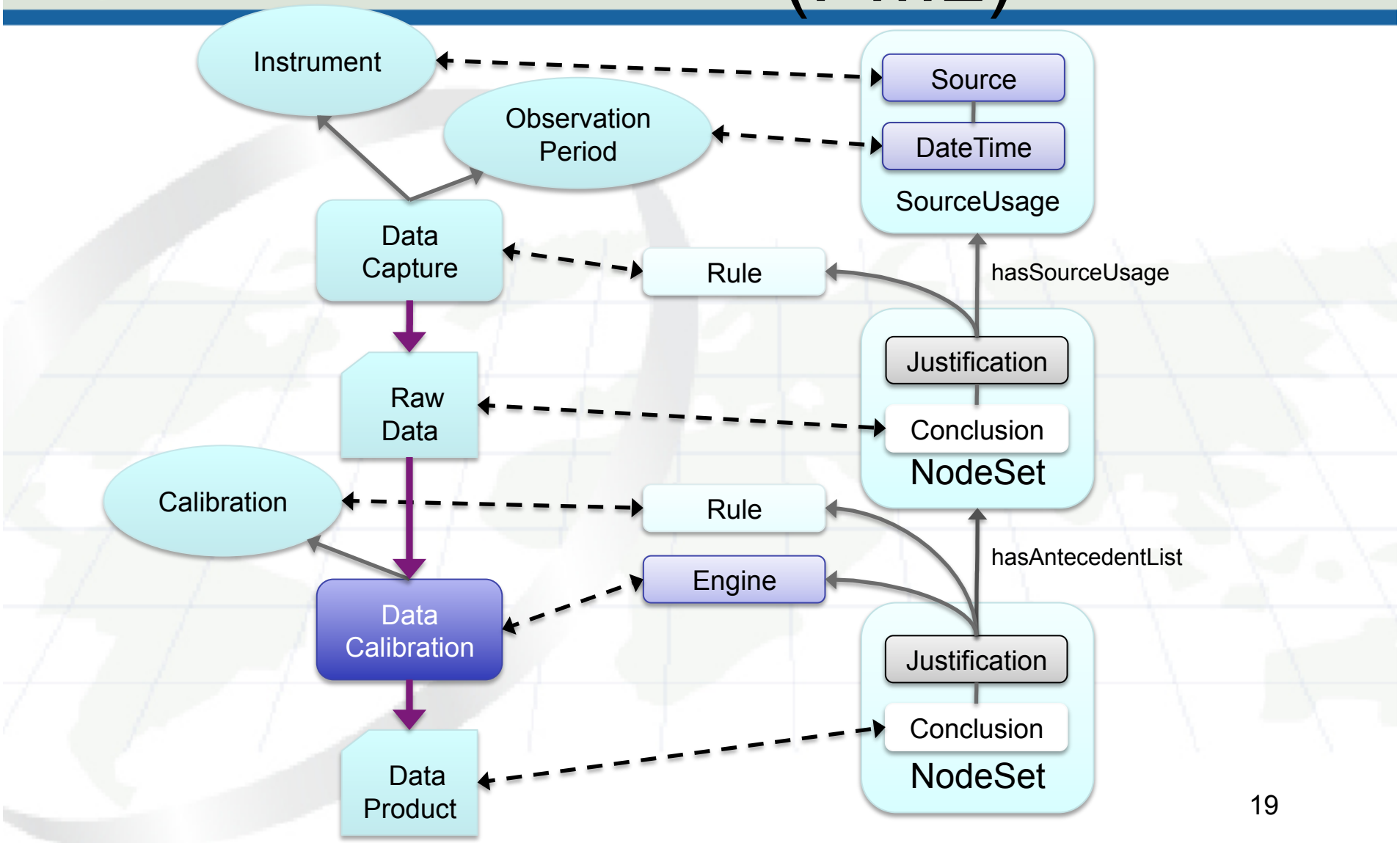
PML NodeSet


pmlj:NodeSet
pmlj:NodeSetList
pmlj:hasConclusion
pmlj:isConsequentOf
pmlj:hasAntecedentList
pmlp:Information
pmlj:InferenceStep
pmlj:hasSourceUsage
pmlj:hasInferenceEngine
pmlj:hasInferenceRule
pmlp:SourceUsage
pmlp:InferenceEngine
pmlp:InferenceRule
pmlp:hasSource
pmlp:Source


site calibration, configuration modes, instrument calibration
Site/Instrument Configuration
Instrument Capture
Instrument Calibration
David Elmore
general comments, weather comments, instrument comments
Image File with Header
Observer Logs
Quick Look
Quick Look GIF
Ship to NCAR
Site Observer
Alice Lecinski
Hawaii

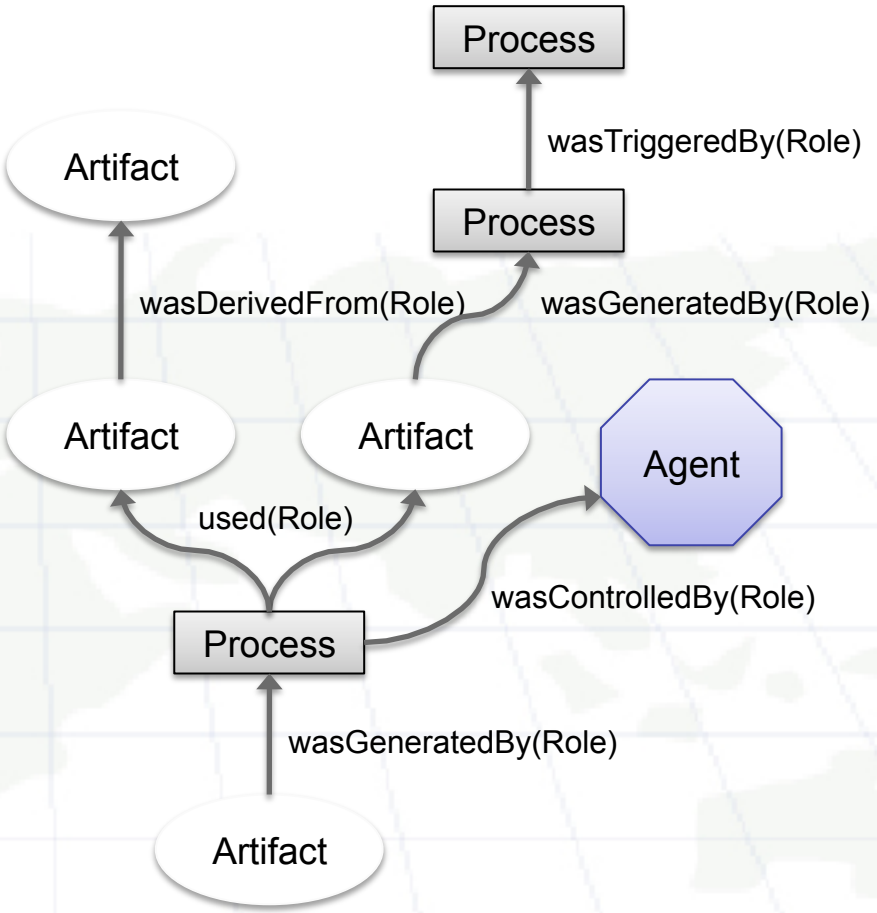

Instrument
Observer
Direct
Direct
File
Bookmark
Preferences
Configuration
Viewers
Help
User:
Sign in
URI
cs.utep.edu/astronomy/solar/eql-examples/extendedQLProduct_1213066454210.owl#answer
Look Up
Query
Justification
Provenance
Zoom In (+)
Zoom Out (-)
Style - Tree
Viewer
Viewer
Generic Web Service
EnhancedQuickLook
Viewer
Item: http://cedarweb.hao.ucar.edu/spcdis/registry/FMT/EQL.owl#EQL
MLSO/HAO/CHIP
North
28 May 2008
DOY 149
00:44:12 UT
East
West
Observation Date: 5/28/2008
Observation Time: 0:42:29:UT
Observation: Low clouds are passing over.
Observer: koon
Text-Transformer:Extended Quick Lookup Viewer
PML Node Set
PML Inference Engine

Concept Alignment (PML)

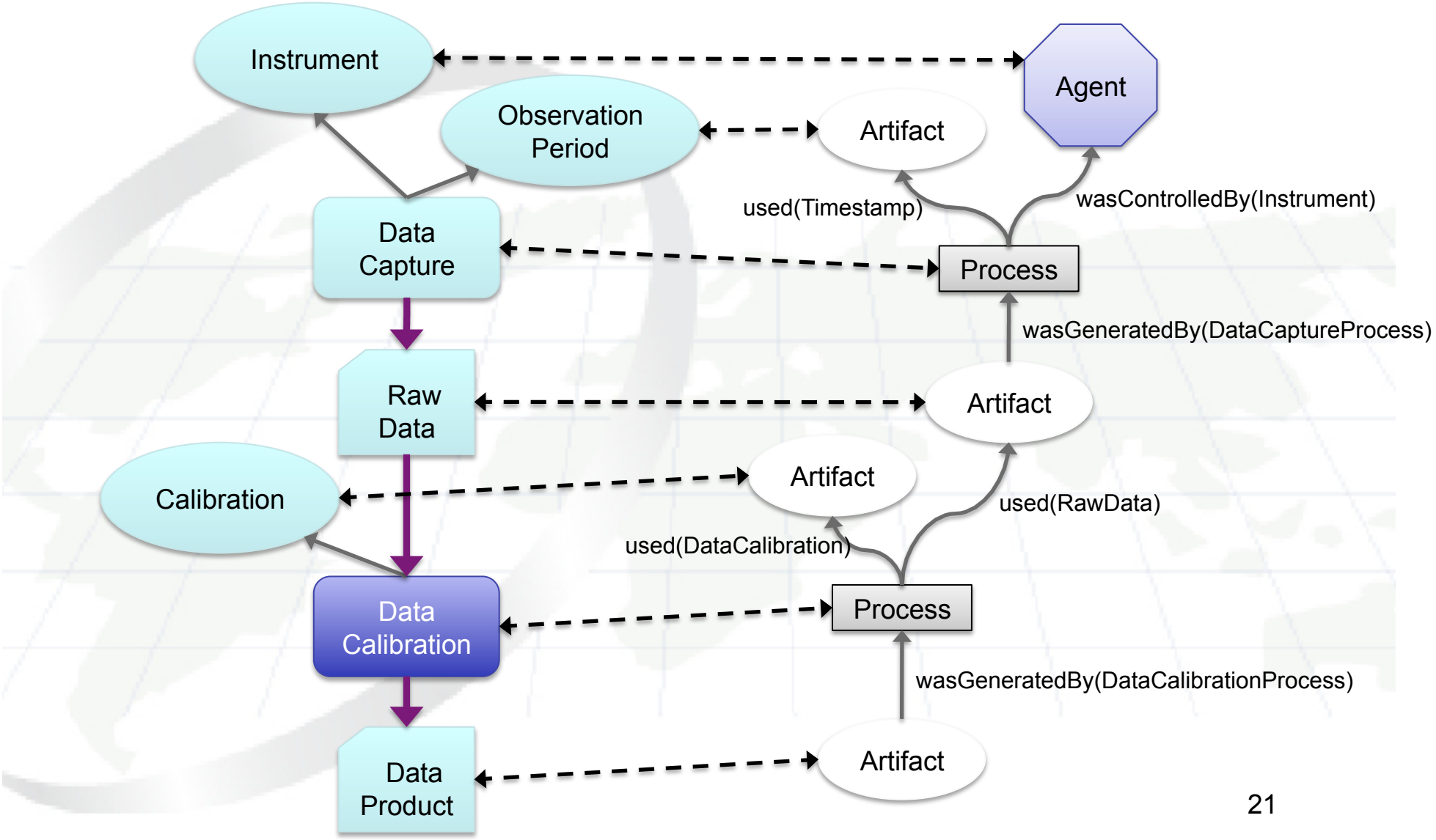


Open Provenance Model

- Agents
  - Catalyst and controlling entity of a process
- Processes
  - Action or Series of actions performed resulting in new artifacts
- Artifacts
  - Immutable piece of state
- Roles
  - Non-semantic flat tags used to provide context in relations



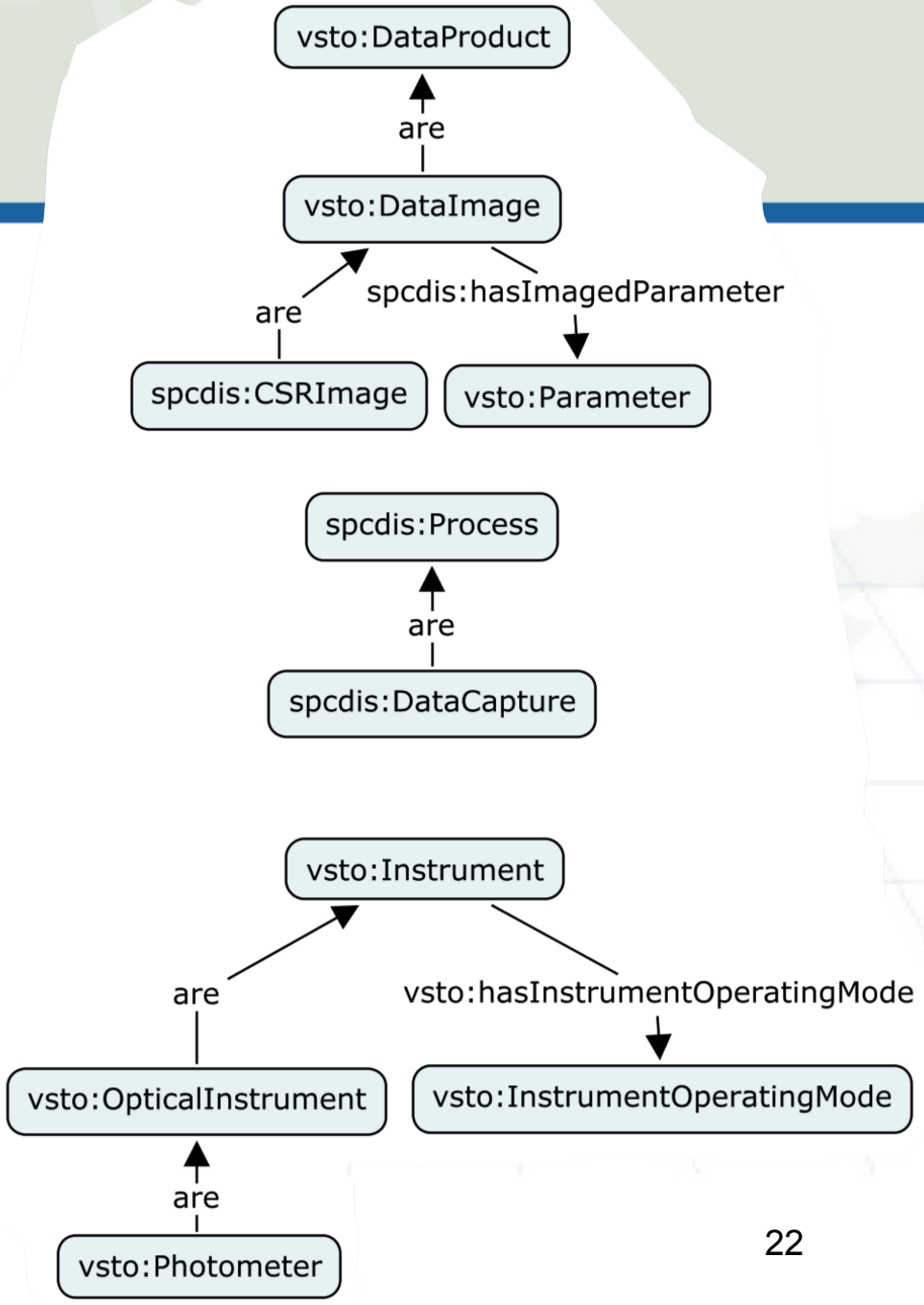
Concept Alignment (OPM)



Domain Models

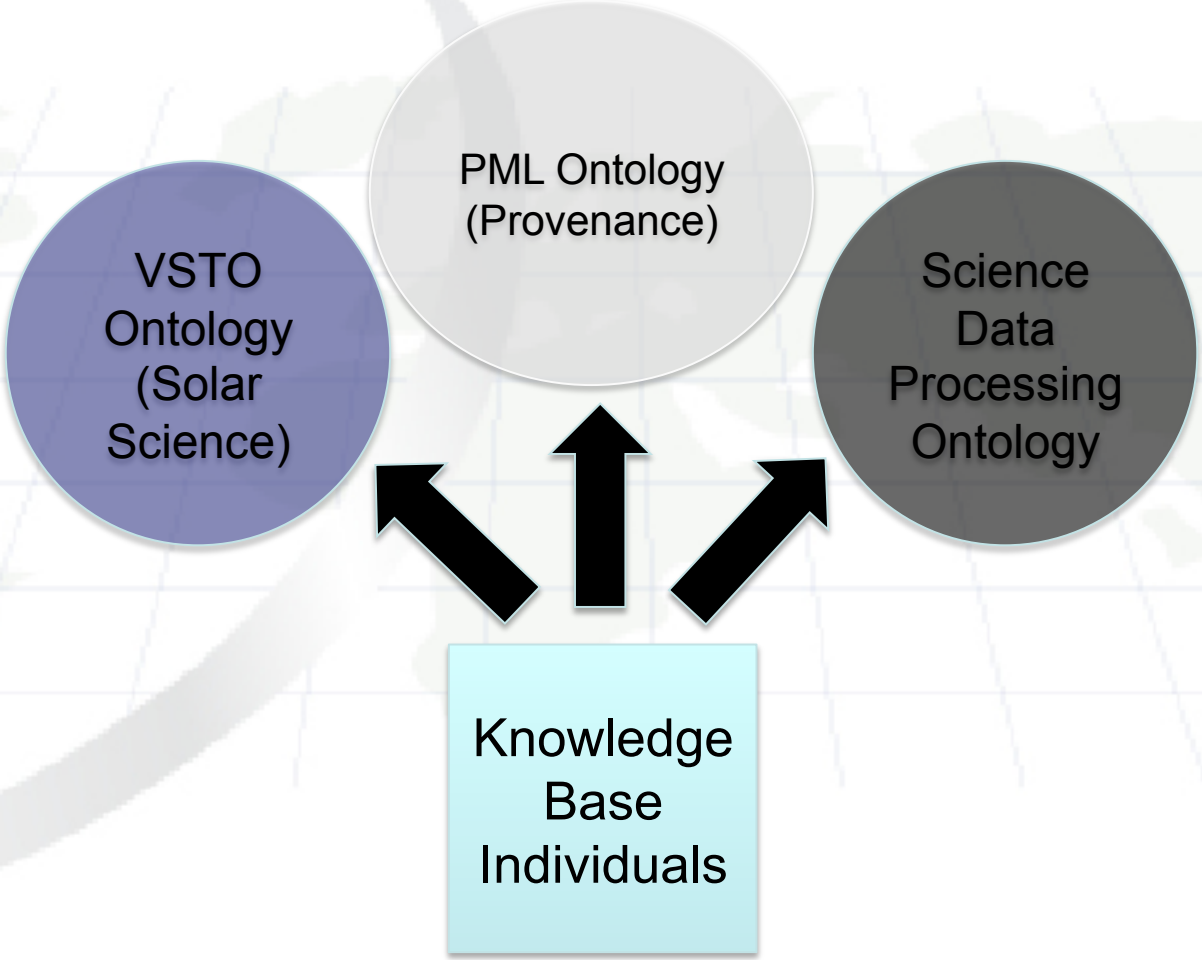
Domain models are independent of the provenance model

- Solar Science Model
  - VSTO ontology
- Science Data Processing Model
  - Devising our own
- ...others as needed

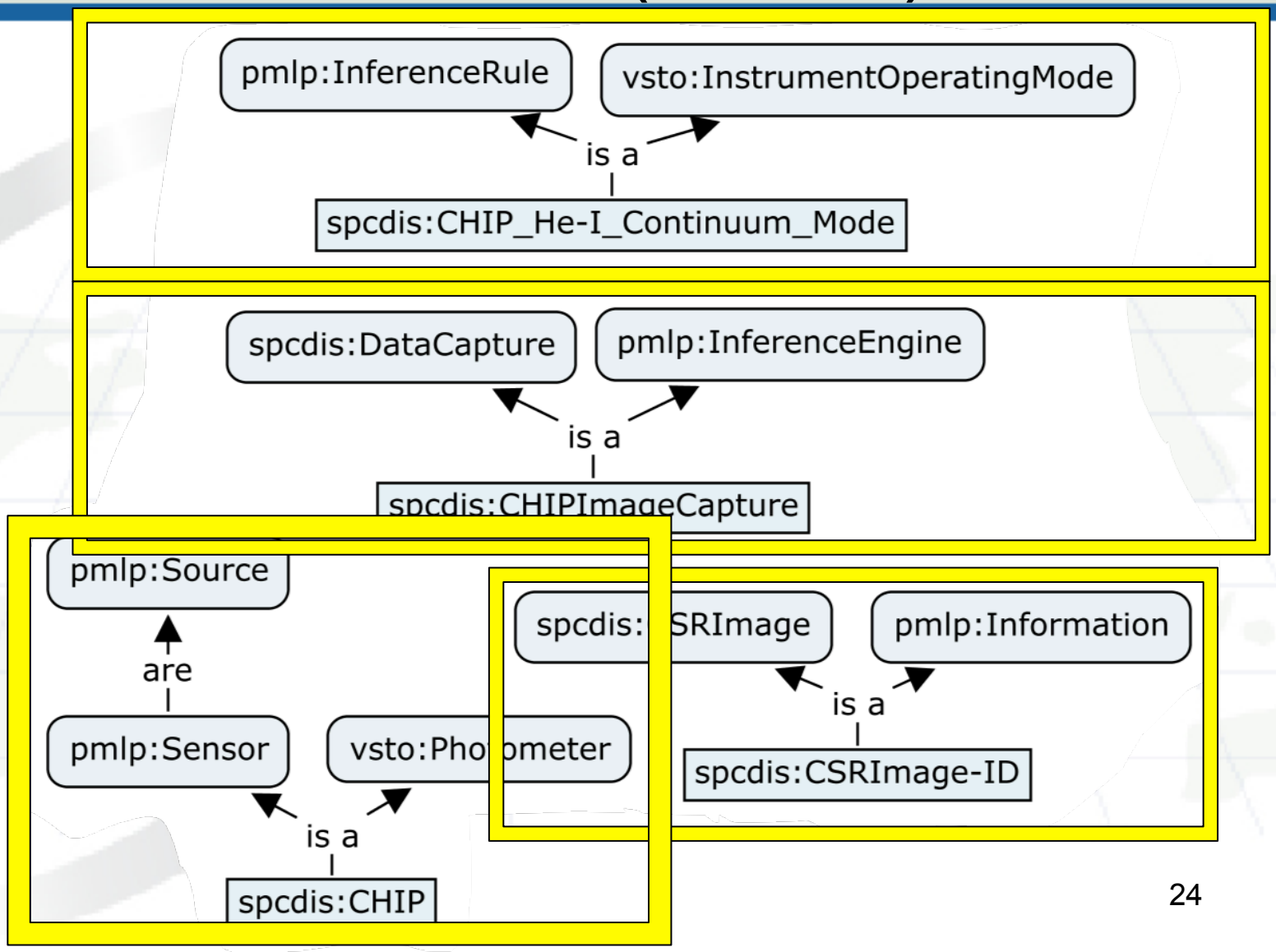


Multi-Model Individuals

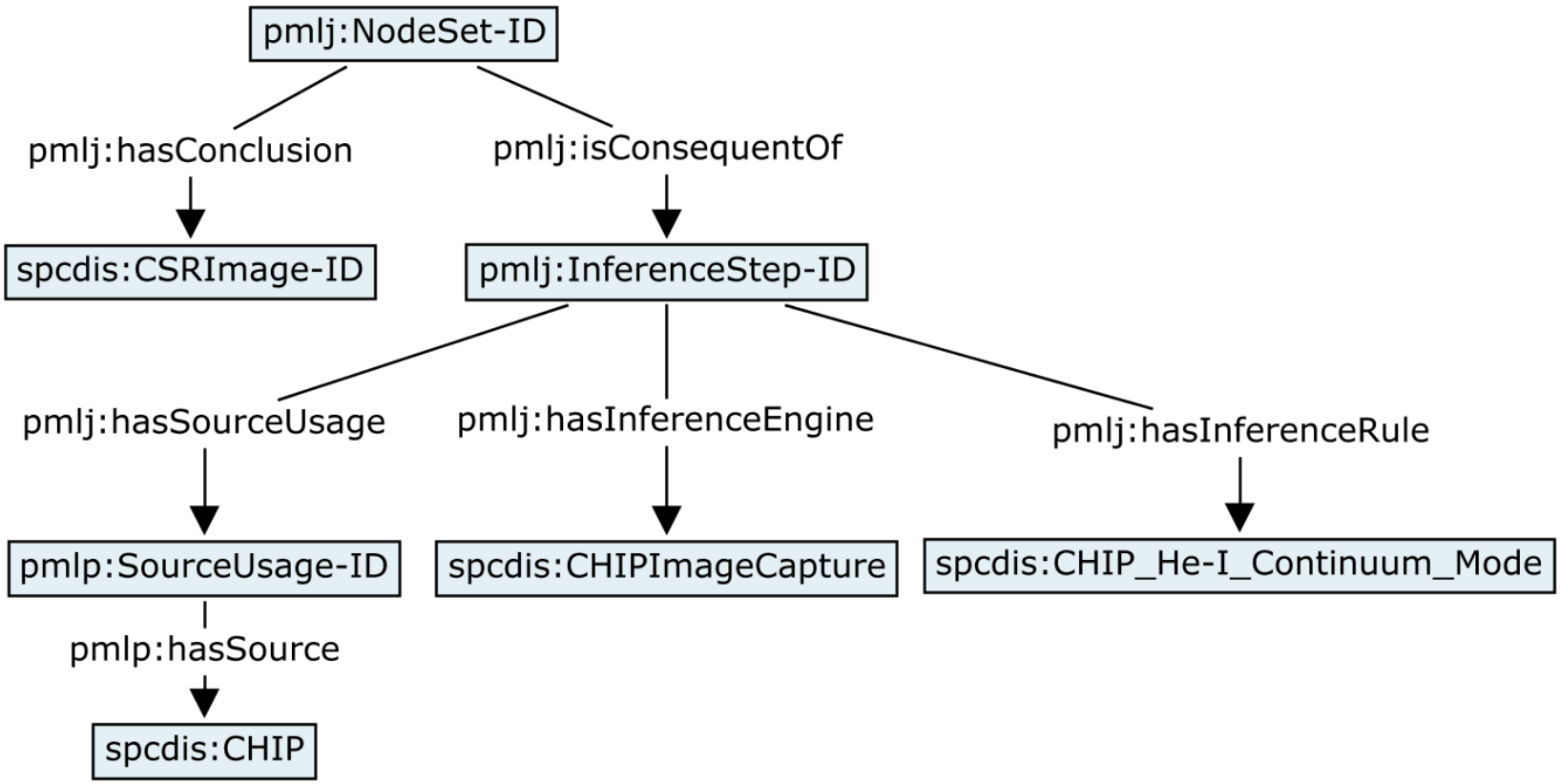
Individuals use OWL's support of multiple inheritance to provide mediation between the models



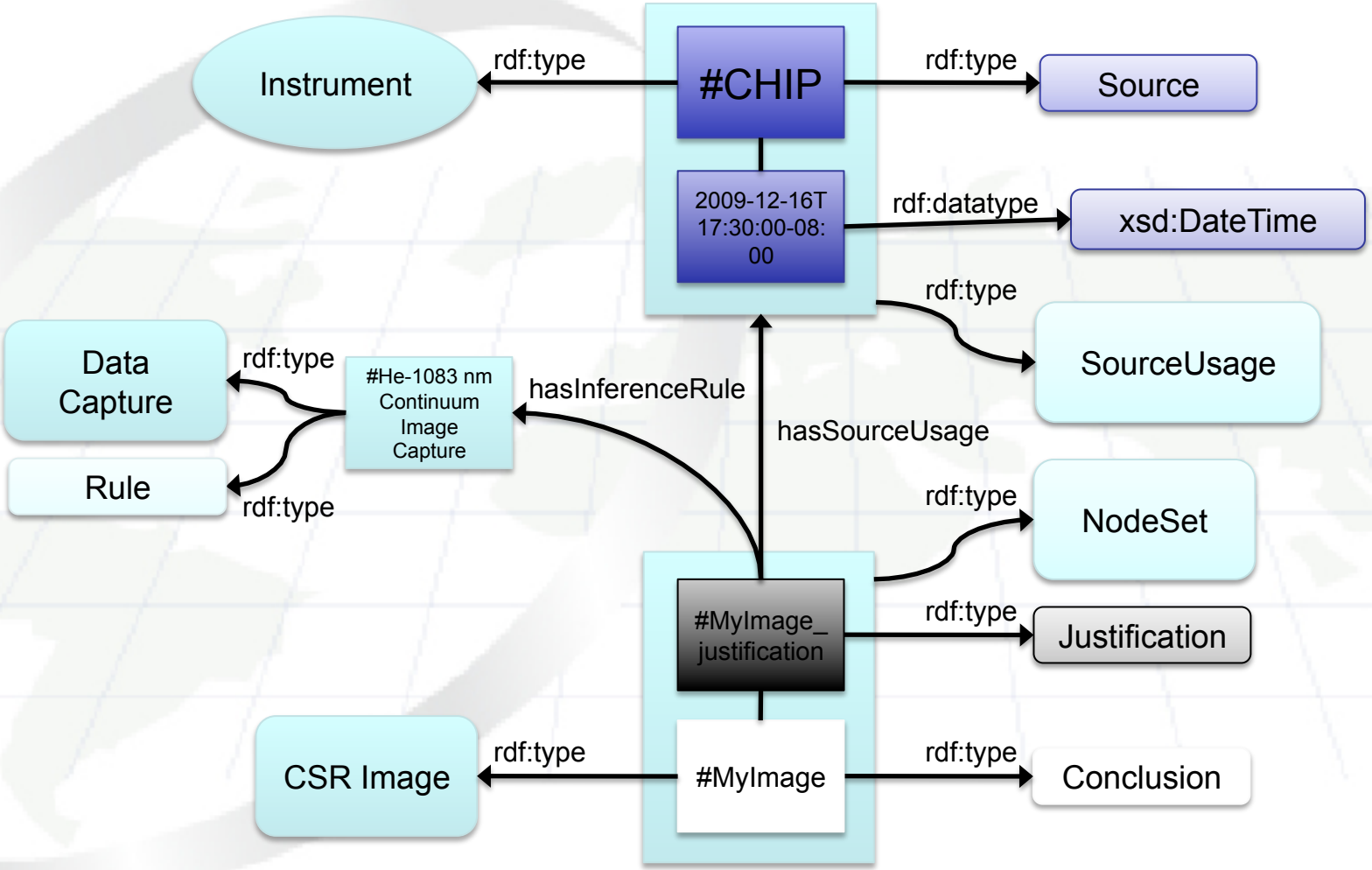
Multi-Model Individuals (cont…)



PML NodeSet using Multi-model individuals

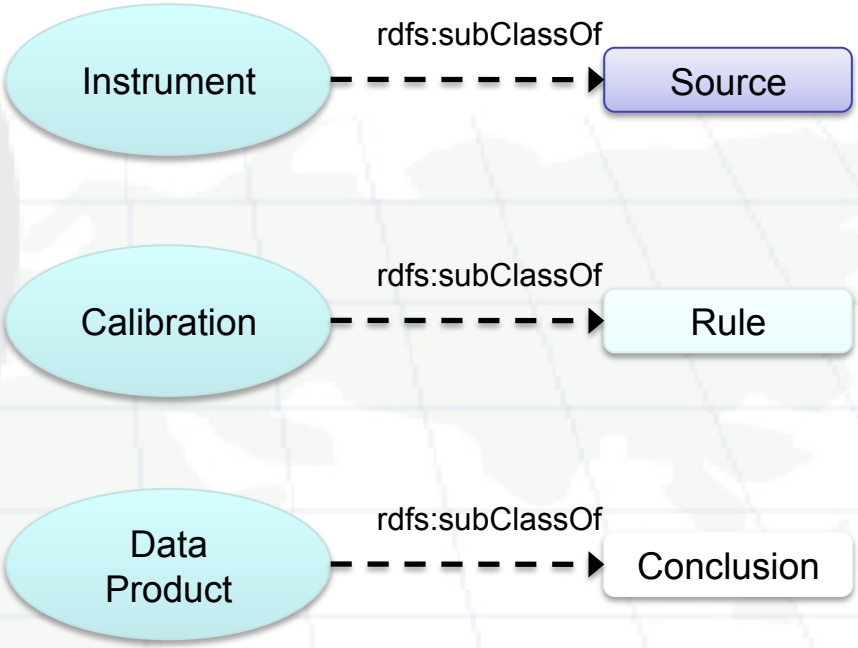


Knowledge Base with Provenance and Domain Models in Alignment



Alignment via Ontology Constructs

- Use ontology constructs to map a relationship between concepts in different domains
- Can be defined in a separate ontology than the models being mapped
- Does not require a change to the source models!
- OWL
  - owl:equivalentClass
  - owl:equivalentProperty
  - owl:sameAs
- RDFS
  - rdfs:subClassOf
  - Rdfs:subPropertyOf



Direct Alignment using Rules*

- Rules provide conditional logic on semantic constructs outside application logic
- Rules can be updated or tweaked without requiring an application update.
- Easily shared and managed
- Provides for more complex mapping than ontology constructs

ex:Instrument(?x)

➔ pmlp:Sensor(?x)

pmlp:Information(?x) ^

pmlp:hasURL(?x,?url) ^

swrlb:endsWith(?url, ".hsh.fts ")

➔ Ex:CHIPIntensityImage(?x)

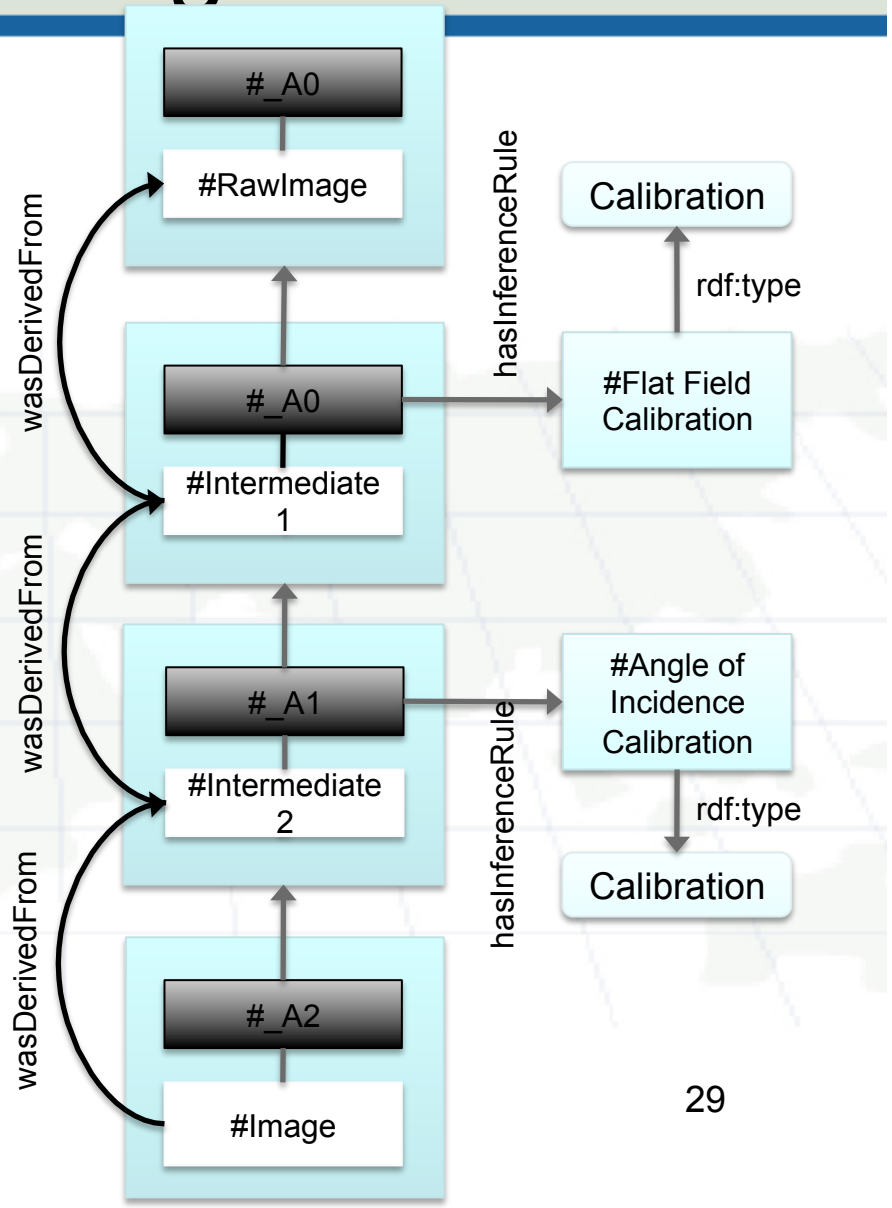
*Many rule systems exist, this slide uses the Semantic Web Rule Language (SWRL)

Querying/Interrogating the Knowledge Base

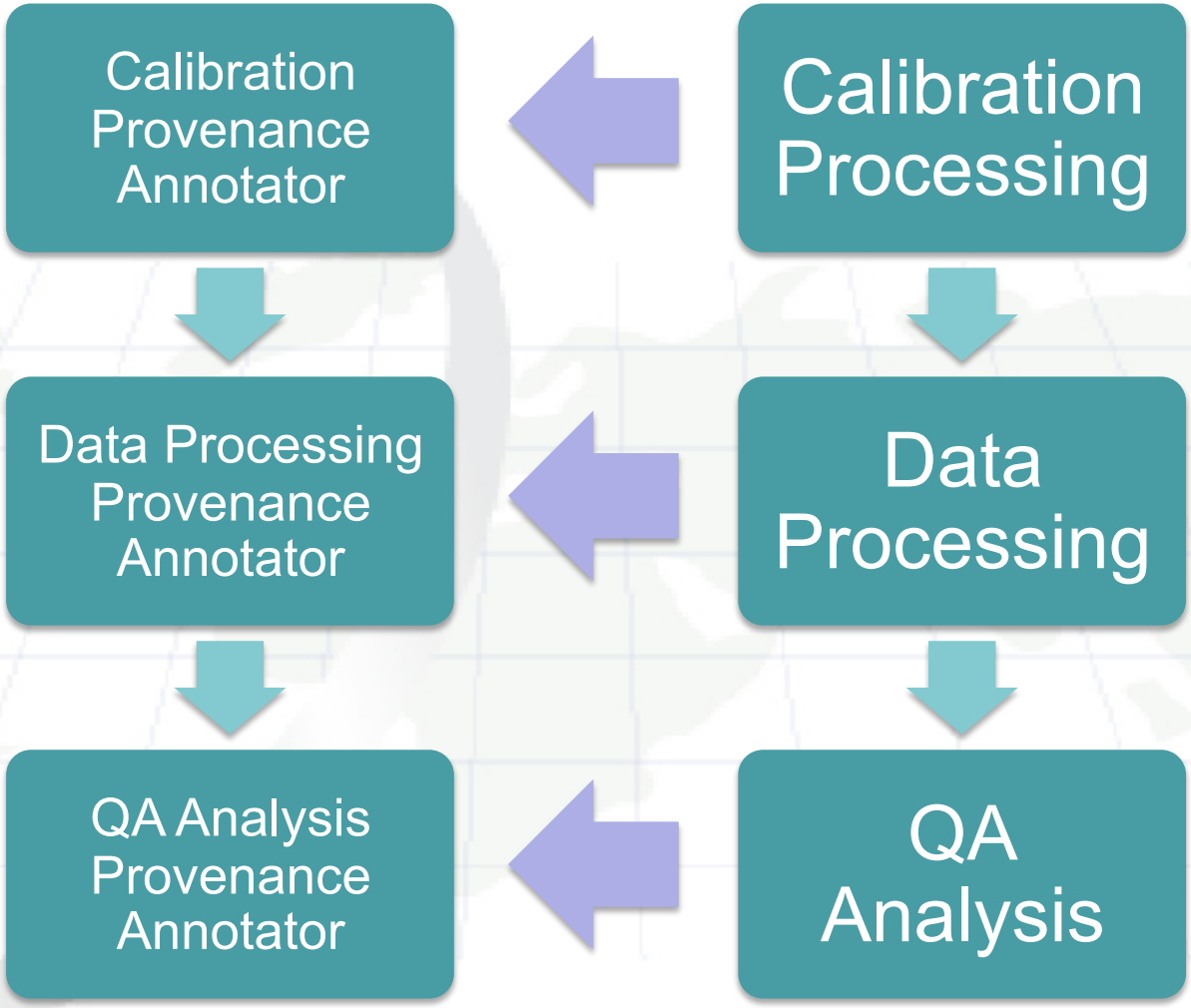
- Back to one use case:

What *calibrations* have been applied to this *image*?

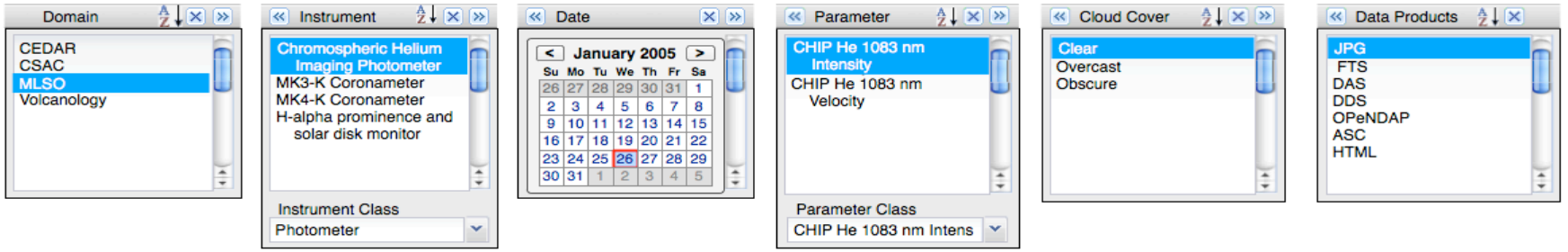
- We construct a query returns any individuals with type Calibration used as the InferenceRule in the justification from any artifact the current artifact was derived from.
- We assume that any calibration applied to an artifact the current artifact was derived from can also be considered as 'applied' to the current artifact, and that the wasDerivedFrom property is transitive

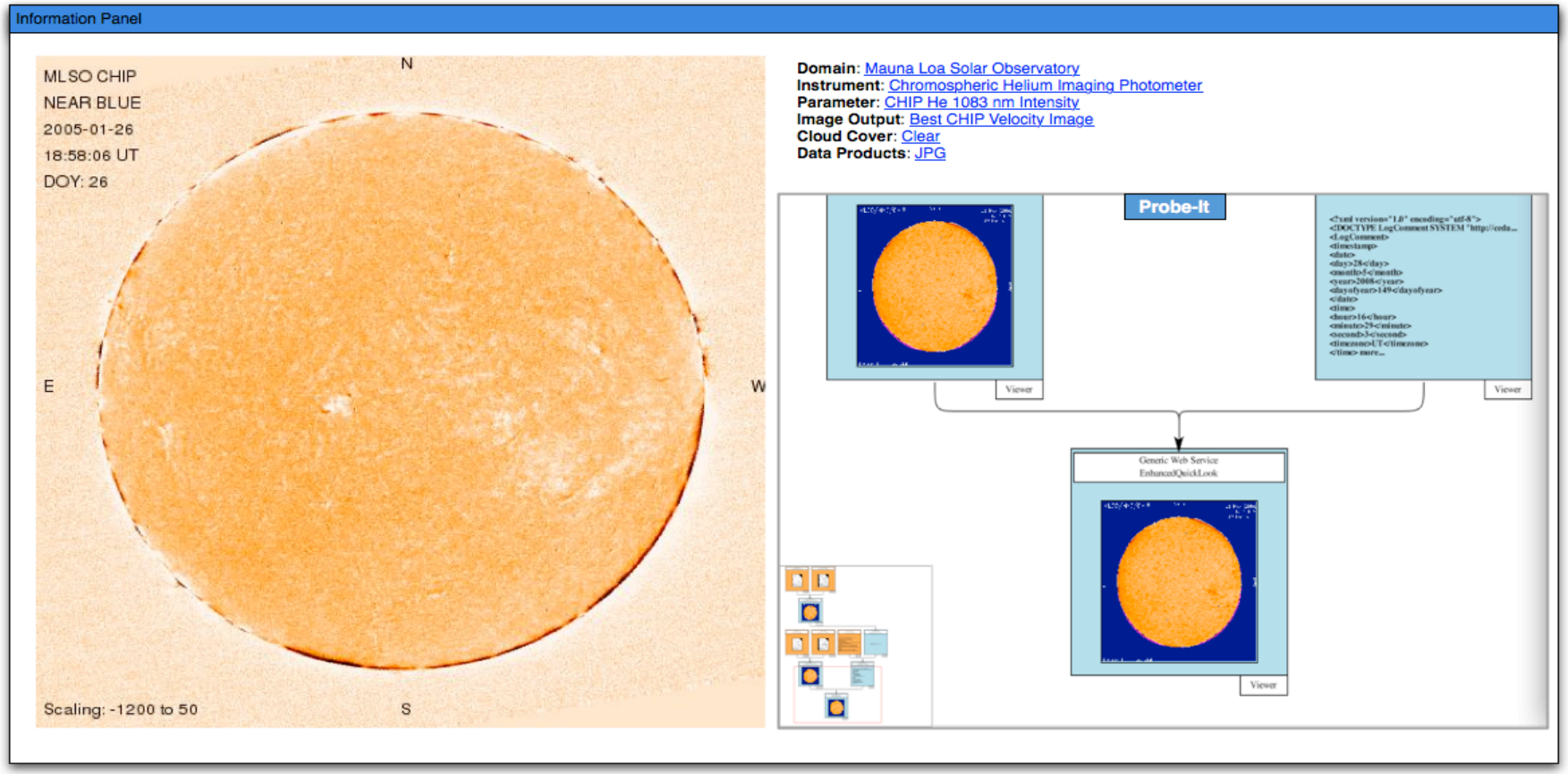


Provenance Capture

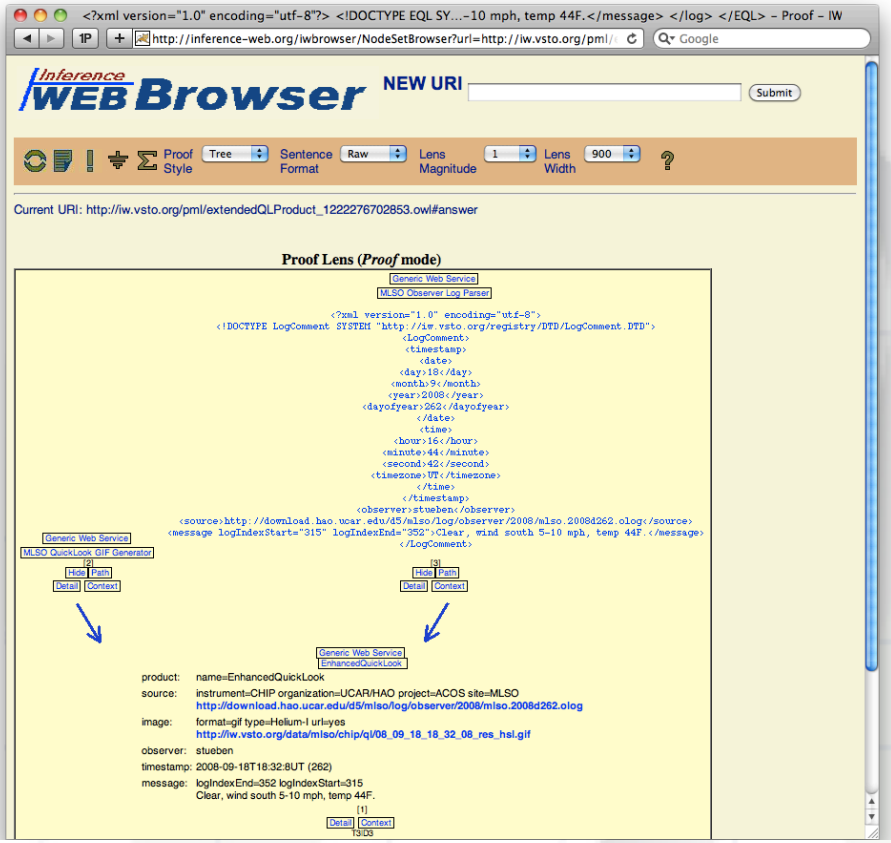


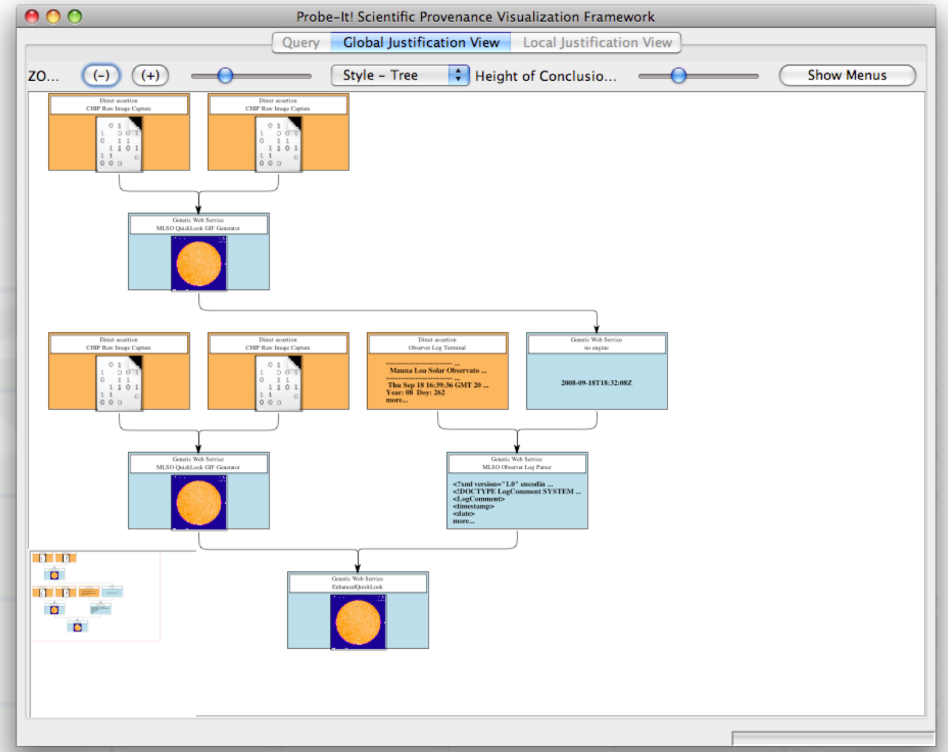
Provenance aware faceted search





Interoperability with Provenance Tools





Inference Web Browser
http://inference-web.org/iwbrowser

Probe-It!
http://trust.utep.edu/probeit/

Multi-sensor Synergy Data Advisor (MDSA)

- Based on NASA sensor assets and multi-option processing - http://giovanni.gsfc.nasa.gov
- Dynamically generated lineage (XML)
- Want to advise for or against certain processing operations on certain data products (both internal and external provenance)

Giovanni Allows Scientists to Concentrate on the *Science*

The Old Way:

Pre-Science

- Find data
- Retrieve high volume data
- Learn formats and develop readers
- Extract parameters
- Perform spatial and other subsetting
- Identify quality and other flags and constraints
- P…rm filtering/masking
- Develop analysis and visualization
- Accept/discard/get more data (sat, model, ground-based)

DO SCIENCE

- Exploration
- Initial Analysis
- Use the best data for the final analysis
- Derive conclusions
- Write the paper
- Submit the paper

Web-based Services:

Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct

Giovanni

Miradr

- Read Data
- Extract Parameter
- Subset Spatially
- Filter Quality
- Reformat
- Reproject
- Visualize
- Explore
- Analyze

The Giovanni Way:

- Minutes
- Days for exploration
- Use the best data for the final analysis
- Derive conclusions
- Write the paper
- Submit the paper

DO SCIENCE

Web-based tools like Giovanni allow scientists to ***compress*** the time needed for pre-science preliminary tasks: *data discovery, access, manipulation, visualization, and basic statistical analysis*.

Scientists have ***more time to do science!***

Inter-comparison of data from multiple sensors

Data from multiple sources to be used together:

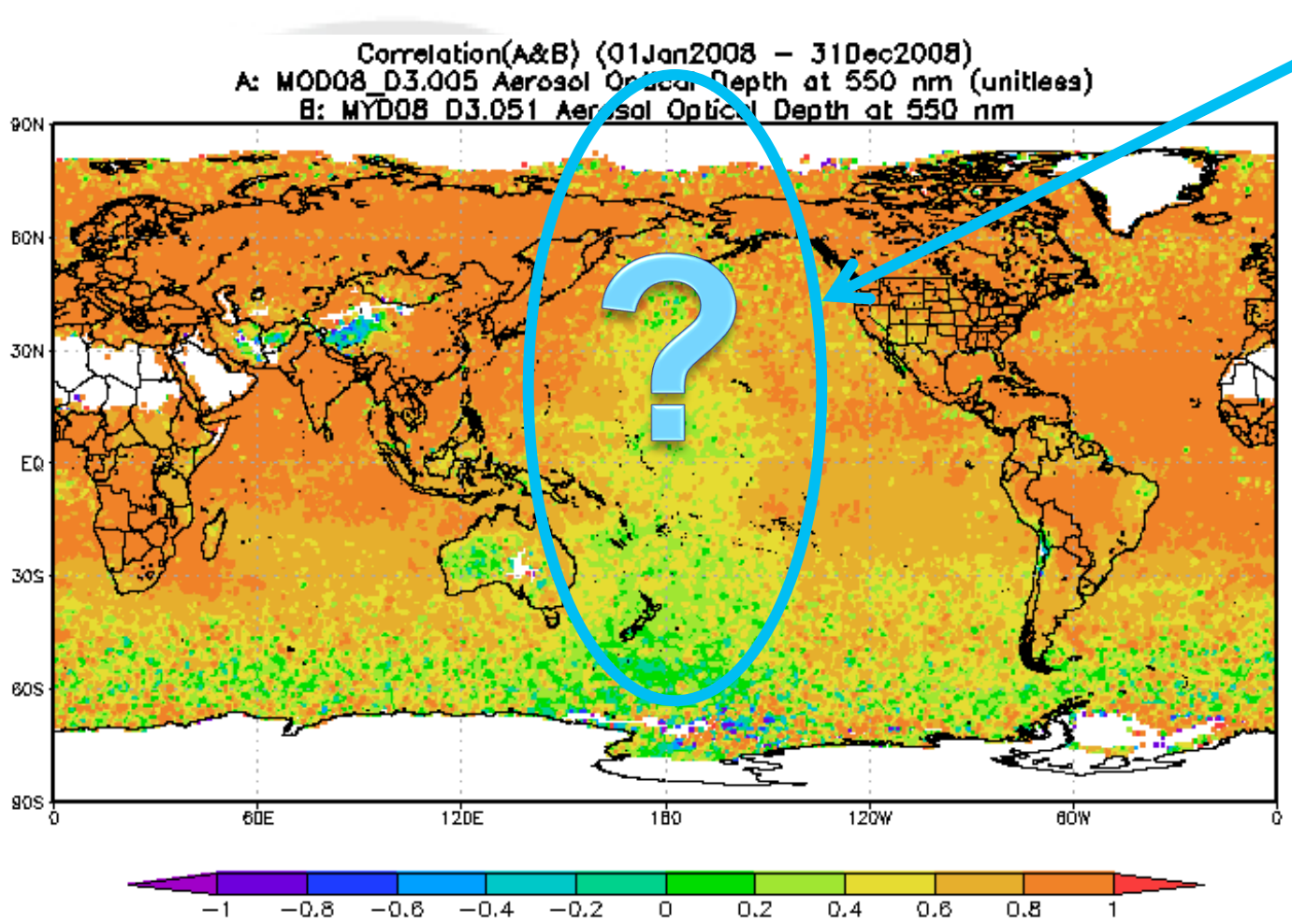
- ACE
- NPP and NPOESS
- Geo-Cape
- European and other countries' satellites
- Models

Harmonization:

- It is not sufficient just to have the data from different sensors and their provenances in one place
- Before comparing and fusing data, things need to be harmonized:
  - Data: format, grid, spatial and temporal resolution
  - Metadata: standard fields, units, scale, quality
  - Provenance: source, assumptions, algorithm, processing steps

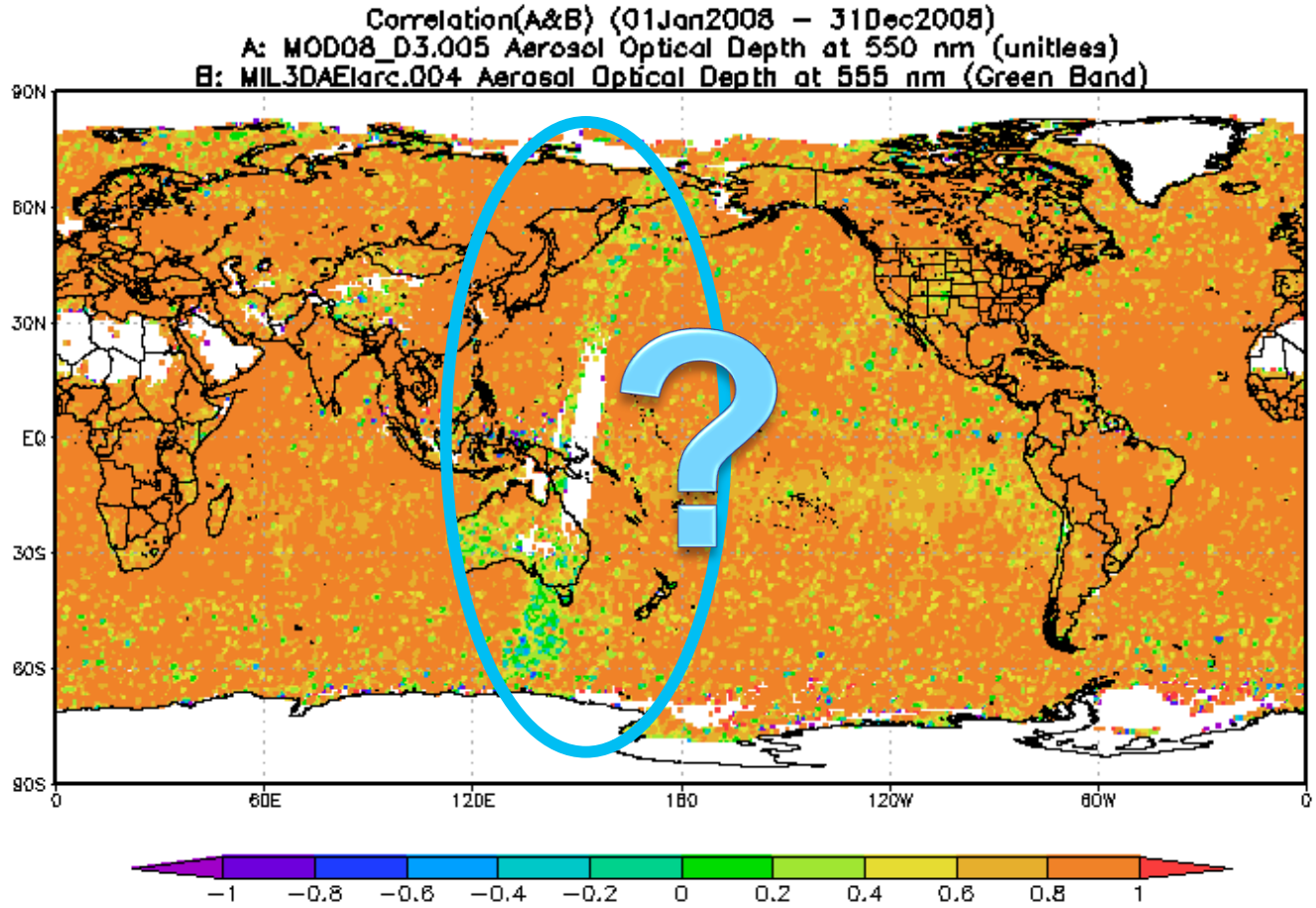
Dangers of easy data access without proper assessment of joint data usage - *It is easy to use data incorrectly*

Why don't MODIS Terra and Aqua Aerosols agree?



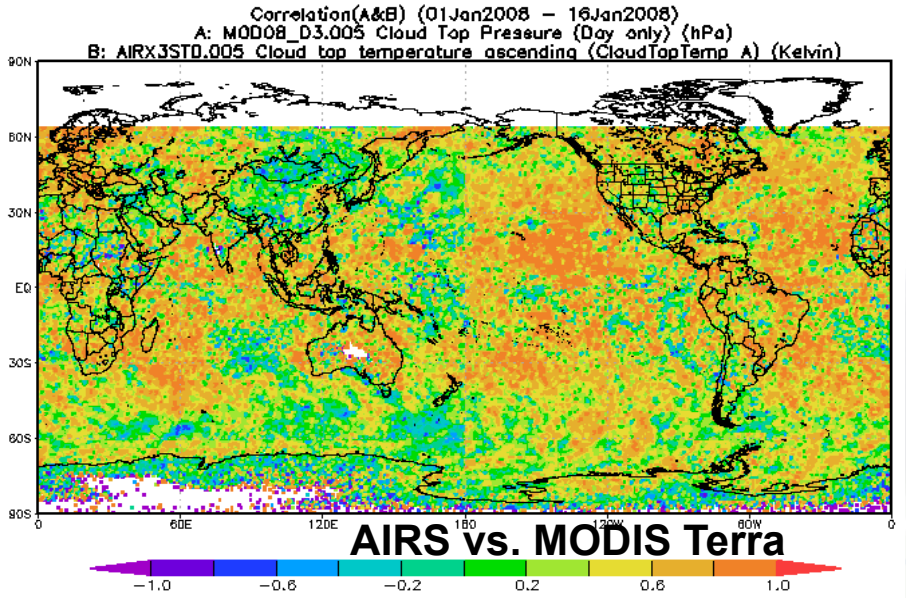
MODIS-Terra vs. MODIS-Aqua: Map of AOD temporal correlation, 2008

Sensitivity Study: Daily AOD MODIS Terra vs. MISR Terra

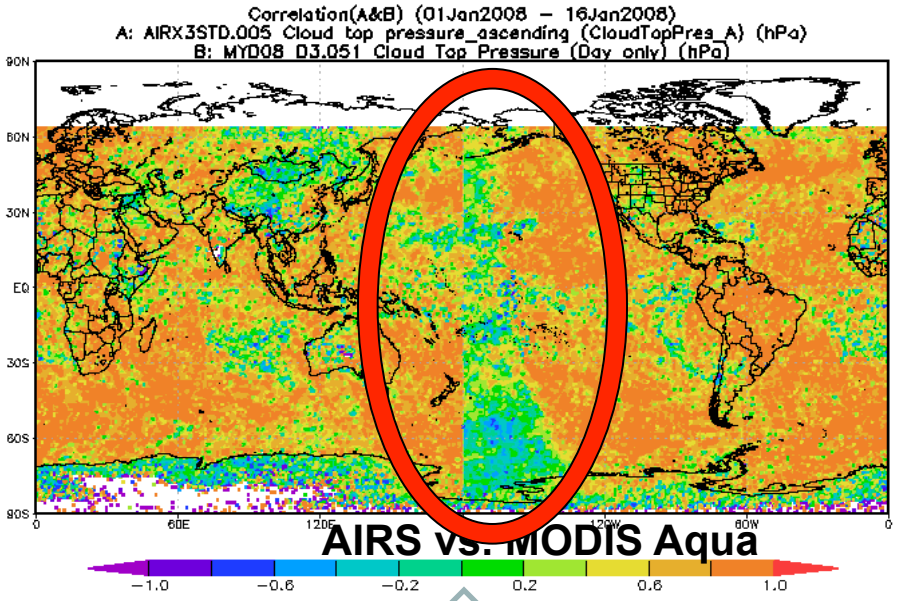


MODIS-Terra vs. MISR-Terra: Map of temporal correlation

MODIS Terra & Aqua vs. AIRS Cloud Top Pressure

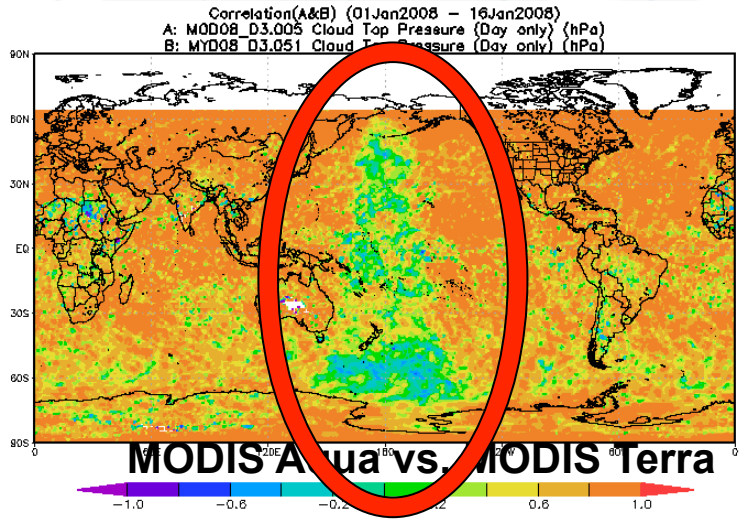


AIRS vs. MODIS Terra



AIRS vs. MODIS Aqua

Correlation maps for Jan 1 – 16, 2008



MODIS Aqua vs. MODIS Terra

Impact:
Findings using aerosol data apply to other geophysical parameters!

Collecting and Delivering Data Provenance

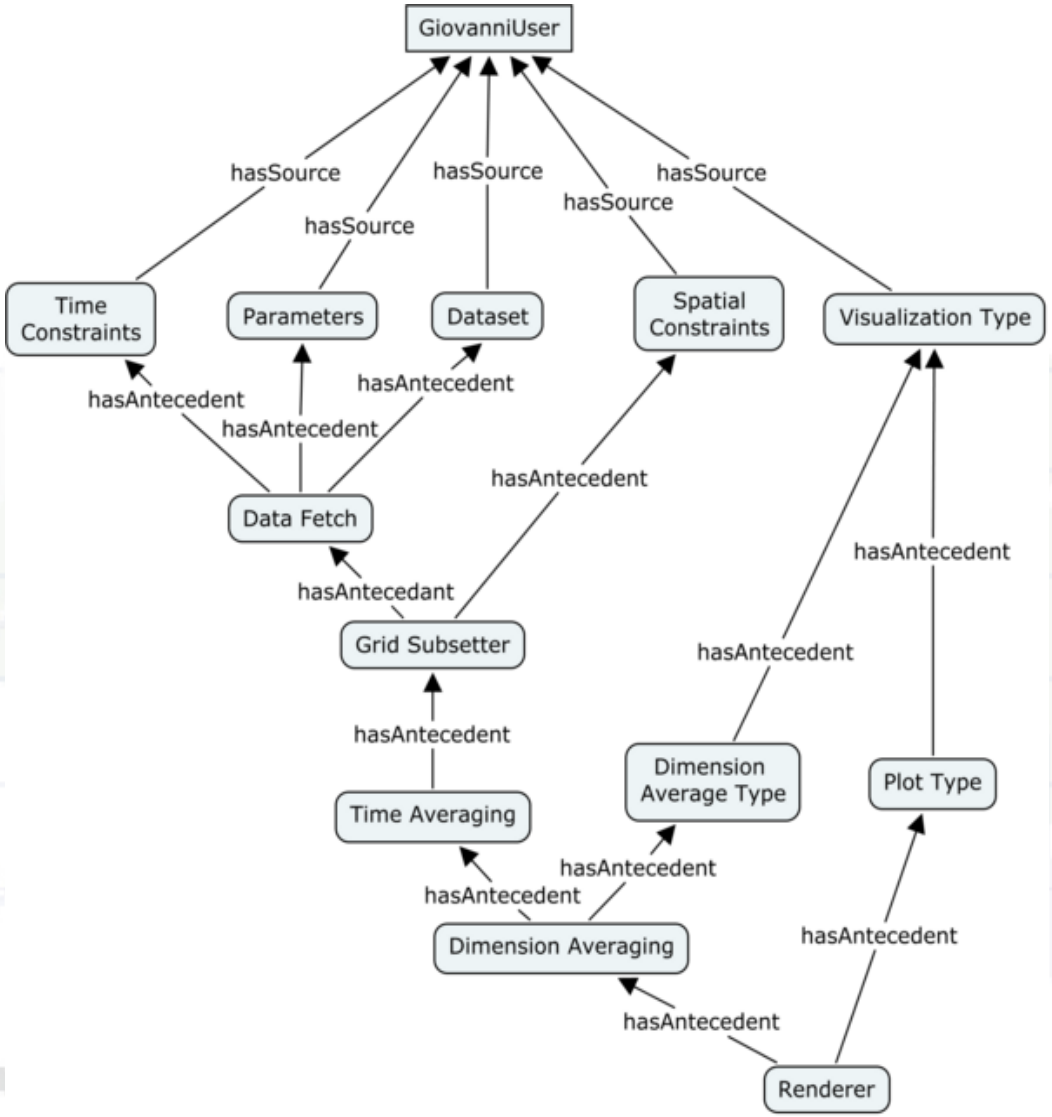
Where to find the knowledge about data and data processing?

- It is scattered in scientific papers, the actual code, unwritten assumptions, folklore, etc.
- Assess sensitivity of the results to variations in processing algorithms/steps…
- Work closely with scientists to guarantee science quality

How to deliver provenance?

- Deliver to users together with the data
- Present to users in a convenient, easy-to-read fashion
- Provide recommendations for different data usage (applications vs. climate studies)

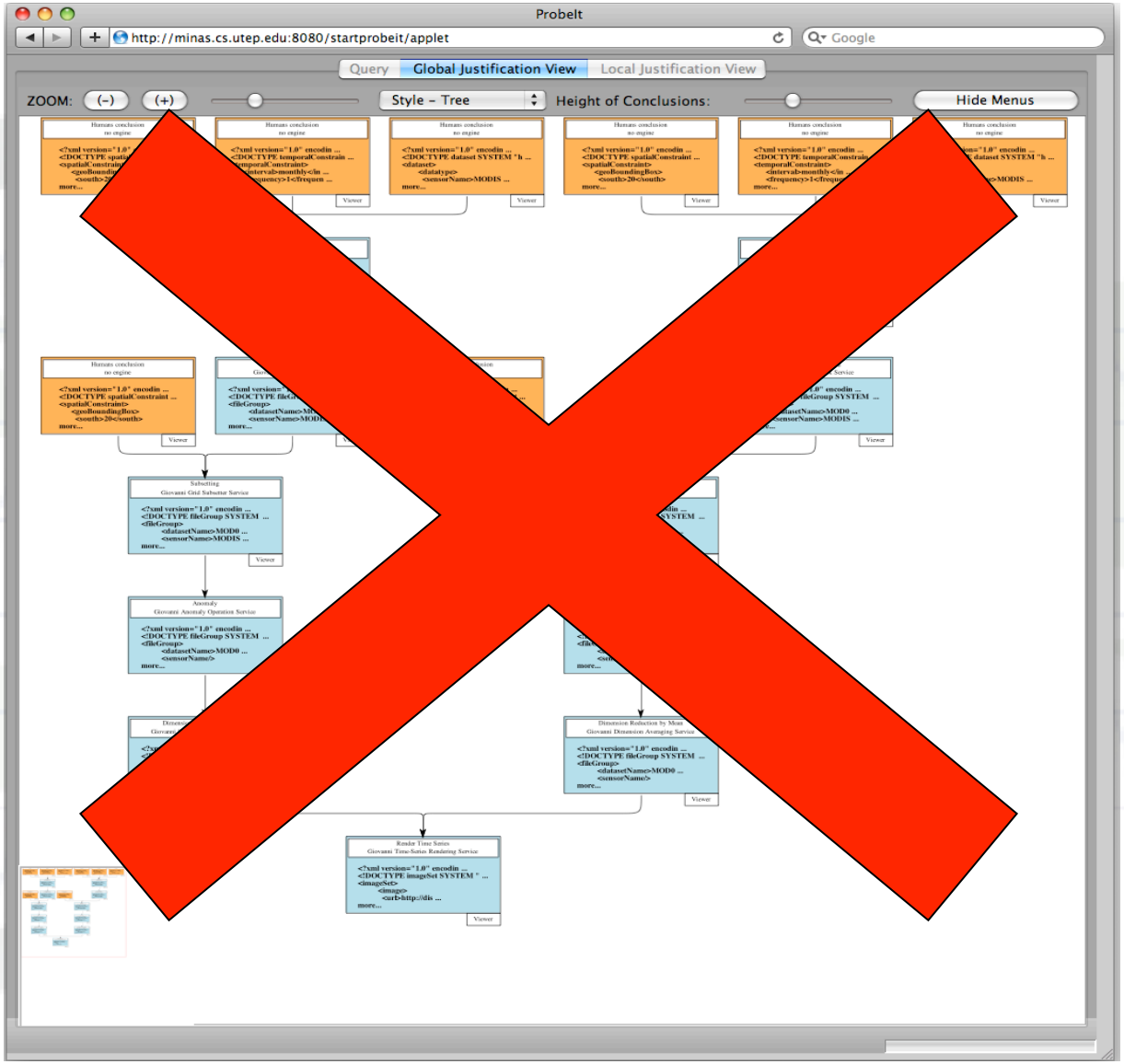
Use Case Processing PML



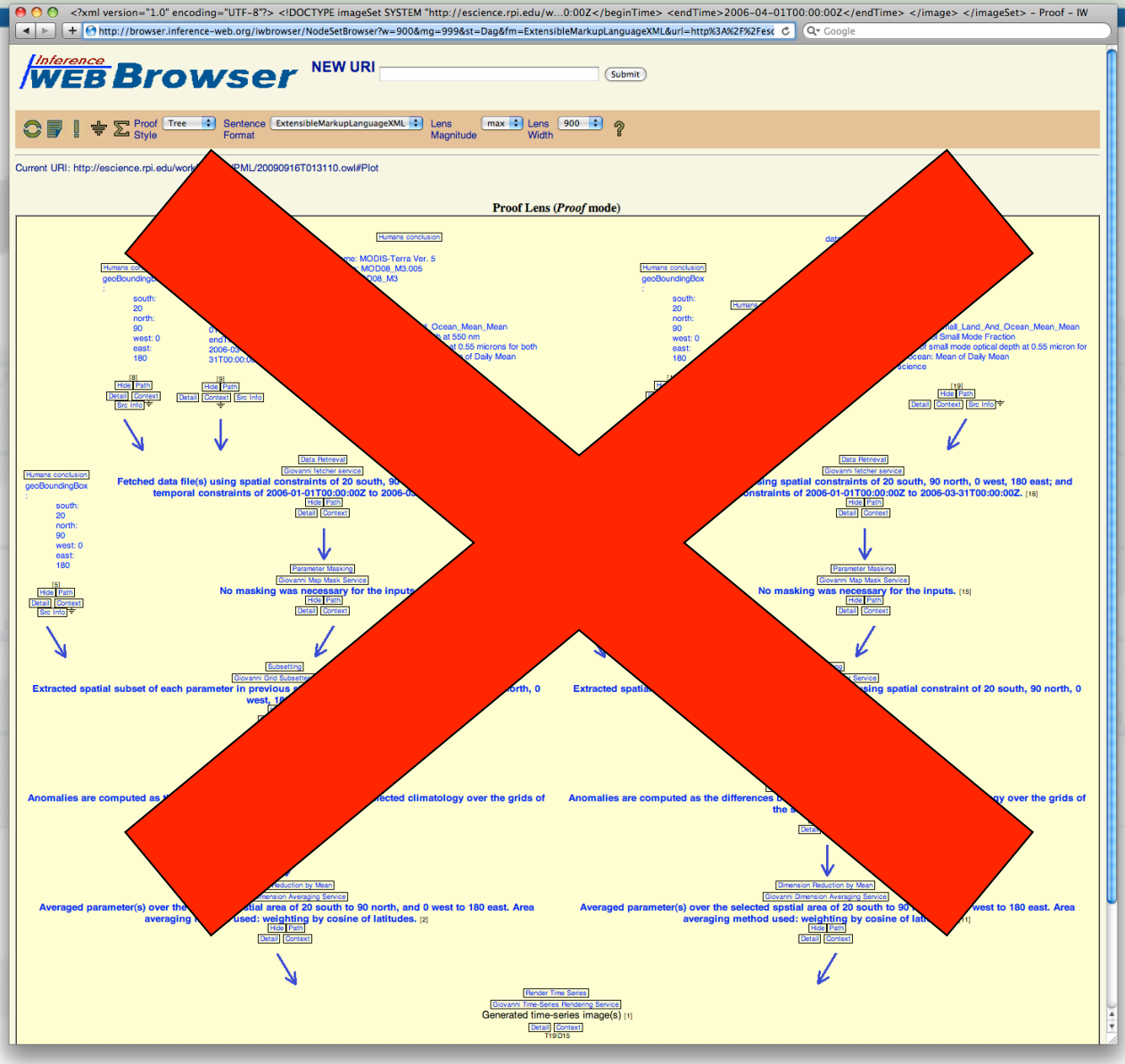
- Proof Markup Language graph for Giovanni processing provenance

- Capability to access PML using IW Browser or Probe-It provenance visualization tools.

Accessing PML using Probe-It provenance visualization tools



IWBrowse


Inference WEB Browser
NEW URI
Submit
Proof Style
Tree
Sentence Format
ExtensibleMarkupLanguageXML
Lens Magnitude
max
Lens Width
900
Proof Lens (Proof mode)
Humans conclusion
geoBoundingBox
south: 20
north: 90
west: 0
east: 180
Data Retrieval
Giovanni fetcher service
Parameter Masking
Giovanni Map Mask Service
No masking was necessary for the inputs.
Subsetting
Dimension Reduction by Mean
Giovanni Dimension Averaging Service
Render Time Series
Giovanni Time-Series Rendering Service
Generated time-series image(s)

However…

- *Normal* people cannot look at these presentations and get answers to their questions, and often *they don't care!*
- Or, if they do, then they want to see a semantic provenance difference – no they don't really know to ask for this but it is the step to establishing *trust* in these systems
- Provenance presentation is still an open challenge…

So … Semantic Advisor

Semantic Advisor

Provides caveats for intercomparison based on user selections

- Parameter, Dataset
- Satellite, Orbit (derived from Dataset)

Service-based Interaction

- Giovanni sends XML with user input
- RPI returns XML with comparison info
- Giovanni renders comparison as table

Features

Leveraging Ontology and Rulesets

- Developed ontology to represent advisories and corresponding semantic rule set to assert advisories when certain conditions hold in the knowledge model.
- Developing web service to ingest user input XML, translate into semantic knowledge based on ontology vocabularies, perform reasoning and rule-based inference on the knowledge, and generate a response XML to return to the Giovanni service.

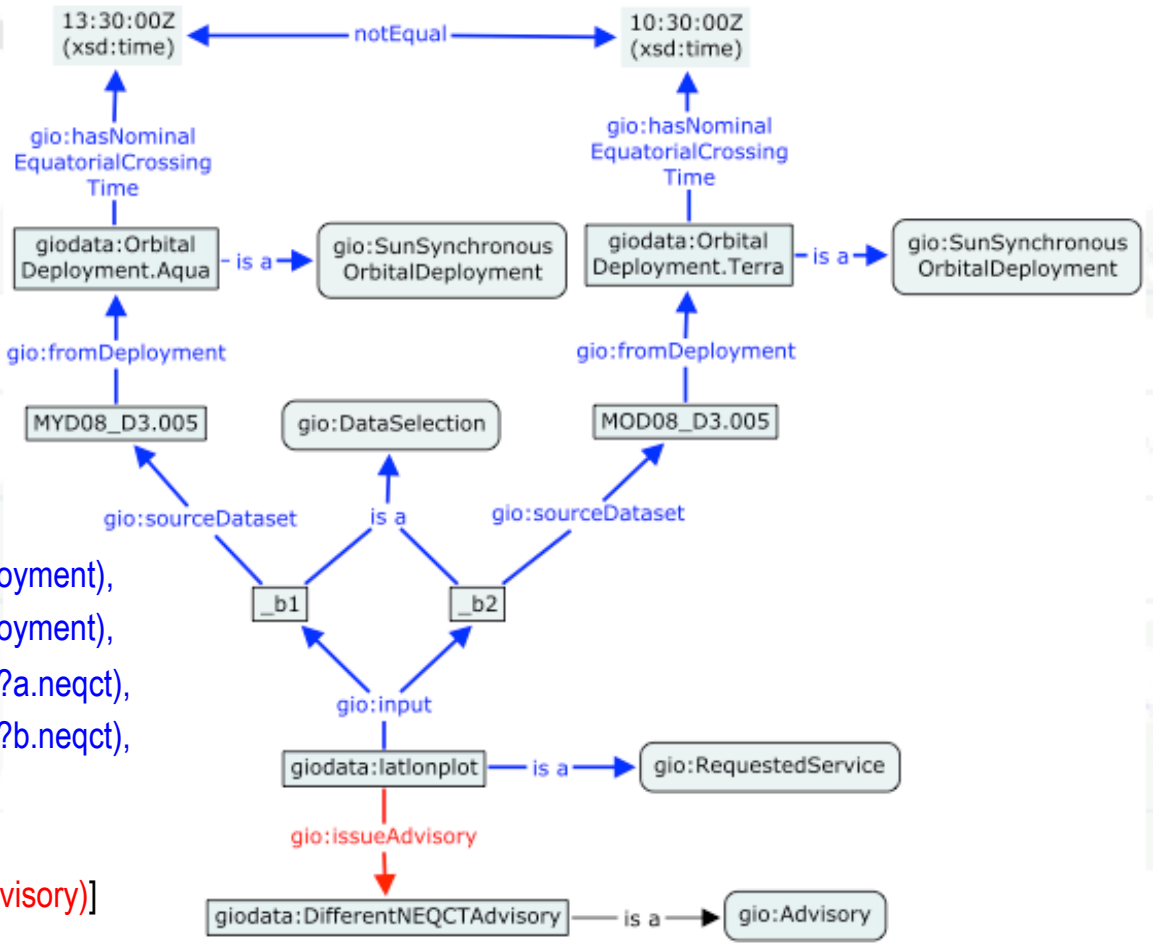
RuleSet Development

Rulesets:

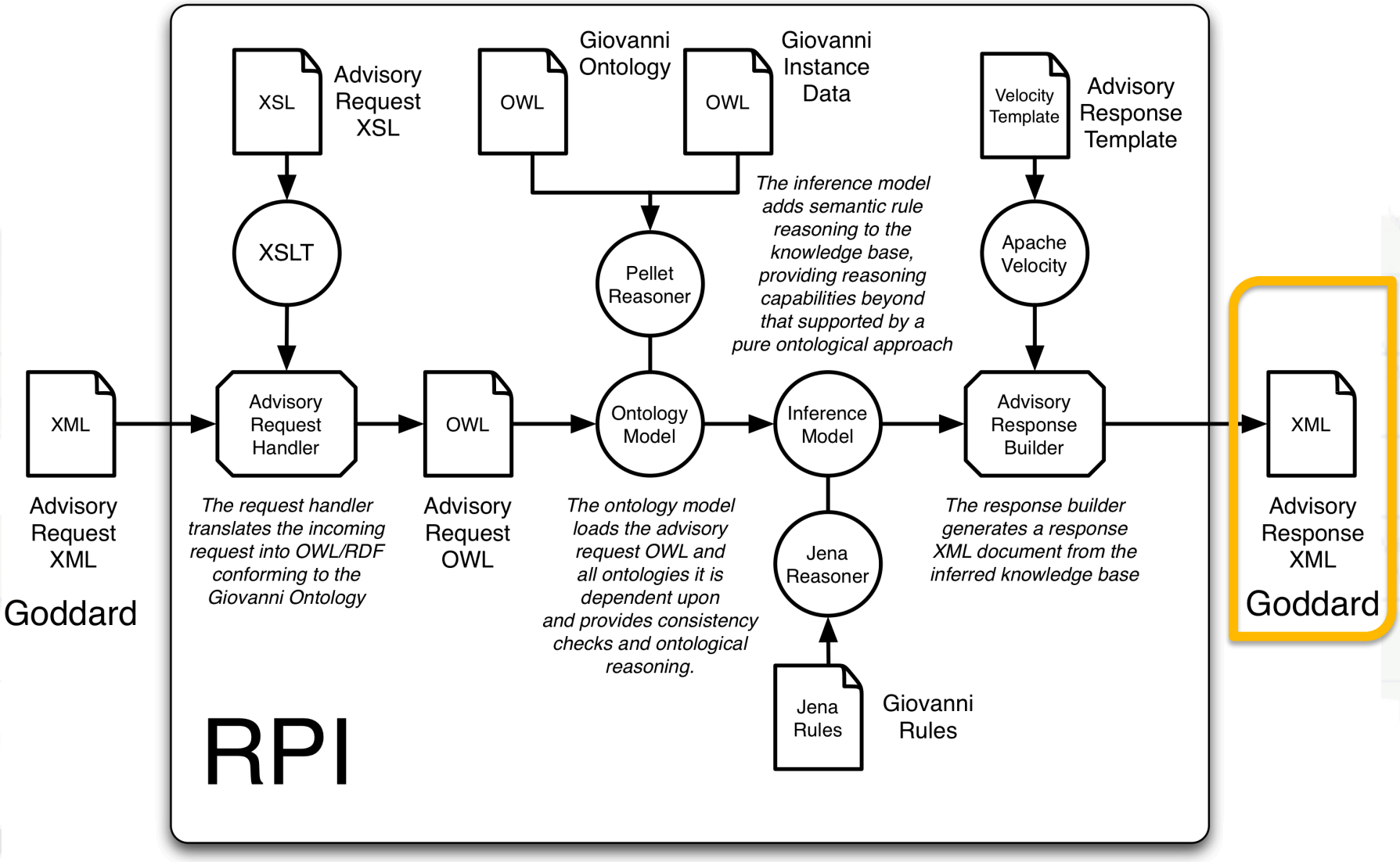
- Rulesets are used to understand the criteria that make up a fitness for purpose or important factors for assessing the concept of a 'set of rules
- Individual rules need to be captured so that the significance of different parameter combinations can be explained / presented to an end user
- Rulesets are needed to indicate to a computer what information important and needs to be propagated in/ to the processing results.

RuleSet Development

[DiffNEQCT:
(?s rdf:type gio:RequestedService),
(?s gio:input ?a),
(?a rdf:type gio:DataSelection),
(?s gio:input ?b),
(?b rdf:type gio:DataSelection),
(?a gio:sourceDataset ?a.ds),
(?b gio:sourceDataset ?b.ds),
(?a.ds gio:fromDeployment ?a.dply),
(?b.ds gio:fromDeployment ?b.dply),
(?a.dply rdf:type gio:SunSynchronousOrbitalDeployment),
(?b.dply rdf:type gio:SunSynchronousOrbitalDeployment),
(?a.dply gio:hasNominalEquatorialCrossingTime ?a.neqct),
(?b.dply gio:hasNominalEquatorialCrossingTime ?b.neqct),
notEqual(?a.neqct, ?b.neqct)
->
(?s gio:issueAdvisory giodata:DifferentNEQCTAdvisory)]



Semantic Advisor

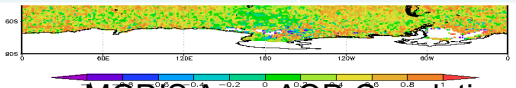


Semantic Advisor

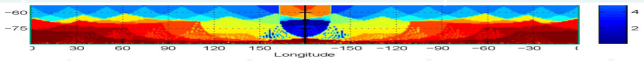
Spatial Area: Longitude (-30, 150), Latitude (-10,60)
Parameters: A: MYD08_D3.005 Aerosol Optical Depth at 550 nm
B: MOD08_D3.005 Aerosol Optical Depth at 550 nm
Temporal Range: Begin Date: Jan 01 2008

Your Selected Options:

	Parameter A	Parameter B	Difference alert
Parameter Name :	Aerosol Optical Depth at 550 nm	Aerosol Optical Depth at 550 nm	
Dataset:	MYD08_D3.005	MOD08_D3.005	← Diff
Data-Day definition	UTC (00:00-24:00Z)	UTC(00:00-24:00Z)	The same but….
Temporal resolution	Daily	Daily	
Spatial resolution	1x1 degree	1x1 degree	
Sensor:	MODIS	MODIS	
Platform:	Aqua	Terra	← Diff
EQCT	13:30	10:30	← Diff
Day Time Node	Ascending	Descending	← Diff
Pre-Giovanni Processes :	*ATBD-MOD-30*	*ATBD-MOD-30*	
Giovanni Processes:	Spatial subset Time average	Spatial subset Time average	



MODIS Terra vs. MODIS Aqua AOD Correlation



Included Overpass time Difference

Continue process to display image

Return to selection page

Final Remarks / Discussion

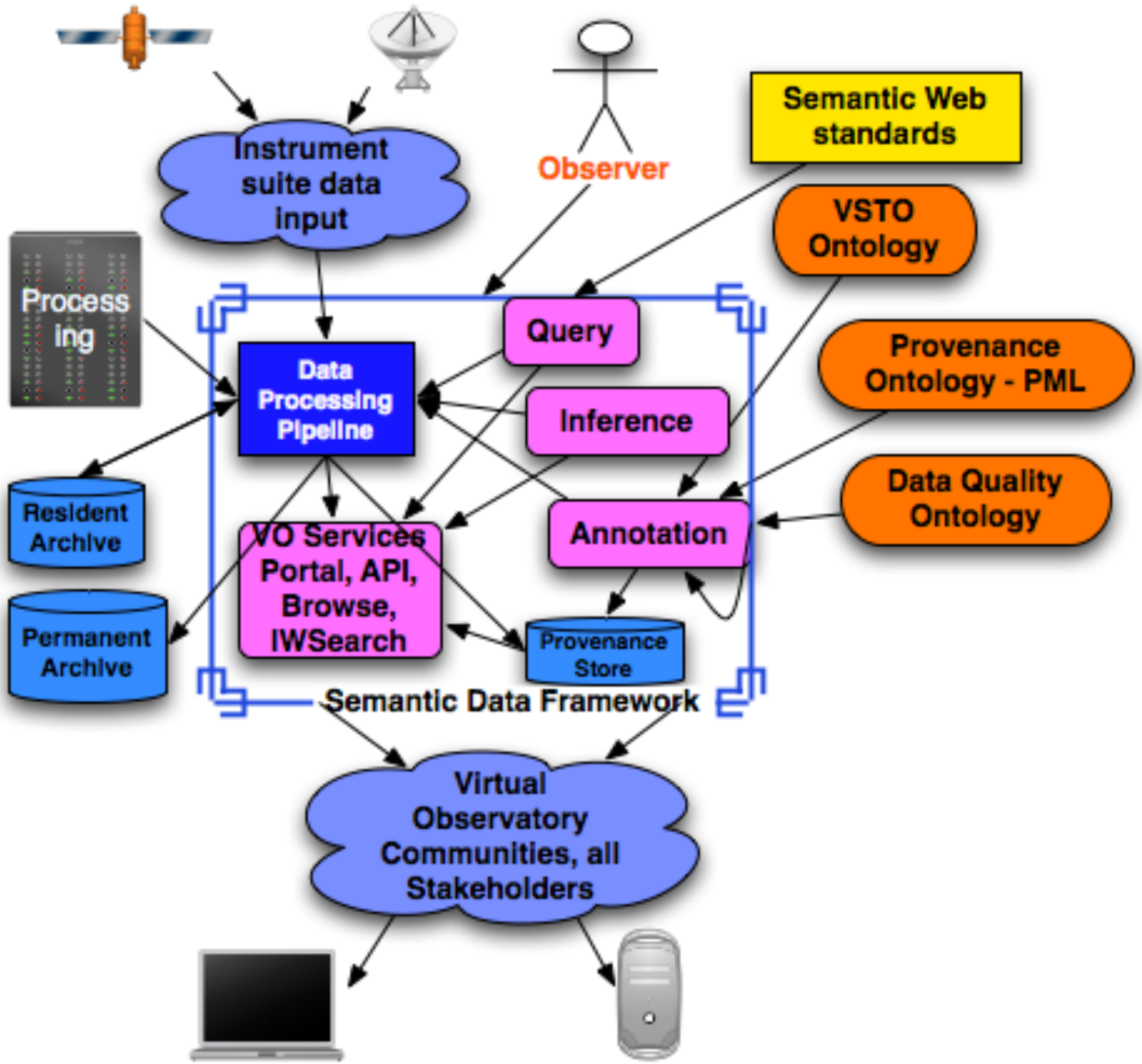
- Integrated Knowledge Provenance and Domain Knowledge Base key to our Use Cases
- PML supports
  - Multi-model individuals (by way of OWL)
  - Causality graphs and justifications
  - Processing history and intent
- Multi-model individuals interoperable with generic PML tools
  - PML is somewhat hard to generate unless you are an expert
  - Tools are needed around this
- Ongoing Steps
  - Investigation of using Rules to infer domain relations from provenance store (and vice versa)
  - Further development on Semantic Faceted Browse
  - Further design on visualization of provenance

Acknowledgements

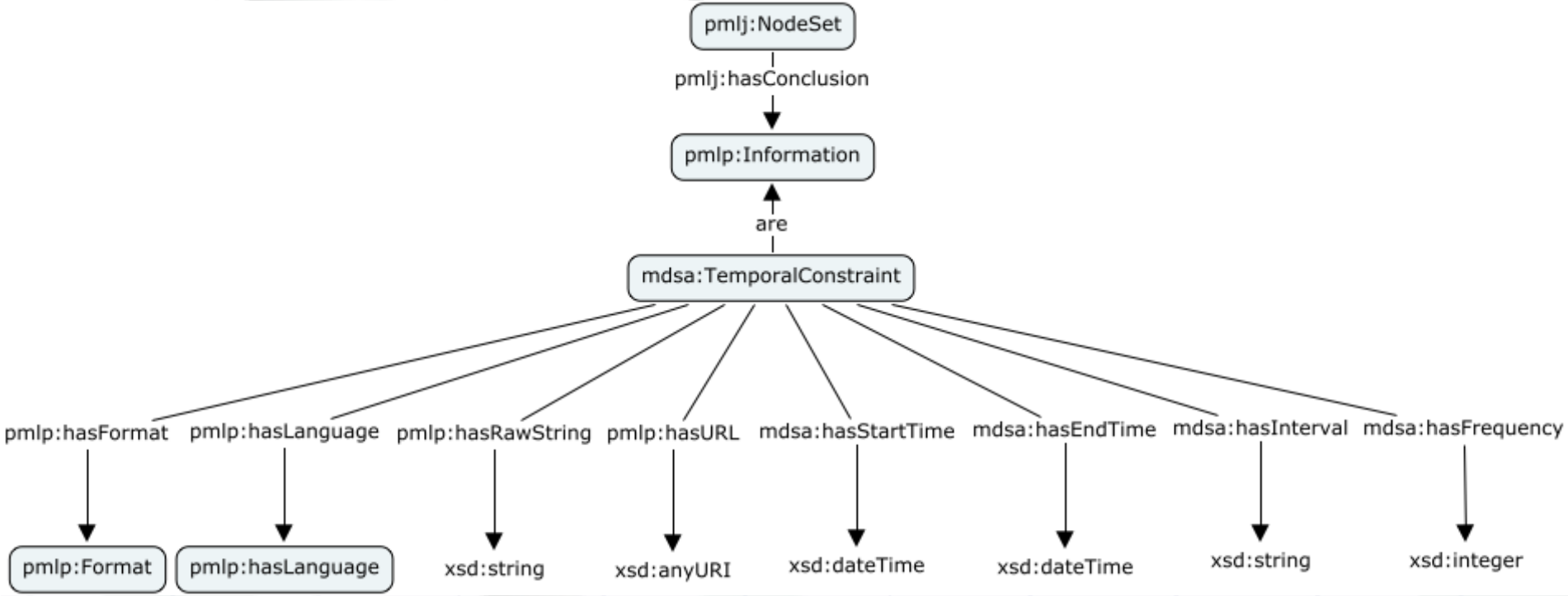
- RPI + Inference Web team
  - Stephan Zednik
  - Deborah McGuinness
  - Patrick West
  - James Michaelis
- NASA/GSFC
  - Gregory Lepkotuh
  - Chris Lynnes
- UTEP/CyberSHARE
  - Paulo Pinheiro da Silva
  - Nicholas del Rio

Links

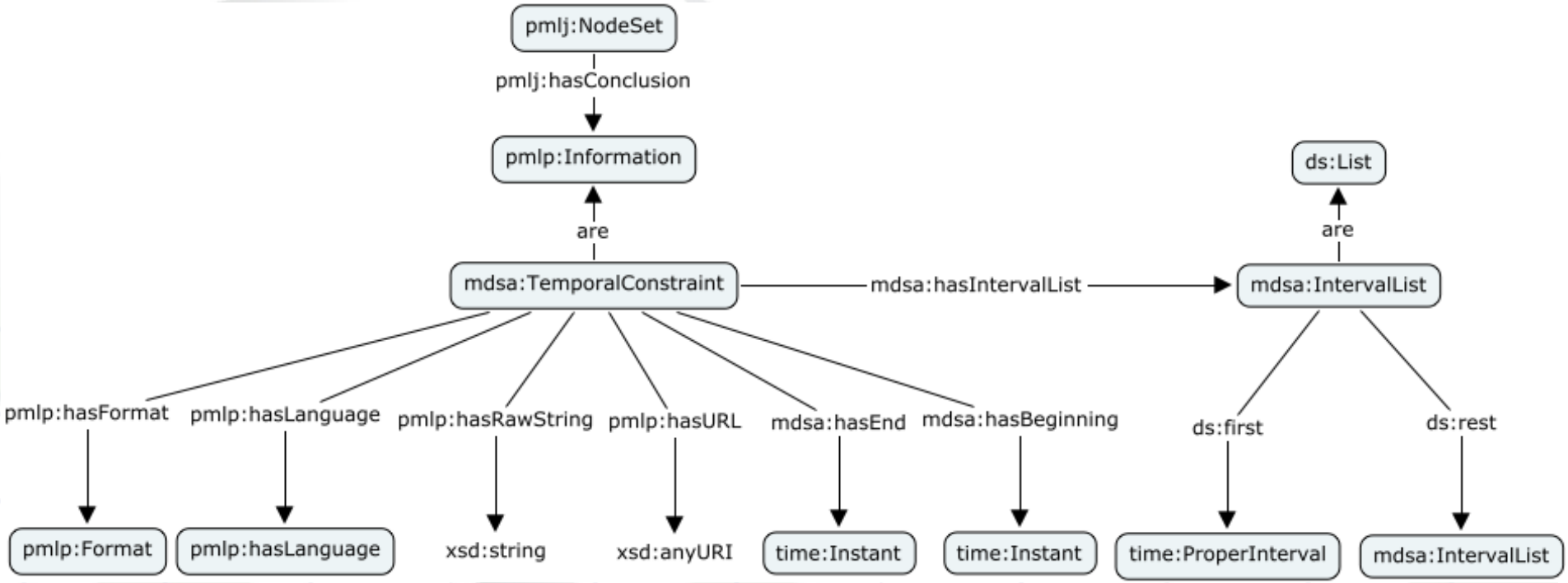
- PML: http://inference-web.org/2007/primer/
- OPM: http://openprovenance.org/
- SWRL: http://www.w3.org/Submission/SWRL/
- Inference Web http://inference-web.org
- Probe-It! http://trust.utep.edu/probe-it/
- SPCDIS http://tw.rpi.edu/portal/SPCDIS
- MDSA http://tw.rpi.edu/portal/MDSA
- Many others…
- Contacts:
  - pfox@cs.rpi.edu
  - zednis@rpi.edu
  - westp@rpi.edu


Instrument suite data input
Observer
Semantic Web standards
VSTO Ontology
Process ing
Query
Data Processing Pipeline
Provenance Ontology - PML
Inference
Data Quality Ontology
Resident Archive
VO Services Portal, API, Browse, IWSearch
Annotation
Permanent Archive
Provenance Store
Semantic Data Framework
Virtual Observatory Communities, all Stakeholders

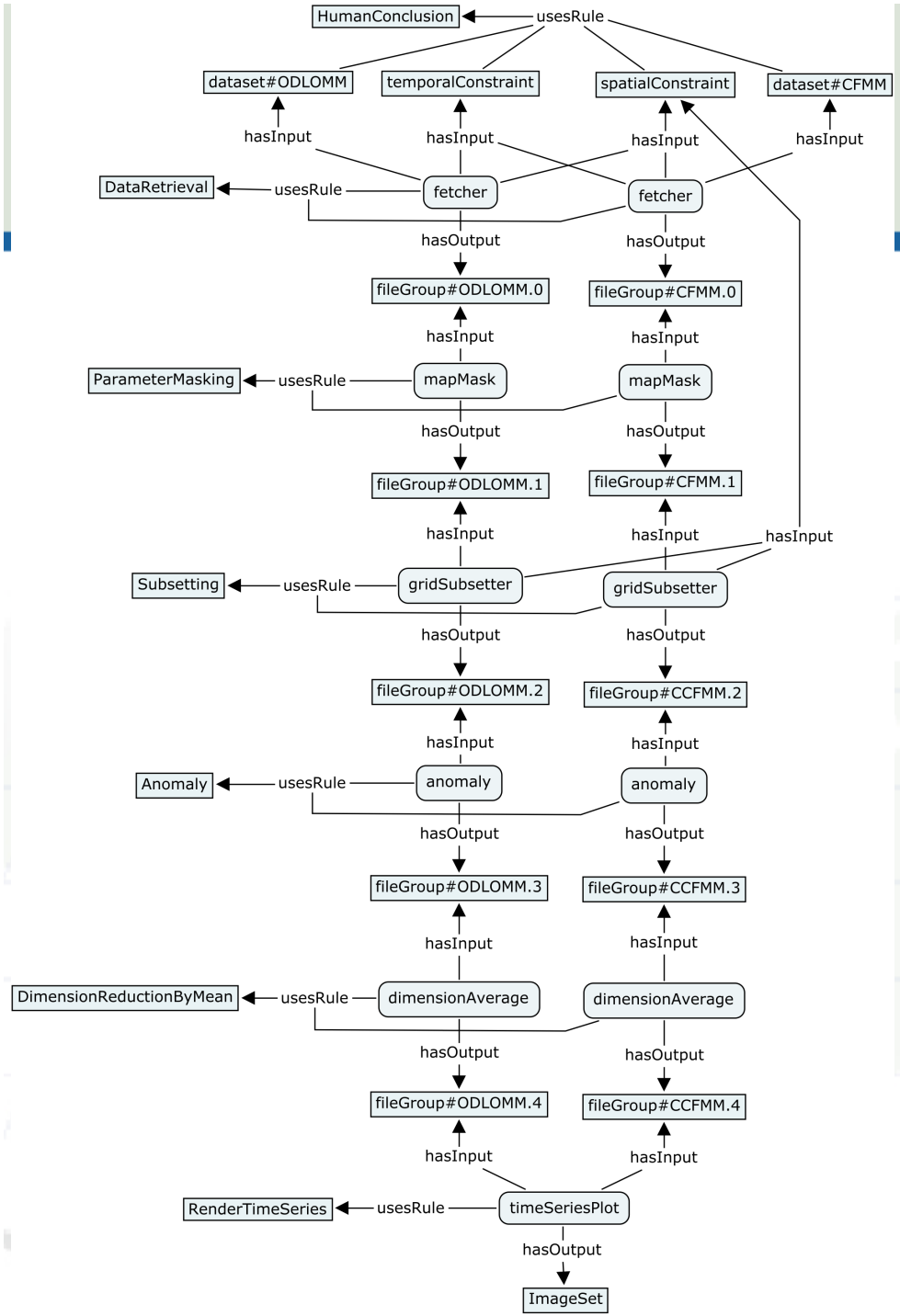
Syntactic Temporal Constraint



Semantic Temporal Constraint



Lineage


HumanConclusion
usesRule
dataset#ODLOMM
temporalConstraint
spatialConstraint
dataset#CFMM
hasInput
DataRetrieval
usesRule
fetcher
fetcher
hasOutput
fileGroup#ODLOMM.0
fileGroup#CFMM.0
ParameterMasking
usesRule
mapMask
mapMask
fileGroup#ODLOMM.1
fileGroup#CFMM.1
Subsetting
usesRule
gridSubsetter
gridSubsetter
fileGroup#ODLOMM.2
fileGroup#CCFMM.2
Anomaly
usesRule
anomaly
anomaly
fileGroup#ODLOMM.3
fileGroup#CCFMM.3
DimensionReductionByMean
usesRule
dimensionAverage
dimensionAverage
fileGroup#ODLOMM.4
fileGroup#CCFMM.4
RenderTimeSeries
usesRule
timeSeriesPlot
ImageSet