ILLUMINATING THE DARKNESS

Exploiting Untapped Data and Information Resources in Earth Science

Dr. Rahul Ramachandran
DAAC Manager, GHRC
NASA/MSFC
rahul.ramachandran@nasa.gov
256-961-7620
Project Team

• Univ. of Alabama in Huntsville
  o Manil Maskey*
  o Xiang Li
• RPI
  o Peter Fox*
  o Stephan Klene
• NASA/GSFC
  o Steve Kempler (Chris Lynnes*)
  o Suhung Shen
  o Chung-Lin Shie
Outline

• Overview of Project
• Use Case Deconstruction
• Initial Results from Data Curation Service
Part 1: Overview
Motivation

• Data preparation steps are cumbersome and time consuming
  o Covers discovery, access and preprocessing
• Limitations of current Data and information
  o Searches on data are boolean searches on instrument or geophysical keywords
  o Underlying assumptions that users have sufficient knowledge of the domain vocabulary
  o Lack support for those unfamiliar with the domain vocabulary or the breadth of relevant data available
Earth Science Metadata: Dark Resources

• *Dark resources* - information resources that organizations collect, process, and store for regular business or operational activities but fail to utilize for other purposes
  
  o Challenge is to recognize, identify and effectively utilize these dark data stores

• Metadata catalogs contain dark resources consisting of structured information, free form descriptions of data and browse images.
  
  o EOS Clearing House (ECHO) holds 3666 data collections, 127 million records for individual files and 67 million browse images.

Premise: Metadata catalogs can be utilized beyond their original design intent to provide new data discovery and exploration pathways to support science and education communities.
Browse Image Example: Understanding regional air pollution from haze

- MODIS 2010 image over India which shows modest level haze pollution is used to drive the search
- How often does Haze occur over Indian subcontinent?

http://rapidfire.sci.gsfc.nasa.gov/cgi-bin/imagery/single.cgi?image=India.A2010345.0510.2km.jpg
Results: Image Retrieval and Metadata

Haze occurs more frequently in Spring than in Summer.

Over half a month in January, haze images were observed in the region.
Goals

• Design a Semantic Middleware Layer (SML) to exploit these metadata resources
  o provide novel data discovery and exploration capabilities that significantly reduce data preparation time.
  o utilize a varied set of semantic web, information retrieval and image mining technologies.

• Design SML as a Service Oriented Architecture (SOA) to allow individual components to be reused and easily integrated into existing NASA’s data and information systems.
Specific Objectives

• Three specific semantic middleware core components
  o Image retrieval service - uses browse imagery to enable discovery of possible new case studies and granule metadata to present analytics results.
  o Data curation service - uses metadata and textual descriptions to find relevant data sets and granules needed to support the analysis of a phenomena or an event.
  o Semantic rules engine - automates data preprocessing and exploratory analysis and visualization tasks.
• Demonstrate value using science use cases

Explore pathways to infuse this technology into existing NASA information and data system
Science Use Cases

• Dust storms, Volcanic Eruptions, Tropical Storms
• Volcanic Eruptions:
  o Emit a variety of gases as well as volcanic ash, which are in turn affected by atmospheric conditions such as winds.
  o Role of Components
    • Image Retrieval Service is used to find volcanic ash events in browse imagery
    • Data Curation Service provides the relevant datasets to support event analysis
    • Rules Engine invokes a Giovanni processing workflow to assemble and compare the wind, aerosol and SO2 data for the vent
Part 2: Use Case Deconstruction

Volcanic Eruptions
**Conceptual Flow and Data Dictionary**

**Phenomena**: As commonly used in weather observing practice, an observable occurrence of particular physical phenomenon.

<table>
<thead>
<tr>
<th>Phenomena</th>
<th>1. Volcanic eruption</th>
<th>2. Hurricane</th>
</tr>
</thead>
</table>

**Event**: Instance of a natural phenomena.

<table>
<thead>
<tr>
<th>Event</th>
<th>1.2008 Chaitén Volcanic eruption, 2. Hurricane Katrina</th>
</tr>
</thead>
</table>

**Physical Manifestation**: feature characteristic, the estimation of which is the purpose of an observation.

<table>
<thead>
<tr>
<th>Physical Manifestation</th>
<th>Volcano: Ash plume, Hurricane: Wind Fields, Eye (Atmospheric Pressure)</th>
</tr>
</thead>
</table>

**Instance** (time and space) of physical manifestation.

<table>
<thead>
<tr>
<th>Instance</th>
<th>1. 2008 Chaitén ash plume, 2. Wind speeds in and around Hurricane Katrina</th>
</tr>
</thead>
</table>

**Measurements (Observable Property)**: How an instrument observes Phenomena.

|-----------------------------------|--------------------------------------------------------------------------------------------------|

**Data Set Variable**: Representation of the measurement in a data file, variables within an actual data file.

<table>
<thead>
<tr>
<th>Data Set Variable</th>
<th>OMSO2e:ColumnAmountSO2_PBL, MOD08:Optical_Depth_Land_and_Ocean_Mean, Precipitation/Visible Frequencies, Pressure</th>
</tr>
</thead>
</table>
Initial Model

```
Initial Model

ex:chaitén_volcanic_eruption_2008 AshPlume
  | is a
  v
dd:Emission

dd:Emission_SO2

dd:ObservableProperty

dd:PhysicalManifestation

dd:VolcanicEruption

dd:DataVariableObservation

om:Observation

om:FeatureOfInterest

om:Process

om:Procedure

om:Method

om:Parameter

om:InnerValue

om:Instant

dd:AerosolOpticalDepth

dd:AtmosphericConcentration

dd:Radiance

dd:Temperature

ex:chaitén_volcanic_eruption_2008_ash_plume

ex:atmospheric_concentration_SO2

ex:surface_temperature

ex:infrared_radiance

ex:visible_radiance

ex:aerosol_optical_dept_thickness
```
Volcanic Eruption: Chaitén 2008

The Chaitén Volcano seen from a commercial flight, October 2008. It was into eruptive phase for the first time in about 9,500 years on the morning of May 2, 2008.

Eruption Time period: May 2 – Nov 2008
Location: Andes region, Chile ( -42.832778, -72.645833)
Browse Images

Band 1-4-3 (true color)  Band 7-2-1  LST

Example: MODIS-Aqua 2008-05-03 18:45 UTC

http://lance-modis.eosdis.nasa.gov/cgi-bin/imagery/realtime.cgi?date=2008124
Example Relevant Data

**Total SO₂ mass:**
e.g. *Chaitén* is 10 (kt) = (kilotons), (1kt = 1000 metric tons)
ftp://measures.gsfc.nasa.gov/data/s4pa/SO2/MSVOLSO2L4.1/
MSVOLSO2L4_v01-00-2014m1002.txt

**Daily SO₂:**
OMI/Aura Sulphur Dioxide (SO₂) Total Column Daily L2 Global 0.125 deg
http://disc.sci.gsfc.nasa.gov/datacollection/OMSO2G_V003.html

**Calibrated Radiances:**
MODIS/Aqua Calibrated Radiances 5-Min L1B Swath 1km
http://dx.doi.org/10.5067/modis/myd021km.006

**Aerosol Optical Thickness:**
MODIS/Aqua Aerosol 5-Min L2 Swath 10km
http://modis-atmos.gsfc.nasa.gov/MOD04_L2/
SeaWiFS Deep Blue Aerosol Optical Depth and Angstrom Exponent Level 2
Data 13.5km
http://disc.gsfc.nasa.gov/datacollection/SWDB_L2_V004.shtml

**IR Brightness Temperature:**
NCEP/CPC 4-km Global (60 deg N - 60 deg S) Merged IR Brightness
Temperature Dataset
Giovanni SO2 Plots

MODIS-Aqua 2008-05-03 18:45 UTC

MODIS-Aqua 2008-05-05 18:30 UTC

http://gdata2.sci.gsfc.nasa.gov/daac-bin/G3/gui.cgi?instance_id=omil2g
Giovanni Infrared Data Plot

MODIS-Aqua 2008-05-03 18:45 UTC

MODIS-Aqua 2008-05-05 18:30 UTC

http://disc.sci.gsfc.nasa.gov/daac-bin/hurricane_data_analysis_tool.pl
Part C: Data Curation
Algorithm for Phenomena

Initial Results
Data Curation Algorithm Approaches

- **Text mining**
  - Pros: Don’t need to explicitly define the phenomena
  - Cons: Dependent of the truth set; Catalog is dynamic and new data may never get classified

- **Ontology Based**
  - Pros: Best precision and recall
  - Cons: Labor intensive to build an explicit model

- **Information Retrieval**
  - **Boolean (Faceted) Search**
    - Pros: Simple to implement
    - Cons: Phenomena can be complex; User may not know all the right keywords
  - **Relevancy Ranking Algorithm**
    - Pros: List most relevant data first
    - Cons: Requires a custom algorithm
Assumptions/Observations

- Catalog metadata (ECHO) is rich and all metadata records have been tagged with appropriate vocabulary terms (GCMD)

- A phenomena can be defined using a bag of keywords using vocabulary terms
  - Information need can be captured by using a broad query

- Keywords (tags) in the metadata and the unstructured text (description) can be used

- Keyword is only used once per metadata record
  - Term frequency does not matter

- Document frequency for keywords can be used
  - Some keywords may occur in many metadata records
Experiment Setup and Approach

- Randomly select 200 sample dataset metadata from ECHO
- Label 200 datasets
  - binary: relevant to phenomena/not relevant to phenomena (Hurricane)
- Compile set of keywords (GCMD) relevant to Hurricane – “bag of words” model

- Filter
  - Spatial filter
  - Temporal resolution
    - “<= daily”
  - 85 datasets filtered out
- Apply algorithms on remaining 115 datasets
  - Jaccard coefficient-based ranking
  - Vector Space Model using Cosine similarity-based ranking
**Algorithms**

**Jaccard Coefficient**

\[ J(A, B) = \frac{|A \cap B|}{|A \cup B|} \]

Where:
- \( A \) - keywords defining a phenomena
- \( B \) - keywords in a given dataset

**Vector Space Model**

- Determine term frequency (\( tf \)): (1 in our case)
- Determine inverse document frequency (\( idf \)): number of metadata records that contain the keyword
- Calculate Cosine similarity
  - Sum (\( tf \times idf \)) for each keyword
90% precision with a 70% recall:
70% of the relevant data are retrieved with 90% precision
Questions