EXPLOITING UNTAPPED INFORMATION RESOURCES IN EARTH SCIENCE

Rahul Ramachandran NASA/MSFC, Peter Fox RPI, Steve Kempler NASA/GSFC and Manil Maskey UAH

MSFC/UAH: Patrick Gatlin, Xiang Li, Amanda Weigel, JJ Miller, Kaylin Bugbee, Ajinkya Kulkarni
GSFC: Chris Lynnes, Suhung Shen, Chung-Lin Shie, Maksym Petrenko
RPI: Stefan Zednik, Anirudh Prabhu
Outline

1. Project Overview
2. Data Curation Service
3. Rules Engine
4. Image Retrieval Service
5. Summary
Part 1: Project Overview
Motivation

- Data preparation steps are **cumbersome** and **time consuming**
  - Covers discovery, access and preprocessing
- Limitations of current Data/Information Systems
  - **Boolean search** on data based on instrument or geophysical or other **keywords**
  - Underlying **assumption** that users have sufficient knowledge of the **domain vocabulary**
  - 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
  - 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.
  - EOS Clearing House (ECHO) holds >6000 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.
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 used by existing systems
Science Use Cases

- Dust storms, Volcanic Eruptions, Tropical Storms/Hurricanes
- Volcanic Eruptions:
  - Emit a variety of gases as well as volcanic ash, which are in turn affected by atmospheric conditions such as winds.
  - Role of Components
    - **Image Retrieval Service** is used to find volcanic ash events in browse imagery
    - **Data Curation Service** suggests 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 event
Find Events: Browse Images

Chaitén Volcano Eruption
Eruption Time period: May 2 – Nov 2008
Location: Andes region, Chile ( -42.832778, -72.645833)
Suggest Relevant Data

**Total SO$_2$ 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
Generate 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
Generate 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
Conceptual Model

- **Phenomena**
  - Event type

- **Physical Feature**
  - Manifestation / Driver of phenomena
  - Has space/time extent
  - Can precede or linger after what is generally thought of as the phenomena event

- **Observable Property**
  - Characteristic/property of physical feature

- **Data Variable**
  - Measurement/estimation of observable feature
Part 2: Data Curation Algorithm for Phenomena

Initial Results
Objectives

- Design a data curation (relevancy ranking) algorithm for a set of **phenomena**
- Provide the data curation algorithm as a stand alone service

- Envisioned Use:
  - Given a phenomenon type (Ex: Hurricane), DCS returns a list of relevant data sets (variables)
    - \(<\text{data of data sets}> = \text{DCS}(\text{Phenomenon Type})\)
  - For a specific phenomenon instance (event: Hurricane Katrina), these curated datasets can then be filtered based on space/time to get actual granules
Overview
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 and map to instances

- **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*
Data Search for Earth Science Phenomena

USER TASK

INFO NEED

QUERY

SEARCH ENGINE

DOCUMENT COLLECTION

RESULTS

REFINE

Study “Hurricane”

All data sets useful in studying “Hurricane”

By pass this step for the end user

How to define a phenomenon?

How to automatically formulate query?

Best relevancy ranking algorithm?
Relevancy Ranking

• Search: Curation problem

• Data curation: **Relevancy ranking** service for a set of Earth science *phenomena*
Relevancy Ranking: Initial Exploratory Experiments

• Approaches tested:
  o O-Rank (Top down approach)
  o Wikipedia Terms
  o Manual Terms Experiment
  o Latent Semantic Index – Dual Set terms
  o Metadata based Ranking

• Key Takeaways:
  o Best results: three approaches where terms describing the phenomenon manually constructed after exploring metadata records
  o Both ontology and automated term construction (Wiki) approaches don’t map well to metadata terms/descriptions
**Follow-on Experiments: Approach**

- **USER TASK**
  - Study “Hurricane”
  - All data sets useful in studying “Hurricane”

- **INFO NEED**
  - How to define a phenomenon?
  - Expert select “bag of words” to define a phenomena
  - Control vocabulary (GCMD) is used for the “words”

- **QUERY**
  - How to automatically formulate query?

- **SEARCH ENGINE**
  - Best relevancy ranking algorithm?
  - Use well known alg: Jaccard Coef, Cosine Similarity, Zone Ranking

- **DOCUMENT COLLECTION**
  - REFINE
  - RESULTS
Assumptions/Observations

• Metadata quality
  o Richness
  o Vocabulary
  o Tags

• Earth Science Phenomena can be defined using a bag of keywords
Experiment Setup

- Datasets: 200 Randomly selected from ECHO
  - Binary Labeling: relevancy to phenomena (Label = Majority: 3 Scientists)

- Labeled Datasets

- Temporal and Spatial Filtering

- Bag-of-keywords
  - 92 keywords: hurricane
  - 103 Keywords: volcano
  - Manually Selected set of GCMD Science Keywords relevant to phenomena

- Jaccard Coeff
- Cosine Similarity
- Zone Ranking
- Ensemble
Top 20 returns (Hurricane)
Next: Find relevant data fields

• Dataset is relevant
  o now what?
  o how do I use the granules for the dataset?

• Need actual data variable name
  o for example: Giovanni uses these fields for visualization

• What we know
  o relevant science keywords (GCMD) – Experts
  o granule data fields and metadata – Auto extract*

• How do we map?
  o manually? May work for few datasets only
    • Hundreds of data variables per granule
  o start with GCMD to CF Standard name
  o most don’t follow CF Standard names
Approach

Dataset

- Extract Science Keywords
- Text processing
- Normalization
- Bag-of-words

Granules

- Extract Variables and Descriptions
- Text processing
- Look up Table
- Normalization
- Bag-of-words

- OPeNDAP, netCDF Libs, ...
- Remove special characters, Tokenize, ...
- Acronym/Abbreviation expansion, CF
- Remove stopwords/Stem/Lemmatize

NLP
Learn Patterns

Suggest Keywords
Assess Metadata

Intersection
Example: GLAS/ICESat L2 Global Thin Cloud/Aerosol Optical Depths Data (HDF5) V033 – Dataset Metadata

John.P.Dimarzio.1@nasa.gov

ICESat Science Investigator-led Processing System (I-SIPS)
757-864-1238 (phone)
David.W.Hancock@nasa.gov

NASA DAAC at the National Snow and Ice Data Center
303-492-6199 (phone)
303-492-2468 (fax)
nsidc@nsidc.org

Science Keywords:
Earth Science  Atmosphere  Clouds
Earth Science  Atmosphere  Aerosols
Example: GLAS/ICESat L2 Global Thin Cloud/Aerosol Optical Depths Data (HDF5) V033

Sample file: GLAH11.033/2006.10.25/GLAH11_633_2117_001_1275_0_01_0001.H5

Data Variables
Example: GLASICESat L2 Global Thin Cloud Aerosol Optical Depths Data (HDF5) V033

Science keyword to variable mapping

- `r_Surface_relh` | Surface Relative Humidity
  - No match
- `r_Surface_temp` | Surface Temperature
  - No match
- `r_Surface_wind` | Surface Wind Speed
  - No match
- `r_cld1_od` | Cloud Optical Depth at 532 nm
  - Score=3 keyword: ATMOSPHERE->CLOUDS->CLOUD OPTICAL DEPTH/THICKNESS
  - Score=2 keyword: ATMOSPHERE->AEROSOLS->AEROSOL OPTICAL DEPTH/THICKNESS

Variable to keyword mapping

- ATMOSPHERE->CLOUDS->CLOUD OPTICAL DEPTH/THICKNESS
  - Score=3 name: `r_cld_ir_OD` | Cloud Optical Depth at 1064 nm
  - Score=3 name: `i_cld1_qf` | Cloud optical depth flag for 532 nm
  - Score=3 name: `i_cld1_uf` | Cloud optical depth flag for 532 nm
  - Score=3 name: `r_cld1_od` | Cloud Optical Depth at 532 nm
  - more with low scores

This approach can be used to assess metadata quality and also suggest keyword annotation!!
Part 3: Rules Engine
What Settings should I use to visualize this event?

Goal: Automate data preprocessing and exploratory analysis and visualization tasks
Strategy

• Service to generate and rank candidate workflow configurations

• Use rules to make assertions about compatibility based on multiple factors
  o does this data variable make sense for this feature?
  o does this visualization type make sense for this feature?
  o does the temporal / spatial resolution of this dataset make sense for this feature?

• Each compatibility assertion type is assigned weights.
  o ex: Strong = 5, Some = 3, Slight = 1, Indifferent = 0, Negative = -1.

• Based on the aggregated compatibility assertions, we calculate the score for each visualization candidate.
# Phenomena Feature Characteristic Mappings

<table>
<thead>
<tr>
<th>Phenomena</th>
<th>East-West Movement</th>
<th>North-South Movement</th>
<th>Temporal Evolution</th>
<th>Spatial Extent of Event</th>
<th>Year-to-Year Variability</th>
<th>May Impact Seasonal Variation</th>
<th>Variation with Atmospheric Height</th>
<th>Global Phenomena</th>
<th>Detection of Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volcano - Ash Plume</td>
<td>Indifferent</td>
<td>Indifferent</td>
<td>Strong</td>
<td>Slight</td>
<td>Strong</td>
<td>Strong</td>
<td>Strong</td>
<td>Strong</td>
<td>Strong</td>
</tr>
<tr>
<td>Flood</td>
<td>Some</td>
<td>Some</td>
<td>Strong</td>
<td>Some</td>
<td>Some</td>
<td>Strong</td>
<td>Some</td>
<td>Slight</td>
<td>Some</td>
</tr>
<tr>
<td>Dust Storm</td>
<td>Strong</td>
<td>Strong</td>
<td>Strong</td>
<td>Strong</td>
<td>Indifferent</td>
<td>Indifferent</td>
<td>Strong</td>
<td>Indifferent</td>
<td>Some</td>
</tr>
</tbody>
</table>
# Service to Characteristic Mappings

<table>
<thead>
<tr>
<th>Service</th>
<th>Visualization</th>
<th>East-West Movement</th>
<th>North-South Movement</th>
<th>Temporal Evolution</th>
<th>Spatial Extent of Event</th>
<th>Year-to-Year Variability</th>
<th>Seasonal Variation</th>
<th>Variation with Atmospheric Height</th>
<th>Global Phenomena</th>
<th>Detection of Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time-averaged Map</td>
<td>Color-Slice Map</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Area-averaged Time Series</td>
<td>Time Series</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>User-defined Climatology</td>
<td>Color-Slice Map</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Vertical Profile</td>
<td>Line Plot</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Seasonal Time Series</td>
<td>Time Series</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Zonal Means</td>
<td>Line Plot</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Hovmoller (Longitude)</td>
<td>Color-Slice Grid</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Hovmoller (Latitude)</td>
<td>Color-Slice Grid</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>
Compute Compatibility

Phenomena: Volcano - Ash Plume

Service - Area Averaged Time Series

<table>
<thead>
<tr>
<th>Temporal Evolution</th>
<th>Detection of Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong</td>
<td>Strong</td>
</tr>
</tbody>
</table>

Area Averaged Time Series: bestFor → Temporal evolution; Detection of events

Next Steps

• Generate rules for compatibility assertions based on
  o data variables
  o temporal / spatial resolution
  o dataset processing

• Explore additional strategies for making compatibility assertions
Part 4: Image Retrieval

Initial Results
Image Retrieval

• Goal: given an image of Earth science phenomenon retrieve similar images

• Challenge: “semantic gap”
  - low-level image pixels and high-level semantic concepts perceived by human
Image retrieval approaches

- Tradition approaches
  - Image features: Color, Texture, Edge histogram...
  - “Shallow” architecture
  - User defines the feature
  - Preliminary experiments

- State of the Art approach
  - Generic
  - No need for domain expert
Deep Learning

• Mimics the human brain that is organized in a deep architecture
• Processes information through multiple stages of transformation and representation
• Learns complex functions that directly map pixels to the output, without relying on human-crafted features

Convolution neural network

Source: Google Research, CVPR 2014
Experiment Setup

- NASA rapid response MODIS imageries
- 600 imageries
- 3 phenomena – Hurricane, Dust, Smoke/Haze
- Train half images with Convolutional Neural Network
- Test
Sample rapid response images

Hurricane  Smoke  Dust
4 layers

- Used number of filters in each layer = 100, 200, 400, 800
- Convolved and Pooled on every layer
- Overall accuracy ~ = as that of 5 layers (slightly better than 6 layers)
## Error Matrix

4 layers, learning rate = 0.003

<table>
<thead>
<tr>
<th>True\Pred</th>
<th>Others</th>
<th>dust</th>
<th>Haze/smoke</th>
<th>Hurricane</th>
</tr>
</thead>
<tbody>
<tr>
<td>Others</td>
<td>173</td>
<td>20</td>
<td>76</td>
<td>7</td>
</tr>
<tr>
<td>dust</td>
<td>29</td>
<td>128</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>Haze/Smoke</td>
<td>79</td>
<td>3</td>
<td>207</td>
<td>7</td>
</tr>
<tr>
<td>Hurricane</td>
<td>30</td>
<td>2</td>
<td>3</td>
<td>163</td>
</tr>
</tbody>
</table>
Accuracy Numbers

Producers Accuracy
• Other: 173/311 = 55.6%
• Dust: 128/153 = 83.7%
• Smoke: 207/309 = 67%
• Hurricane: 163/177 = 92.1%

Users Accuracy
• Other: 173/276 = 62.7%
• Dust: 128/180 = 71.1%
• Smoke: 207/296 = 69.9%
• Hurricane: 163/198 = 82.3%

Overall accuracy ~ 70.6%
Summary

• Build three specific semantic middleware core components
  o Image retrieval service - uses browse imagery to enable discovery of possible new case studies and also presents exploratory analytics.
  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 a topic.
  o Semantic rules engine - automates data preprocessing and exploratory analysis and visualization tasks.

Explore pathways to infuse these components into existing NASA information and data system