ILLUMINATING THE DARKNESS: EXPLOITING UNTAPPED DATA AND INFORMATION RESOURCES IN EARTH SCIENCE

Manil Maskey* UAH, Rahul Ramachandran NASA/MSFC, Peter Fox RPI, Steven Kempler NASA/GSFC

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

Earth Science Technology Forum, 2016
Earth Science Metadata: Dark Resources

• **Dark resources**: not used beyond **intended** purpose
  - Challenge: recognize, identify and effectively utilize for other **purposes**

• **Metadata catalogs**:
  - contain dark resources
  - structured information
  - free form descriptions: data and browse images

• **NASA’s Common Metadata Repository**
  - > 6000 data collections
  - 270 million records for individual files
  - 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 Earth science and education communities.
Project Goals

• Design a Semantic Middleware Layer (SML) to exploit 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.
  o automate

• Design SML as a Service Oriented Architecture
Data Curation Service

• Relevancy ranking algorithm for a set of *phenomena*
• Stand alone service

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

USER TASK

INFO NEED

Query

SEARCH ENGINE

DOCUMENT COLLECTION

RESULTS

REFINE

Study “Hurricane”

All data sets useful in studying “Hurricane”

How to define a phenomenon?

How to automatically formulate query?

Best relevancy ranking algorithm?
Rules Engine: What settings should I use to visualize this event?

Goal: Automate data preprocessing and exploratory analysis and visualization tasks
Compute Compatibility

Use rules to make **assertions** about **compatibility** based on multiple factors

---

**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>

<table>
<thead>
<tr>
<th>Area Averaged Time Series</th>
<th>Temporal evolution; Detection of events</th>
</tr>
</thead>
<tbody>
<tr>
<td>bestFor →</td>
<td></td>
</tr>
</tbody>
</table>

---

Service to generate and rank candidate workflow configurations

---


Volcanic ash image - By Boaworm (Own work) [CC BY 3.0](http://creativecommons.org/licenses/by/3.0), via Wikimedia Commons
User needs to decide:
- Variable(s)
- Time
- Space
- Plot type

http://giovanni.sci.gsfc.nasa.gov/giovanni/
Giovanni – Dark Data Edition

Rules Service: highlights suitable plots based on selected event & variables

Curation Service: event type filters relevant variables

Selected event & its time  Event Client

Selected Date Range (UTC): 2015-07-31 00:00:00 to 2015-11-24 23:59:59
Valid Range: 2000-03-01 to 2016-01-01

Select Variables:
- Events (all)
- Events (by products)
  - Hurricane (14)
  - Volcano (14)
- Events (by variables)
  - Hurricane (2)
  - Volcano (2)

Number of matching Variables: 14 of 905  Total Variables included in Plot: 1

Variable Filter:
- Sea and Lake Ice
- Snow
- Temperature
- Extremes
- Volcanoes
- Water Oe.

Event Types:
- Landslides
- Manmade
- Severe Storms
- Snow
- Temperature
- Extremes

Event:
- Calbuco Volcano, Chile
- Cotopaxi Volcano, Ecuador
- Manam Volcano
- Masaya Volcano, Nicaragua
- Momotombo Volcano, Nicaragua
- Mount Etna Volcano, Italy
- Ruang Volcano, Indonesia, July 2015

Done  Clear Event Selection

Help  Reset  Feedback
Plot Data  Go to Results
Giovanni - Dark Data Edition

Event Discovery
- Dust
- Volcano
- Etna
  - Calbuco
- Fire
- Hurricane
- Flood

EONET

Event Analysis Workflow

Event Analysis
Calbuco 2015-04-22
- Turbulent Fluxes - 0.9
- Wind stress - 26
  - Time Series
  - Hovmöller
- Heat Flux - 21
  - Vertical Profile
- MODIS Aerosols - 0.8
  - AOD - 15
  - Zonal Mean

Data Curation Service
Rules Service
Giovanni Analysis Service
Serendipitous Discovery

**Data Curation Goal:** map dataset keywords to granule variables

**Application of Data Curation for Operational Use:**
Data Curation Algorithm can be used to assess
- Metadata quality for both dataset and granules
  - Find incorrect/incomplete keyword annotations
- Automatically suggest science keywords
Operational Use: Prototype Variable Mapping

Mapping Scores Generated by Algorithm

Opportunity to develop this prototype and infuse into operational use at DAACs to improve metadata quality
SML Components

GIOVANNI DATA ANALYSIS TOOL

OTHER CLIENTS

SEMANTIC MIDDLEWARE

IMAGE RETRIEVAL/CLASSIFICATION

DATA CURATION SERVICE

RULES ENGINE

PROCESSING ENGINE

BROWSE IMAGES

METADATA DATABASE

ONLINE DATA REPOSITORIESS
Image Classification

- Goal: label images in archives with a known Earth science phenomenon

- Challenge: “semantic gap”
  - low-level image pixels and high-level semantic concepts perceived by humans
“Deep” Architecture

• Features are key to recognition
• What about learning the features?
• Deep Learning
  o Hierarchical Learning
  o Mimics the human brain that is organized in a deep architecture
  o Multiple stages of representation

Convolutional Neural Network (CNN)
- Applicable to Images
- Supervised
Transfer Learning

- CNN requires large number of parameters
- Transfer learning
  - Use internal representation learned from one classification task to another
- Faster learning
- Better accuracy
Applications: Searching for Events

- Detection of phenomena in Browse Imagery

<table>
<thead>
<tr>
<th>True/Pred</th>
<th>Dust</th>
<th>Hurricane</th>
<th>Smoke</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dust</td>
<td>287</td>
<td>8</td>
<td>32</td>
<td>33</td>
</tr>
<tr>
<td>Hurricane</td>
<td>0</td>
<td>379</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Smoke</td>
<td>12</td>
<td>12</td>
<td>443</td>
<td>9</td>
</tr>
<tr>
<td>Other</td>
<td>33</td>
<td>9</td>
<td>23</td>
<td>211</td>
</tr>
</tbody>
</table>

Confusion Matrix

Overall Accuracy = 87.88%

Hurricane – True Positive
Dust – True Positive
Smoke – True Positive
**Applications: Enabling new science**

- Dust climatology – Collaboration with Sundar Christopher, UAH Atmospheric Science Professor

<table>
<thead>
<tr>
<th>True\Predicted</th>
<th>Dust</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dust</td>
<td>1379</td>
<td>379</td>
<td>1758</td>
</tr>
<tr>
<td>Other</td>
<td>260</td>
<td>4932</td>
<td>5192</td>
</tr>
<tr>
<td></td>
<td>1639</td>
<td>5311</td>
<td>6950</td>
</tr>
</tbody>
</table>

**Confusion Matrix**

Based on GIBS

**Validation Accuracy = 91%**
Applications: Improving forecast operations

- Hurricane intensity estimation - Collaboration with Dan Cecil, NASA/MSFC Atmospheric Scientist

<table>
<thead>
<tr>
<th>True\Predicted</th>
<th>td</th>
<th>ts</th>
<th>h1</th>
<th>h2</th>
<th>h3</th>
<th>h4</th>
<th>h5</th>
<th>no_cat</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>td</td>
<td>3168</td>
<td>335</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>3510</td>
</tr>
<tr>
<td>ts</td>
<td>489</td>
<td>4823</td>
<td>159</td>
<td>5</td>
<td>11</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>5496</td>
</tr>
<tr>
<td>h1</td>
<td>9</td>
<td>484</td>
<td>1158</td>
<td>92</td>
<td>20</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1770</td>
</tr>
<tr>
<td>h2</td>
<td>3</td>
<td>76</td>
<td>214</td>
<td>513</td>
<td>145</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>960</td>
</tr>
<tr>
<td>h3</td>
<td>6</td>
<td>40</td>
<td>33</td>
<td>155</td>
<td>689</td>
<td>55</td>
<td>0</td>
<td>0</td>
<td>978</td>
</tr>
<tr>
<td>h4</td>
<td>1</td>
<td>18</td>
<td>17</td>
<td>12</td>
<td>142</td>
<td>810</td>
<td>32</td>
<td>0</td>
<td>1032</td>
</tr>
<tr>
<td>h5</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>27</td>
<td>59</td>
<td>216</td>
<td>0</td>
<td>306</td>
</tr>
<tr>
<td>no_cat</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>32</td>
<td>54</td>
<td>14106</td>
</tr>
</tbody>
</table>

Overall Accuracy : 81 % (Top 2 Probabilities 95.73%)

Data: NRL Images, HURDAT
Ongoing Work

Browse Image-based Event Explorer

Arabian Dust Event, 07/24/2013

GIBS Tile

SML
Data Curation
Rule Engine
Giovanni
Journal Publications

• Submitted:

• In Progress:
  o Relevancy Algorithm to Curate Earth Science Data for Different Phenomena – to be submitted to Computers and Geoscience

• Planned:
  o Dust climatology (collaboration with Sundar Christopher)
  o Hurricane intensity estimation (collaboration Dan Cecil)
Thanks to NASA ESTO for their support
Contact

Manil Maskey
manil.maskey@nasa.gov

Rahul Ramachandran
rahul.ramachandran@nasa.gov