Developing a Knowledge Base for NASA Earth Science and Hydrologic Applications

Amanda Weigel¹,², Patrick Gatlin¹,³, Rahul Ramachandran¹,³, JJ Miller¹,², Manil Maskey¹,³, Jia Zhang¹,³, Emily Berndt³

NASA/MSFC Data Science Informatics Group¹
University of Alabama in Huntsville²
NASA Marshall Space Flight Center³
Carnegie Mellon University⁴
What common challenges do Earth science and hydrologic data users encounter?

- Data **discovery**
- Data **use**
- Identifying **key resources** about the data.
  - Accessing introductory material (for unfamiliar users).
- Determining what **methods** to use – data processing, quality control and analysis.
Data Use Challenges

To address these challenges, what difficulties are presented?

How can data and resources be linked in order to improve the data spin-up time?

How can we work to educate unfamiliar users?
Data Use Challenges

To address these challenges, what difficulties are presented?

How can data and resources be linked in order to improve the data spin-up time?

How can we work to educate unfamiliar users?

- Applied: I better stop the car!
- Context: The traffic light I am driving towards has turned red
- Meaning: South facing traffic light on corner of Pitt and George Streets has turned red
- Raw: Red, 192.234.235.245.678, v2.0

© 2011 Angus McDonald
What is a Knowledge Base?

• Think “Google Search”.
• Developed by Google in 2012 to enhance the results of its search engine by systematically linking information.
• Aggregates structured and detailed information about a defined topic.
• Enables users to resolve their query without having to navigate and assemble information manually.
• Why not apply it to Earth science and hydrologic data and information?
Project Objectives

1. Identify key science information and develop an information model.
2. Extract key information from scientific literature (e.g. hypothesis, conclusions, methods, datasets, variables, etc.).
3. Link scientific knowledge to datasets, resources, services and scientists.

Google, The Knowledge Graph
Technical Approach

Terminology

What is an **entity**?
A thing with distinct and independent existence.
Examples: Variables, datasets, instruments, platforms etc.

What is a **relationship**?
The connection between two entities.
Example: “Snow water equivalent (SWE) is retrieved by AMSR-2”.

![Diagram showing the relationship between SWE and AMSR-2]
Technical Approach

Information Model

Development based on the Global Changing Information System (GCIS) information model

The information model defines entities and relationships pertinent to NASA Earth science and hydrologic data, publications and resources.
Technical Approach

Key Challenge

Knowledge base construction uses both structured and unstructured content (e.g., journal articles).

Structured Content

- Metadata, tables, controlled vocabularies

Unstructured Content

- Journal articles, ATB documents, user guides
Technical Approach

How do we extract entities from unstructured content?

• Natural language processing (NLP) – Computers analyzing, understanding and deriving meaning from human language.

• Semantic Entity Recognition (SER) – NLP technique used to identify entities in text.

• Use NLP (SER) techniques to identify entities within the unstructured text.

• Apply to journal publication text to extract and identify data, models, methods, people, and institutions (i.e., entities).

• Generate a truth set – Dictionary of known models, science keywords, CMR NASA Earth science data catalogue.
Technical Approach

Refine Algorithm

Accuracy Threshold?

SNL/ SER

Vocabularies, Dictionaries, CMR Catalogue

Information Table

Compare Results

SNL/ SER

AGU Publications

SME

Manual Entity Extraction

Training Set

Refine Algorithm

AGU Publications

SME

Manual Entity Extraction

Training Set

Apply Information Model

Identify Relationships

New Knowledge

CUAHSI Hydroinformatics Conference
Preliminary Results

Example of extracted and populated information from unstructured sources

<table>
<thead>
<tr>
<th>Paper Title</th>
<th>Satellite observations of desert dust-induced Himalayan snow darkening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors</td>
<td>Ritesh Gautam, Toppei J. Yasunari, Ritesh Gautam, N. Christina Hsu, William K. M. Lau, Toppei J. Yasunari</td>
</tr>
<tr>
<td>Missing Acronym</td>
<td>BC (black carbon), HTP (Himalaya-Tibetan Plateau), WH (western Himalaya), SF (snow cover fraction), IGP (Indo-Gangetic Plains), WH (western Himalaya), AOD (Aerosol Optical Depth), NT (noon), TOA (top-of-atmosphere), μ (micron), RTM (radiative transfer model), SSA (single scattering albedo)</td>
</tr>
<tr>
<td>Id</td>
<td>55</td>
</tr>
<tr>
<td>Title</td>
<td>Satellite observations of desert dust-induced Himalayan snow darkening</td>
</tr>
<tr>
<td>Year</td>
<td>2013</td>
</tr>
<tr>
<td>Date</td>
<td>16-Mar-2013</td>
</tr>
<tr>
<td>Document Id</td>
<td>55</td>
</tr>
<tr>
<td>Flag</td>
<td>0</td>
</tr>
<tr>
<td>Subject Category</td>
<td>Geology</td>
</tr>
<tr>
<td>Channel Name</td>
<td>GEOPHYSICAL RESEARCH LETTERS</td>
</tr>
<tr>
<td>Key Words</td>
<td>HYDROLOGICAL CYCLE; SPECTRAL ALBEDO; TIBETAN PLATEAU; MONSOON; AEROSOLS; PRODUCTS; GLACIERS; IMPACTS; CLIMATE; COVER</td>
</tr>
<tr>
<td>Author Keyword</td>
<td>Dust, Snow, Remote Sensing</td>
</tr>
</tbody>
</table>

3 Extracted Figure/Table

(a) Satellite image of a major dust outbreak over South Asia, on 9 June 2003 from Terra/MODIS, indicating visibly dust-laden snow surface in the western Himalaya (WH); (b) zoom-in over WH on 15 June 2003; (c) daily AOD variations over the foothills, south of the WH snow cover; (d) MODIS spectral surface reflectance on 15 June indicating the VIS-NIR gradient for WH (30°N–34°N, 76°E–40°E), with error bars of ±1 for representing pixel-level variability
Data Use Challenges

To address these challenges, what difficulties are presented?

How can data resources be linked in order to improve the data spin-up time?

How can we work to educate unfamiliar users?
Challenge - Publications and technical documents often prove difficult for new and unfamiliar users to digest.

NASA Global Hydrology Resource Center (GHRC) Data Active Archive Center (DAAC)

NASA Short-term Prediction Research and Transition Center (SPoRT)

What resources are available to introduce data, methods and concepts?
- GHRC DAAC Data Recipes
- GHRC DAAC Micro Articles
- NASA SPoRT Quick Guides
NASA GHRC DAAC Data Recipes

What is a Data Recipe?
Tutorials or step-by-step instructions to help users learn how to discover, visualize and use data, information, software and techniques.

Types of Data Recipes
• Using netCDF data in ArcGIS
• GHRC tool tutorials
• Python notebooks and scripts
• Data format conversions and georeferencing

Discover GHRC Data Recipes Here
NASA GHRC DAAC Micro Articles

What is a Micro Article?
A short, interesting document that brings together data and key science concepts

Creates a knowledge base for users by curating around GHRC’s data and science thematic areas

Types of Micro Articles
- Instruments
- Phenomena
- Events or Case Studies
- Publications

Discover Micro Articles Here
What is a Quick Guide?

Short, easy to use resources that highlight key aspects of a data product or tool.

Intended to assist forecasters in quickly recalling information during times of operation.

Available Forms

- Download/print
- Interactive web browser
- Interactive through personal display system

Discover SPoRT Quick Guides here
Other Resources

How do these benefit Earth science and hydrologic applications?

• These resources provide introductory information that is easy to read and understand without overwhelming users.
• Each point to additional documents for more detailed information.
• Each contain information on commonly used data, models, and software.
• They link directly to data, helping users understand a dataset and how to apply it towards research or applications.
• Populating this information within a knowledge graph allows users to search and discover information on data and methods for a broad user community.
Next Steps

• Investigate generating easy to understand resources in a structured format to allow more seamless integration within the knowledge graph.
• Continue refining SER for Earth Science
• Continue building and evaluate a training set for SER (working with graduate students and SMEs)
• Scale efforts to to all Earth Science related journal titles in the Wiley Online Library
• Begin mining graphs to obtain new information
  • Prediction of relationship between entities (i.e., Network Link Prediction)
  • Automatic generation of new content (e.g., MicroArticles)
Benefits to NASA Earth Science and Hydrology

• Addresses the challenge in navigating the increasing volume of data and information.
• To provide an operational knowledge base to enhance NASA’s Earth science research.

Beneficial Applications

• Hypothesis formulation and testing:
  • Automate the search for and compilation of background information.
  • Given a topic, what hypotheses have been tested?
  • What data/tools are being used to test a hypothesis?
  • Common paths to knowledge discovery.

• Mission development/review:
  • What kinds of instruments/parameters are needed to specify science objectives?
  • Impact of a mission by linking it with publications and dataset distribution.
Thank you, questions?

amanda.m.weigel@nasa.gov

NASA/MSFC Data Science Informatics Group