Building a Knowledge Graph for Earth Science

Rahul Ramachandran\textsuperscript{1}, Jia Zhang\textsuperscript{2} and Tsengdar Lee\textsuperscript{3}

\textsuperscript{1} NASA Marshall Space Flight Center
\textsuperscript{2} Carnegie Mellon University Silicon Valley
\textsuperscript{3} NASA Headquarters

Presented at the NITRD Open Knowledge Network Workshop
Oct 4 – 5, 2017
Q: This data set is used for predicting crop yields in CropYieldModelA under the future climate scenario RCP8.5.
Watson: What is temperature and precipitation data set in IPCC climate projection archive at URL... By the way, the technical document is available at URL...

Q: How can I grow potato to survive, now that I am alone on Mars?
HAL 9000: According to Journal of Aerospace Engineering, User Manual of Water Sequestration System, and Journal of Horticultural Research, Your best chance is to...
Many Data Sets – Which one should I use?

• Graph on the left, made from metadata in GCMD, represents data sets organized by their locations (data centers).
• Additional information may be used to associate the data sets to their use cases.
Knowledge Graphs link key entities in a specific domain with other entities via relationships.

Researchers mine these graphs to make probabilistic recommendations and to infer new knowledge.

The goal of this project is to develop an end-to-end automated methodology for incrementally constructing Knowledge Graphs for Earth Science.
Why NASA Needs a Knowledge Graph

• Nearly all of the information contained within research papers is unstructured.
  • Difficult to search

• This project aims to build a traversable graph from unstructured text, tables, and figures within research papers.

• Addresses the challenge in extracting useful knowledge from the increasing volume of data and information.
### Example use case for KG Application

#### I Want to Study a Phenomenon: Dust

<table>
<thead>
<tr>
<th>What sensors would be useful?</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODIS</td>
</tr>
<tr>
<td>OMI</td>
</tr>
<tr>
<td>TOMS</td>
</tr>
<tr>
<td>CALIPSO</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What do I need to detect dust? (Physical Property)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radiance, Emissivity, Brightness Temperature</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What do I need for retrieval? (Physical property)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optical Depth</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Which datasets do I need?</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODIS/Terra Calibrated Radiances 5-Min L1B Swath 1km V006, ....</td>
</tr>
</tbody>
</table>

Overall Methodology

- Consists of two stages
  - Development of Heuristic algorithms to perform Semantic Entity Identification (Phenomena, Dataset, Instrument, Variable (Physical Property), Workflow...) to assist a human expert in building training data [Steps 0-2]
  - Use Deep Learning Algorithms to improve results [Steps 3-7]
Semantic Entity Identification (SEI) Algorithm

• Goal: extract entities to build a training dataset
• Use existing taxonomies where available
  • GCMD (variables, instruments, etc…)
  • SWEET (phenomena)
• Data set identification
  • Build profile based on extracted entities
  • Use profile to search NASA catalog for the most relevant dataset
• Use heuristics to prevent noise
  • Ignore “Background” sections of paper
  • Heuristic algorithms are brittle and need to be replaced with Deep Learning algorithms
• Provide an easy to use tool for human experts to label results from the heuristic
• Build training data set for Deep Learning algorithm
• Can we use reCAPTCHA to scale up?
Preliminary Results from SEI and Key Findings

Based on selected set of papers focused on study of dust from satellites

Key takeaways:
- Semantic entity identification is a difficult problem
- Use of existing taxonomies can help for specific instances (instruments/platforms)
- However, quality of the taxonomy can impact results for other instances (Physical Property, Phenomena)
- Dataset profile approach is dependent on the metadata quality in a data catalog
Other interesting findings

• Dataset names are very rarely mentioned in the text. When they are, key information is usually missing.

• The way the data is described in the metadata catalog does not match the way the data is described in the papers.
  • Example: MODIS Level 1B science keywords in the metadata are Spectral/Engineering > Infrared Wavelengths and Spectral/Engineering > Visible Wavelengths versus dust, brightness temperature, etc...

• Terminology used in the papers can be inconsistent.
  • Synonyms for the same phenomena
  • Dust, dust storm, wind borne minerals, mineral dust
How Big Data May Help

Steve Papa: Data volume is cumulative, analysis possibilities are combinatorial (video)

• Problem with BI is the idea of semantic layer. Problem with semantic layer is you have to define the relationships for everything up front. This is directly at odds with the idea that with the information you can’t know all the questions you want to ask up front. In fact every time you find anything interesting it begs for more questions which means the semantic layer is out of date.

• [Company] creates a crowd source semantic layer. Everyone uses a crowd source semantic layer everyday. It’s called Google type-ahead search. Every time someone searches for something that has not been searched before you are going to benefit from what you are looking for from that type-ahead. Every incremental query grows the knowledge base. We can apply the similar idea to enterprise schema where someone comes up with a new query, there is a new relationship there someone can discover.
Data & Challenges

Current Access to Data
- NASA technical library
- AGU/Wiley published literatures

Challenges:
- Data considered proprietary. Difficult to gather and difficult to share. The research is hard to scale.
- NLP is still far from NLU. We need significant advancement in NLU.