### **Portland State University**

### **PDXScholar**

Computer Science Faculty Publications and Presentations

**Computer Science** 

2013

## Taming the Metadata Mess

Veronika Margaret Megler Portland State University, vmegler@gmail.com

Follow this and additional works at: https://pdxscholar.library.pdx.edu/compsci\_fac



Part of the Databases and Information Systems Commons

# Let us know how access to this document benefits you.

### Citation Details

Megler, Veronika Margaret, "Taming the Metadata Mess" (2013). Computer Science Faculty Publications and Presentations. 131.

https://pdxscholar.library.pdx.edu/compsci\_fac/131

This Poster is brought to you for free and open access. It has been accepted for inclusion in Computer Science Faculty Publications and Presentations by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: pdxscholar@pdx.edu.

# Taming the Metadata Mess



# V.M. Megler, PSU

# Supervised by David Maier, PSU





# **Abstract**

The rapid growth of scientific data shows no sign of abating. This growth has led to a new problem: with so much scientific data at hand, stored in thousands of datasets, how can scientists find the datasets most relevant to their research interests? We have addressed this problem by adapting Information Retrieval techniques, developed for searching text documents, into the world of (primarily numeric) scientific data. We propose an approach that uses a blend of automated and "semi-curated" methods to extract metadata from large archives of scientific data, then evaluates ranked searches over this metadata. We describe a challenge identified during an implementation of our approach: the large and expanding list of environmental variables captured by the archive do not match the list of environmental variables in the minds of the scientists. We briefly characterize the problem and describe our initial thoughts on resolving it.

# **Prior Work**

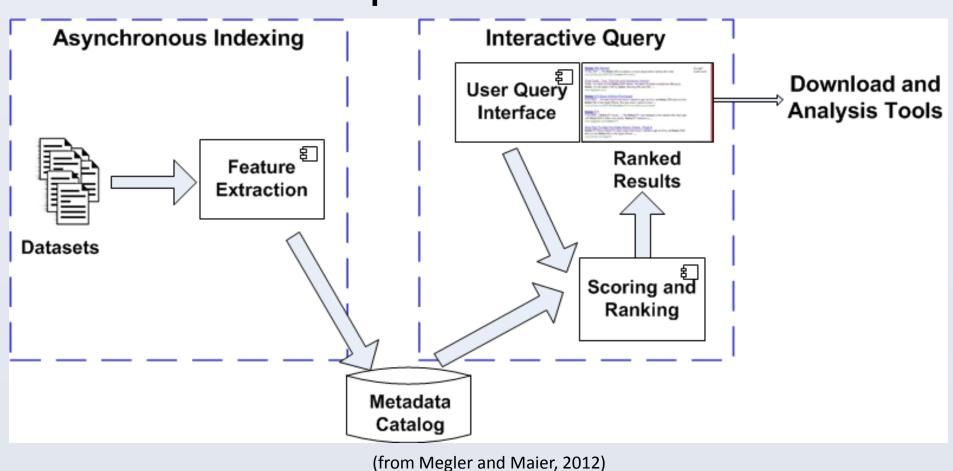
# Addressed the problem of finding relevant data in a "big data" archive (Megler and Maier, 2011)

- Many datasets, dataset shapes and sizes, physical locations, formats, tools
- ➤ "Misremembered" datasets → lost data
- > Example information need:

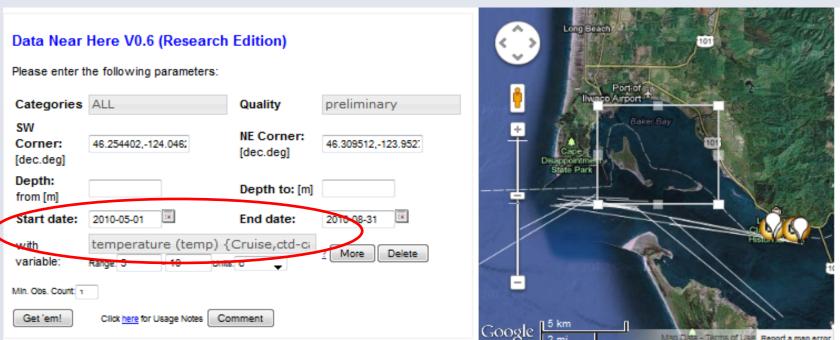
"observations collected near [lat = 45.5, lon = -124.4] in mid-2010, with temperature between 5-10C"

Solution: Build search engine for scientific data

# IR Architecture Adapted to Scientific Data Search



# Ranked Search Over Data: Location, Time, Variables



here we	re 24 results	returned; all are listed, ar	nd 24 initially s	shown on map	o. Temp was f	found in 2	24 entri	es.					
Display	Туре	Collection	Quality	Start Time	End Time	From Depth	To Depth	temp	Observations	Data Location	Score	DNH	
1	Cruise	Cruise, May-June 2010, Wecoma, 2010-07-16, Segment 3	preliminary	2010-07-16 05:16 PDT	2010-07-16 05:29 PDT	-5	-5	9.89:12.14 c	14	Download	98	DNH	Ш
2	Cruise	Cruise, April 2010, Wecoma, 2010-04-17, Segment 4	preliminary	2010-04-17 04:06 PDT	2010-04-17 04:26 PDT	-5	-5	10.60:10.85 c	21	Download	97	DNH	
3	Cruise	Cruise, April 2010, Wecoma, 2010-04-17, Segment 11	preliminary	2010-04-17 18:52 PDT	2010-04-17 23:59 PDT	-5	-5	10.88:11.21 c	244	Download	96	DNH	

Figure: "Data Near Here" Search Interface (from Megler & Maier, 2011)

- Build metadata catalog to represent archive contents
- Individual datasets scanned once, summarized into a "feature" per data
- Features stored in catalog
  Similarity search is performed
- Search results ranked on distance based similarity to query terms

over catalog's contents

# Data Near Here V0.5: Dataset Details Dataset Summary Genery Carlot for Coastal Margin Disservation and Prediction Cass space of Personance Daily Forecuriner, 2004-06-28 Type Chair Formatic CSV Causily war, data Time: Start 2009-05-28 16.09 PDT Coepfin Mar. On (free surface) Deptin Mar. On (free surface) For for Time: End 2009-05-28 16.09 PDT Coepfin Mar. On (free surface) For for Time: End 2009-05-28 16.09 PDT Coepfin Mar. On (free surface) For of Values Data Location Deprinad Last Updated (2011-12-01-00.12 PBT Cirk Nere for this observation Units Datatype Minimum Maximum Number Conductivity Units Datatype Minimum Num

- Search result leads to "dataset summary"
- Displays dataset & variable information from metadata catalog

# Emerging problem: Many names for same environmental variable

"Semantic diversity"

Figure: Example Dataset Summary Page

>Similar problems in other areas, e.g. units

Category	Example	Desired Result	Possible Technical Approach		
Minor variations and misspellings	air_temperature, air_temperatrue, airtemp	Make them the same	Translate current to desired name		
Synonyms	C, degC, Centigrade	Make them the same	Translate current to desired name		
Abbreviations	MWHLA	Use full/canonical variable name	Translate current to desired name		
Excessive variables	Quality assurance variables: qa_level	Exclude from search Show in detailed dataset views	Mark variables Exclude from search		
Ambiguous usages	temp: temporary or temperature?	Identify and expose variables. Allow curator to:     clarify where possible     hide variable     leave as is	Provide interface to specify options		
Source-context naming variations	Temperature: air_temperature or water_temperature depending on source context	Specify context of variable Make context accessible to user	Link to multiple taxonomies		
Concepts at multiple levels of detail	Fluorescence, vs. fluores375, fluores400	Collapse or expose as needed	Allow variables to be grouped Support hierarchical menus		

**Motivation** 

# The Metadata Wrangling Process Components Configure: directories, file types, naming **Archive Datasets** Often exists as a conventions translation table Scan archive Perform known transformations "The mess that's left" Add external Perform metadata Discover discovered transformations **Working Catalog** transformations External Metadata Publish Generate hierarchies Configure: levels, Metadata aggregation Catalog **Example Process Archive Datasets** (1) 3 Scan archive Perform known transformations 2,6 Add external Perform metadata Discover discovered transformations **Working Catalog** transformations External 7 Metadata **Publish** Generate (5) hierarchies Set of composable components Compose into "metadata processing chain" Metadata Details of process different for each archive Catalog

# Major curatorial activities

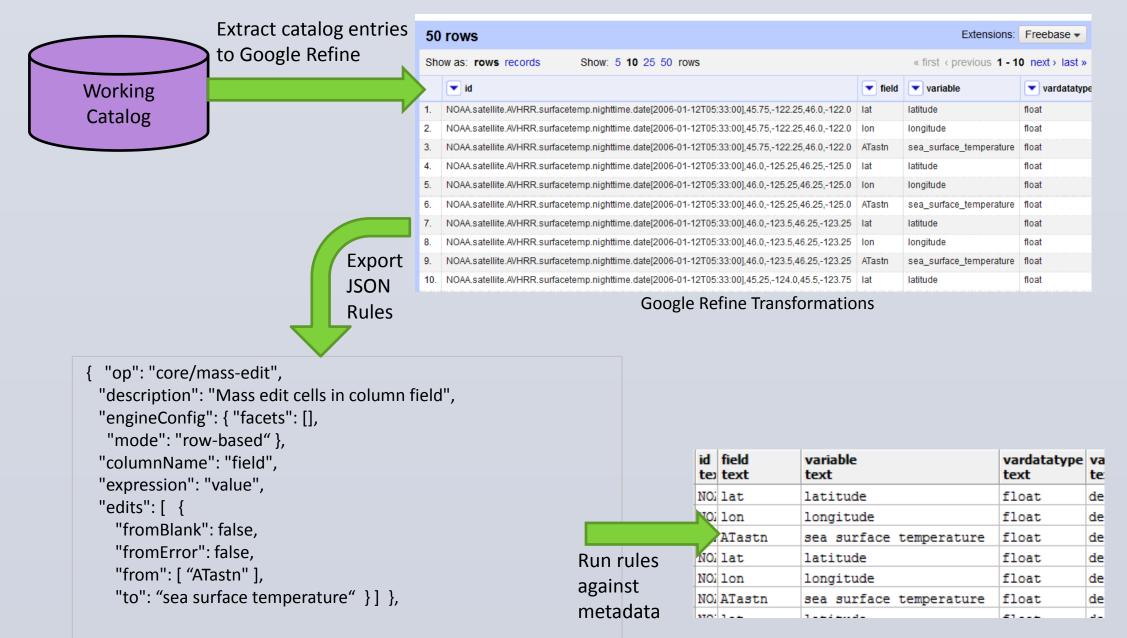
- 1. Creating metadata wrangling process for archive from composable components
- 2. Running & rerunning process
- 3. Improving process

E.g., modifying a hierarchy; adding entries to a synonym table; specifying an additional directory to scan

4. Validating process results

E.g., verifying that all files in a directory are of the same type; checking that all harvested variables names occur in the current synonym table as preferred or alternate terms; determining that expected datasets show up

# Discovering Transformations with Google Refine



# For More Information

Contact V.M. Megler at: <a href="mailto:vmegler@cs.pdx.edu">vmegler@cs.pdx.edu</a> or David Maier at: <a href="mailto:maier@cs.pdx.edu">maier@cs.pdx.edu</a>

With thanks to the scientists at Center for Coastal Margin Observation and Prediction (CMOP). This work is supported by NSF award OCE-0424602