The Problem of Semantics in the Metadata Mess

V.M. Megler
David Maier
Portland State University

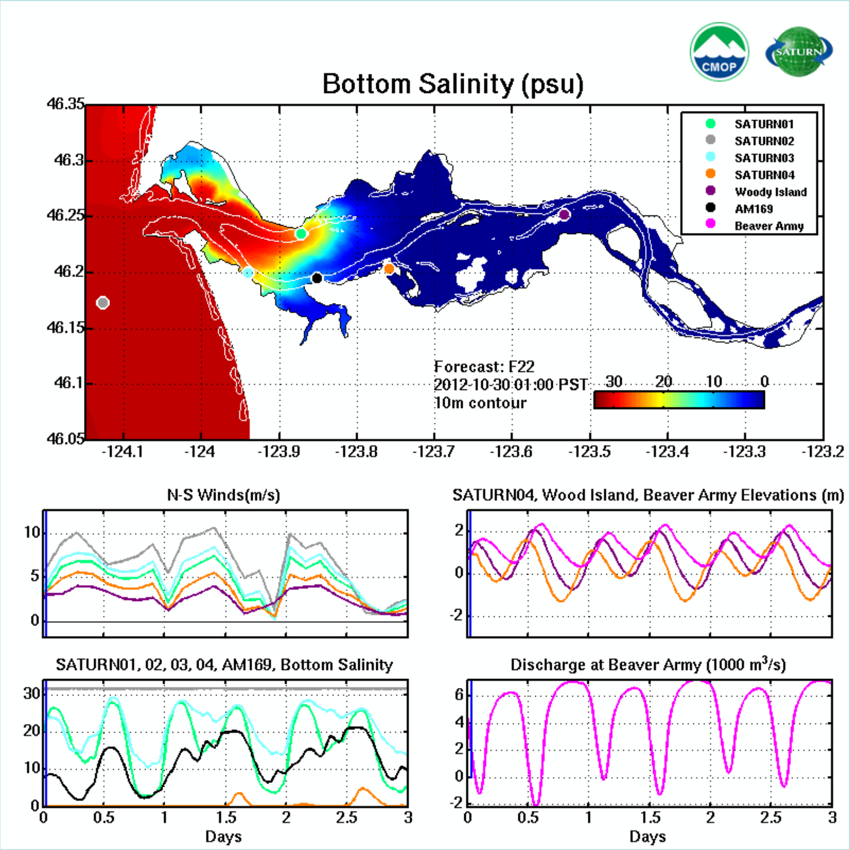


Figure: CMOP's "Virtual Columbia River"

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

Agenda

- Our “Big Data” Search Engine
- The Metadata Mess
- Reducing Semantic Diversity
- “Metadata Wrangling”
- Current State

Our “Big Data” Search Engine

- Problem: finding relevant data in a “big data” archive
  - Many datasets, dataset shapes and sizes, physical locations, formats, tools (Megler and Maier, 2011; 2012; 2013)
  - Example information need:

    “observations collected near [lat = 45.5, lon = -124.4] in mid-2010, with temperature between 5-10C”
- Solution: a data search engine that operates over big data archives

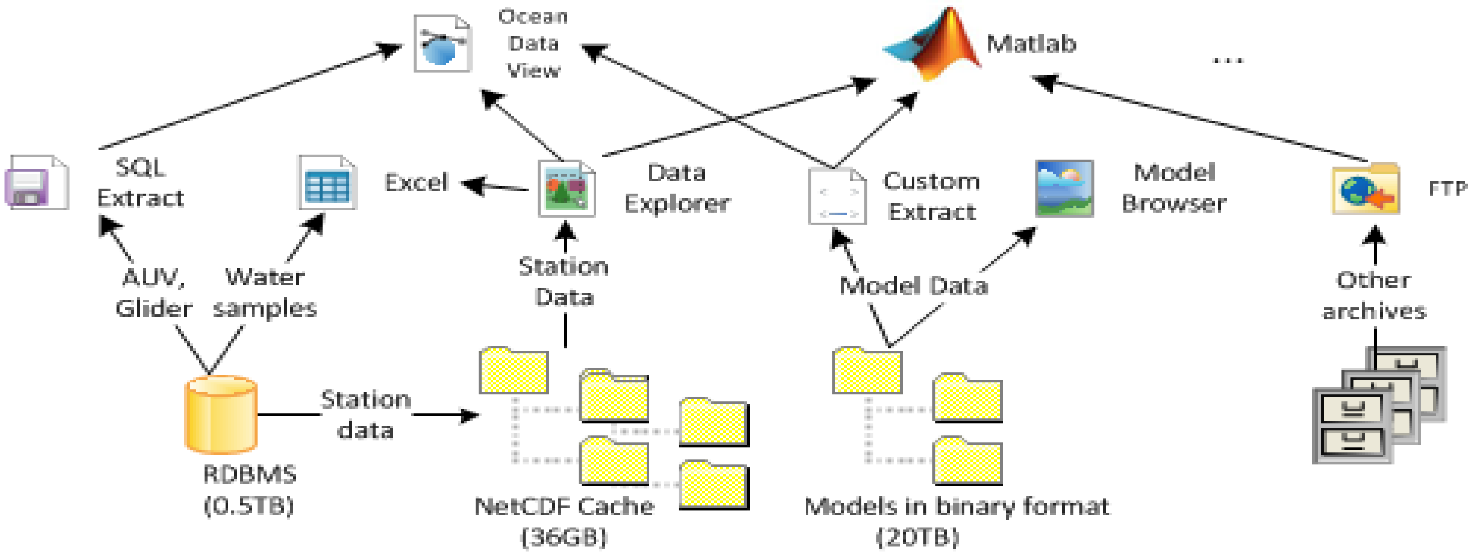


Figure: Heterogeneity of Data Formats and Data Access Tools in One Scientific Archive

IR Architecture Adapted to Scientific Data Search

- Approach:
  1. Scan (heterogeneous) data; extract summary features
  2. Search over features, with real-time response
     - Return ranked results, with links to data and tools

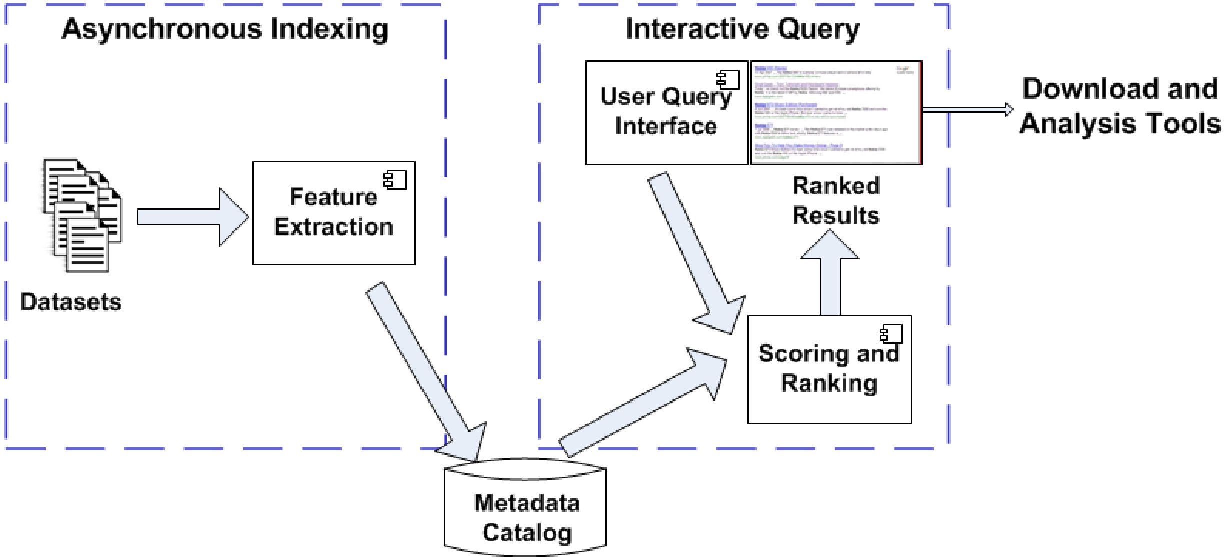
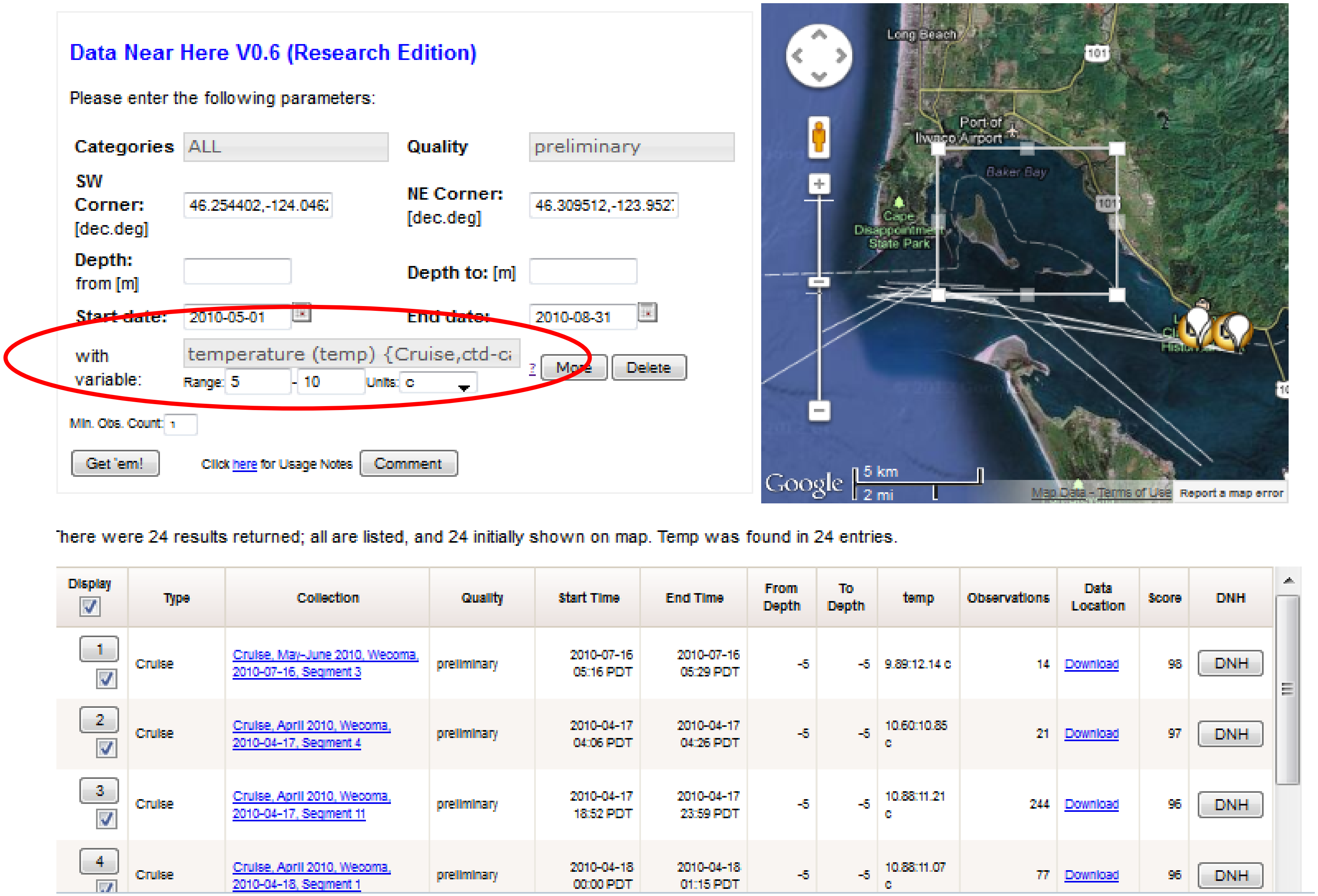


Figure: Information Retrieval Architecture, adapted to data search (from Megler and Maier, 2012)

Ranked Search Over Data: Location, Time, Variables



Data Near Here V0.6 (Research Edition)

Please enter the following parameters:

Categories	ALL	**Quality**	preliminary
SW Corner: [dec.deg]	46.254402,-124.046	**NE Corner:** [dec.deg]	46.309512,-123.952
Depth: from [m]		**Depth to:** [m]	
Start date:	2010-05-01	**End date:**	2010-08-31
with variable:	temperature (temp) {Cruise,ctd-c Range: 5 - 10 Units: c		

Min. Obs. Count: 1

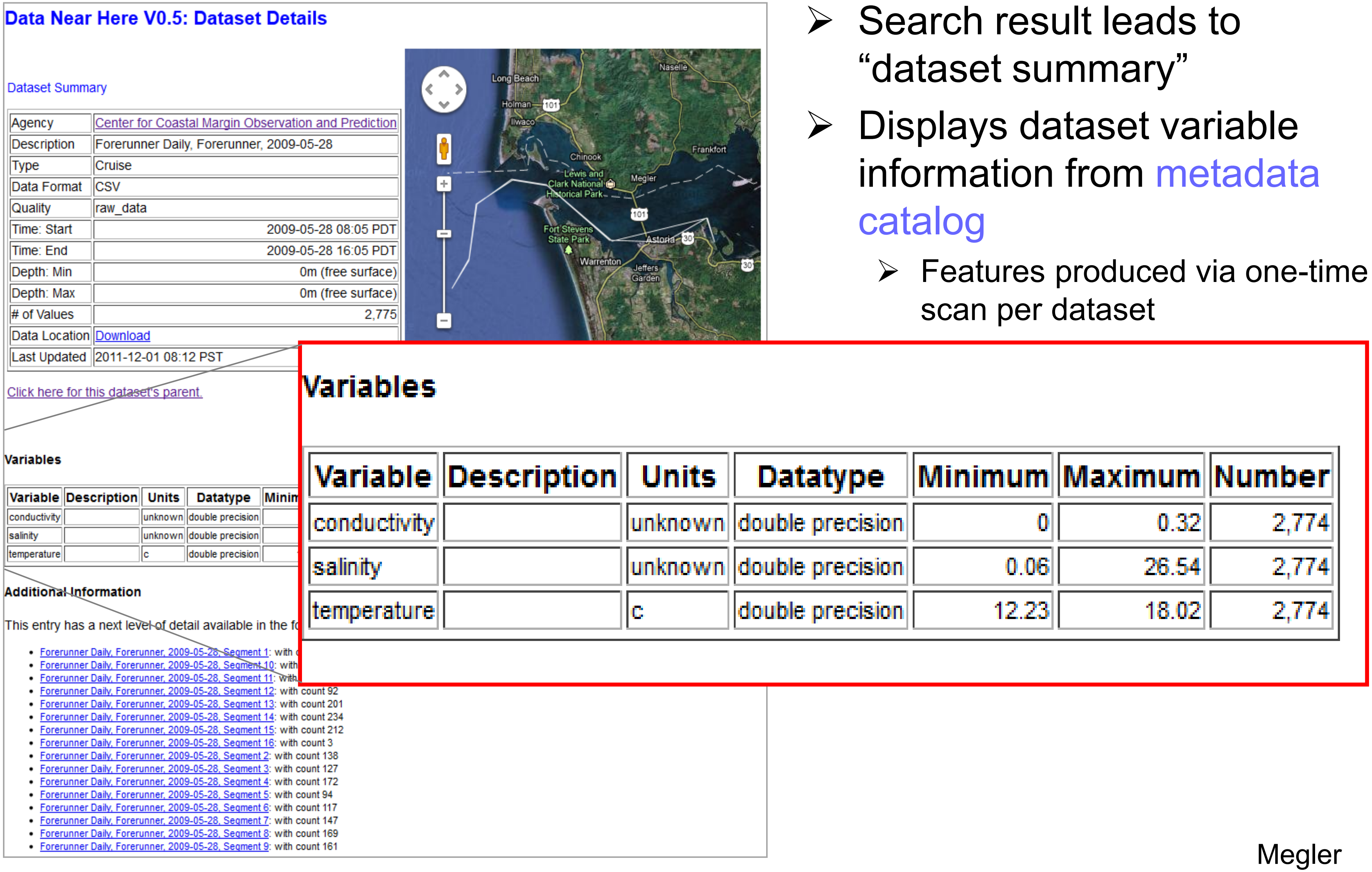
here were 24 results returned; all are listed, and 24 initially shown on map. Temp was found in 24 entries.

Display ☑	Type	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
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
4 ☑	Cruise	Cruise, April 2010, Wecoma, 2010-04-18, Segment 1	preliminary	2010-04-18 00:00 PDT	2010-04-18 01:15 PDT	-5	-5	10.88:11.07 c	77	Download	96	DNH

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

Detailed Search Result: Variable Information

- Search result leads to “dataset summary”
- Displays dataset variable information from metadata catalog
  - Features produced via one-time scan per dataset



Data Near Here V0.5: Dataset Details

Dataset Summary

Agency	Center for Coastal Margin Observation and Prediction
Description	Forerunner Daily, Forerunner, 2009-05-28
Type	Cruise
Data Format	CSV
Quality	raw_data
Time: Start	2009-05-28 08:05 PDT
Time: End	2009-05-28 16:05 PDT
Depth: Min	0m (free surface)
Depth: Max	0m (free surface)
# of Values	2,775
Data Location	Download
Last Updated	2011-12-01 08:12 PST

Click here for this dataset's parent.

Variables

Variable	Description	Units	Datatype	Minimum	Maximum	Number
conductivity		unknown	double precision	0	0.32	2,774
salinity		unknown	double precision	0.06	26.54	2,774
temperature		c	double precision	12.23	18.02	2,774

Additional Information

This entry has a next level of detail available in the fo

- Forerunner Daily, Forerunner, 2009-05-28, Segment 1: with c
- Forerunner Daily, Forerunner, 2009-05-28, Segment 10: with
- Forerunner Daily, Forerunner, 2009-05-28, Segment 11: with
- Forerunner Daily, Forerunner, 2009-05-28, Segment 12: with count 92
- Forerunner Daily, Forerunner, 2009-05-28, Segment 13: with count 201
- Forerunner Daily, Forerunner, 2009-05-28, Segment 14: with count 234
- Forerunner Daily, Forerunner, 2009-05-28, Segment 15: with count 212
- Forerunner Daily, Forerunner, 2009-05-28, Segment 16: with count 3
- Forerunner Daily, Forerunner, 2009-05-28, Segment 2: with count 138
- Forerunner Daily, Forerunner, 2009-05-28, Segment 3: with count 127
- Forerunner Daily, Forerunner, 2009-05-28, Segment 4: with count 172
- Forerunner Daily, Forerunner, 2009-05-28, Segment 5: with count 94
- Forerunner Daily, Forerunner, 2009-05-28, Segment 6: with count 117
- Forerunner Daily, Forerunner, 2009-05-28, Segment 7: with count 147
- Forerunner Daily, Forerunner, 2009-05-28, Segment 8: with count 169
- Forerunner Daily, Forerunner, 2009-05-28, Segment 9: with count 161

Motivation for This Work

Emerging problem: Many names for same environmental variable*

- E.g.: temperature, temp, water_temperature
- "Semantic diversity"
- Similar problems in other areas, e.g. variable units

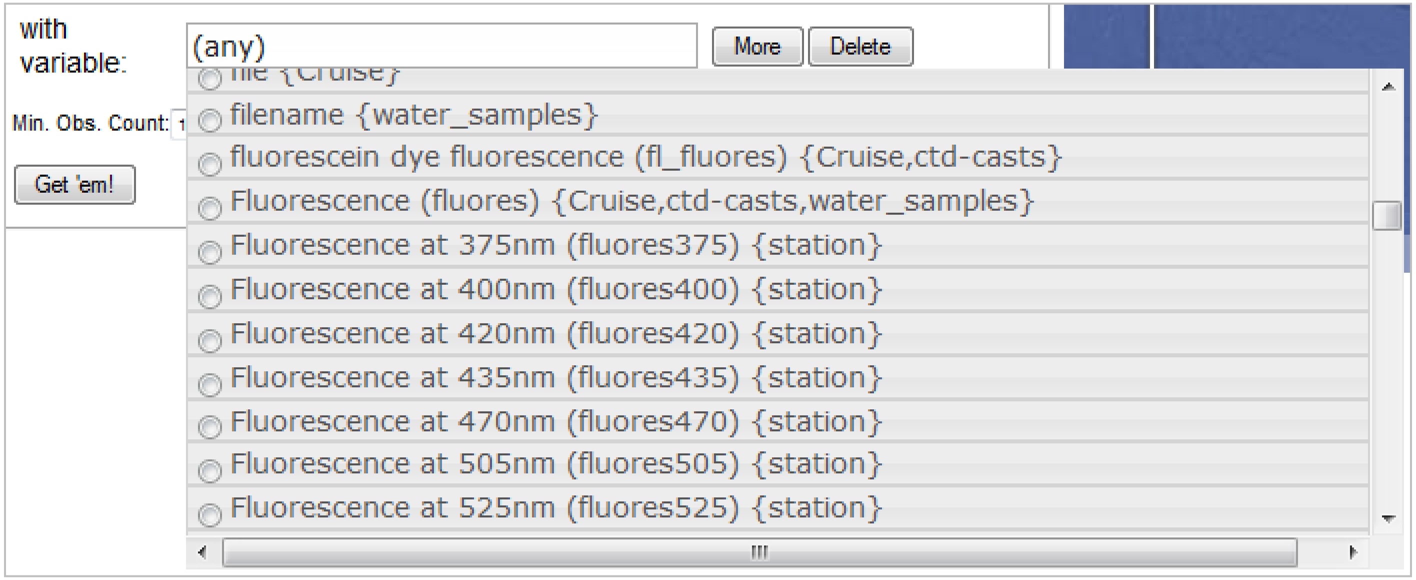
* Hereafter called "variables"

The Metadata Mess

- Working assumption: each named column in a (publicly available) file / dataset represents a valid variable

- Result: Ever increasing number of variables (over 300 at CMOP)

- Problem:
  - Hard for searchers to navigate, locate desired variable
  - Not what the archive wants to expose – “metadata mess”



○ file {Cruise}
○ filename {water_samples}
○ fluorescein dye fluorescence (fl_fluores) {Cruise,ctd-casts}
○ Fluorescence (fluores) {Cruise,ctd-casts,water_samples}
○ Fluorescence at 375nm (fluores375) {station}
○ Fluorescence at 400nm (fluores400) {station}
○ Fluorescence at 420nm (fluores420) {station}
○ Fluorescence at 435nm (fluores435) {station}
○ Fluorescence at 470nm (fluores470) {station}
○ Fluorescence at 505nm (fluores505) {station}
○ Fluorescence at 525nm (fluores525) {station}

Figure: Variable List as Exposed in Search Tool

Characterizing the Metadata Mess

- Archive curator’s goal: to present the metadata he wishes he had

- Sources of the mess:
  - Poor, unenforced or multiple naming standards
  - Data from multiple or external sources or systems
  - Changes in systems, standards and personnel over time
  - Many researchers, from different fields
  - Changing research foci

- Can’t we repair the archive?
  - Datasets must be modified or regenerated – not practical
  - May require changing code, systems – expensive, limited payoff
  - Names may be set by vendors or external data providers
  - Time-consuming, error-prone – and problems recur
  - Change is constant

The Metadata Mess (2)

- Alternative approach: compensate for the mess

- How?

  - Reduce semantic diversity

    Perfection not needed

  - Provide transformation layer from “what is” to “what should be”

Categories of Semantic Diversity

Category	**Example**
Minor variations and misspellings	*air_temperature*, *air_temperatrue*, *airtemp*
Synonyms	*C*, *degC*, *Centigrade*
Abbreviations	*MWHLA*
Excess variables	Quality assurance variables: *qa_level*
Ambiguous usages	*temp*: *temporary* or *temperature*?
Source-context naming variations	*temperature* may mean *air_temperature* or *water_temperature*, depending on source context
Concepts at multiple levels of detail	*Fluorescence*, vs. *fluores375*, *fluores400*

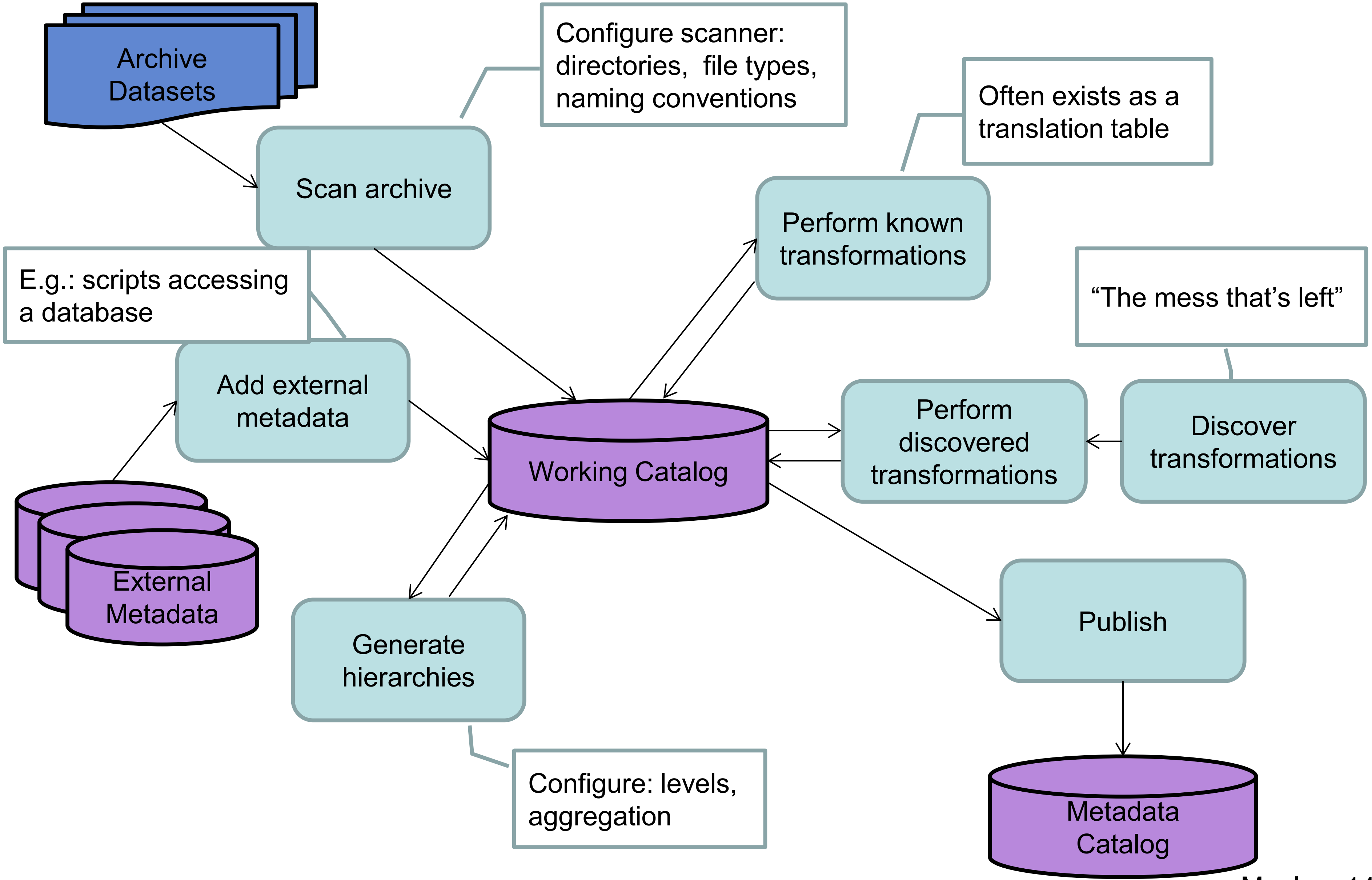
Semantic Diversity: Overall Approach

- Principles:
  - No one approach sufficient
  - All approaches must be simple; robust; tolerant of continued growth and ambiguity
  - “Refunds and exchanges available”
    - Provide defaults
    - Improve results via overrides, modifications, adjustments
    - Be non-destructive: re-doable metadata processing
- “Semi-curated” model
  - Curator performs some work for each new type of data indexed
  - Curator can review, adjust and override currently-used defaults and prior decisions

Reducing Variable-Name Diversity: Possible Approaches

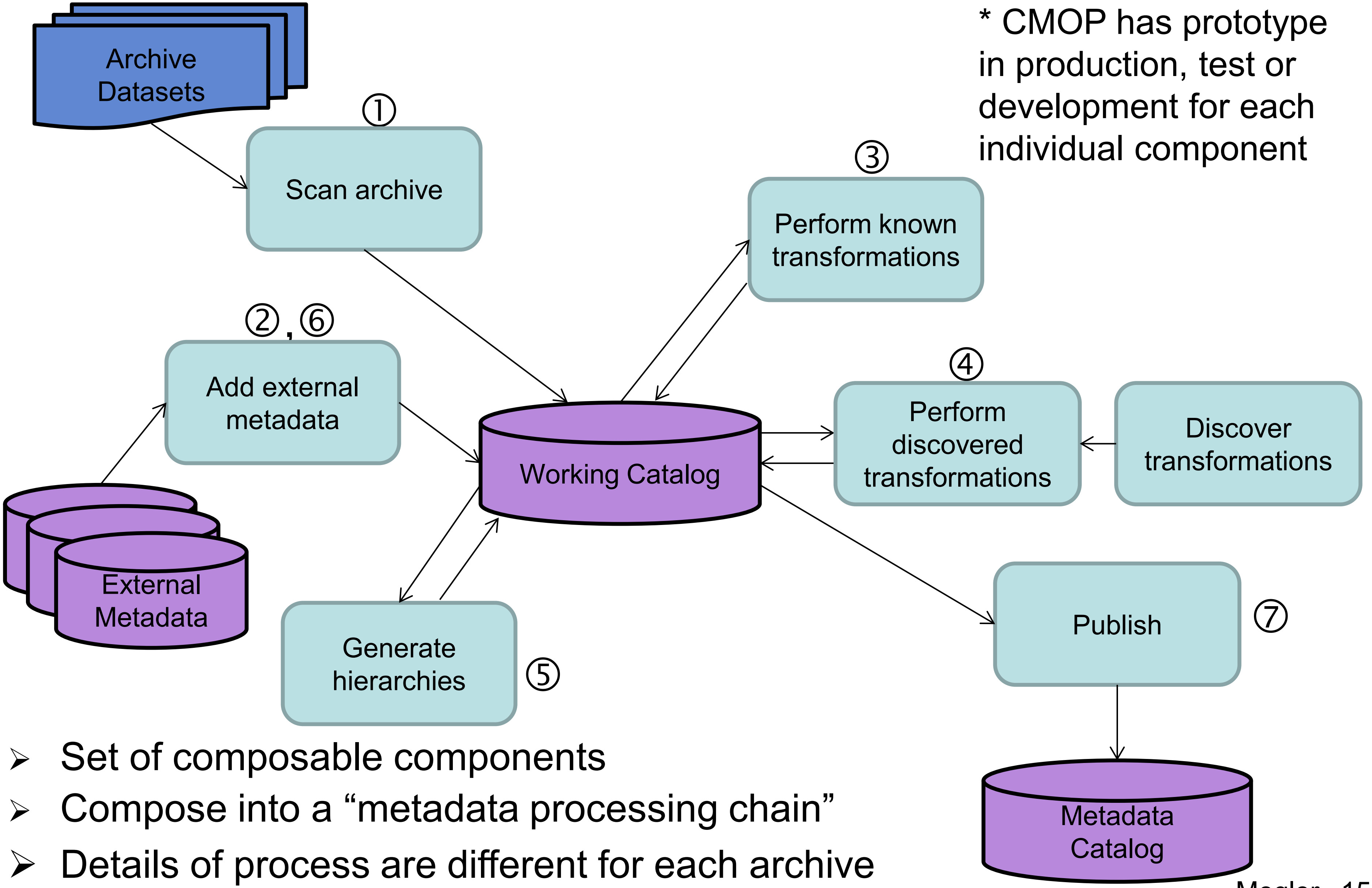
Category	Example	Desired Result	Possible Technical Approach
Minor variations and misspellings	*air_temperature, air_temperatrue, airtemp*	Make them the same	Translate current to desired name
Synonyms	*C, degC, Centigrade*	Make them the same	Translate current to desired name
Abbreviations	*MWHLA*	Use full/canonical variable name	Translate current to desired name
Excess 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

Components of "Metadata Wrangling"



Archive Datasets
Scan archive
Configure scanner: directories, file types, naming conventions
E.g.: scripts accessing a database
Add external metadata
External Metadata
Working Catalog
Perform known transformations
Often exists as a translation table
"The mess that's left"
Perform discovered transformations
Discover transformations
Generate hierarchies
Configure: levels, aggregation
Publish
Metadata Catalog

Metadata Wrangling Process



* CMOP has prototype in production, test or development for each individual component

- Set of composable components
- Compose into a "metadata processing chain"
- Details of process are different for each archive

Current State

- Diversity of variable names is an issue – even within a single archive

  Even larger issue when searching over federated archives

- Metadata wrangling is an ongoing activity

- We have:
  - Analyzed the problem for our archive (CMOP) and data included from other archives
  - Suggested possible approaches to address
  - Experimented with components of the process (scanner; hierarchy generator; scripts to add metadata; discovering & applying transformations)

- Giving a data curator tools to manage what she exposes – to manage her metadata mess – we enable easier use of her data archive.
- By combining this work with our search engine, we allow more effective discovery, access and use of the archive's contents.

References

[1] J. P. Ahrens, B. Hendrickson, G. Long, S. Miller, R. Ross, and D. Williams, “Data-Intensive Science in the US DOE: Case Studies and Future Challenges,” *Computing in Science Engineering*, vol. 13, no. 6, pp. 14 –24, Dec. 2011.

[2] B. Howe, H. Green-Fishback, and D. Maier, “Scientific Mashups: Runtime-Configurable Data Product Ensembles,” in *Scientific and Statistical Database Management*, 2009, pp. 19–36.

[3] E. Perlman, R. Burns, Y. Li, and C. Meneveau, “Data exploration of turbulence simulations using a database cluster,” in *Proc. of the ACM/IEEE conf. on Supercomputing*, 2007, pp. 1–11.

[4] E. Stolte and G. Alonso, “Efficient exploration of large scientific databases,” in *Proc. of VLDB*, 2002, p. 633.

[5] S. L. Pallickara, S. Pallickara, M. Zupanski, and S. Sullivan, “Efficient metadata generation to enable interactive data discovery over large-scale scientific data collections,” in *2nd IEEE International Conference on Cloud Computing Technology and Science*, 2010, pp. 573–580.

[6] A. Rajasekar and R. Moore, “Data and metadata collections for scientific applications,” in *High-Performance Computing and Networking*, 2010, pp. 72–80.

[7] V. M. Megler and D. Maier, “Finding Haystacks with Needles: Ranked Search for Data Using Geospatial and Temporal Characteristics,” in *Scientific and Statistical Database Management*, 2011, vol. 6809.

[8] V. M. Megler and D. Maier, “When Big Data Leads to Lost Data,” in *PIKM 2012: 5th Workshop for Ph.D. Students at CIKM*, 2012.

[9] D. Maier, V. M. Megler, A. Baptista, A. Jaramillo, C. Seaton, and P. Turner, “Navigating Oceans of Data,” in *Scientific and Statistical Database Management*, 2012, vol. 7338, pp. 1–19.

[10] V.M. Megler, “Managing the Metadata Mess”, in *ICDE 2013: Workshop for Ph.D. Students at ICDE*, 2013.

[11] P. Lord and A. Macdonald, “e-Science Curation Report,” 2003.

[12] J. K. Batcheller, “Automating geospatial metadata generation – An integrated data management and documentation approach,” *Computers & Geosciences*, vol. 34, no. 4, pp. 387–398, 2008.

[13] T. Hey and A. E. Trefethen, “The Data Deluge: An e-Science Perspective,” in *Grid Computing: Making the Global Infrastructure a Reality (eds F. Berman, G. Fox and T. Hey)*, John Wiley & Sons, Ltd, Chichester, UK, 2003, pp. 809–824.

[14] P. Cornillon, J. Gallagher, and T. Sgouros, “OPeNDAP: Accessing Data in a Distributed, Heterogeneous Environment,” *Data Science Journal*, vol. 2, no. 0, pp. 164–174, 2003.

[15] J. Parsons and Y. Wand, “Attribute-based semantic reconciliation of multiple data sources,” *Journal on Data Semantics I*, pp. 21–47, 2003.

[16] E. Rahm and P. A. Bernstein, “A survey of approaches to automatic schema matching,” *the VLDB Journal*, vol. 10, no. 4, 2001.

Major Curatorial Activities

1. Creating process
2. Running (or 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 certain directory were indeed 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 do indeed show up.

Managing “the Mess that’s Left”

- “Discovered transformations” – discovered by reviewing results so far
  - Experimenting with Google Refine*

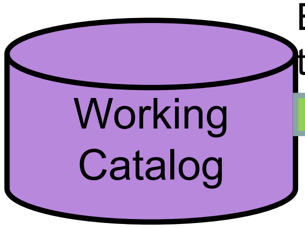
- Archive curator:
  1. Accesses list of variables (along with sample datasets they appear in)
  2. Reviews list
  3. Generates set of variable-name transformations and rules
  4. Applies rules and checks results for validity
  5. Exports rules and “applies”

- Transformation Engine:
  - Reruns at intervals: as new datasets are scanned
  - Applies rules to existing metadata

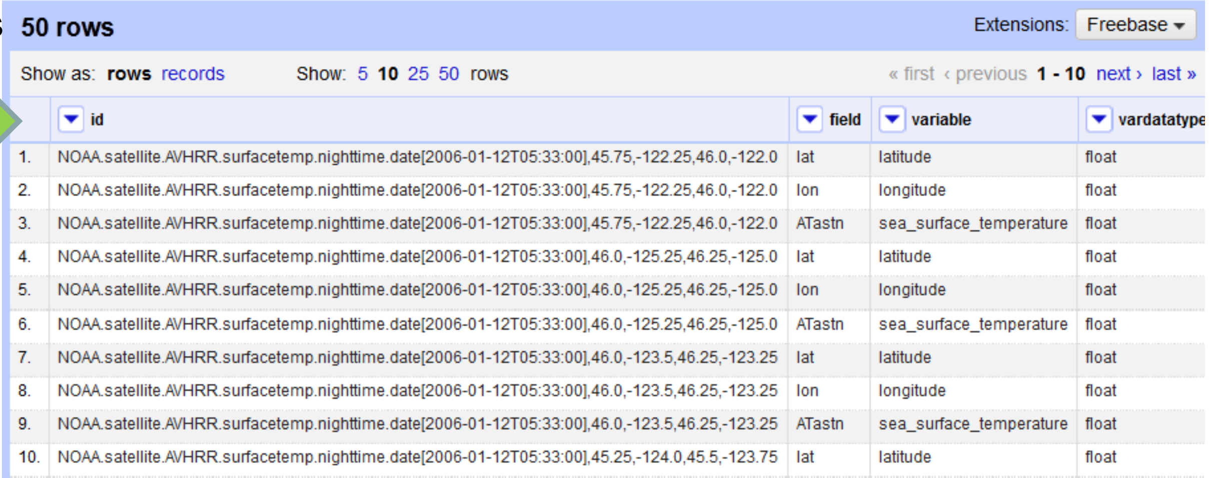
- Search engine:
  - Searches over “cleaned” metadata

*http://code.google.com/p/google-refine/

Discovering Transformations with Google Refine



Extract catalog entries to Google Refine



50 rows

Extensions: Freebase

Show as: rows records   Show: 5 10 25 50 rows   « first ‹ previous 1 - 10 next › last »

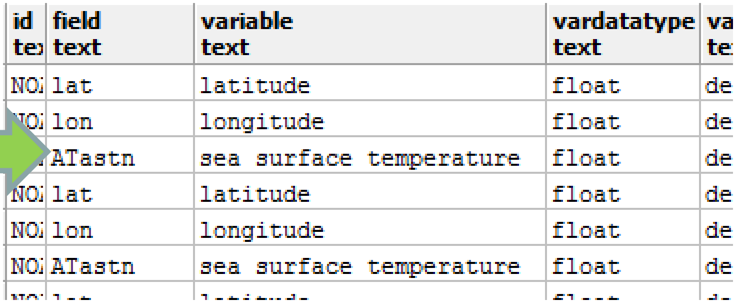
	id	field	variable	vardatatype
1.	NOAA.satellite.AVHRR.surfacetemp.nighttime.date[2006-01-12T05:33:00],45.75,-122.25,46.0,-122.0	lat	latitude	float
2.	NOAA.satellite.AVHRR.surfacetemp.nighttime.date[2006-01-12T05:33:00],45.75,-122.25,46.0,-122.0	lon	longitude	float
3.	NOAA.satellite.AVHRR.surfacetemp.nighttime.date[2006-01-12T05:33:00],45.75,-122.25,46.0,-122.0	ATastn	sea_surface_temperature	float
4.	NOAA.satellite.AVHRR.surfacetemp.nighttime.date[2006-01-12T05:33:00],46.0,-125.25,46.25,-125.0	lat	latitude	float
5.	NOAA.satellite.AVHRR.surfacetemp.nighttime.date[2006-01-12T05:33:00],46.0,-125.25,46.25,-125.0	lon	longitude	float
6.	NOAA.satellite.AVHRR.surfacetemp.nighttime.date[2006-01-12T05:33:00],46.0,-125.25,46.25,-125.0	ATastn	sea_surface_temperature	float
7.	NOAA.satellite.AVHRR.surfacetemp.nighttime.date[2006-01-12T05:33:00],46.0,-123.5,46.25,-123.25	lat	latitude	float
8.	NOAA.satellite.AVHRR.surfacetemp.nighttime.date[2006-01-12T05:33:00],46.0,-123.5,46.25,-123.25	lon	longitude	float
9.	NOAA.satellite.AVHRR.surfacetemp.nighttime.date[2006-01-12T05:33:00],46.0,-123.5,46.25,-123.25	ATastn	sea_surface_temperature	float
10.	NOAA.satellite.AVHRR.surfacetemp.nighttime.date[2006-01-12T05:33:00],45.25,-124.0,45.5,-123.75	lat	latitude	float

Google Refine Transformations

Export JSON Rules

```
{   "op": "core/mass-edit",
   "description": "Mass edit cells in column field",
   "engineConfig": { "facets": [],
     "mode": "row-based“ },
   "columnName": "field",
   "expression": "value",
   "edits": [   {
       "fromBlank": false,
       "fromError": false,
       "from": [ “ATastn" ],
       "to": “sea surface temperature“  } ]  },
```

Run rules against metadata



id te:	field text	variable text	vardatatype text	va te:
NO:	lat	latitude	float	de
NO:	lon	longitude	float	de
	ATastn	sea surface temperature	float	de
NO:	lat	latitude	float	de
NO:	lon	longitude	float	de
NO:	ATastn	sea surface temperature	float	de