2013

Taming the Metadata Mess

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

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

Taming the Metadata Mess

Portland State UNIVERSITY

V.M. Megler, PSU

Supervised by David Maier, PSU

CMOP

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

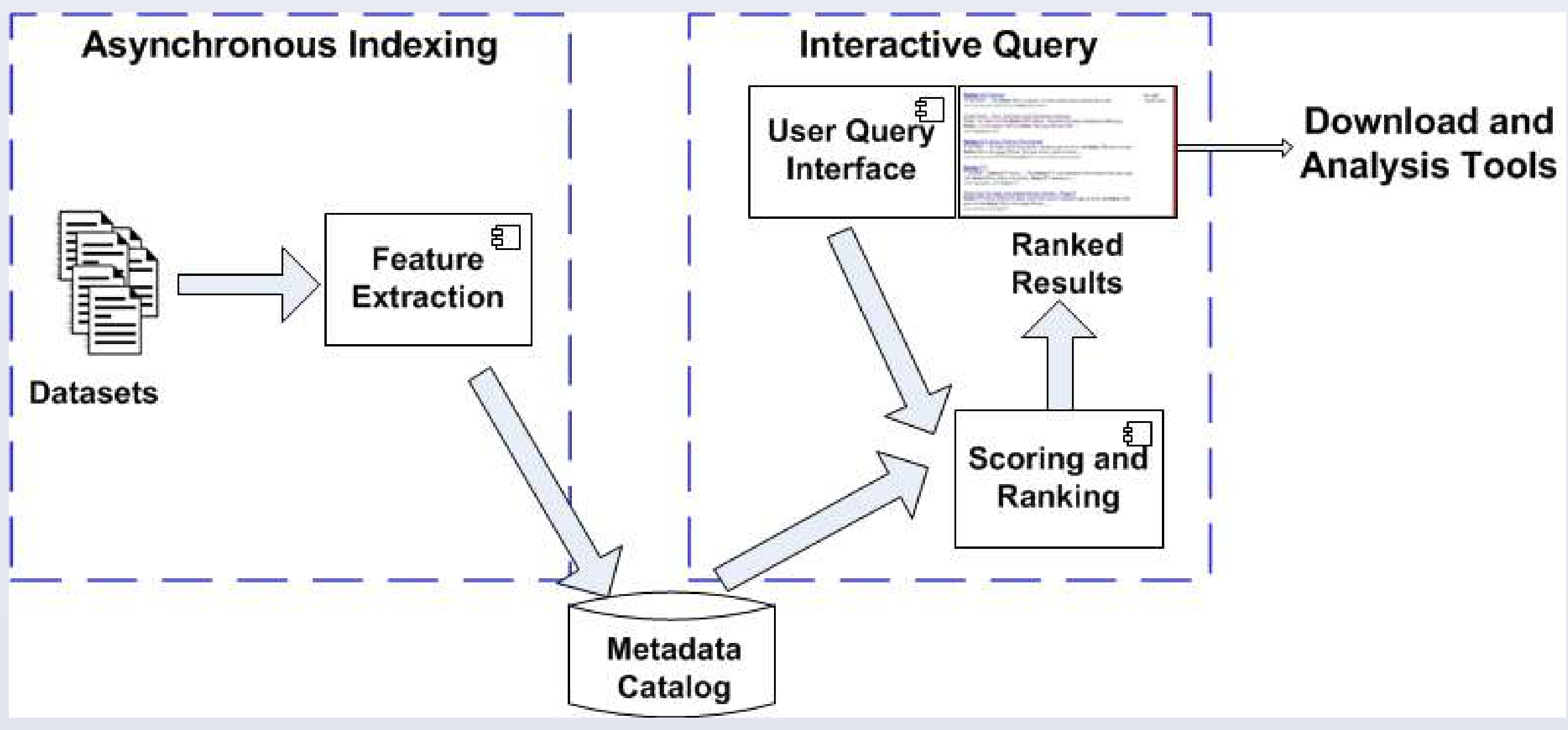
Addressed the problem of finding relevant data in a "big data" archive (Megler and Maier, 2011)

- ➢ 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"

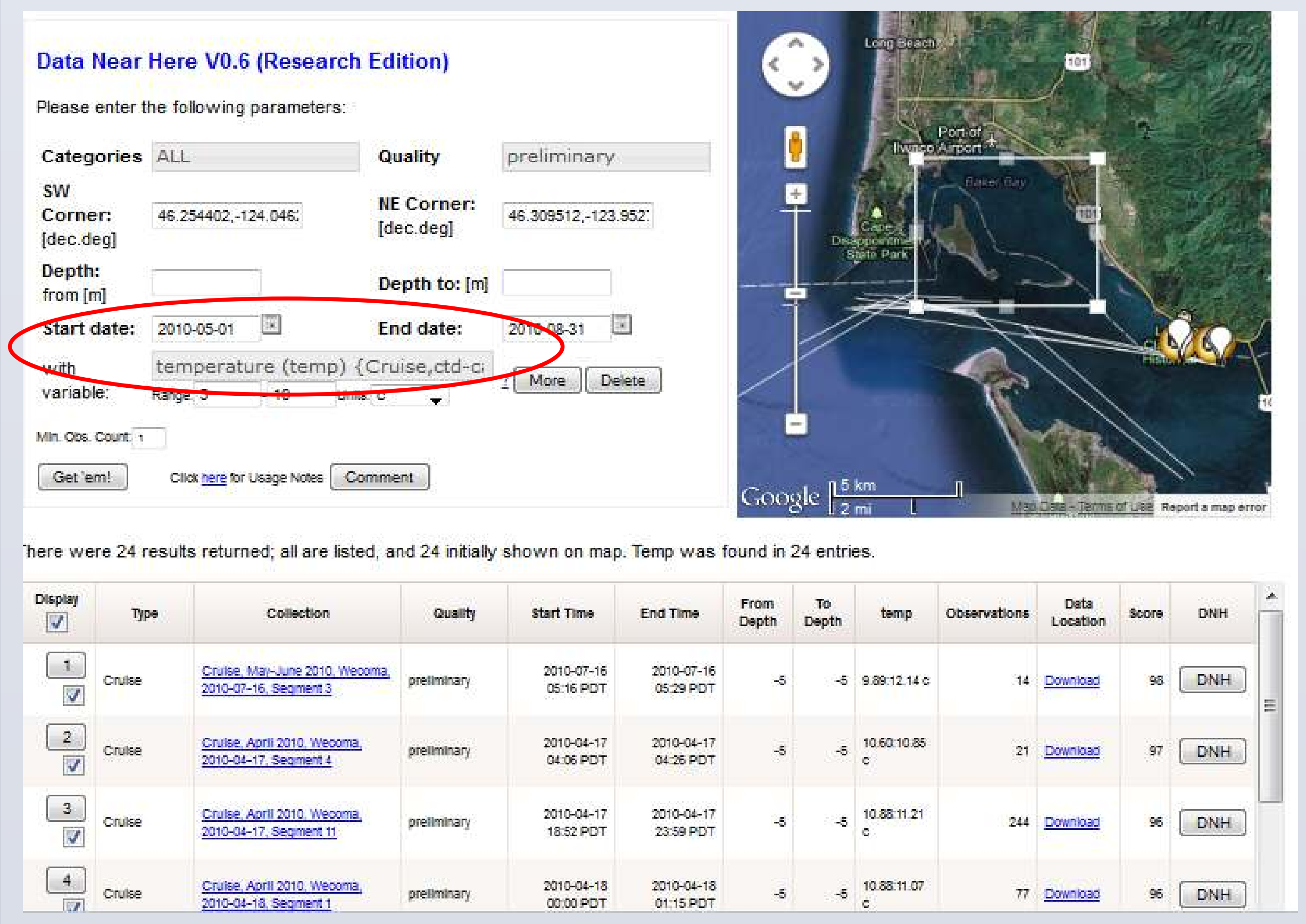
Solution: Build search engine for scientific data

IR Architecture Adapted to Scientific Data Search



(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

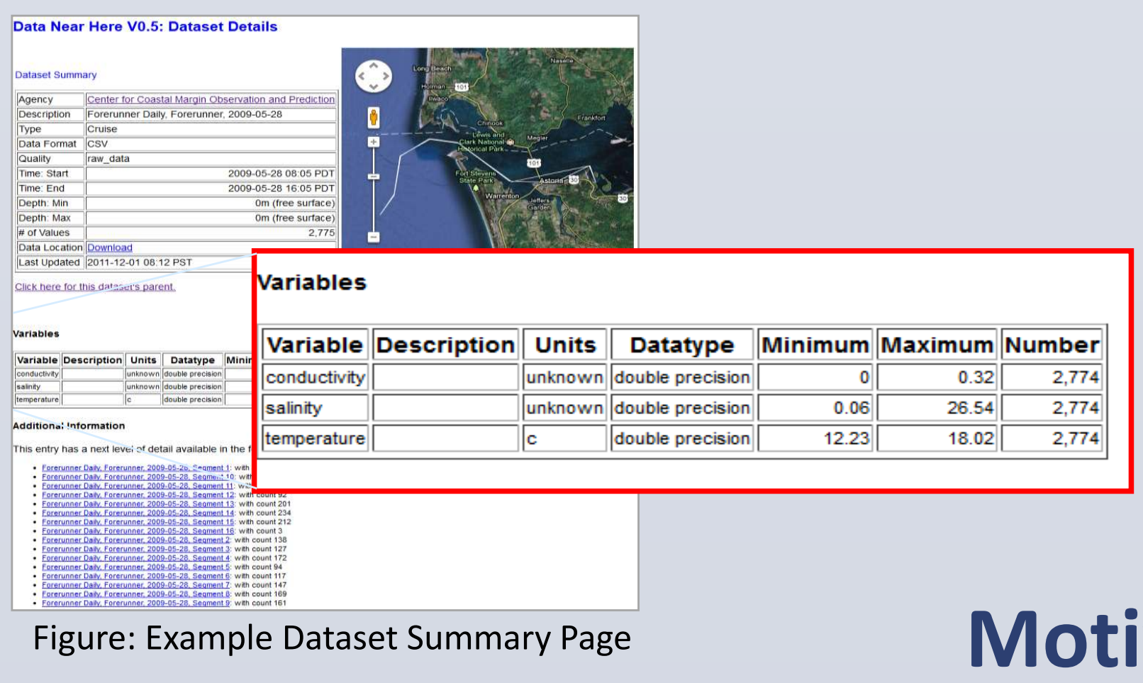
Min. Obs. Count: 1

There 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)

- 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 over catalog's contents
- Search results ranked on distance based similarity to query terms



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 f

- Forerunner Daily, Forerunner, 2009-05-28, Segment 1: with
- 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

Figure: Example Dataset Summary Page

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

Motivation

Emerging problem: Many names for same environmental variable

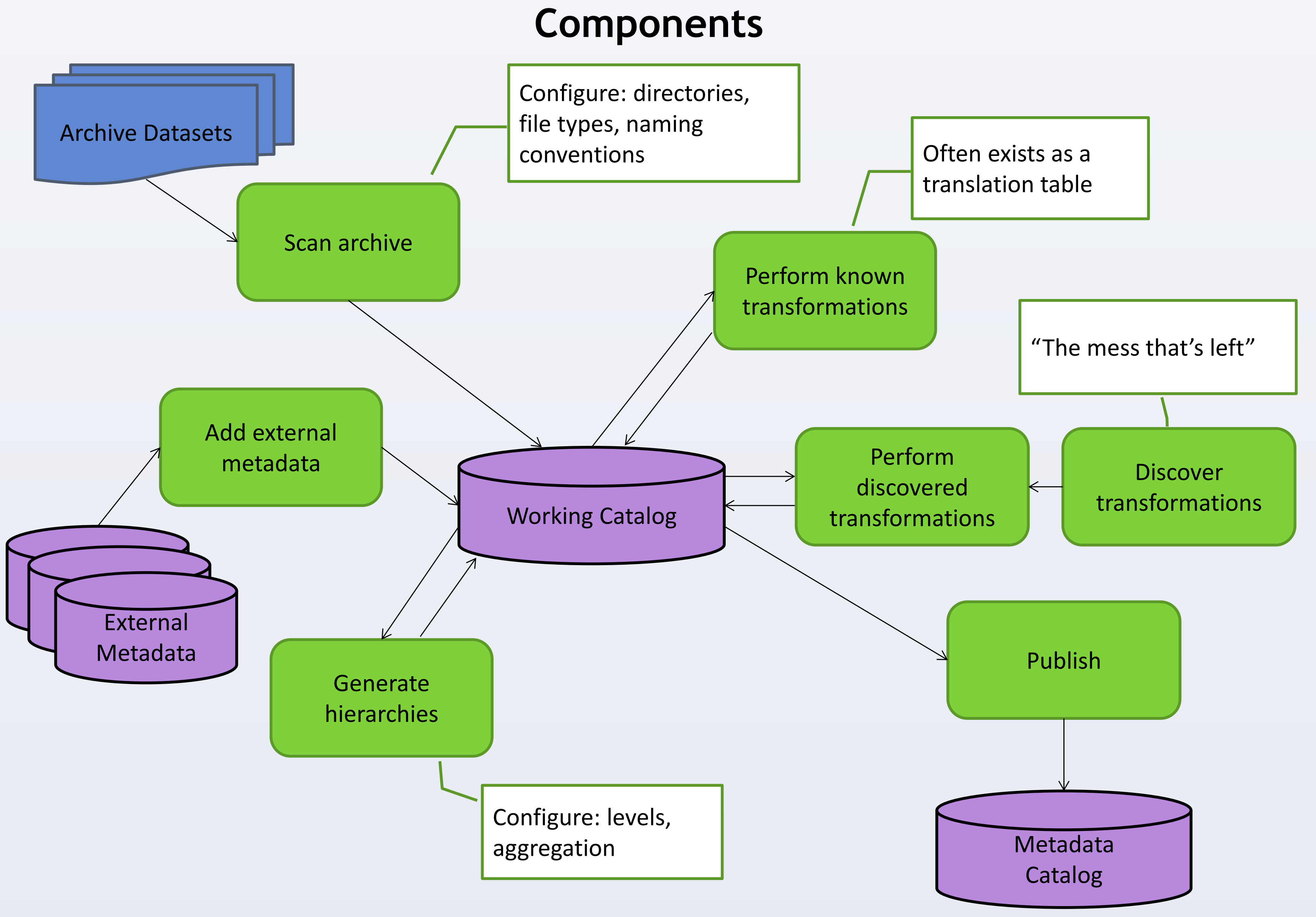
- ➢"Semantic diversity"
- ➢Similar problems in other areas, e.g. units

Table: Categories of Semantic Diversity, and Possible Approaches

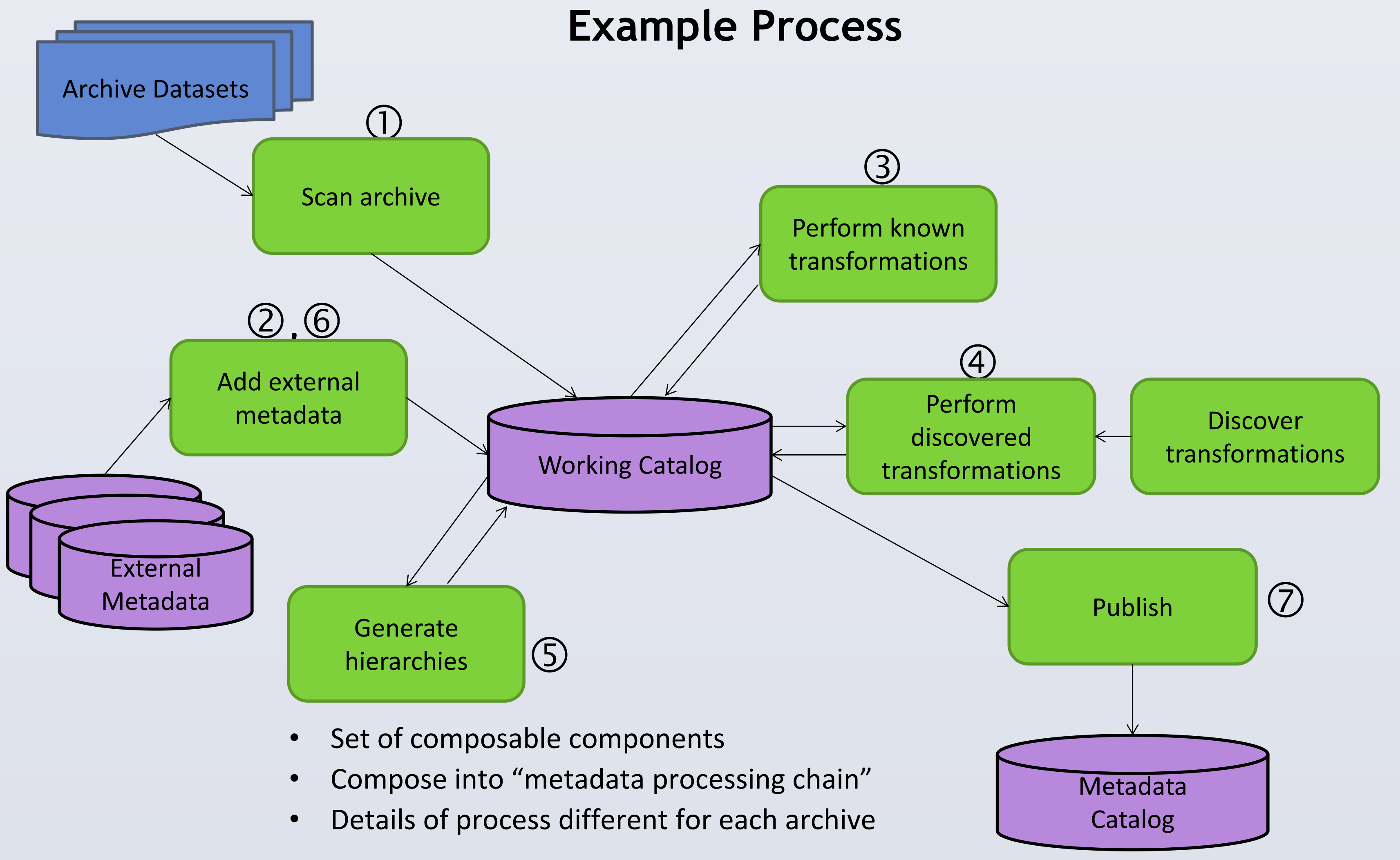
Category	Example	Desired Result	Possible Technical Approach
Minor variations and misspellings	*air_temperature*, *air_temperatrue*, *airtemp*	Make them the same	Translate current to desired name
Synonyms	*C*, *degC*, *Centigrade*	Make them the same	Translate current to desired name
Abbreviations	*MWHLA*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

The Metadata Wrangling Process

Components


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

Example Process



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

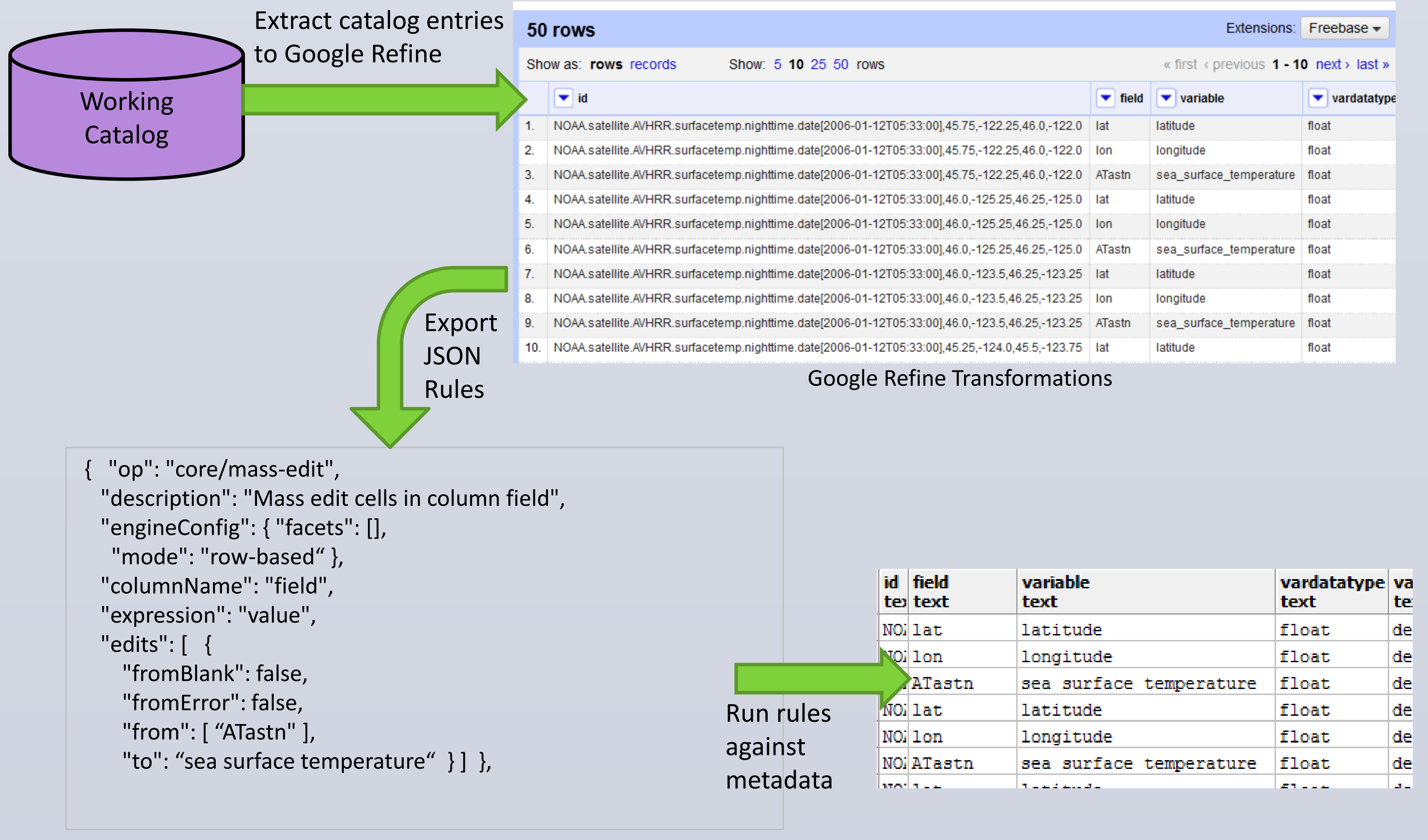
Major curatorial activities

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


Extract catalog entries to Google Refine
Working Catalog
Export JSON Rules
Run rules against metadata
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

```
{   "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“  } ]  },
```

id text	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

For More Information

Contact V.M. Megler at: vmegler@cs.pdx.edu or David Maier at: maier@cs.pdx.edu

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