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

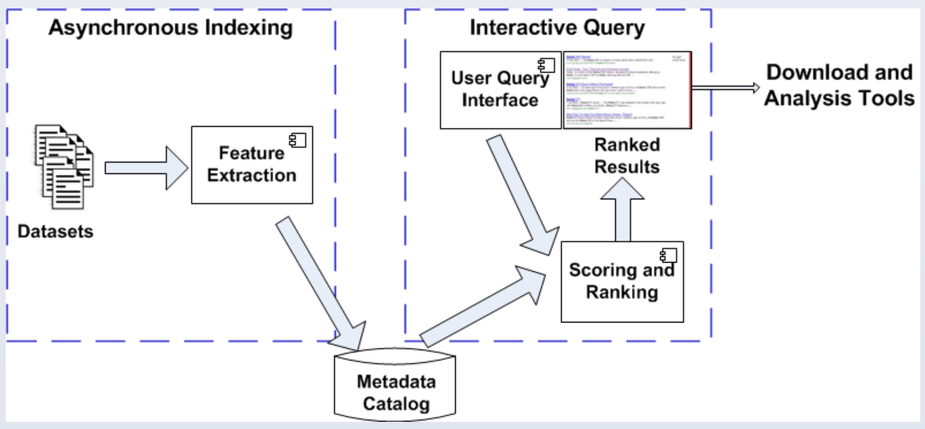
Addressed the problem of finding relevant data in a "big data" archive
(Megler and Maier, 2011)

- 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



(from Megler and Maier, 2012)

Ranked Search Over Data: Location, Time, Variables

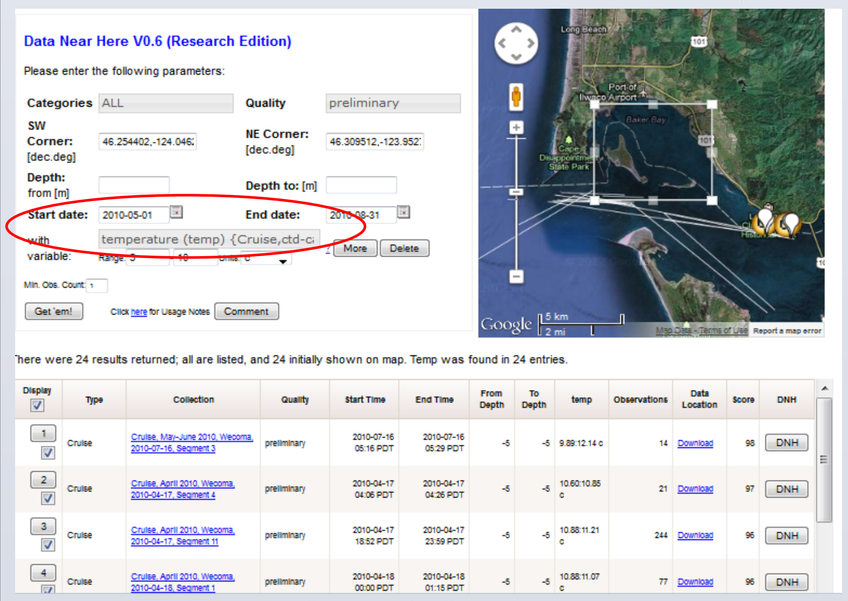


Figure: "Data Near Here" Search Interface (from Megler & Maier, 2011)

- 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

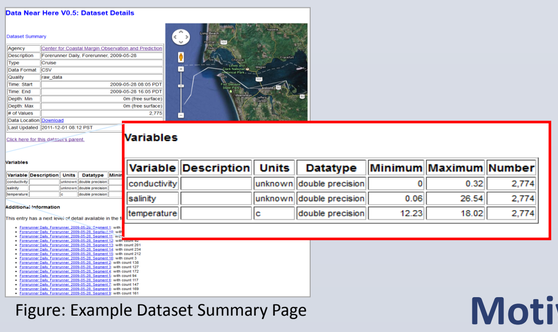


Figure: Example Dataset Summary Page

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

Motivation

Emerging problem: Many names for same environmental variable

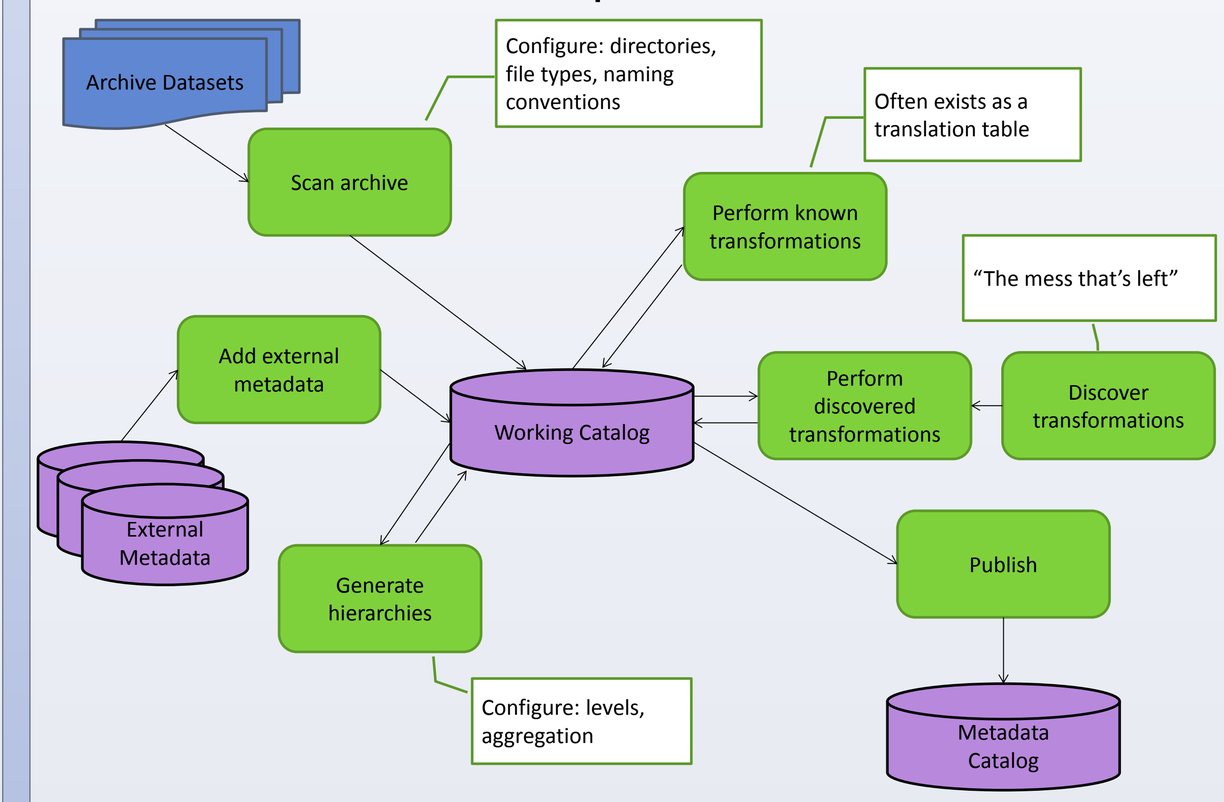
- "Semantic diversity"
- Similar problems in other areas, e.g. units

Table: Categories of Semantic Diversity, and Possible Approaches

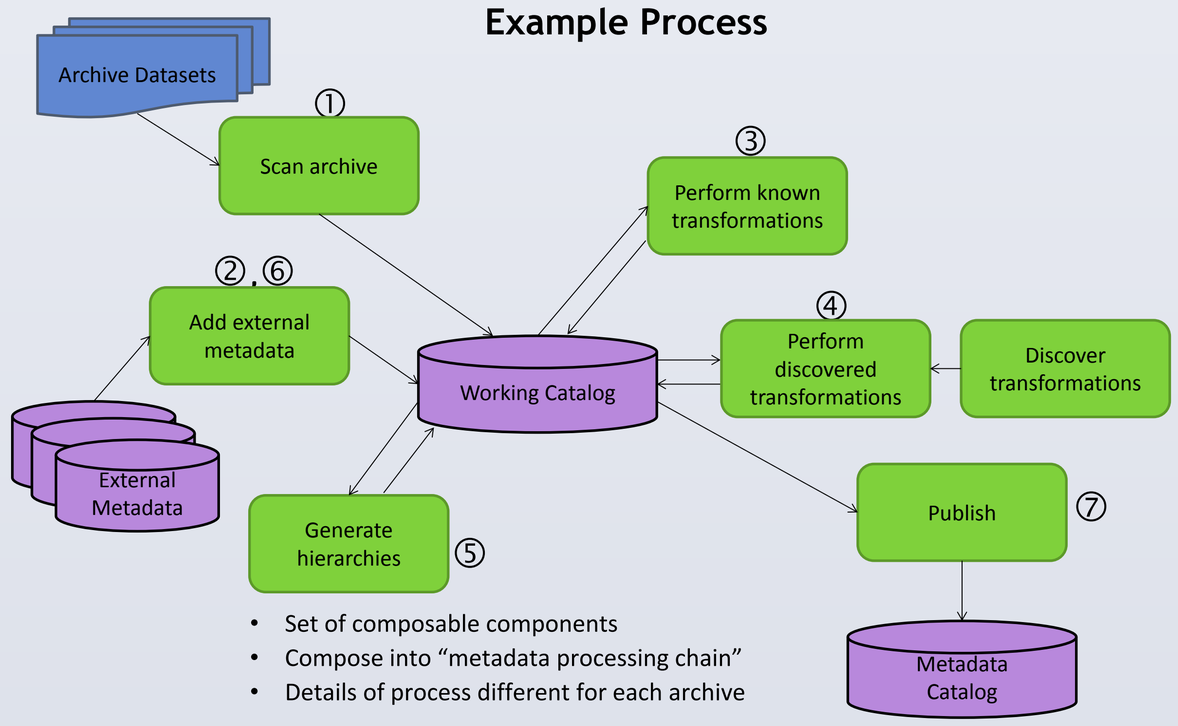
Category	Example	Desired Result	Possible Technical Approach
Minor variations and misspellings	*air_temperature*, *air_temperatrue*, *airtemp*	Make them the same	Translate current to desired name
Synonyms	*C*, *degC*, *Centigrade*	Make them the same	Translate current to desired name
Abbreviations	*MWHLA*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



Example Process



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

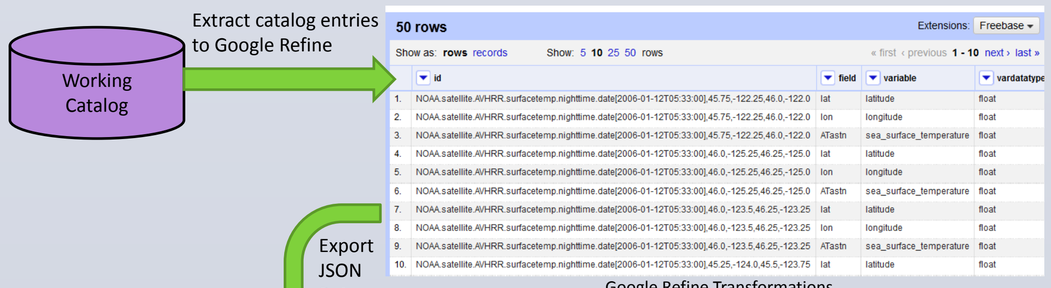
Major curatorial activities

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



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" } ] },
```

Run rules against metadata

id text	field text	variable text	vardatatype text	va te
NO…	lat	latitude	float	de
NO…	lon	longitude	float	de
NO…	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.