2013

Taming the Metadata Mess

Veronika Margaret Megler

Portland State University, vmegler@gmail.com

Let us know how access to this document benefits you.

Follow this and additional works at: https://pdxscholar.library.pdx.edu/compsci_fac

Part of the Databases and Information Systems Commons

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. For more information, please contact 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 
- Example information need: “observations collected near [lat = 45.5, lon = -124.4] in mid-2010, with temperature between 5-10°C”

Solution: Build search engine for scientific data

IR Architecture Adapted to Scientific Data Search

Ranked Search Over Data: Location, Time, Variables

Motivation
Emerging problem: Many names for same environmental variable
- “Semantic diversity”
- Similar problems in other areas, e.g., units

The Metadata Wrangling Process

Components
- Archive Datasets
- Scan archive
- Perform known transformations
- “The mess that’s left”
- Perform discovered transformations
- Publish

Example Process
1. Scan archive
2. Perform known transformations
3. Perform discovered transformations
4. Publish

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: 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.