Taming the Metadata Mess

V.M. Megler[#1]

Supervised by David Maier[#2]

[#]*Department of Computer Science, Portland State University*
Portland, Oregon, USA

{[1]vmegler, [2]maier}@cs.pdx.edu

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.**

I. Introduction

The last decade has seen massive changes in scientific data archives; along with rapid increases in size, many of these archives now include a mission of sharing their data with other researchers, educators and the interested public. While the general attitude seems to be that 'more data is better,' growth in holdings can actually introduce impediments to science [1]. As an archive grows and ages, the diversity of its contents tends to increase: over time, data is stored in changing data formats and data structures, stored in multiple physical locations, collected from multiple sources, stored using changing naming conventions. The result is increasing heterogeneity.

In an archive consisting of thousands of datasets with this level of diversity, it can be daunting to find specific datasets containing desired data. This challenge crosses industries and fields of study; in scientific research, it is seen as a constraint on discovery [1]. Efforts to address this challenge have resulted in archives making their data available for direct access and download via internal and external interfaces. Within some fields of study, gateways have developed that aggregate metadata from multiple archives [1]. Still, scientists have difficulty locating data that meets their research needs. In our work with one scientific archive, the Center for Coastal Margin Observation and Prediction (CMOP)[1], the scientists brought this issue of finding relevant data to our attention as one of their highest priority problems with their computing infrastructure (CMOP RIG meeting, July 15, 2010, private communication). The cost of time spent in searching for data, and the potential failure to gain value from relevant data that was collected but cannot be located, provides the motivation for this research.

Our scientists have tools available to them for finding relevant data, but these tools do not always meet their needs. The tools fall into three major categories.

- Data-access approaches, such as selecting from a series of hierarchical menus to discover if the eventual result contains the desired data. Data-archive portals often support such an approach. The scientist must know which selections to choose, or try all of them; as the number and diversity of available datasets and options increases, the scientist has more difficulty identifying the correct path to the desired dataset.
- Visualizing individual datasets. The scientists have powerful analysis and visualization tools available to them to use in visualizing a dataset (e.g., [2–4]); however, as the number of datasets increases, visualizing each dataset is not feasible.
- Text-based search of metadata associated with datasets [5], [6]. These searches depend on the archive providing appropriate metadata, and on the user's information needs being expressible in the same terms found in the metadata.

The author's dissertation addresses the scientists' need by adapting techniques well-known in the fields of Information Retrieval (IR) and web search to scientific data. While we initially focus on scientific data – which includes a large number of archives across many fields of study – we believe the techniques likely have wider applicability.

The contributions made in the author's research so far are the following. We have:

- Defined a new problem: scientific data search as an Information Retrieval problem [7].
- Formulated an approach: applying Information Retrieval techniques to scientific datasets [7], [8].
- Implemented a prototype, "Data Near Here" [7], [9].
- Provided evidence of utility via two user studies (results not yet published).
- Developed an initial formulation of a model and componentized architecture [8], to generalize this work. The thesis will include an expanded formulation.

In Section II, we briefly outline the prototype we built to test these ideas. Our experiences with the prototype have

[1] http://www.stccmop.org

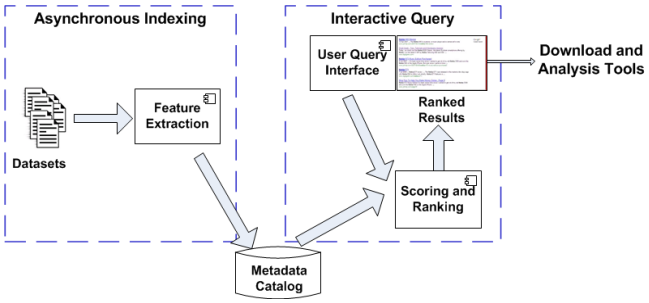

Fig. 1. High-Level Architecture for Searching over Data

demonstrated that the concept of ranked search and techniques from IR can be applied fruitfully to scientific data. In addition, we have identified challenges to address as we generalize and scale up this initial work to include more data sources and eventually more archives.

One challenge, described in this paper, is the issue of "semantic diversity" of the environmental variables and column names found in the thousands of datasets we currently catalog. We describe this challenge in Sections III and IV, and suggest approaches to solving it in Section V. Section VI touches on related work and we conclude in Section VII.

II. Searching Over Data

We adapt the well-known architecture used by Internet search engines in searching text documents for use over datasets, as shown in Figure 1. Features are extracted from each dataset, and stored in an index. Our index takes the form of a metadata catalog; the entries in the catalog are created by a single offline scan of datasets. We use these features as a summary of the dataset. In our work so far, the main features we use are the column name, units, data type and range of the data in each column. We provide a user query interface and a scoring-and-ranking engine. The scientist can represent his information need as a query, and the scoring-and-ranking engine returns the datasets most similar to the query in a list ranked by similarity.

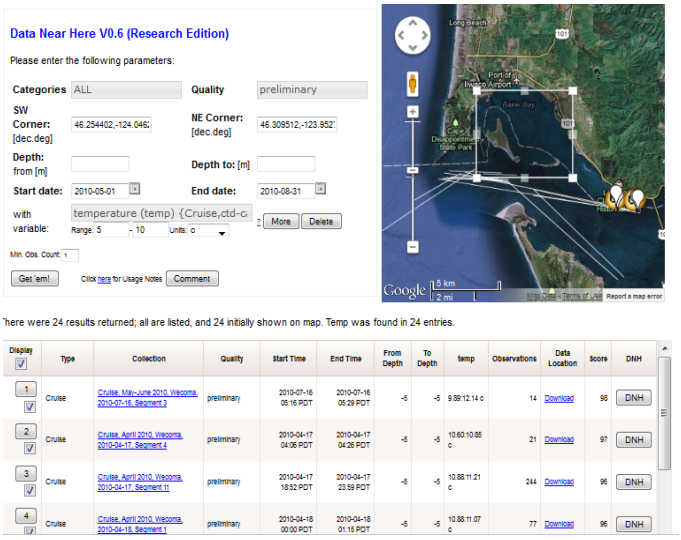

Fig. 2. User Interface for Data Near Here. No full matches were found; several partial matches to a query with time, space and a variable with limits are listed, and more are shown on the map.

As our test case, we work with an archive of observational data collected by CMOP and partner institutions from instruments in the Columbia River and off the coasts of Oregon and Washington over a period of approximately 15 years. This data consists of environmental variables (hereafter called variables) such as salinity, oxygen concentration and nitrogen, and is available for download by the public via a number of interfaces. As is common for such archives, the observations are stored in datasets and databases, with each dataset consisting of a set of named columns for each represented variable within that dataset. We developed a prototype of the system described in Figure 1, called "Data Near Here," which is being tested in production at CMOP [7], [9]. At present, the prototype is being used internally; after some period of validation it will be opened to the public. Figure 2 shows a screenshot of the interface, with a sample query and ranked datasets returned for that query.

III. Collecting Metadata

Creation of metadata for scientific datasets is an acknowledged and ongoing problem. Relying on manually generated metadata is considered a prescription for failure, as annotation by scientists is considered burdensome and is often ignored [10–12]. One group noted that the users wanted more metadata than providers were interested in providing, and that providers stopped providing access to data when more metadata was required of them [13]. Automatic metadata generation has been identified as a potential solution; the kind of metadata to be generated is assumed to be a domain-specific problem.

To address this need, we envision a "semi-curated" model for data archives. That is, we expect the curator to perform some work for each new type of data indexed, such as each new file format. After that, additional data sources of the same type should be handled by performing some minor configuration, such as adding a line to a configuration file. We harvest metadata automatically wherever possible; currently, we gather metadata from the file system, from the headers and data within each dataset, and from a metadata store in an RDBMS in which CMOP stores details such as the data source and general description for a category of data.

IV. The Metadata Mess

When scanning datasets to extract metadata, we assumed that each named column in a dataset represented a valid variable, since these datasets are publicly available. Therefore, whenever a file header contained a column heading and associated units, we captured that information and treated it as a valid instance and combination of variable and units.

As we scanned more of the archive's datasets, including historic datasets and those extracted from or contributed by other institutions, the number of distinct variables represented rose to over 300. However, the number of distinct variables collected is, in the minds of center scientists, far smaller, perhaps on the order of one or two dozen.

We investigated sources for this order-of-magnitude

discrepancy between the number of column names found and the number of environmental variables in the minds of the scientists. One source is the difference in focus amongst scientists; oceanographers had in mind one set of variables, while a microbiologist, depending on her research project, had a different set, with only some overlap with the oceanographers – or even other microbiologists. Another source is "metadata mess": the effects of collecting data from multiple sources, each with a slightly different name for the same variable, or variables with slightly different meanings.

We analysed the list of variables to see how individual variables related to the variables in the minds of the scientists. We found certain categories of variability, listed in Table 1; for each category, we identified a desired result and a possible technical approach (further described in Section V), also shown in the table. A survey of the full list of variables found a substantial number of each of these cases. (Units for variables exhibit similar problems but at smaller scales.)

The first five categories shown are aspects the archive should ideally handle internally or otherwise shelter users from. The categories of minor variations and misspellings, synonyms, use of abbreviations, and excessive variables can all be addressed by translating the current name to a corrected or canonical name. It might be desirable to "repair" the datasets involved; however, while it may be practical to regenerate a subset of the datasets, current processes or project traceability and provenance may depend on datasets retaining the existing variable names. Data coming from an outside source must be reimported with the corrected variable names. In practice, only a subset of data can be corrected, and even then after some time these errors will once again manifest.

The last two categories, source-context naming variations and concepts at multiple levels of detail, exist for both the archive and the archive user. For source-context naming variations, it may be appropriate for an oceanography archive to standardize on "temperature" for water temperature. However, for data search engines to be useful, both the archive and the user must be able to specify which context they intend, so that the appropriate interpretation can be made.

The case of multi-level concepts is the most complex. For example, fluorescence may be measured at different wavelengths and stored as separate variables in a dataset: fluores375, fluores400, etc. For a microbiologist studying the data, each of these wavelengths is a separate variable. For the oceanographer, all wavelengths may be thought of as a single variable called "fluorescence". Likewise, ocean modelers often regard surface_temperature as a variable distinct from water_temperature, since it represents a boundary condition of inputs from external influences (wind, sun). In essence, such situations are manifestations of property precedence, as described by Parsons and Wand [14], where attributes that appear different at one level can be regarded as the same at a more abstract level. We also note that a scientist may move through several phases of detail when searching for data. She may begin with a more general query, while trying to assess what data is available: is there any fluorescence information available, and if so, what kinds? On finding some, she becomes progressively more selective.

An inspection of an unrelated research archive of traffic data found that their data exhibits the same categories, leading us to believe this problem may be common.

V. Taming The Metadata Mess

Our approach is guided by two principles. First, since an archive is likely to have all these cases, no single approach will be sufficient. Second, given a limited staff for maintenance and a constantly changing environment, all approaches must be simple, robust, and tolerant of continued growth and ambiguity; that is, if only partly applied, they should provide benefit to the section of the archive to which

TABLE I
Characterization of Sources of Variable-Name Diversity

Category	Example	Desired Result	Possible Technical Approach
Minor variations and misspellings	air_temperature, air_temperatrue, airtemp	Make them all the same	Translation of current variable name to desired name
Synonyms	"C", "degC", "Centigrade"	Make them all the same	Translation of current variable or units name to desired name
Abbreviations	"MWHLA"	Use full or canonical variable name	Translation of current variable name to desired name
Excessive variables	Quality assurance or statistical variables, such as calibration variables: "qa_level"	These variables should not be part of allowable search criteria; however, users may still want to know whether they're in the dataset.	Remove from search, but allow variable to be seen in the detailed dataset information
Ambiguous usages	"temp": does this mean temporary or temperature?	Identify and expose the variables. Allow the owner or curator to clarify where possible, choose to not expose the variable, or leave as is	Provide owner or curator with an interface that allows them to specify these different options
Source-context naming variations	"temperature" may mean "air temperature" or "water temperature", depending on the context of the source	Specify context of variable, and make the context accessible to user	Link to multiple taxonomies: see discussion
Concepts at multiple levels of detail	Fluorescence, vs. fluores375, fluores400	Allow the multiple variables to be "collapsed" or exposed as needed	Hierarchical menus of variables

they have been applied, while not limiting access to the rest.

We are experimenting with using a "metadata wrangling" tool to manipulate the metadata index. At a time of his choosing (that is, not synchronized with the dataset-scanning process), the archive curator can review a list of the variables along with representative datasets in which they were found. To address the first five problems, we are experimenting with using Google Refine[2] as a tool for the curator to specify a set of variable-name transformations and rules, including constraints on the datasets to which they should be applied, to generate a "cleaner" list of variables. (Excessive variables are addressed by transforming the variable name to "null".) We found it straightforward to apply such a set of rules in the specified order to our metadata catalog, to generate a "display name" for any variable. The rules can be applied automatically on a regular basis to newly scanned datasets, ensuring that new instances of the problems are transformed. Changes to the search engine are minimal: we expose and search over the new variable names. We keep the display name separate from the original field name within the dataset, and search using the revised name. Detailed dataset displays show both the revised and original name, thus allowing traceability. Any untransformed variables are left as-is and are unaffected by this approach, achieving our second goal.

The approach of transforming individual variable names does not solve the problem of specifying source context and multi-level concepts. This problem is an area for future research. We may allow the data curator to link variables to one or more taxonomies. It is likely that at any time, some variables will match an agreed-to domain taxonomy; some variables will match a local-archive naming standard; some newer variables will not yet have a standard, and some variables will match none of these cases. Any solution must handle this level of heterogeneity. In addition, the search engine must be able to search over this combination.

VI. Related Work

There are obvious similarities between the problem described here and the field of semantic interoperability or semantic reconciliation [14]. As Parsons and Wand note, most approaches can be categorized as schema-based or attribute-based. Schema-based approaches assume that data elements can be mapped into a well-defined semantic data model; however, data archives may not have such a model. We are not attempting to formally map or reconcile the schemas; our approach is more akin to Internet search approaches that attempt to locate relevant items despite misspelled words. For example, some search engines will recognize American versus English spellings of a word ("color" and "colour") as being the same. We apply that concept here.

For larger archives or collections, minor variations and synonyms could be addressed by some form of automatic name matching, while abbreviations could be addressed by description matching, using methods described by Rahm and Bernstein [15]. We believe that for many archives, this level of formality is unrealistic and not required for useful search.

We will test our approaches by analysing the reduction in variable diversity achieved, and by working with CMOP scientists to validate the usefulness of the result.

VII. Conclusion

We briefly described our work and challenges we identified while developing a search engine for a data archive. Even within the context of a single archive, the diversity of variable names is an issue. We present an initial analysis and some approaches to addressing the problem. By giving a data curator tools to manage what he exposes – to manage his metadata mess – we can enable easier use of the data archive. By combining this work with our search engine, we can allow more effective use of the data archive's contents. This work is a step in that direction.

Acknowledgment

This work is supported by NSF award OCE-0424602. We thank CMOP staff and scientists for their support, and Basem Elazabbi and Kristin Tufte for their assistance with the metadata examination.

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*, Hawaii, 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] P. Lord and A. Macdonald, "e-Science Curation Report," 2003.

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

[12] S. Weidman and T. Arrison, "Steps Toward Large-Scale Data Integration in the Sciences: Summary of a Workshop." National Research Council of the National Academies, 19-Aug-2009.

[13] 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.

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

[15] E. Rahm and P. A. Bernstein, "A survey of approaches to automatic schema matching," *the VLDB Journal*, vol. 10, no. 4, pp. 334–350, 2001.

[2] http://code.google.com/p/google-refine/