5-2013

Data Near Here: Bringing Relevant Data Closer to Scientists

Veronika M. Megler
Portland State University

David Maier
Portland State University

Data Near Here: Bringing Relevant Data Closer to Scientists

By V. M. Megler and David Maier

Abstract: Large scientific repositories run the risk of losing value as their holdings expand, if it means increased effort for a scientist to locate particular datasets of interest. We discuss the challenges that scientists face in locating relevant data, and present our work in applying Information Retrieval techniques to dataset search, as embodied in the Data Near Here application.

Keywords: Scientific archives, ranked search, information retrieval, data discovery

1. Introduction

Consider two users of the data archive at an environmental observatory. Joel is an oceanographer looking for simultaneous low oxygen and low pH (high acidity) in a river estuary, as it may indicate that upwelling ocean water is entering the river system. He is interested in data from any time period with these conditions. Lynda is a microbiologist looking for data that includes temperature observations, from near where she collected water samples in early August 2011.

Joel and Lynda face common challenges in finding relevant data in the multi-terabyte archive:

- There are many sources of data: fixed sensor stations, cruise flow through, cruise casts, collected water samples, underwater robots, simulation results. Even if Joel or Lynda knows where all the different data sets are stored, it would be tedious in the extreme to examine each individually to see if it satisfied his or her information need.
- The specific information sought might not exist – coverage is sparse for some data sources. In such cases, data similar to what is desired might still be useful. Joel might find value in a dataset with low oxygen but medium pH. For Lynda, if there is no temperature data from her sampling location in early August, data at that location from late July, or data from early August but from a bit further away from her sampling location, might still be useful.
- Even if one of the scientists believes there is a dataset matching his or her information need, that belief can be mistaken. The scientist might misremember the time or place it was gathered or what measurements it contained, or just be wrong about its existence.
- There are differences between the users about what facets of the data sets characterize the information need, and how important a close match on each facet is. Joel is interested in a fairly broad geographic area, has no restriction on time, but has specific requirements for oxygen and acidity values. Lynda, in contrast, has tighter constraints on area, a time period of interest of a week or so, but no constraints on observed values, just that the dataset has temperature.
- The most relevant data for Joel or Lynda might be hosted in an archive elsewhere.
- Joel and Lynda have access to state-of-the-art visualization and analysis tools, but they find themselves spending more and more time on locating data, and having less for applying those tools.

Both Lynda and Joel work at the same ocean observatory; they are involved in defining the data that the observatory collects, to meet their current research needs. We expect them to have good knowledge about the observatory's data archive, its contents and its structure. However, they are not involved in all collection activities, and memories fade even for those they are involved in [1]. In our experience,

researchers like Joel and Lynda can spend an inordinate time just locating and selecting suitable data sets, before even starting the data analyses that might lead to scientific insights.

Joel and Lynda asked us if we could help. “Data Near Here” is the result.

2. Archives and Gateways

“Big Data” collections, such as observatory archives, represent a large, continuing investment of funds and people. One expects the value of such sources to grow as their holdings increase. Yet there is real danger that each expansion of an archive makes each individual dataset within it more difficult to locate, thus compromising that value. Growth in data must be accompanied by improvements in tools that help scientists like Joel and Lynda easily find the data they need.

In the oceanographic community, efforts to make data sharing easier date back to at least the early 1990s [2]; but the challenge is not limited to that community. Barros et al. [3] describe a digital-library approach for uploading, storing and browsing spreadsheets of ecological observations. Ahrens et al. [4] describe the efforts by the climate-change community to share observations and modeled data via a centralized gateway. As shown in Figure 1 (after their paper), the gateway consists of: a metadata collection, contributed by data providers; modules for data discovery, data integration and access of data products; and a web portal with client-tool access capabilities, which exposes the data to the user community. The authors use the climate-change case study to draw attention to the growth of data and the challenges in making the data useful to researchers, and the similarity of these issues to other scientific domains. They note, “when datastreams aren’t optimally exploited, scientific discovery is delayed or missed.” A similar challenge and conceptually similar solution for ecology is presented by Reichman et al. [5] and separately by Baker and Chandler [6].

Should Joel’s or Lynda’s information needs expand past the geographic region covered by their home observatory, there is much in these efforts that can help them access and analyze datasets – once they have found relevant ones. However, knowing whether relevant data resides in a specific archive is currently dependent on knowing which observatories operate in which areas, and how to locate, access and navigate the relevant portal or gateway. Their problem of finding relevant data (corresponding to data-discovery in Figure 1) is repeated, at a higher level.

3. An Information-Retrieval Approach

The evolution of data-access approaches loosely mirrors the evolution of text-access capabilities on the Internet. Initially, web pages were available for anyone who knew the pages’ URLs. Then, some users created themed directories of pages (such as the first version of Angie’s List); other users navigated these lists and hierarchies to find the pages relevant to them. Simple search capabilities were then added to ease this task. We now have large-scale search engines that index these directories and the pages they catalog. In response to a user query, the search engine identifies and returns – based on the terms the user has searched for – a list of possibly relevant pages, along with a snippet from each. The user can further examine pages from this list to find the ones suited to his or her information need. The search engine acts as a filter, presenting a smaller list of pages than the user would otherwise have to browse; it also orders the list by some notion of relevance, hoping to present the best pages near the beginning.

The scientific community has moved from the “direct-link-to-datasets” phase to the directory approach, and is now transitioning to simple search capabilities for data. Observatories make their data available on

the Internet; each has a website or portal through which a scientist can navigate to find datasets of interest. Some portals offer a text search capability. The user enters words representing his or her interest, and the system searches metadata contributed by the data provider for those terms, treating the metadata much as a text document. However, creating such metadata is a labor-intensive and oft-neglected part of research projects [1], [2]; and such searches are successful only if the metadata contains words that match those for which the scientist searches.

In some geospatially aware portals, the user can enter a geographic area of interest and an additional search capability is used: the geospatial extent of each dataset is compared to the user's area, generally using geometric relationships (such as *contains* or *intersects*). Such a geometric comparison may even be coupled with the results of a text search. Even this combination, though, only begins to address the scientists' needs in searching for data.

We wondered: Can we adapt techniques from Internet search and Information Retrieval (IR) to help Joel and Lynda find relevant datasets? Can we move beyond word-matching and geometric comparisons, and estimate the relevance of a dataset to a scientist's information need? Further, can we do so with interactive response times, for the "big data" found in an observatory?

Internet-based IR approaches separate off-line indexing of a collection of web resources (HTML pages and other documents) from interactive, on-line query. The indexing process is a *feature-extraction* task: each web page is scanned and relevant features extracted. A feature may be the count of a word in the page, or it may be an image name, a title string, or a link found in the page. The features associated with each web page are stored in an index. During interactive query, the user's query terms are compared to each web page's features, and a score computed for the page. Scores are interpreted as a measure of similarity to the query, and hence relevance of a page. The highest-scoring web pages are returned in a list.

We adapt this approach in an architecture for a dataset search engine, shown in Figure 2. As with Internet-search architectures, we first perform offline feature extraction from the datasets. As each dataset is processed, its features are added to a metadata catalog. We then provide an interactive-query component that operates against the metadata catalog to return a ranked list of datasets.

Of course, to realize our approach, we must be able to determine dataset features, express a scientist's information need as a query, and to score and rank the datasets based on their relevance to the scientist's query.

4. Searching for Data Near Here

To make this discussion more tangible, we describe an implementation of these ideas at an ocean observatory, the Center for Coastal Margin and Prediction (CMOP, www.stccmop.org). Figure 3 shows the combined query interface and results page for our application, which implements the user-interface and results-page components shown in Figure 2. We call the application "Data Near Here" (DNH) [7], [8].

Joel and Lynda use the query interface to specify query terms that represent an information need. Joel specifies low oxygen, low pH and the area of the river estuary, while Lynda specifies the date range of interest (first week of August 2011), an area enclosing the locations of her water samples, and her need

for temperature values. They enter their queries using a combination of selections, entry boxes and the map interface (shown in the top section of Figure 3).

Each query is sent to a scoring-and-ranking service, as shown in Figure 2. Results return within a few seconds, in the form of a list of datasets, ranked by score. A "snippet" is provided for each dataset that also includes links to the data or appropriate data-access tools, as seen in Figure 3. Datasets with known geospatial extents are also drawn on the map; the map in Figure 3 shows some cruise legs (diagonal lines) and some single-location vertical profiles (markers) where temperature data was collected. Joel and Lynda can use the links to further explore the details for any dataset and validate its relevance to their research interests. They can move readily from data discovery to accessing data products, and on to data integration (Figure 1).

5. A Notion of Dataset Similarity

DNH relies on a concept of "dataset similarity", used in the scoring-and-ranking component shown in Figure 2, that we've yet to explicitly articulate here. We believe such a concept exists in the heads of the scientists, and we have begun teasing out some its aspects.

A scientist can generally describe the kind of data he or she is looking for in a quantitative way. Joel and Lynda provided such descriptions in our initial scenarios; they use similar descriptions for existing datasets they currently work with. Note that each scientist is, in essence, providing a (partial) summary description of the dataset he or she would ideally like to find; but does not give every detail about the dataset's contents, nor does he or she enumerate the individual observations in the dataset.

Joel and Lynda can tell us if an individual dataset meets his or her information need or not; and further, whether it is an exact match, a "close" match, or "not close at all." We also observe that each often describes the match separately for each part of his or her information need: Lynda may say that a dataset "is in the right area, and has temperature values, but it's not close to the time I want." Joel may say, "The oxygen values aren't in the range I'm looking for." These assessments provide a hint on estimating the similarity of a dataset to a query.

Cognitive science has long recognized that people frequently use distance as a metaphor for similarity, carrying over to non-spatial phenomena and dimensions. Lakoff [9] notes that interpreting time as distance is a common metaphor (for example, "far in the future" or "they are close in age"). While testing shows that people are inaccurate in estimates of absolute distance, it also shows that they are relatively consistent in ordinal rankings. So distance seems a reasonable metaphor to use for dataset similarity. Thus the "near" in Data Near Here connotes more than spatial proximity.

Consider a dataset with a temperature column whose values range between 6 and 9C. If Lynda asks for data with temperatures between 5 and 10C, then all temperature data in that dataset falls within that range. Given three other datasets, with temperatures between 8 and 12C, between 11 and 15C, and between 16 and 22C, we have a good idea how Lynda would rank them in order of closeness to her query term. We used this insight to develop a scoring function that computes, for a numeric query term expressed as a desired range, the distance between that range and the dataset's values. The shorter the distance, the higher the score and the more similarity to the desired range. We apply the same thinking to time intervals and geographic regions.

In prior work we developed a scoring function that can compare data ranges and knowledge of which datasets contain the desired variable and what that variable's data range is [7]. Using this scoring function, the scoring and ranking component in Figure 2 can compute a score for each dataset. For multiple query terms, we scale each distance score by the magnitude of the desired range in the corresponding query term. This scaling produces unit-less scores, which we can combine across query terms, say by averaging. This approach balances query terms against each other: the request for a specific, short time period can be weighed against a relatively large geospatial area. With a combined score for each dataset, we can rank the datasets by similarity to the scientist's query.

6. Creating Dataset Summaries

To use this approach in a search system, we need to quickly compute the relevance scores of many datasets against a scientist's query. Comparing each dataset to a query directly (Figure 4a) to calculate a score does not scale as the amount of data increases. CMOP, for example, has tens of thousands of datasets, and some of them contain millions of observations. It could take hours to answer one query over this archive. If instead we can estimate the relevance from a compact dataset summary, we are better positioned to provide interactive response.

The previous discussion provides clues on creating a useful dataset summary. A column in a dataset can be summarized by the variable name, data type, units (if known) and the bounds of its values; in IR terms, we can consider this information a *feature* for this dataset. Taking the temperature example above, we can summarize the column as ⟨"temperature", float, 6C, 9C⟩. We can simplify spatial data to a geometric feature such as a box or polyline. Such a feature for each column in a dataset, table or spreadsheet, perhaps combined with external information such as file name and file type, can constitute a dataset summary. We pre-compute this summary for each dataset by performing a one-time scan of the dataset in its original location and format (using the feature extraction component shown in Figure 2, and in Figure 4b), and store the summary in an RDBMS along with a pointer to the original data. At query time we apply the scoring formula to our collection of dataset summaries to quickly evaluate estimated scores for a large number of datasets.

While we use column information extensively in producing dataset summaries, our architecture does not restrict us to that form. We can accommodate alternative representations of dataset components, as long as query terms and the similarity functions are adjusted to match. An example is latitude and longitude, which we summarize jointly by a 2-D geometric footprint. For instance, a cast at a single location is abstracted as a point, whereas a cruise track is summarized as a polyline that approximates the vessel's trajectory. A query term for this component of a dataset can also be a 2-D, and we currently use a distance-based similarity function [7].

7. Hierarchies of Scale

Multiple scientists might use the same datasets, but have very different scales of data of interest. Lynda could be looking for data for a fairly short time period, since a different time in the tidal cycle is likely to change her results. In Figure 5, the most interesting portion of data for Lynda is the cruise segment from July 28, 10-12 a.m., since it is closest in time and space to her query. Even though another part of the cruise track intersects her water sample, it is not close enough in time to be very relevant. For Joel, in contrast, the most relevant data is for the whole day of August 1st – a much larger portion of the dataset. In both cases, only a fraction of the entire two-month cruise dataset is relevant to each scientist. These

relevant subsets can be hard to locate and are easily overlooked in a multi-million-observation dataset. What is a "meaningful unit" for one scientist may not be for another. The unit of dataset creation might be different still, driven by observation logistics or processing convenience.

We address this diversity among the "meaningful units" for multiple scientists, and between those and the unit of dataset creation, by allowing multiple summaries to exist simultaneously, representing different subsets of a single, larger dataset. Each subset summary carries with it the relationship it has to contained subsets and to the overall dataset, in a hierarchy. While summaries at all levels are considered when processing a query, a scientist also can explicitly browse up and down the hierarchy via the relationship links. In the example in Figure 5, a dataset for a two-month cruise is broken up into individual days (a temporal split), and then into individual cruise segments or *legs* (a geographic split). Lynda can find and download just the two hours she is interested in, while Joel can navigate up to the whole day that he wants. We can also compose multiple smaller datasets into a larger meaningful unit. For example, for fixed observing stations, we compose each year's months into a year summary, and the years into a lifetime summary. The linkage between datasets in the hierarchy and physical structures is flexible. A dataset can correspond to a file, a portion of a file, a set of files, or a query against a database.

The primary purpose of the hierarchical partitioning is to match the different scales of interest of scientists, although it can also help with search performance. We do not constrain the granularity of the summaries nor the levels in a hierarchy; in particular, we do not require the hierarchies to be uniform, balanced, nor even non-overlapping. We may have different granularities at the same level of a hierarchy, such as the legs of a mobile mission as it navigates through an estuary, and larger subsets for more homogeneous data collected in the open ocean. We treat datasets at different granularity uniformly during search. It is possible for a smaller dataset subset and a containing larger subset to both appear in a query result, though likely with different relevance scores.

8. Making Metadata

How do we merge and partition datasets, to set up the metadata hierarchies? And, how much work is required to create all this metadata?

Metadata creation is an ongoing issue for scientific data collections, including for gateways such as those shown in Figure 1. One group notes that users want more metadata than providers are interested in providing, and that providers stop providing access to data when more metadata is requested from them [2]. Automatic metadata generation is ideal, since the manual metadata annotation by scientists required in most systems is considered burdensome and is often ignored, incomplete, or incorrect.

Hill et al. [10] differentiate *contextual metadata*, which is externally provided (e.g., by a scientist), from *inherent metadata*, derived from automated analysis of the data (e.g., a count of items included in a collection). It is the capturing and searching of inherent metadata that has been our initial focus, though we use contextual metadata in limited forms, such as the quality level assigned to the data (as can be seen in Figure 3). Our dataset summaries are primarily built from information available within the data itself (data ranges) and from the file header and operating system for individual files, or from the database catalog in the case of RDBMS data. Among the main reasons we opted for inherent metadata are uniformity and coverage across repository holdings, the ability to regenerate it as we refine and extend the features we wanted to capture, and, simply, success in using it.

Our collection methodology is "semi-curated", limiting human involvement in metadata gathering as much as possible. In general, the data owner or curator must configure or code certain options once for each new kind of data cataloged. Once an extractor exists for a specific kind of data, new datasets of the same type can be processed automatically. For example: the first data we processed from a science cruise required a new extraction program, as it was our first instance of a mobile collection platform. That extraction program automatically creates metadata as part of normal data handling of observations collected during additional cruises. Adding the next kind of mobile data, for autonomous unmanned vehicles (AUVs), required a few days of additional work due to differences in how that data was stored. Now, additional AUV missions are handled automatically. Handling a third type of mobile observation collection was also quick. In common with the approach used by Internet search engines in crawling the web, we perform the maximal possible preprocessing of metadata during extraction, to minimize computation during search. Since metadata creation is infrequent compared to search, extraction-processing speed is not critical.

Defining hierarchies is similar in terms of the effort involved. When a new kind of data is added to the catalog, we consult scientists on what hierarchy strategy might make sense for that kind of data, and whether an existing hierarchy strategy can be reused. Once a partitioning strategy has been decided and coded, it can be automatically applied as broadly as desired. For example, when we added support for AUVs, we asked if we should treat them the same as existing mobile platforms, or whether there was a reason to segment the data differently. (There was not, and the existing mobile platform extractor was reused with minor modifications.) Right now, deciding what hierarchy to use for specific data collections is an art, although we see common patterns emerging that are likely possible to automate. We note that having subsets of data at multiple levels requires us to pre-compute metadata summaries at all levels for efficiency. However, we generally can collect the metadata for all levels in a single pass of the dataset, so hierarchies do not significantly increase the cost of generating metadata.

We assume that the data will be heterogeneous in format and content. Further, we leave the data in its original format and location, but provide direct access to the data from the summary where a suitable tool can be appropriately parameterized (for example, a parameterized URL for a data download program). Each novel format will require an extractor that can understand that format; where the content volume and meaning warrants separate processing, extraction may be specialized further (for example, mobile versus non-mobile stations, both stored in NetCDF files). We currently have three extractors running to cover the majority of CMOP's observational data holdings. One extractor runs against several thousand NetCDF datasets for fixed-location observing stations, and uses a single time-based policy for defining hierarchies. A second extractor runs against an RDBMS that stores observations from the three kinds of mobile platforms: science cruises, AUVs, and gliders. This extractor uses a mix of temporal and geographic policies for its hierarchy. A third extractor runs against the water-sample collection. This dataset is small but has high value to the scientists, thus making the building of a custom extractor worthwhile.

9. Towards Universal Data Search

DNH has been deployed at CMOP, with strongly positive response from researchers. We are working towards complete coverage of CMOP's current data holdings, while trying to keep up with new observation capabilities. (In some instances, new sources are incorporated with no extra human action, such as for a deployment of an additional sensor gathering data of a previously seen type.) There have been requests to make external datasets searchable through DNH, and we are working on accommodating

some of these requests. (See Sidebar A.) This interest in expanding the coverage of DNH leads us to speculate how broadly it could be applied.

We see little difficulty in deploying another instance of DNH at a different ocean observatory. The main additional work would be creating metadata extractors for datasets unlike those at CMOP, and making decisions about how to hierarchically decompose dataset collections. But what about using DNH for other scientific disciplines? In its efforts to ease data sharing, the oceanographic community differentiated between oceanography-specific issues and discipline-neutral problems, such as data access [2]. In the same way, we have differentiated between oceanography-specific aspects of our implementation (the environmental variables, the details of the hierarchy, the exact similarity function) and the discipline-neutral approaches and ideas. Thus, we believe that the overall architecture – and significant portions of the implementation – of DNH would readily carry over to other scientific repositories with a size and scope similar to CMOP. Other domains would likely require new or modified similarity functions (see Sidebar B), and possibly additional ways to specify query terms and display results.

Can we push the DNH approach further, to be a universal portal for locating relevant datasets across all scientific disciplines – the dataset equivalent of web search? There are challenges to realizing this vision. While CMOP is multidisciplinary, we have had success with a more-or-less-fixed mapping from query terms to dataset features. A broad-spectrum portal might have to do this mapping dynamically, on a per query basis, or at least have a way of selecting among different domain-specific mappings. The issue is not just which physical phenomenon (temperature, pressure, velocity) is being measured or modeled, but also what entity is manifesting it. (Does "temperature" mean "water temperature" or "air temperature" or "star-surface temperature"?) Additionally, in a large portal, scaling (both in numbers of datasets and numbers of queries) and performance are critical. We have begun investigating performance enhancements for DNH [11].

10. Conclusion

Some big-data projects archive output from a single instrument (telescope, collider), and scientists can fairly readily locate the data they need. But in other cases, such as ocean observatories, big data means many datasets, of multiple types, possibly in more than one archive, and navigating among them to find relevant data is a challenge. We have been applying IR techniques – with good success – to providing ranked search of (mostly) numeric data. Our work required adaptation of those techniques, particularly in selection of features and similarity measures. We have also introduced containment hierarchies over datasets to match different scales of interest among scientists. While Data Near Here was initially targeted at internal researchers of one observatory archive, we are already extending it to datasets from other archives of interest to those users. We think it could be possible to extend DNH to community scales, or perhaps even to all of science.

Acknowledgements

We thank the scientists of CMOP, who have been generous with their time (and unstinting in their advice), and the CMOP cyber-team, who have been invaluable in getting DNH deployed and integrated with other data tools [4]. We also thank Ben Sanabria, Justin Corn and Basem Elazzabi, who have helped with various parts of DNH testing and implementation. This work is supported by NSF grant OCE-0424602.

11. References

[1] W. K. Michener, "Meta-information concepts for ecological data management," *Ecological informatics*, vol. 1, no. 1, pp. 3–7, 2006.

[2] 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.

[3] E. G. Barros, A. H. F. Laender, M. A. Gon\ccalves, and F. A. R. Barbosa, "A digital library environment for integrating, disseminating and exploring ecological data," *Ecological Informatics*, vol. 3, no. 4, pp. 295–308, 2008.

[4] 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.

[5] O. J. Reichman, M. B. Jones, and M. P. Schildhauer, "Challenges and opportunities of open data in ecology," *Science(Washington)*, vol. 331, no. 6018, pp. 703–705, 2011.

[6] K. S. Baker and C. L. Chandler, "Enabling long-term oceanographic research: Changing data practices, information management strategies and informatics," *Deep Sea Research Part II: Topical Studies in Oceanography*, vol. 55, no. 18, pp. 2132–2142, 2008.

[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, pp. 55–72.

[8] 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.

[9] G. Lakoff, *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics Into Being*, 1st ed. New York NY: Basic Books, 2000.

[10] L. L. Hill, G. Janee, R. Dolin, J. Frew, and M. Larsgaard, "Collection metadata solutions for digital library applications," *J. of the American Soc. for Information Science*, vol. 50, no. 13, pp. 1169–1181, 1999.

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

Short Bios:

V.M. Megler is currently a PhD candidate in Computer Science at Portland State University, having received an M.Sc. there in 2012 and a B.Sc. from Melbourne University, Australia. Megler's most recent industry position was as Executive IT Architect at IBM, publishing more than 20 industry technical papers on applications of emerging technologies to industry problems. Current PhD research centers on applying Information Retrieval techniques to scientific data; general research interests include applications of emerging technologies, scientific information management and spatio-temporal databases. V.M. Megler can be reached at vmegler@cs.pdx.edu.

David Maier is Maseeh Professor of Emerging Technologies in the Department of Computer Science at Portland State University. His research interests include scientific information management, data stream systems, superimposed information, and declarative cloud programming. Maier has a PhD in Electrical Engineering and Computer Science from Princeton University. He is an ACM Fellow, a Senior Member of IEEE and a member of SIAM. Contact him at maier@cs.pdx.edu.

FIGURES

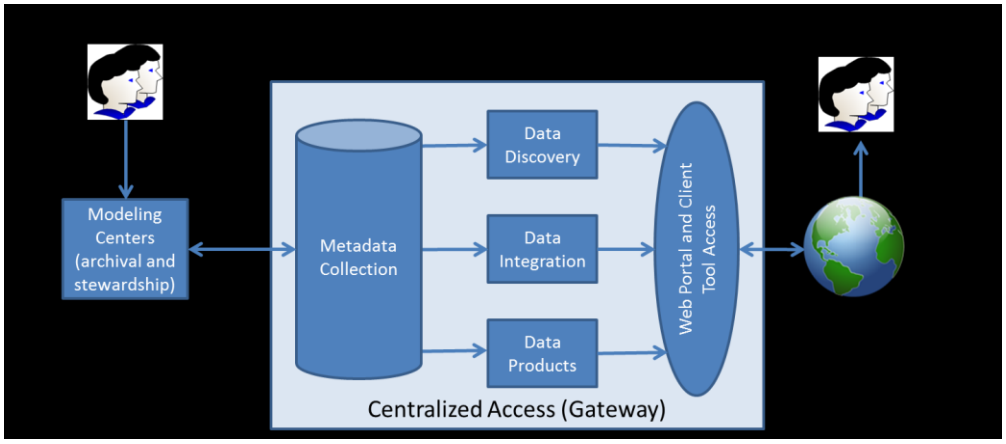


Figure 1. Efforts in the climate-change community to allow users to discover, integrate and download data from a variety of data providers via a centralized gateway. After Ahrens et al. [4]

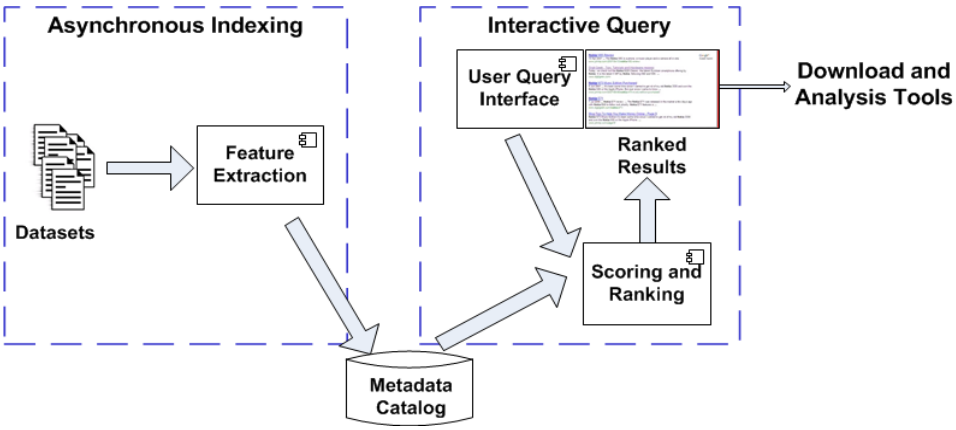
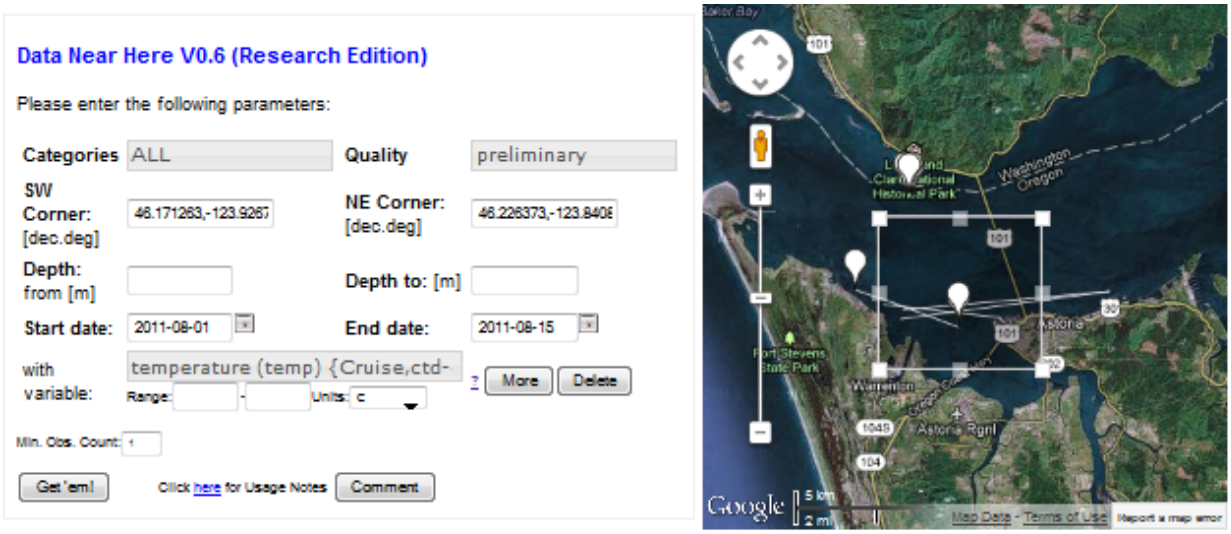


Figure 2. High-Level Architecture for Dataset Similarity (from [10])



There were 50 results returned; all are listed, and 25 initially shown on map. Temp was found in 50 entries.

Display	Type	Collection	Quality	Start Time	End Time	From Depth	To Depth	Temp	Observations	Data Location	Score	DNH
1	Cruise	Cruise, July-August 2011, Wecoma, 2011-08-02, Segment 3	preliminary	2011-08-02 19:47 PDT	2011-08-02 23:59 PDT	-5	-5	15.01:19.20 c	253	Download	100	DNH
2	ctd-casts	July-August 2011, Wecoma, cast072 (Binned)	preliminary_data	2011-08-04 20:27 PDT	2011-08-04 20:41 PDT	-14	-2.5	10.55:17.20 c	25	Download	100	DNH
3	ctd-casts	July-August 2011, Wecoma, cast097 (Binned)	preliminary_data	2011-08-09 09:50 PDT	2011-08-09 09:56 PDT	-9.5	-0.5	17.76:98.98 c	21	Download	100	DNH
4	Cruise	Cruise, July-August 2011, Wecoma, 2011-08-01	preliminary	2011-08-01 00:00 PDT	2011-08-01 23:59 PDT	-5	-5	11.87:19.31 c	1,409	Download	99	DNH
5	Cruise	Cruise, July-August 2011, Wecoma, 2011-08-09, Segment 2	preliminary	2011-08-09 10:05 PDT	2011-08-09 10:49 PDT	-5	-5	17.27:19.13 c	44	Download	99	DNH

Figure 3. The query-entry-and-results page for the Data Near Here application. Lynda enters her query using a combination of drop-down selection lists, text entry boxes, and the movable rectangle on the map. The query results are returned in a list below the query interface, ordered by estimated relevance, and are also shown on the map. The diagonal lines on the map represent cruise legs, while markers represent datasets captured at a single geospatial location (possibly representing many depths). Note that markers wholly outside (but near) the query rectangle have been returned; these markers represent datasets that are close in time and have temperature information.

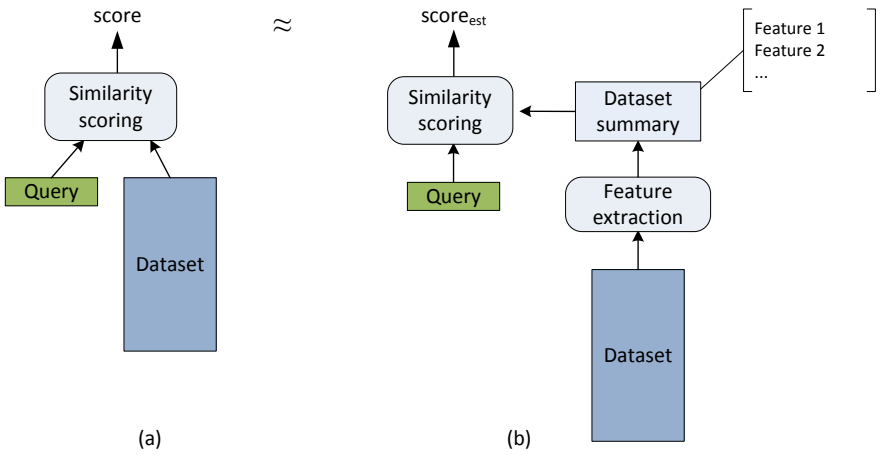


Figure 4. Calculating Dataset Similarity to a Query. (a) Desired Dataset Similarity Calculation. (b) Estimate of Dataset Similarity from Dataset Summary.

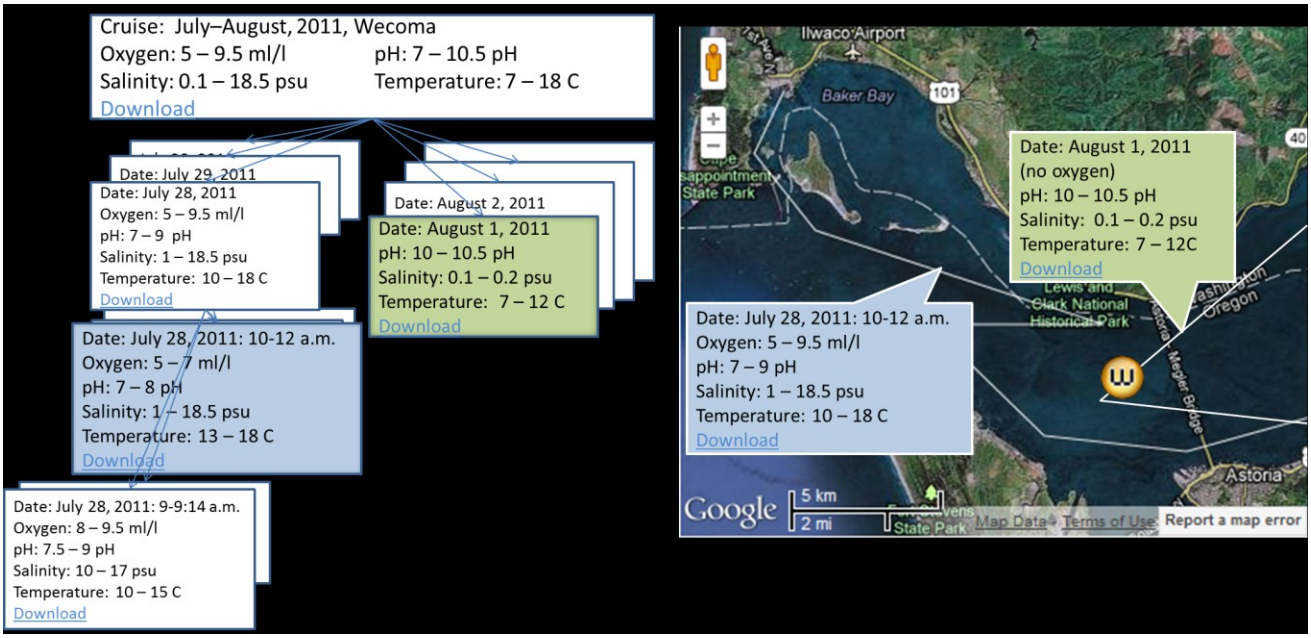


Figure 5. A dataset containing several million environmental observations (taken at 3 millisecond intervals) during a specific 2-month science cruise, segmented into a hierarchy. The white line on the map shows the cruise track, and the marker "w" shows the location of Lynda's water sample. The most detailed level is a single simplified segment or leg of a cruise, often covering a few hours; these segments are aggregated by day, and lastly into an entire cruise dataset. The most relevant portion to Lynda is shown shaded on the left in the hierarchy, while the most relevant portion to Joel is shown shadowed on the right.

SIDEBAR A: Adding Other Archives

Within hours of showing our first users how to search with Data Near Here, we got requests to include other data in its catalogs, including some from sources outside of CMOP. Note that we need not host the data to provide our search service – only to access it to build summaries for our catalog. Some of these sources are easy to incorporate – for example, when an archive collects similar kinds of measurements, such as temperature and salinity, but at another location. For others, we have to create new feature extractors, but the data is comparable enough we do not need to change similarity functions or query terms.

One area that becomes a bit of a challenge is variable names. Even between CMOP data sources, we see some heterogeneity in naming ("salinity", "qa_salinity", "water_salinity"). For the most part, our current users share a consistent set of concepts, so we handle this diversity by a fixed mapping of feature names to query terms. However, if we were to extend DNH to be a *joint* portal, providing data set search across a community of observatories (for use by the combination of their researchers), we would expect to find heterogeneity on the concept side as well. Different users might expect a given query term ("temperature") to map to different variables ("water_temperature", "air_temperature"). Thus the matching of query terms to variables would not be fixed, but would depend on the querier (and perhaps the query). We would like to figure out the matching with the minimum of effort on the user side, avoiding profiles or additional dialogues with the query interface, but figuring out the matching automatically will be challenging. There may be clues in the query as a whole to exploit. For example, hints could come from the search range (30-35C is probably an air temperature) and other query terms ("salinity" would indicate water temperature), for example.

SIDEBAR B: Extending DNH to Other Domains

What would it take to extend Data Near Here to other scientific domains? First off, we would need to understand what constitutes "distance" for those other disciplines. With that knowledge, we would need to formulate feature-extraction and similarity functions that allow us to quickly estimate those distances "well enough", plus set up suitable hierarchies if scientists' interests span multiple scales. Some domains would be more challenging than others:

- Domains using other kinds of environmental and geophysical data should be relatively easy targets, using summary data, query terms and similarity functions like what we have now. For domains that deal with many species, such as metagenomic surveys that sequence thousands of microbes, we would need some new notions of summary and similarity. Here we have a categorical field with an internal structure, namely a taxonomy. Similarity between a query for one species and a data set with another could be based on the number of levels to a common ancestor (1 to the genus level, 2 to the family level), aggregated over the collection of species represented in the dataset.
- Neuroimaging. One difference here is distance between brain regions might be better characterized topologically rather than by geometrically; another is that time, when it appears, is often relative (300ms after stimulation, 2-week-old individual) rather than absolute. Other types of features might be relevant, such as genotypes or diagnoses associated with the individual who was imaged.

- Genomics: Genomic data might require a different notion of similarity than one based on closeness in some metric space. An example we have investigated is gene networks based on correlation of expression over the course of some disease, such as influenza. Here closeness of networks or modules might mean there are similar correlation weights between the same genes, or many genes respond in the same pattern over the course of the disease. If it happens that query terms resemble summary features, as it does in the current DNH prototype, then an existing dataset of interest can be used as query to find similar datasets. Such a "Data Like This" feature is currently supported in DNH (see the "DNH" button in Fig. 3); from our initial discussions, it seems it would be particularly valued in the genomics domain.