Data Near Here: Bringing Relevant Data Closer to Scientists

Large scientific repositories risk losing value as their holdings expand, because of the increased difficulty in locating particular datasets of interest. To address this issue, the Data Near Here application applies information retrieval techniques to dataset search.

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, which may indicate that upwelling ocean water is entering the river system. He's 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 multiterabyte archive:

- There are many sources of data—fixed sensor stations, cruise flow through, cruise casts, collected water samples, underwater robots, and simulation results—so even if Joel and Lynda know where all the different datasets are stored, it would be extremely tedious to examine each dataset individually to see if it satisfied their particular information needs.
- The specific information sought might not exist—coverage is sparse for some data sources. In such cases, data similar to what a scientist desires 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.
- Scientists sometimes mistakenly believe a dataset exists containing the information they need—for example, a scientist might misremember the time or place he or she gathered the information or what measurements it contained or might just be wrong about its existence.
- Not all users have the same specificity requirements—researchers vary in their need for a close match on different facets of the information contained in a dataset. For instance, 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 and a target time period of a week or so, but she has no constraints on observed values beyond 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 locating data, with less time to apply 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

1521-9615/13/$31.00 © 2013 IEEE
Copublished by the IEEE CS and the AIP

V.M. Megler
Portland State University
David Maier
Portland State University

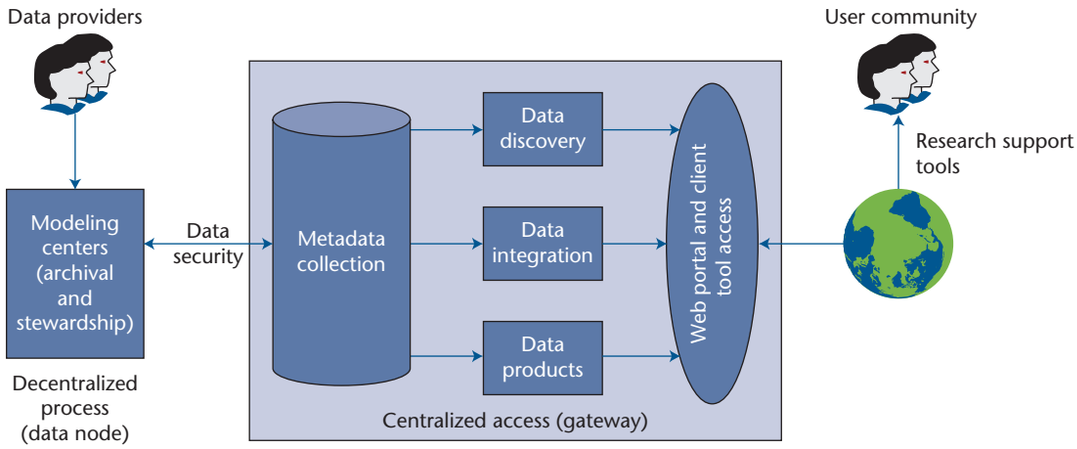


Figure 1. Efforts in the climate-change community to help users discover, integrate, and download data from a variety of data providers via a centralized gateway (modeled after the work of James Ahrens and his colleagues).[4] A centralized gateway integrates and downloads data from various providers. The gateway consists of metadata collections, 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.

knowledge about the observatory's data archive, its contents, and its structure. However, researchers aren't involved in all collection activities, and memories fade even for those they are involved in.[1] Researchers like Joel and Lynda can spend inordinate time just locating and selecting suitable datasets before even starting the data analyses that might lead to scientific insights.

Joel and Lynda asked us if we could help. The Data Near Here (DNH) application is the result.

Archives and Gateways

"Big Data" collections, such as observatory archives, represent a large, continuing investment of funds and people. You would expect the value of such sources to grow as their holdings increase. Yet archive expansion makes each individual dataset within it more difficult to locate, thus compromising that value. To help scientists like Joel and Lynda easily find the data they need, research tools must improve as the data expands.

In the oceanographic community, efforts to make data sharing easier date back to at least the early 1990s,[2] but the challenge isn't limited to that community. Evandrino Barros and his colleagues[3] describe a digital library approach for uploading, storing, and browsing spreadsheets of ecological observations. James Ahrens and his colleagues[4] describe the efforts by the climate-change community to share observations and modeled data via a centralized gateway. As Figure 1 shows, 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 that exposes the data to the user community. The authors use the climate-change case study to draw attention to the challenge of making data accessible and useful to researchers as it continues to expand, and the similarity of these issues to other scientific domains. They note that "when datastreams aren't optimally exploited, scientific discovery is delayed or missed." O.J. Reichman and his colleagues[5] and Karen Baker and Cynthia Chandler[6] present a similar challenge and solution for ecology research databases.

Should Joel's or Lynda's information needs expand past the geographic region covered by their home observatory, these efforts can help them access and analyze datasets—once they've found the relevant ones. However, Joel and Lynda now must know which outlying observatories hold relevant data and how to find and navigate the relevant portals or gateways. They thus repeat the problem of finding relevant data (corresponding to Figure 1's data-discovery approach) at a higher level.

An Information Retrieval Approach

The evolution of data-access approaches somewhat mirrors the evolution of text-access capabilities on the Internet. Initially, webpages were available for anyone who knew their URLs. Then, some users created themed directories of pages (such as the first version of Angie's List) and other users navigated these lists and hierarchies to find the pages relevant to them. Adding simple search capabilities eased this task. We now have

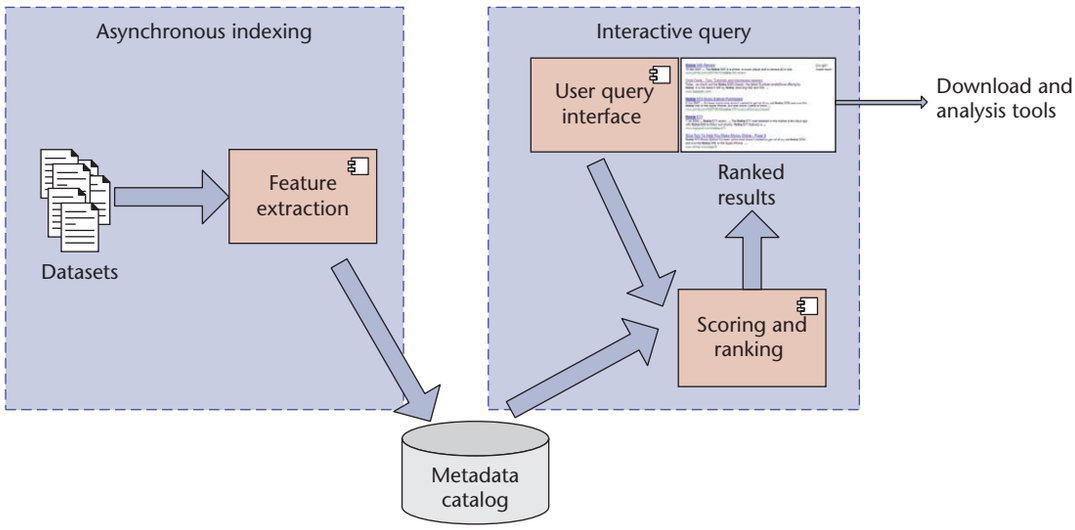


Figure 2. High-level architecture for dataset similarity (based on work by V.M. Megler and David Maier[7]). First, we perform offline feature extraction. Then, as we process each dataset, we add its features to a metadata catalog. Next, we provide an interactive query component that operates against a metadata catalog to return a ranked list of datasets.

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 user's query terms—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 needs. 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 use an additional search capability: 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 might 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 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 information retrieval approaches separate offline indexing of a collection of Web resources (HTML pages and other documents) from interactive, online query. The indexing process is a *feature-extraction* task that scans each webpage and extracts relevant features. One feature could, for example, count the number of times a word appears on the page, or it could be an image name, a title string, or a link found in the page. The metadata catalog stores features associated with each webpage in an index. During interactive query, scoring-and-ranking component compares the user's query terms to each webpage's features, and computes a score for the page. The component ranks the pages by score, as we interpret the score as a measure of similarity to the query and hence, as a measure of a page's relevance. Then it returns the highest-scoring webpages in a list.

We adapted this approach from an architecture for a dataset search engine (see Figure 2). As with Internet-search architectures, we first perform

offline feature extraction from the datasets. As we process each dataset, we add its features to a metadata catalog. We then provide an interactive query component that operates against the metadata catalog to return a ranked list of datasets.

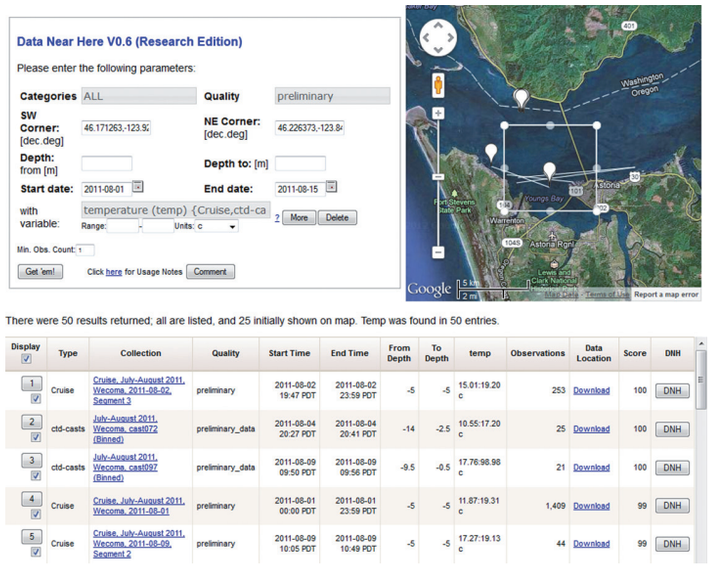
Of course, to realize our approach, we must be able to determine dataset features, express a scientist's information need as a query, and score and rank the datasets based on their relevance to that query.

Searching for Data Near Here

To make this discussion more tangible, we describe how we implemented 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. This is the DNH application.[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 target date range (the 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).

Our application sends each query to a scoring-and-ranking service, as Figure 2 shows. Results return within a few seconds, in the form of a list of datasets ranked by score. The application provides a "snippet" for each dataset that also includes links to the data or appropriate data-access tools (see Figure 3). DNH maps datasets with known geospatial extents; the map in Figure 3 shows some cruise legs (diagonal lines) and some single-location vertical profiles (markers) where researchers collected temperature data. 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).



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 (DNH) application. Lynda enters her query using a combination of drop-down selection lists, text-entry boxes, and the movable rectangle on the map. DNH returns query results in a list below the query interface, in order of estimated relevance, and shows the results 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 DNH returned markers wholly outside (but near) the query rectangle; these markers represent datasets that are close in time and have temperature information.

A Notion of Dataset Similarity

DNH uses a concept of "dataset similarity" to score and rank information (see Figure 2). We believe this concept exists in scientists' minds, and we've begun teasing out some of its aspects.

Usually, scientists can quantitatively describe the data they need. Joel and Lynda provided quantitative descriptions in our initial scenarios; they use similar descriptions for existing datasets that they currently work with. In essence, the scientists provide a (partial) summary description of the dataset they would ideally like to find, but don't give every detail about the dataset's contents, nor do they enumerate the individual observations in the dataset.

Joel and Lynda can tell us if an individual dataset meets their information needs, and further, whether it is an exact match, a "close" match, or "not close at all." We also observe that each scientist often describes the match separately for each part of his or her information need: Lynda might say that a dataset "is in the right area, and has temperature values, but it's not close to the time I want," while Joel might 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. George Lakoff notes that interpreting time as distance is a common metaphor (for example, "far in the future" or "they are close in age").[9] Although tests show that people are inaccurate in

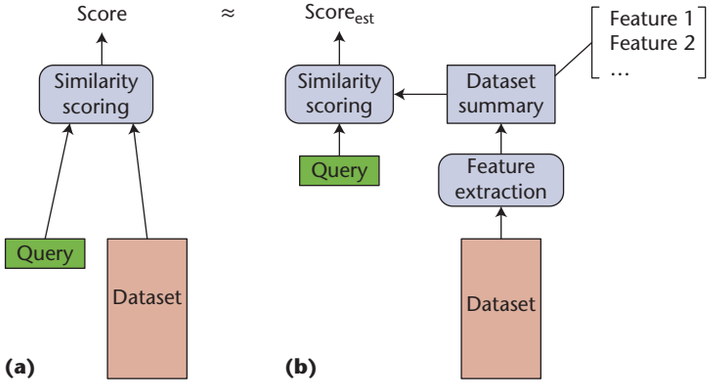


Figure 4. Calculating dataset similarity to a query. (a) Desired dataset similarity calculation, and (b) estimate of dataset similarity from a dataset summary.

estimates of absolute distance, they also show that they are relatively consistent in ordinal rankings. Thus, the "near" in "Data Near Here" connotes more than spatial proximity and works as a metaphor for dataset similarity.

Consider a dataset with a temperature column whose values range between 6 and 9°C. If Lynda asks for data with temperatures between 5 and 10°C, then all temperature data in that dataset falls within her desired range. Given three other datasets, with temperatures between 8 and 12°C, between 11 and 15°C, and between 16 and 22°C, 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. In this scenario, we would say that 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 assess which datasets contain the desired variable and how close that variable's data range is.[8] 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 unitless scores, which we can combine across query terms, say, by averaging. This approach balances query terms against each other: DNH weighs the request for a specific, short time period 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.

Creating Dataset Summaries

To use this approach in a search system, we must 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 doesn't 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 in 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. DNH can summarize a column in a dataset by the variable name, data type, known units, and the bounds of its values; in information retrieval terms, we can consider this information a *feature* for this dataset. Taking the aforementioned temperature example, we can summarize the column as `<"temperature", float, 6C, 9C>`. We can simplify spatial data to a geometric feature, such as a polygon or polyline. Such a feature for each column in a dataset, table, or spreadsheet—perhaps combined with external information such as the file name and file type—can constitute a dataset summary. We precompute 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 Figures 2 and 4b), and store the summary in a relational database management system (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 estimate scores for a large number of datasets.

Although we use column information extensively in producing dataset summaries, our architecture doesn't 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. For example, we summarize latitude and longitude jointly with a 2D geometric footprint. We also abstract a cast at a single location as a point, whereas we summarize a cruise track as a polyline that approximates the vessel's trajectory. A query term for this dataset component can also be in 2D, and we currently use a distance-based similarity function.[10]

Hierarchies of Scale

Multiple scientists might use the same datasets, but have very different scales of target data.

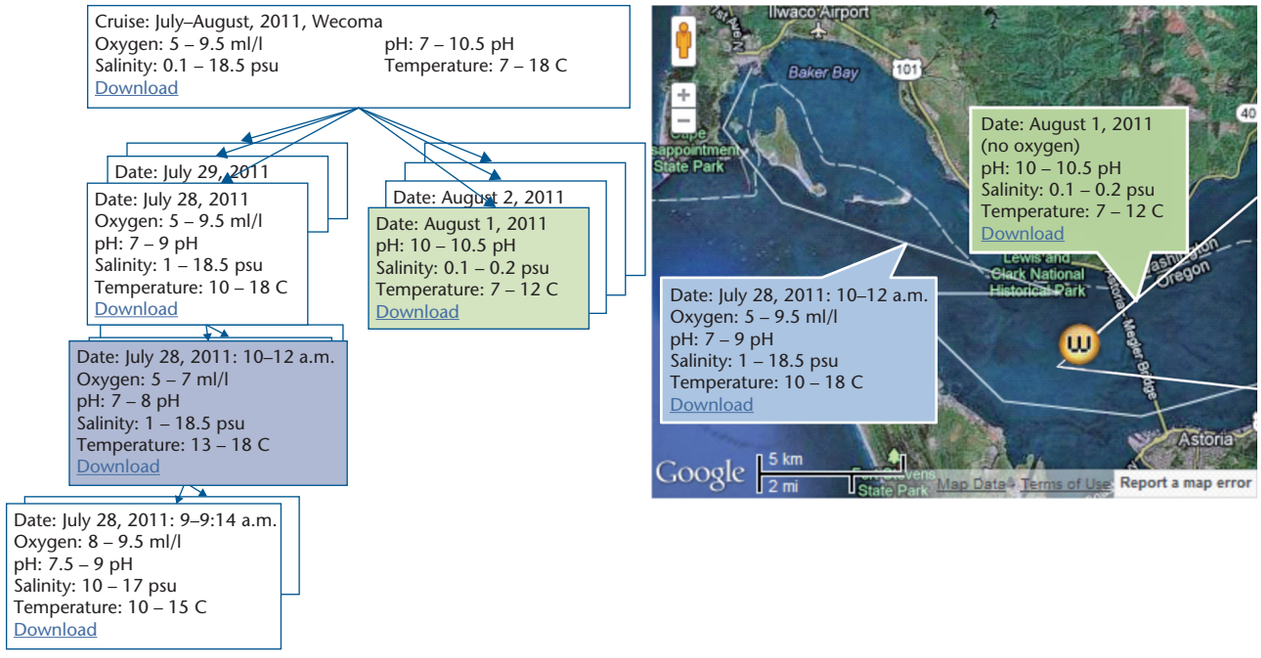


Figure 5. A dataset containing several million environmental observations (taken at 3-millisecond intervals) during a single two-month science cruise, segmented into a hierarchy. The white line on the map shows the cruise track, and the marker "w" shows the location of Lynda's water sample. The most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

Lynda could require data for a fairly short time period, because different tidal cycle times are likely to change her results. In Figure 5, the cruise segment on 28 July from 10 p.m. to midnight interests Lynda most, because it's closest in time and space to her query. Even though another part of the cruise track intersects her water sample, it isn't close enough in time to be very relevant. For Joel, in contrast, the most relevant data is for the whole day of 1 August—a much larger portion of the dataset. In both cases, only a fraction of the entire two-month cruise dataset is relevant; these relevant subsets can be hard to locate and are easily overlooked in a multimillion-observation dataset. What is a "meaningful unit" for one scientist might not be for another. The unit of dataset creation is likely to be different from their interests, as it is driven by observation logistics or processing convenience.

We address this diversity among the "meaningful units" for multiple scientists, and between those units and the unit of dataset creation. To accomplish this, DNH allows multiple summaries to exist simultaneously, representing different subsets of a single, larger dataset. Each subset summary knows its hierarchical relationship to contained subsets and to the overall dataset. Although DNH considers summaries at all levels 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, DNH breaks up a dataset for a two-month cruise into individual days (a temporal split), and then into individual cruise segments or *legs* (a geographic split). Lynda can find and download only the two hours she needs, while Joel can navigate to the whole day of his choosing. We can also compose multiple smaller datasets into a larger, meaningful unit. For example, for fixed observing stations, we can compose each year's months into a yearly summary, and the years into a lifetime summary. The links among datasets in the hierarchy and physical structures are flexible. A dataset can correspond to a file, a portion of a file, a set of files, or a database query.

Primarily, the hierarchical partitioning matches the different scales of interest to scientists, although it can also help with search performance. We constrain neither the summaries' granularity nor the hierarchy's levels; in particular, we don't require the hierarchies to be uniform, balanced, or even nonoverlapping. We can have different granularities at the same level of a hierarchy; the legs of a mobile mission as it navigates through an estuary might have relatively fine granularity

compared to the more homogeneous data collected in the open ocean. During search, we uniformly treat datasets at different granularity. It's possible for small and large dataset subsets to appear in a query result, although it's likely that they'll have different relevance scores.

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 want to produce, and that providers refuse access to data when users request more metadata.[2] Also, most systems require scientists to manually annotate metadata. Scientists often consider this burdensome and ignore or poorly execute the annotations. This makes automatic metadata generation ideal.

Linda Hill and her colleagues[10] differentiate *contextual metadata*, which is externally provided (for example, by a scientist), from *inherent metadata*, derived from automated data analysis (such as a count of items included in a collection). Our initial focus is on capturing and searching inherent metadata, though we use contextual metadata in limited forms, such as the data's quality level (see Figure 3). We build our dataset summaries primarily 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. We opted for inherent metadata for several reasons: uniformity and coverage across repository holdings, the ability to regenerate it as we refine and extend the features we want to capture, and, simply, our success in using it.

Our collection methodology is semicurated, 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, DNH can automatically process new datasets of the same type. For example, the first data we processed from a science cruise required a new extraction program, because it was our first instance of a mobile collection platform. That extraction program now automatically creates metadata as part of normal data handling of observations collected during cruises. Adding another kind of mobile data—for autonomous unmanned vehicles (AUVs)—required a few days of additional work because of differences in how that data was stored. Now, DNH handles additional AUV missions automatically. DNH then was quickly modified to process a third type of mobile observation collection for gliders. Just as an Internet search engine crawls the Web, we perform maximal preprocessing of metadata during extraction to minimize computation during search. Because metadata creation is infrequent compared to search, extraction-processing speed isn't critical.

Defining hierarchies requires a similar amount of effort. When researchers add a new kind of data to the catalog, we consult scientists on what hierarchy strategy might make sense for the data and whether we can reuse an existing strategy. Once we've decided on and coded a partitioning strategy, DNH can automatically apply it as broadly as a user desires. For example, when we added support for AUVs, we asked if we should treat AUVs the same as existing mobile platforms, or whether there was a reason to segment the data differently. (There was not, and we reused the existing mobile platform extractor 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 possible to automate. We should point out that having subsets of data at multiple levels requires us to precompute metadata summaries at all levels for efficiency. However, we generally collect the metadata for all levels in a single pass of the dataset, so hierarchies don't 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 DNH places parameters appropriately for a suitable tool (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, we can further specialize extraction—such as for mobile versus nonmobile stations stored in Network Common Data Form (NetCDF) files.

Currently, three extractors cover most of the Center for Coastal Margin Observation and Prediction's (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

Adding Other Archives

Within hours of showing our first users how to search with Data Near Here (DNH), we received requests to include other data in the application's catalogs, including some from sources outside of the Center for Coastal Margin and Prediction (CMOP; www.stccmop.org). Note that we don't host the data to provide our search service—we only need 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 must create new feature extractors, but the data is comparable enough that we don't need to change similarity functions or query terms.

One area—variable naming—presents a special challenge. Even among CMOP data sources, we see some heterogeneity in naming (for example, *salinity*, *qa_salinity*, and *water_salinity*).[1] For the most part, our current users share a consistent set of concepts, so we handle this diversity through a fixed mapping of feature names to query terms. However, if we were to extend DNH to be a joint portal, providing dataset search across a community of observatories (for use by their various researchers), we'd 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* or *air_temperature*). Thus, the matching of query terms to variables wouldn't be fixed, but would depend on the inquirer (and perhaps the query). We'd like to figure out the matching with minimal user effort, avoiding profiles or additional dialogs with the query interface, but figuring out automatic matching will be challenging. There might be clues in the query as a whole that we could exploit. For example, hints could come from the search range (30 to 35°C is probably an air temperature) and other query terms (*salinity* would indicate water rather than air temperature).

Reference

1. V.M. Megler, "Taming the Metadata Mess," *Conf. 29th IEEE Int'l Conf. Data Eng.*, PhD symp., in press, IEEE, 2013.

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 valuable to the scientists, thus making the building of a custom extractor worthwhile.

Towards Universal Data Search

We deployed DNH at CMOP, with a strongly positive response from researchers. We're working towards complete coverage of CMOP's current data holdings, while trying to keep up with new observation capabilities. (In some instances, DNH incorporates new sources with no extra human action, such as deploying an additional sensor gathering data of a familiar type.) We've received requests to make external datasets searchable through DNH, and we're working on accommodating some of these requests (see the "Adding Other Archives" sidebar). This interest in expanding DNH's coverage led us to speculate about its broader applications.

Deploying another instance of DNH at a different ocean observatory would be relatively straightforward. The primary additional work would be creating metadata extractors for new dataset types and deciding how to hierarchically decompose dataset collections. However, 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 differentiated between oceanography-specific aspects of our implementation (the environmental variables, the hierarchy's details, and the similarity function) and the discipline-neutral approaches and ideas. Thus, we believe that DNH's overall architecture—and significant portions of its implementation—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 the "Extending DNH to other Domains" sidebar), 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—that is, to be the dataset equivalent of Web search? We recognize the challenges of realizing that vision. Although CMOP is multidisciplinary, we've 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 be able to select from among different domain-specific mappings. The issue isn't just which physical phenomenon (temperature, pressure, or velocity) scientists are measuring and modeling, but also what entity is

Extending DNH to Other Domains

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

Environmental and Geophysical Data

Domains using other kinds of environmental and geophysical data should be relatively easy targets, using summary data, query terms, and similarity functions similar to what we have now. For domains that deal with many species, such as metagenomic surveys that sequence thousands of microbes, we would need new notions of summary and similarity. In such domains, we have a categorical field with an internal structure—namely, a taxonomy. We could base similarity between a query for one species and a dataset with another 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 that the distance between brain regions might be better characterized topologically rather than geometrically; another is that time, when it appears, is often relative (for example, 300 ms after stimulation, or a two-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. For example, we've investigated gene networks based on correlation of expression over the course of some disease, such as influenza. Here, closeness of networks or modules might mean that there are similar correlation weights between the same genes, or many genes responding 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, we can use an existing dataset of interest as a query to find similar datasets. DNH currently uses a "data like this" feature (see the "DNH" button in Figure 3 in the main text); from our initial discussions, it seems that this would be particularly valuable in the genomics domain as well.

manifesting it. (For example, 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've begun investigating such performance enhancements for DNH,[11] and we think it could be possible to extend DNH to community scales—or perhaps even to all of science. CISE

Acknowledgments

We thank the scientists at the Center for Coastal Margin Observation and Prediction (CMOP), who have been generous with their time (and unstinting in their advice), as well as the CMOP cyber team, who have been invaluable in getting Data Near Here (DNH) deployed and integrated with other data tools.[8] We also thank Ben Sanabria, Justin Corn, and Basem Elazzabi, who helped with various parts of DNH testing and implementation. This work is supported by US National Science Foundation grant OCE-0424602.

References

1. W.K. Michener, "Meta-Information Concepts for Ecological Data Management," *Ecological Informatics*, vol. 1, no. 1, 2006, pp. 3–7.
2. P. Cornillon, J. Gallagher, and T. Sgouros, "OPeNDAP: Accessing Data in a Distributed, Heterogeneous Environment," *Data Science J.*, vol. 2, 2003, pp. 164–174.
3. E.G. Barros et al., "A Digital Library Environment for Integrating, Disseminating and Exploring Ecological Data," *Ecological Informatics*, vol. 3, no. 4, 2008, pp. 295–308.
4. J.P. Ahrens et al., "Data-Intensive Science in the US DOE: Case Studies and Future Challenges," *Computing in Science & Eng.*, vol. 13, no. 6, 2011, pp. 14–24.
5. O.J. Reichman, M.B. Jones, and M.P. Schildhauer, "Challenges and Opportunities of Open Data in Ecology," *Science*, vol. 331, no. 6018, 2011, pp. 703–705.
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, 2008, pp. 2132–2142.
7. V.M. Megler and D. Maier, "Finding Haystacks with Needles: Ranked Search for Data Using Geospatial and Temporal Characteristics," *Scientific and Statistical Database Management*, 2011, vol. 6809, pp. 55–72.
8. D. Maier et al., "Navigating Oceans of Data," in *Scientific and Statistical Database Management*, vol. 7338, 2012, pp. 1–19.

9. G. Lakoff, *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being*, Basic Books, 2000.
10. L.L. Hill et al., "Collection Metadata Solutions for Digital Library Applications," *J. Am. Soc. for Information Science*, vol. 50, no. 13, 1999, pp. 1169–1181.
11. V.M. Megler and D. Maier, "When Big Data Leads to Lost Data," *Proc. 5th Ph.D. Workshop on Information and Knowledge*, ACM, 2012; http://doi.acm.org/10.1145/2389686.2389688.

V.M. Megler *is a PhD candidate in computer science at Portland State University. Her research centers on applying information retrieval techniques to scientific data, as well as on emerging technologies, scientific information management, and spatiotemporal databases. Megler has an MS in computer science from Portland State University. Contact her at vmegler@cs.pdx.edu.*

David Maier *is the 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 the Society for Industrial and Applied Mathematics (SIAM). Contact him at maier@cs.pdx.edu.*