Searching the Web for Earth Science Data: Semiotics to Cybernetics and Back

Bruce R. BARKSTROM
Asheville, NC 28804, USA

ABSTRACT

This paper discusses a search paradigm for numerical data in Earth science that relies on the intrinsic structure of an archive's collection. Such non-textual data lies outside the normal textual basis for the Semantic Web. The paradigm tries to bypass some of the difficulties associated with keyword searches, such as semantic heterogeneity. The suggested collection structure uses a hierarchical taxonomy based on multidimensional axes of continuous variables. This structure fits the underlying 'geometry' of Earth science data better than sets of keywords in an ontology. The alternative paradigm views the search as a two-agent cooperative game that uses a dialog between the search engine and the data user. In this view, the search engine knows about the objects in the archive. It cannot read the user's mind to identify what the user needs. We assume the user has a clear idea of the search target. However he or she may not have a clear idea of the archive's contents. The paper suggests how the user interface may provide information to deal with the user's difficulties in understanding items in the dialog.

Keywords: Earth science data, Navigational Search, Recall and Precision

1. INTRODUCTION

We are accustomed to using search engines to find information on the Web. Once we bring up the browser, we type in a keyword-based query. Then we receive a list of references that point to HTTP links that may contain the information we want. Examining the items in the list to see if they satisfy our search goals is the lengthy part of the search.

This approach fits reasonably well with the use of databases. There is also substantial experience with algorithms that rank text-based results to try to match user search targets. However, it may not work well for numerical data from Earth science measurements.

One difficulty is the large range of scientific experience and vocabulary in the communities using Earth science data. This diversity contributes to semantic heterogeneity that makes building ontologies difficult [7]. The complexity of natural language semantics also makes it difficult to build automated ontologies that reliably aid text-based searches [11].

User experience also changes search behavior. An early study of library user behavior, [13], found that users who were unfamiliar with a scientific discipline used the catalog metadata to search for items in the library. Researchers with more experience went directly to the library stacks. There they searched for their target material by finding books near an expected location in the library shelves.

This paper suggests there may be some more effective approaches to information searches for Earth science data than keyword queries.

One alternative to keywords (and controlled vocabularies) is to use icons to suggest choices users can make to access useful information. For example, the U.S. National Oceanic and Atmospheric Administration (NOAA) administers a Weather Forecast Web site, [15]. This site provides Web pages with non-verbal icons that link to weather information.

Another NOAA Web site provides information for disaster first-responders, [14]. This site is interesting because it guides users from named storms to digital photos of storm damage largely by providing non-verbal icons. A critical aspect of this site is that it creates a stable decision structure for user choices. Users navigate from their initial entry into the Web site to the desired data items. The site contains images from one to two major storm damage areas per year. Thus, it has only a small selection of storms. It does not try to provide images from every storm.

In contrast, the current design of search engines attempts to provide access to 'everything.' However, the search engine has no way of "reading the user's mind." It is hard for the search engine to make the search precise. It acts as a single agent in charge of finding the user's targets in an extraordinarily large search space.

A more appropriate model may be to emulate the training librarians have in starting a user search with a "reference interview." This approach converts the single-agent search game into a two-agent cooperative game, [16]. The user knows his or her search goal. However, he or she does not know the specific contents of the archive. The search engine has concrete knowledge of the content. It needs to engage the user in a dialog to identify his or her search targets.

During this dialog, the search engine can provide a list of possible links the user might choose. It can also supply hints about why these might be useful choices. Such hints can use links to Web pages that provide definitions or explanatory tutorials. Having a static decision structure for searches also helps users returning to a site later. A user is likely to find it easy to remember a bookmarked location. A jumble of keywords built after the search engine fails to return a useful target creates a very frustrated user.

The next section of the paper discusses the design considerations that provide structure to collections of Earth science data. Then it expands on the difficulties caused by use of keyword queries when users come from diverse communities that are semantically heterogeneous. After that discussion, it considers how using hierarchical taxonomies of search categories can aid in searches. This discussion uses Formal Concept Analysis to link Web pages together. Since users may not understand the words or icons of the search engine, section 5 suggests ways of providing access to ancillary material.

2. ON THE STRUCTURE OF COLLECTIONS OF EARTH SCIENCE DATA

Not many years ago, library scientists had a natural tendency to think that most objects in their collections were like books. These

are discrete objects. Librarians had to create order from the naturally unordered objects. The classic approach was to create ordering principles from metadata given by title, author, and subject. For nonfiction books, librarians shelved them by subject. For fiction, they shelved them in alphabetic order by author.

Serials provided an exception to this approach. The objects in a serial are typically journal issues that contain individual, loosely correlated articles or papers. For a set of journal issues, the volumes were typically ordered by date of publication.

In contrast, producers of Earth science data design collections of discrete files that are correlated with each other. The correlations arise from sampling patterns intended to extract particular information from measurements.

The sampling patterns exist in a multidimensional space that usually includes a time dimension. It often includes two or three spatial dimensions: longitude, latitude, and a vertical coordinate. The axes may include other dimensions such as the direction from which the measuring device sees a target area. Remote sensing instruments need to include the wavelength or frequency of light the observing instrument uses for making measurements. Rather than organizing collections based on textual metadata, Earth scientists organize their collections in multidimensional geometries. This difference is important because it provides a distance metric for ordering collections.

The sociology of the disciplinary group to which the data producers belong influences the sampling patterns. For example, some producers belong to a discipline that usually conducts long-term observations from a network of ground stations. They are likely to produce multiple time series, with one series from each network site. Another example comes from communities that use remote sensing instruments on satellites. Satellite orbits strongly influence the sampling patterns in data files from this type of platform. Some of their data files may be time-ordered images strung along the orbit. Other data files may contain global spatial distributions organized by the sequence of months in the observations. A third example comes from Earth scientists who make measurements from ships or aircraft. They have sampling patterns that follow the trajectory of the measurement platform.

There is a multi-century history of mathematical developments in physics and chemistry that directly tie measurements to the physical reality of the measured quantities. Algorithms for data analysis use these mathematics to interpret the measurements in terms of the physics of the observed phenomena. For example, they incorporate interpolations between and extrapolations from the measurement points.

Data producers derive their algorithms from long-accepted physical principles, such as conservation of energy and momentum. They are not just statistical data analysis approaches, although Earth scientists may use those. Conventional “big data analytics” provides correlations and unexpected anomalies. However, Earth scientists are fond of noting that “correlation does not prove causation.” Thus, they prefer analysis techniques based on first principle derivations from the mathematics that describe the fields being measured.

3. AN ANALYSIS OF KEYWORDS AND ICONS FOR DATA SEARCHES (SEMIOTICS)

The social organizations within which Earth scientists work have several important properties. First, they provide a set of customs, such as experimental protocols. Second, they mediate social interactions using role models and power relationships. Third, they provide a dialect that the community uses for information exchange. Often, a small community's dialect derives from the dialect of a larger disciplinary group. These dialects can become highly specialized. They also evolve over time.

Other communities have similar properties. The dialects of Earth science researchers are not be the same as the dialect of K-12 students. Likewise, the dialects of IT developers or library scientists will differ from the researcher dialects.

It is not easy to produce a consistent set of meanings across a broad subject range because words have multiple meanings. An unabridged dictionary has four to five definitions for each word. [9] provides a memorable quantitative example of this fact. The authors developed a standard vocabulary of about two hundred terms from a cookbook. When they used these terms in a controlled vocabulary, users required two to three iterations to find the `proper' term. The error rate on a single try was frustratingly high.

The term 'semantic heterogeneity' describes some of the difficulties associated with dialect consistency. [4] provides a list of about forty categories for this phenomenon, such as misspellings and case sensitivity.

Non-textual Earth science data has an additional difficulty: the text defining terms may occur in documents separated from the data. Ontologists developing keyword lists may have to rely on short phrases in data producer documentation. Unfortunately, an ontologist's dialect may not include terms commonly used in the vocabulary of data producers. The latter use abbreviations and specialized nomenclature that they may not have time to incorporate in their documentation. Thus, the ontologist may only find abbreviations of key concepts.

The term “fractional cloud cover” is a good example of this difficulty. It is an abbreviated phrase for one kind of field in a satellite data product catalog. This term names the fraction of a particular area covered by clouds. However, when researchers use this term they distinguish different meanings based on which instruments and which algorithms produce the measurements. The fractional cloud cover from a geosynchronous imager is not the same as fractional cloud cover from a ceilometer at an airport. Researchers compare simultaneous numerical values of this parameter over a particular area as part of data validation. They argue about which data source is closest to the `true' value. An ontologist might assume that all a user needs to know is which data files contain “fractional cloud cover.” However, researchers would justifiably note that this assumption might seriously mislead a naive user.

An ontologist might not note the need for metadata distinguishing the instruments providing the data or the algorithms that calculate the numerical values. Without this information, data users may have difficulty distinguishing data from different sources. Library catalogers might place these distinguishing metadata into subject categories using rules such as those from the Library of Congress. Because of the highly technical nature of this information, library catalogers may not have created subject categories for it.

[8, p. 318] notes that

> “For many modeling systems (like object-oriented programming systems, library catalogs, product taxonomies, etc.) a large part of the modeling process is the way items are placed into classes. This process is usually done by hand and is called categorization or cataloging. The usual way to think about such a system is that something is placed

intentionally into a class because someone made a decision that it belongs there."

Because the Semantic Web is built on the notion that "Anyone can say Anything about Any topic," Semantic Web ontologies aim for consistent classifications that may allow items in a collection to belong to many classes. This breadth of coverage makes ontology development by non-specialists in a discipline a very difficult endeavor.

Algorithm derivations create other sources of difficulty. In many fields of the Earth sciences, the derivations use mathematical notations that fall outside of normal text-based semantics. The notations are notoriously difficult to standardize. A person familiar with meteorology would recognize *u*, *v,* and *w* as symbols for the velocity components of the wind. Should an ontology developer include these symbols in the list of keywords? Alternatively, should the developer use much longer keyword phrases, such as "Vertical velocity of air, positive for upwelling?" Perhaps the ontology should include both of these alternatives.

Professional science researchers expect data to have clear statements of uncertainty. These statements help users distinguish between different sources of information. The international standard for the guidelines in such statements is [10]. In simple versions of uncertainty quantification, this standard expects data providers to give lower and upper bounds for each measured value. The (unknown) true value might lie within these bounds with a certain probability. Different users might want to use different probabilities for this range. The correct mathematical answer is to supply a function that would let the user specify his or her probability choice. Of course, a function lies outside the usual results sets returned to a query.

In summary, keyword searches for particular objects of Earth science data suffer from

- Semantic heterogeneity issues
- Divergent vocabularies used by different user communities
- Difficulties connecting sources of keyword terms with numerical data files.
- The need for non-textual material in qualifying searches for objects.

Given these difficulties in developing an ontology that covers many disciplines, it is reasonable to consider a much narrower approach. A search engine might concentrate on making sure that users can discover only objects within the archive that holds the collection of Earth science data.

It may also make sense to use non-textual signs or icons rather than keywords. Using the icons as anchor points for links is a simple application of this idea. Using the icons to guide choices the user can make in traversing a set of choices is an advanced version. This approach uses the icons to lead the user from little knowledge to an end state that selects objects of user interest.

In making this suggestion, we move from text-based queries to a semiotic view. The field of semiotics has many approaches [6]. This paper adopts a position similar to Umberto Eco's. It views semiotics as the study of signs in the context of social groups. [8] provides a rather engaging discussion about how signs become tokens of communication within a social community.

4. WEB-BASED SEARCHES OF HIERARCHICAL TAXONOMIES (CYBERNETICS)

Natural Collection Sequencing Principles for Earth Science Data

In the book-based library observed by Morse [13], experienced researchers were likely to search for books in the library shelves rather than in the library's catalog. For non-fiction books, librarians arranged the book shelving by a subject sequence. These familiar sequences include the Dewey Decimal system or the one from the Library of Congress.

Electronic cataloging can sort objects for presentation to users in any selected order. Google and other search engines use algorithms such as page ranking based on links between Web pages. Of course, page ranking is not necessarily based on a stable sorting structure that can help users search for related objects. As a result, users must undertake the burden of examining the links presented by a conventional search engine. Personal experience suggests that the rankings do not align with an easily remembered subject list. Furthermore, keyword ambiguity can lead to placing popular references ahead of search topics that lead to much less popular technical subjects.

One can hardly expect a user to examine all 1,325,934 'hits' to find the ones that match his or her search topics. Typically, users will search a few pages and then try a new search with rearranged keywords. Alternatively, they may abandon their search.

Collections of Earth science data have an additional source of search confusion. The content of the file metadata is often the same for files in sub-collections. Thus, catalogers cannot use this metadata to distinguish one file in a sub-collection from another. Producers of earth science data organize their collections based on the instruments that make the measurements, as well as their sampling patterns. The fields involved in these patterns include time, as well as latitude and longitude. Some sampling patterns include altitude (for atmospheric data) or depth beneath the surface (for oceanic data). Data may have other attributes that distinguishes one sub-collection from another.

Data production architects use different metadata to sequence their production flows. For example, solar constant (or irradiance) data needs only the time sequence of observations. Solar constant producers reduce all their measurements to the average distance from the Sun to the Earth. They do not use geolocation (longitude and latitude). As a second example, some of the data files from NASA's Earth Radiation Budget Experiment place measurements in twenty-four hour intervals. For data files from a single Sun-synchronous satellite, each file has observations from the entire Earth. The only way to distinguish one file in such a sub-collection from another is by the date of the observation. As a third example, geologists order rock core samples from a single well by depth. For an archive of rock samples, the coarse ordering of the samples uses the geolocation of the well from which they come. Once the producer has selected a well, the archive sequences the samples from it by the depth from which the sample came.

As a practical matter, these intrinsic orderings provide sequences of files that experienced users recognize. Researchers learn these ordering principles as part of their professional education. Search engine designers can order their results to display these 'natural' sequences. That makes searching for particular data objects much easier and more efficient. This approach also lets users perform interpolation searches for particular objects.

Hierarchical Taxonomies

It is natural to think of metadata as attributes curators attach to objects to distinguish one from another. In a classic library cataloging scheme, a librarian places cataloging metadata in the bibliographic record for each book. The most familiar of the fields in this metadata are the subject, title, and author for each book. For each object in a collection of Earth science data, a data producer can record the kind of attributes from section 3 of this paper. The question is how can the provider or curator organize the metadata to aid a search for that object?

An attractive approach is to organize the metadata into a hierarchical taxonomy. This approach is similar to the one used in the biological community to identify plant or animal species, e.g. [5]. A biological taxonomic hierarchy has several levels. At the top level is the biological classification that distinguishes between plants and animals. At deeper levels of the hierarchy, the attributes that distinguish categories are much finer grained. In botany these might involve distinguishing between the color of berries or the shape of leaves.

In this approach, attributes provide answers to the question "how can a cataloger distinguish one set of objects from another set?" Searching is easier if each set has only one distinguishing attribute. If the distinctions require several attributes for each object, searching is more complex.

Web Site Traversal

[2] provides a discussion of the way in which the links between data files and computational processes lead to a layered graph. This data structure contains files that have layers with similar content. It also suggests there is a distinction between primary data product files and files that contain ancillary data. This ancillary data includes such content as calibration coefficients or documentation. As we'll suggest later, a search engine may need to provide a way for users to obtain explanations when they're confused. The search engine can respond by showing links to aids that help in understanding the data or the documentation.

[3] discusses algorithms that convert the layers of data product files into a network of linked Web pages. The key to this work is creating a formal context, a table that connects objects and attributes. The rows in the table belong to the objects in a particular layer. The columns belong to the metadata attributes. In this treatment, the nodes in the layered graph are Web pages. The algorithm shows how to construct links between Web page nodes.

A navigational search identifies a path between the root of the graph and the leaves that contain the target objects. With a hierarchical classification, the search moves from a layer with few details through layers of increasing detail. In each layer, the path starts on a node with no target objects. It ends on a node in the next layer.

Discussion of This Approach

Navigating from the root of the graph to an object that meets the user's needs involves a sequence of user choices. This is different from evaluating a list of results sets from a keyword query.

We can call the usual approach the "Delphic Oracle" scenario. The user approaches the oracle with a list of keywords. The Oracle accepts the list, vanishes behind a screen, and emerges with boxes containing text output with a list of possible query answers. The Oracle is not responsible for evaluating which of the list members satisfies the user's need.

The approach we're suggesting here is more like a librarian's "reference interview." In that interaction, the librarian engages the library patron in a dialog. This dialog helps the librarian understand the context and the user's search target. Some users have very specific targets; others have only a diffusely specified one. Users with poorly specified targets need more help clarifying what they want.

Such a dialog divides the work more evenly between the user and the search engine. The search engine should contain a detailed understanding of the contents of the repository. The user does not have that understanding. However, the user knows (at least roughly) the target of his or her search.

It is unlikely that the search engine will actually read the user's mind. In the Oracular approach, the search engine acts as a single agent responsible for answering very complex questions. The reference interview approach becomes a two agent cooperative game with limited scope, [16].

The dialog is thus formalized by rounds of interaction between the search engine and the user. The search engine's side of the dialog starts with the engine showing the user some choices. When the user selects one of these choices, he or she links to a new Web page. In doing so, the user learns something about the contents of the repository. Hopefully, the choice moves his or her search toward the target. The user's choice provides the search engine with more information about the user's target. The engine automates the presentation of the user's choices. The user's choice provides feedback to the search engine. This feedback regulates the interaction between the engine and the user. It is a cybernetic view of the search process.

We can use information theoretic terms to provide a more quantitative view of the search process. Initially the search engine might quantify the probability of a user selecting one target from N objects as $1/N$. In the usual information theoretic treatment, the entropy, H, of this distribution is $\log_2(N)$. At the end of a successful search, the search engine's uncertainty reduces the entropy to 0. The information gain of the search engine is $\log_2(N)$.

The user's search path through the Web pages may require him or her to make D choices. The average information gain per choice is $\log_2(N)/D$. Roughly speaking, if H/D is large, the reduction in uncertainty for each choice is large as well. This quantification suggests that the interaction can have a steep learning curve.

5. ADDING USER SEARCH AIDS (SEMIOTICS AGAIN)

Providing a navigational search structure is a key feature of our suggested approach. In addition, the user experience needs to deal with at least four other attributes that we describe in the following subsections. We note that we can use signs or icons to help users select their choices. Thus we return to semiotics after touching on cybernetics.

Providing 'Hints' to Help Users Understand Choices

In the middle of navigating through the Web site a user may have difficulty making a selection. The user could be unfamiliar with keywords or icons.

In this case, the user interface can present `hints' about the objects users can choose when they make a selection. For example, in the NOAA ERIC site, the Web page for selecting a storm lists the name of the storm. It also provides the dates on which it caused damage and the locality of the storm's damage. Somewhat deeper in the hierarchy, the site has a map with the location and the path

of the storm. That map also shows a grid of boxes within which the aircraft obtained images.

Linking to Definitions and Explanatory Material

If nomenclature really confuses users, the Web site should provide additional aids. These could include pop-up menus with definitions, as well as links to longer explanatory material or tutorials. The aids might request information about the user's scientific experience or familiarity with the objects at the next level.

Linking to Ancillary Documentation

One goal of the user interface is to provide the user with enough information to understand what he or she is receiving. Links to ancillary documentation can help fulfill this goal.

It is important to match the scientific level of this documentation with the level of user experiences.

If the user is a K-6 student, he or she may be looking for material to use in reports for class assignments. The material provided might include simple explanations and images without copyright restrictions.

If the user is a professional researcher familiar with the scientific context, the interface may provide access to more technical information. This may include instrument designs and blueprints, as well as calibration and characterization documentation. The user may also need documentation or source code for the algorithms used in data reduction. Algorithms contain mathematical descriptions of the logic the producers used to convert instrument measurements to a more highly processed form.

The current emphasis on data transparency and repeatability creates a high standard for descriptions of the assumptions and logic the producers used. [12] provides useful background on the difference between a visual presentation and one that exposes the logical basis for the material. Visual presentations could include videos as well as PowerPoint-style sound bites. However, this style is unlikely to provide an adequate exposition for understanding the logical structure of an algorithm.

Obtaining Data and Code for Checking Repeatability

Finally, the user interface must provide adequate documentation for users to access the data from the archive. Access may require example data and scripts that users can adapt to their local computing environment.

Additional Comments

These suggestions for what a Web site needs to provide access and understanding of a repository's holdings are nontrivial. Overall, they probably require a much greater emphasis on increasing the collaboration between data producers and archive curators. This is likely to require coordination between producers, curators, and funding agencies to obtain the necessary resources.

6. CONCLUDING COMMENTS

This paper provides an outline of an approach for designing a search engine for Earth science data. [14] provides a working example of this approach. The approach makes heavy use of a taxonomic hierarchy of objects in the archive (or repository) that holds the collection. It does not use a search engine that attempts to answer queries for a very broad range of topics. Rather it helps the user navigate through the taxonomy to find objects that the user needs. Section 5 of the paper suggests ways in which this approach can deal with misunderstandings between the user and the search interface. Similar approaches to explanatory material can also aid in helping a less-focused user clarify his or her understanding of search aids.

7. REFERENCES

[1] Allemang, D. and Hendler, J., **Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL**, Second Edition, Amsterdam, NL: Morgan Kaufman (Elsevier), 2011.

[2] Barkstrom, B. R., "A mathematical framework for earth science data provenance tracing", **Earth Sci. Inform.**, Vol. 3, No 3, 2010, pp. 167-196.

[3] Barkstrom, B. R., "Applying User-Guided Dynamic FCA to Navigational Searches", **Proc. Tenth Int. Conf. On Concept Lattices and Their Applications**, La Rochelle, FR, Oct. 15-18, 2013, pp. 275-280, available from <ceur-ws.org/Vol-1062/paper_short1.pdf

[4] Bergman, M. K., "Sources and Classification of Semantic Heterogeneities", **Web Blog: AI3-Adaptive Information, Adaptive Innovation, Adaptive Infrastructure**, 2006. URL: <http://www.mkbergman.com/874/brown-bag-lunch-sources-and-classification-of-semantic-heterogeneities/>.

[5] Britton and Brown, **An Illustrated Flora for the Northern United States and Canada in Three Volumes**, New York, NY: Dover, 1970.

[6] Chandler, D., **Semiotics, the Basics**, Second Edition, New York, NY, Routledge: 2007: [**Semiotics for Beginners** is similar and available at: <http://visual-memory.co.uk/daniel/Documents/S4B/>

[7] Davis, E. and Marcus, G., "Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence,", **Comm. ACM**, Vol. 58, 2015, pp. 92-103.

[8] Eco, U., **Kant and the Platypus***:* **Essays on Language and Cognition**, McEwen, A., translation from the Italian, San Diego, CA: Harvest Book, Harcourt, 1997.

[9] Furnas, G. W., Laundauer, T. K., Gomez, L. M., and Dumais, S. T., "The Vocabulary Problem in Human-System Communication", **Comm. ACM**, Vol.. 30, No. 11, 1987, pp. 964-971.

[10] GCGM/WGI, **Evaluation of measurement data — Guide to the expression of uncertainty in measurement [GUM]**, New York, NY: JCGM, 2010 URL:<http://www.bipm.org/utils/common/documents/jcgm/ JCGM_100_2008_E.pdf>

[11] Green, S., Heer, J., and Manning, C. D., "Natural Language Translation at the Intersection of AI and HCI", **Comm. ACM**, Vol. 58 2015, pp. 48-53.

[12] Lamport, L., "Document Production: Visual or Logical?", **Notices of the Amer. Math. Soc.**, June 1987, pp. 621-624.

[13] Morse, P. J., **Library Effectiveness**, Cambridge, MA: The MIT Press, 1968.

[14] National Geodetic Survey, Emergency Response Imagery, 2016, available at <http://www.storms.ngs.noaa.gov/eri_page/index.html>

[15] National Weather Service, NWS Forecast, 2016, available at <http://www.weather.gov>

[16] Russell, S. and Norvig, P., **Artificial Intelligence: A Modern Approach**, Third Ed., New York, NY: Pearson (Prentice-Hall), 2010.