Multiple Database Personalities: Facilitating Access to Research Data

Sherry K. Pittam, F. Joe Hanus,
Dept. of Botany and Plant Pathology, Oregon State University
Corvallis, Oregon, 97331
pittams/hanusj@nacse.org

and

Mark Newsome, Cherri Pancake,
Dept. of Computer Science, Oregon State University
Corvallis, Oregon 97331
newsome/pancake@research.cs.orst.edu

ABSTRACT

We are moving from an era of largely analytical research into a time where a key research agenda will be to provide insights into relationships and interactions based on information gleaned from data repositories worldwide. Scientists have voiced the need for easy access to existing databases.

As research scientists, we realize that while the today's network infrastructure will readily support access to such data, it doesn't ensure usability. If a database is to accommodate users beyond the research group that created it, we must find ways of giving it additional "personalities" to suit different audiences. Unfortunately, most Web-to-database software targets professional programmers, not scientists.

In this paper, we show that when Web interface software is responsive to the skill levels and preferences of scientists, it can be surprisingly easy to create Web interfaces that expose research data in different ways. We describe how a group of lichenologists exploited HyperSQL, a scientifically-oriented Web-to-database tool, to create database interfaces for two audiences.

The first interface, the *Synoptic Key of the Lichen*, is rather terse, assuming that end-user is an experienced scientist. Using the same database, they constructed *LichenLand*, an interface intended for secondary school students. It uses colorful annotations and simple explanations to emphasize learning through discovery.

KEYWORDS: User Interface Design, World Wide Web, SQL, Database, Biology, Lichen, HyperSQL

1. THE NEED FOR INTERDISCIPLINARY ACCESS TO RESEARCH DATA

A major socioeconomic revolution is taking place that is firmly rooted in the creation, distribution and use of research information [9, 10]. We are moving from an era of largely analytical research into a time where research will provide insights into relationships and interactions based on information warehoused in ever-growing databases worldwide. The need to interpret and synthesize meaningful conclusions from extremely large-scale data banks is driving the emergence of a new breed of scientists, those who can derive principles from data archives [2]. The traditional approach of lone investigators scrutinizing just their own experimental or observed data in pursuit of phenomena or underlying structure is too limited in scope.

The problem for the individual and the scientist no longer is how to acquire and store raw data, but how to access and make sense of large volumes of data gathered for different purposes. Key data reside in databases managed through diverse database software, on diverse platforms, and with entirely different data structures, each maintained by a single agency or individual. The scientists who use each database are expert in its particular domain and are aware of its value and limitations [8]. For outside researchers, however, there may be no indication of which data are more important, most reliable, or even most recent. At one workshop, Robbins described the main features of the data management landscape within molecular biology, referring to the technical and sociological constraints that make it impractical to merge all genomic data into a single, consistent repository. Other participants projected that autonomous organizations will continue to be the dominant mode for managing genomic information into the foreseeable future [13]. This situation extends to other biological disciplines as well. The ability to interact with multiple, remote databases will be vital in expediting research, since it will eliminate the need for data to be re-collected by each researcher. As resources, facilities, and funding for research become scarcer, it will grow in economic importance as well.

Scientists have vocalized, loudly and clearly, the urgency of developing easy access to databases that have already been populated at a considerable investment of time and dollars (e.g., [7, 14]). Many scientists find themselves effectively shut out of even major databases, due to the user-hostile nature of current query languages and interfaces. Formats and nomenclature are idiosyncratic to each database. The researcher often must develop his/her own programs or query scripts simply to understand what data is available.

The problems are exacerbated when data spans multiple disciplines, as is the case for data that might establish significant ecological interrelationships. Now the researcher must cope not only with difficult access and lack of guidance about the database, but also with the distinct traditions of nomenclature, field methodology, data organization, etc., within each disciplinary domain. Yet the need for providing access across broad cultural horizons is clear. Many of the most compelling research problems in the biological and environmental sciences will not be approachable until scientists can be effective in accessing, interrelating, and synthesizing the results of distributed, multi-disciplinary databases. Further, education of the next generation of scientist and policy-makers will not be effective unless it can be based on a manageable subset of realistic data.

2. WHY DATA NEEDS “MULTIPLE PERSONALITIES”

Modern computing and communication systems provide the infrastructure to send bits anywhere, anytime in mass quantities. But connectivity alone is not an answer. Of itself, connectivity cannot assure that useful communication occurs across disciplines or cultures, or that appropriate knowledge is integrated from different sources and domains [6].

As research scientists, we realize that we need access to information held in the data banks of colleagues in other disciplines, and that while connectivity provides access it doesn't guarantee usability. It certainly doesn't imply usability by other communities of professionals outside the scientific community, such as policy-makers charged with determining long term environmental, health and medical objectives, business leaders who are attempting sustainable use of natural resources, or educators who are using “real-life” data in the classroom. As one software developer notes, “We do not want to build user interfaces so simple that the user who needs to undertake a more complex task, for example to issue a sequence of requests some of which depend on the outcome of previous requests, cannot do so at all. A simple point-and-click interface cannot easily express these more complex objectives” [15].

A data resource must have multiple “personalities” in order to accommodate multiple user communities. As Hammond explained, “The ideal [medical data] system permits us to have a system in which all who need data can have exactly what they need. It's complicated by the fact that there are 40 different types of people that probably have a legitimate need for access to health care information” [15]. It's important that data be presented differently according to the context in which it will be used. At the same time, the development of multi-personality databases cannot place an undue burden on the research scientist who maintains the data and who is making it available to others.

The Web has been cited frequently for its role in making the Internet easier to use and more broadly meaningful [15]. In fact, because of the ubiquity of Web clients and the ease with which Web pages can be produced, the Web makes an excellent medium for building multiple database interfaces that are tailored to the needs of widely differing groups.

From the perspective of research scientists, however, Web-based access to databases has been only poorly supported and existing interfaces are extremely non-intuitive. A number of developments are essential if the concept of multiple personalities for research data is to be realized. In particular, it is essential that Web-to-database interface software meet the following design criteria:

1) Interface features must make it possible to minimize the amount of text that must be entered by users. Since the likelihood of inappropriate or misspelled values is much higher among an inter-disciplinary or diverse-level audience, query input mechanisms should be based on recognition rather than recall (e.g., point-and-click from scrollable lists).
2) Additional details must be available at all times to accommodate varying levels of user expertise. Online context-sensitive annotation, instruction, or help should not be more than a click away.
3) The expertise necessary to develop a Web interface should be within the skill range of a scientific researcher. Many of the most up-to-date repositories of key information are in the hands of small research groups that do not include computer scientists or database professionals.
4) It must not be necessary to make substantive modifications to the database in order to give it multiple personalities. Because the primary role of research databases is to advance research within a specific domain, database owners have neither the interest nor the resources to restructure, normalize, or otherwise transform their data simply to permit access to outsiders.
5) Mechanisms should be in place to make the interface largely self-maintaining. Input values for scrollable lists or drop-down menus, for example, should be obtained dynamically from the database at time the query screen is displayed — not hard coded into the Web forms.
6) Security and access privileges should be maintained at the database and operating system level, and not be dependent on Web security mechanisms.
7) In order to safeguard data integrity, it must not be necessary to maintain the Web interfaces on the same machine as the database. This is particularly important when the data includes sensitive information (e.g., distributions of endangered species or personal health data).

8) It should be possible for a skilled user to build his/her own specially tailored interface to a remote database, with nothing more than read-access to the database.

No commercial software products satisfy these requirements. Indeed, several current tools fail to satisfy any of them at all [11]. In response to what we saw as an important emerging need, three of us developed a powerful Web-to-database interoperability layer. HyperSQL [12] is an interpreter that functions as a gateway to remote databases (Figure 1). Its scripting language makes it possible to layer forms- and hypertext-based query interfaces on top of an existing Sybase, Oracle, or ODBC (Windows 95/NT) compatible relational database. and works with any Web browser, the interface can be tried out immediately. HyperSQL provides a number of pre-built "query components" — such as menus and scrollable lists that automatically display all values the database has stored for a given field — making it possible to design interfaces that eliminate the possibility of spelling errors or invalid choices. The interface is self-maintaining because the SQL and HTML are generated dynamically when the user fills out an interface form, so; results automatically reflect the latest information from the database. A special form of hyperlink, called a "querylink," is capable of performing additional queries to access related information elsewhere in the database.

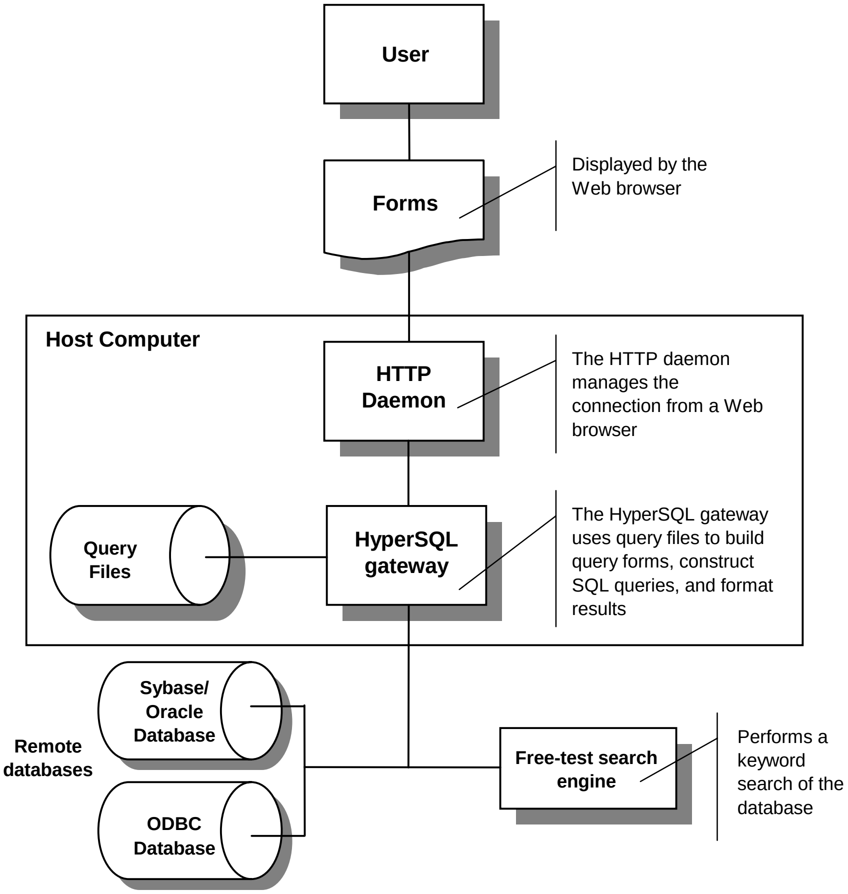

User
Forms
Displayed by the Web browser
Host Computer
HTTP Daemon
The HTTP daemon manages the connection from a Web browser
Query Files
HyperSQL gateway
The HyperSQL gateway uses query files to build query forms, construct SQL queries, and format results
Sybase/ Oracle Database
Remote databases
ODBC Database
Free-test search engine
Performs a keyword search of the database

Figure 1. HyperSQL architecture, illustrating the relationships between databases, the HTTP daemon, and the HyperSQL gateway. HyperSQL has key advantages over database-specific Web interfaces. It is location independent (can be on the database, browser, or a third computer) and requires no modification of either the browser or the database. HyperSQL users can build Web interfaces to remote databases without help from the database owner.

Interfaces created with HyperSQL automatically generate SQL queries, establish communications with the database, and format the results as HTML for the Web browser. HyperSQL also supports password access to restrict control to sensitive data.

To build an interface, a small set of HyperSQL commands is typed into a text file. Because HyperSQL is simple to use

3. MULTIPLE PERSONALITIES FOR LICHEN DATA: AN EXAMPLE:

The use of HyperSQL to support multiple database personalities began when lichenologists from the Dept. of Botany and Plant Pathology downloaded the software to create a Web interface to their data related to lichens of the Pacific Northwest (Figure 2). (A lichen consists of two

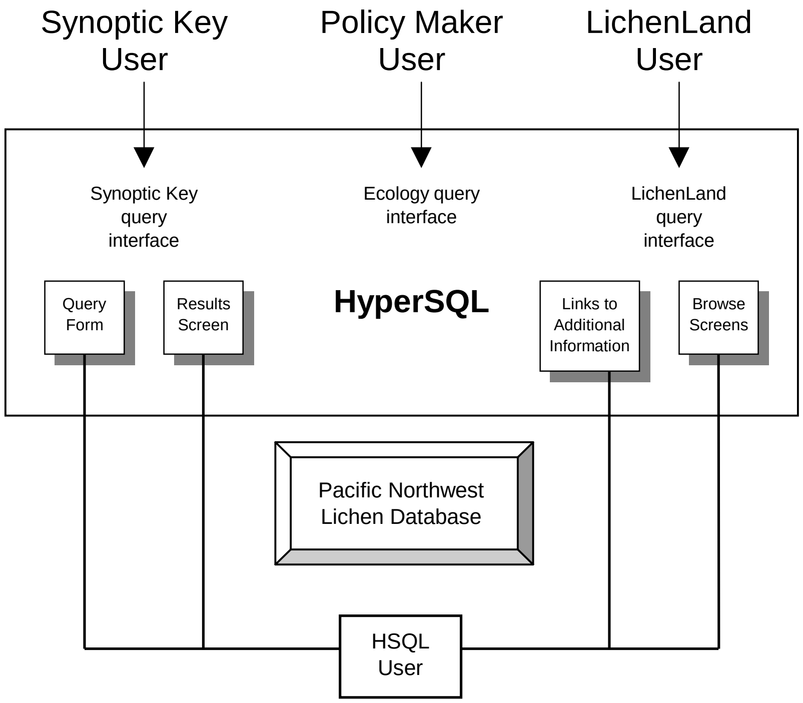


Figure 2. Giving databases multiple personalities. The HyperSQL user can, without SQL coding, create user forms and interfaces quickly and easily. HyperSQL makes it practical to develop Web-based interfaces tailored to specific groups of users.

mutually dependent organisms, fungi and algae, that live as one in a symbiotic relationship.) The relational database, managed with Sybase or Oracle software, houses an extensive collection of literature, image, taxonomic, chemical, and ecological characteristics (Figure 3), accessed via an identification "key."

Traditionally, the use of a key to identify a biological specimen requires making a series of comparisons based on structured questions (e.g., Is it bigger than a tennis ball? If yes, is it slick or is it fuzzy?). This classic structure, called a dichotomous key because of its reliance on yes/no decisions, is the basis for most published keys used classifying biological organisms, from microbes to mastodons. The drawback is that if even a single decision cannot be made (e.g., you cannot determine if the color is mauve or taupe), the process fails. For each organism, there is only a single correct path through the key.

The Web interface to the lichen database was built on a more human-friendly paradigm. Synoptic keys use a checklist of compiled characteristics, each with a set of potential values. The user identifies an organism by checking off all known characteristics. Organisms not fitting the pattern are eliminated, resulting in a list of one or possibly more organisms that match the user's observations.

This approach has some important advantages:

- Identification can be made on the basis of incomplete information (as long as it is sufficient to distinguish the organism).
- Different users can arrive at the same identification after following different paths through the checklist.
- Even when the user can't pinpoint enough characteristics for a conclusive identification, he/she will have narrowed the scope of possibilities in a rational way.
- Results returned from queries using the synoptic key contain images and text, and can consist also of audio and animation objects.

HyperSQL was used to build the Web-based Synoptic Key of the Lichens (Figure 4). This interface is intended for use by professional botanists, foresters, and ecologists. In addition to characteristics based on the appearance of lichens, it permits identification on the basis of chemical and optical analyses. Each characteristic is presented as a drop-down menu of values, which are retrieved dynamically from the database to ensure that information is up-to-date. The user may select a specific value, or leave it specified as a "wildcard." Any number of characteristics can be specified, in any order, before the user submits the query. If the criteria are sufficient to make identification, the database responds with the information, and with querylinks that point to a variety of related information

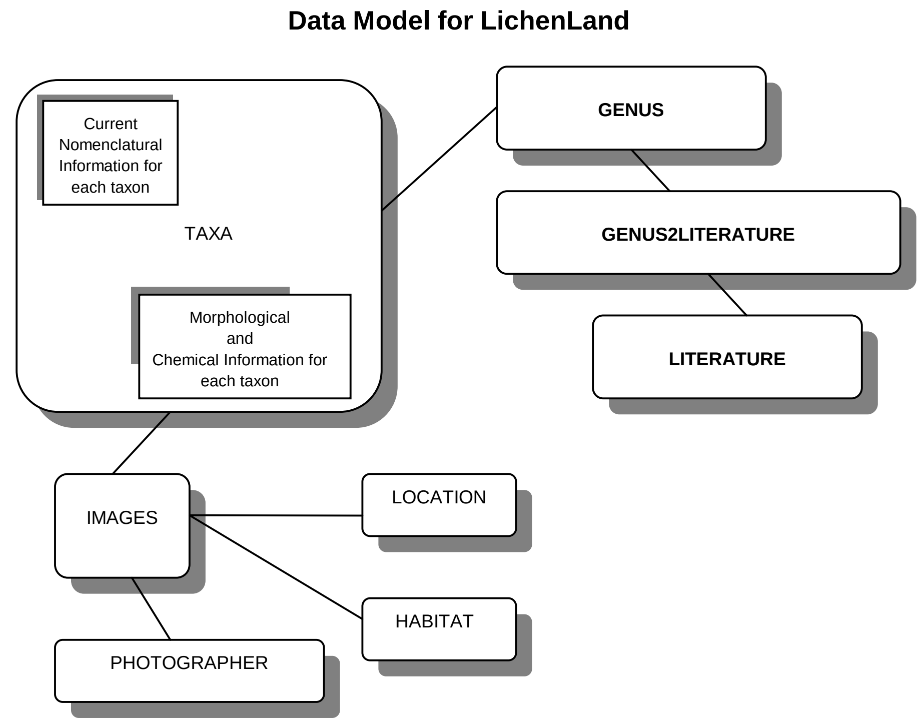

Data Model for LichenLand
Current Nomenclatural Information for each taxon
TAXA
Morphological and Chemical Information for each taxon
GENUS
GENUS2LITERATURE
LITERATURE
IMAGES
LOCATION
HABITAT
PHOTOGRAPHER

Figure 3. Database structure for LichenLand and the professional-level Synoptic Key for the Lichens. The taxonomic data resides in a single table, with key links to the images and genus tables; these in turn contain links to the remaining tables in the database. All data resides in this single database — only the user interfaces are different.

(literature citations, distribution information, images, etc.). When a positive identification is not possible due to insufficient characteristics, the user is able to see how many organisms are in the possible result set, and even proceed to view the images of all possible organisms although the user of this interface is more likely to return to the characteristics page and specify additional attributes. Response to the interface has been extremely favorable, since users have a great deal of latitude in choosing which characteristics to address first.

Recognizing that the database content but not the highly technical presentation would be useful to other audiences as well, the database owners proceeded to create a second interface with a completely different personality. LichenLand is intended for secondary school students. In this case, it has an icon-driven top page that emphasizes learning through discovery (Figure 5). Each cartoon illustrates a taxonomic trait, and is linked to an illustrated discussion of what that trait means and what the choices are for that trait. Again, it is not necessary to specify values for any particular number of traits; if a positive identification isn't possible, the user is offered suggestions about which traits to try next. In many cases, identifications can be made entirely by comparison of the specimen with pictures. The interface makes it possible for even a completely untrained user to experience success in taxonomic identification. Hyperlinks and querylinks also make it easy to learn about the relationships between lichens and their ecological habitats. As with the other interface, response to LichenLand has been overwhelmingly favorable. A number of educators have written to us describing how they are using it to supplement high school coursework in biology and general science.

It is important to note that the underlying database is identical in both cases. It is the Web interface — and the specific queries generated in response to user actions — that creates the illusion of different databases.

4. CONCLUSIONS

Ease-of-use is the dominating characteristic, both of the multiple-personality interfaces described here, and of the interoperability software used to create them. HyperSQL was developed in direct collaboration with biological scientists, in order to ensure that it would meet their needs [11]. The proof of this is that the lichenologists were able to download and install the HyperSQL software without assistance, learn its use, and construct the two interfaces in under two weeks even though they had no particular expertise in database interfaces, Web interfaces, or SQL.

The two Web interfaces to the lichen database are successful largely because they are organized to enhance the ease with which users learn about and apply synoptic keys. In some cases, the same features also simplify the job of the database owners. For example, each interface uses drop-down lists to ensure that only valid values are specified as search criteria to the database. This improves usability by eliminating all possibility of typographic or orthographic errors. (In fact, it simply is not possible to specify invalid information using either interface, since all user input involves selection of value choices, hyperlinks, or querylinks.) Since the lists are generated dynamically when the initial screen is loaded into the browser, the freshness of values is assured, without the need for explicit maintenance of the interfaces.

Although at the present time, both interfaces are completely open to public access, HyperSQL's facilities for maintaining password protection over portions of the database will soon be employed. We anticipate adding information on species that are endangered or threatened, so it will be necessary to ensure that locational information is available only to appropriate researchers.

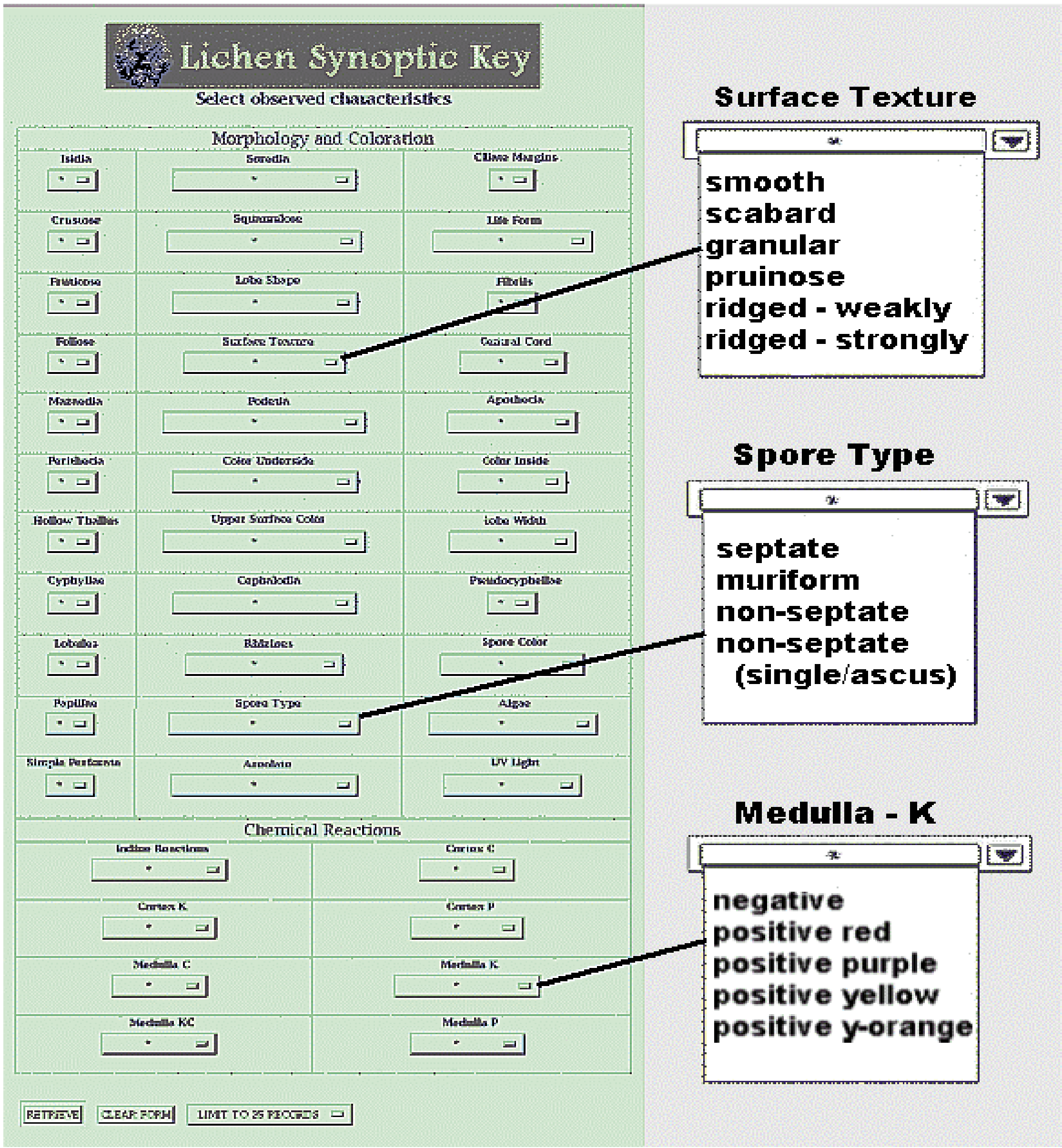
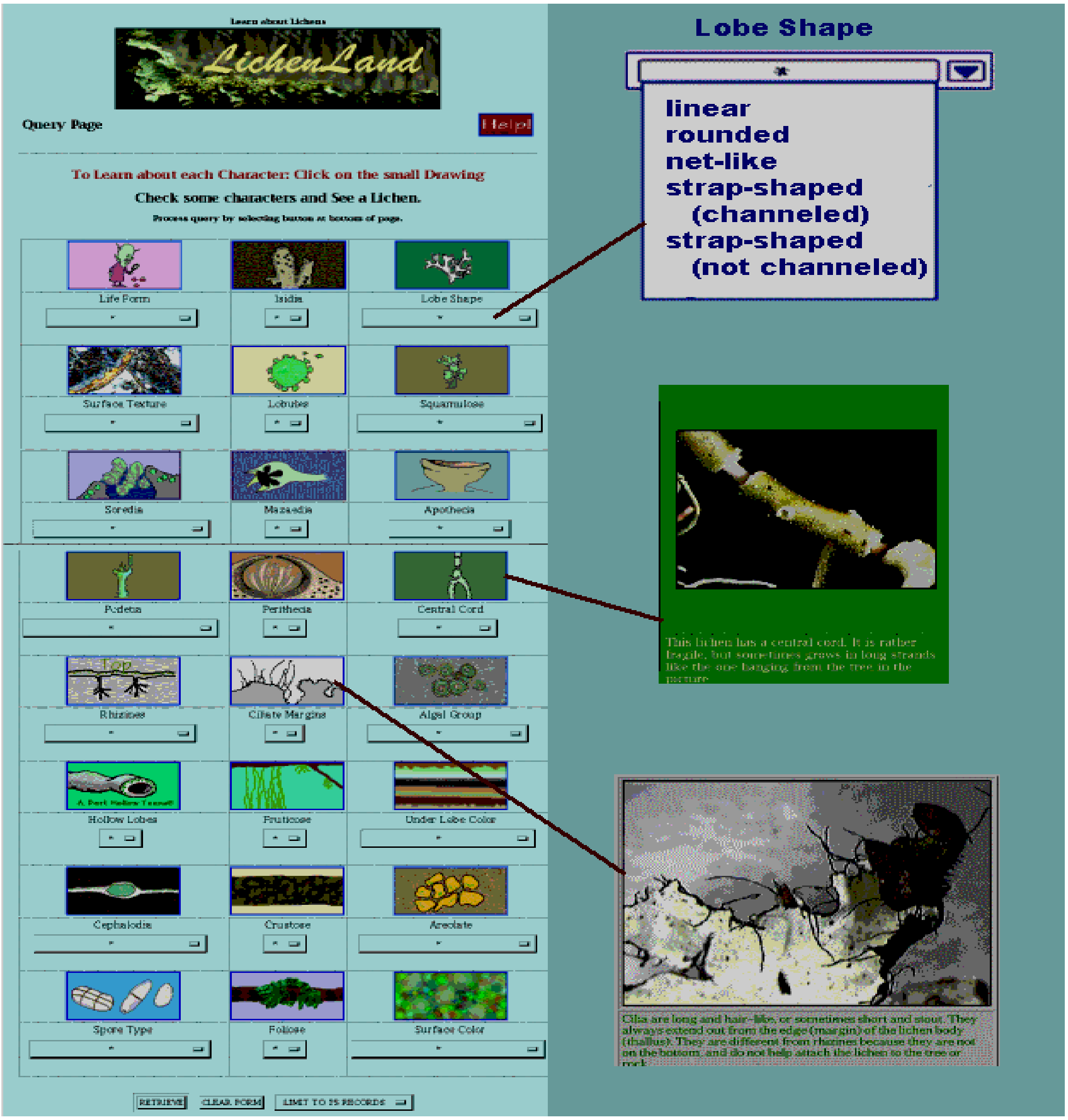


Figure 4. Synoptic Key of the Lichens. This interface targets professional botanists, foresters, and ecologists. It is a streamlined, uncluttered interface listing characteristics that might need to be determined through chemistry or microscopy. Not that each interface (Figs. 4 and 5) provides a unique "personality" targeting a specific user community.



To Learn about each Character: Click on the small Drawing

Check some characters and See a Lichen.

Process query by selecting button at bottom of page.

This lichen has a central cord. It is rather fragile, but sometimes grows in long strands like the one hanging from the tree in the picture

Cilia are long and hair-like, or sometimes short and stout. They always extend out from the edge (margin) of the lichen body (thallus). They are different from rhizines because they are not on the bottom, and do not help attach the lichen to the tree or rock

***Figure* 5.** Lichenland. Designed for the novice, this interface provides a colorful, graphically oriented interface. Each characteristic is linked to a page of descriptive material containing images and text to teach about the trait. HyperSQL produces the drop-down lists, and because they are generated on the fly, the interface does not need to be reconstructed each time data in the tables are modified.

The project demonstrates that when interface software is responsive to the skill levels and preferences of non-computer scientists, it can be surprisingly easy to create Web interfaces that expose research data in entirely new ways. Adding multiple personalities made it possible for disparate groups to access and apply the lichen database in ways that are appropriate for each. The evolution of better ways to access information will require careful study of the people who are to use the data, as well as the development of new mechanisms that provide appropriate pathways through unfamiliar data. It is software like HyperSQL, responsive to the needs of both data providers and their end-users, which will make such pathways possible.

5. REFERENCES

[1] Davis, F.W. 1995. Information Systems for Conservation Research, Policy, and Planning. *Bioscience*, special biodiversity supplement. Summer 1995, pp. 36-41.

[2] French, J.C., Jones, A.K., and Pfaltz, J.L. 1990. *Scientific Database Management*. Technical Report TR-90-22, Department of Computer Science, University of Virginia.

[3] Fayyad, U., Haussler, D., and Stolorz, P. 1996. Mining Scientific Data. *Communications of the ACM*, 39(11:51-57.

[4] Gray, J. 1996. Evolution of Data Management. *IEEE Computer*, 29(10):38-46.

[5] David Messerschmitt. 1996. *In*: NII 2000 Steering Committee, Computer Science and Telecommunications Board Commission on Physical Sciences, Mathematics, and Applications National Research Council. *The Unpredictable Certainty: Information Infrastructure Through 2000*. National Academy of Science.

[6] Kyng, M. 1991. Designing for Cooperation: Cooperating in Design. *Communications of the ACM*, 34(12):65-73.

[7] Lubchenco, J. 1995. The Role of Science in Formulating a Biodiversity Strategy. *Bioscience*, special biodiversity supplement, Summer 1995, pp. 7-9.

[8] Masys, D.R. 1989. Toward Global Data Interfacing. In *Biomolecular Data: A Resource in Transition*. R. Colwell, ed. New York, Oxford University Press, pp. 254-259.

[9] Meleis, H. 1995. The Future of Information Networks and Distributed Applications. *Proceedings of the International Symposium on Autonomous Decentralized Systems*, IEEE Computer Society Press, Los Alamitos, CA, pp. 144-152.

[10] Meleis, H. 1996. Toward the Information Network, *IEEE Computer*, 29 (10):59-67

[11] Newsome, M. A Browser-based Tool for Designing Query Interfaces to Scientific Databases. Ph.D. Dissertation, Department of Computer Science, Oregon State University, 1996.

[12] Newsome, M., Pancake, C., and Hanus, J. HyperSQL: Web-based Query Interfaces for Biological Databases. *Proceedings of the 30th Hawaii International Conference on System Sciences*, Maui, Hawaii, Jan 7-10, 1997.

[13] Robbins, R.J. 1994. Genome Informatics I: Community Databases. Report of the Invitational DOE Workshop on Genome Informatics, 26-27 April 1993. *Journal of Computational Biology*, 1(3):173-190.

[14] Tiedje, J.M. 1994. Microbial Diversity: Of Value to Whom? *ASM News,* 60(10):524-525.

[15] NII 2000 Steering Committee, Computer Science and Telecommunications Board Commission on Physical Sciences, Mathematics, and Applications National Research Council. National Academy of Science. 1996. *The Unpredictable Certainty: Information Infrastructure Through 2000*, Chapter 6: *Public Policy and Private Action*. National Academy Press, Washington, D.C. 1996. Copyright 1995 by the National Academy of Sciences. http://www.nap.edu/readingroom/books/unpredictable/

[16] NSF Workshop on Distributed Heterogeneous Knowledge Networks, co-convened by the National Center for Atmospheric Research and the San Diego Supercomputing Center, funded by the National Science Foundation, Boulder, Colorado May 8-9, 1997. http://www.scd.ucar.edu/info/FORMS/WorkShopHome.Html