Chrysalis: User Agents in the Construction of Floristic Digital Libraries

J. Alfredo Sánchez[1], Cristina A. López[2], John L. Schnase[3]
[1]*Laboratory of Interactive and Cooperative Technologies, Universidad de las Américas-Puebla, alfredo@cca.pue.udlap.mx,*
[2]*Laboratory of Interactive and Cooperative Technologies, Universidad de las Américas-Puebla, cristina@cbi.mobot.org*
[3]*Center for Botanical Informatics, Missouri Botanical, schnase@mobot.org*

Abstract

User agents are the basis for an emerging style of human-computer interaction. This paper describes an agent-assisted approach to the construction of floristic digital libraries, which consist of very large botanical data repositories and related services. In the proposed environment, termed *Chrysalis*, authors of plant morphologic descriptions can enter data into a digital library via a web-based editor. An agent that runs concurrently with the editor suggests potentially useful morphologic descriptions based on similar documents existing in the library. Benefits derived from the introduction of Chrysalis include reduced potential for errors and data inconsistencies, increased parallelism among descriptions, and considerable savings in the time regularly spent in visually checking for parallelism and manually editing data.

Keywords: agents, agent-based interfaces, digital libraries, FNA, Chrysalis.

1. Introduction

Among the emerging fields of study in computer science, Digital Libraries (DL) has become one of the most vigorous multidisciplinary research areas. A host of conferences, workshops and electronic publications (see, for example, [ACM. 1995], [Fox and Marchionini 1996], [Schnase *et al*. 1994], and [Shipman et al. 1995]) attest to the enthusiasm and the creativity of a research community whose goal can be regarded as freeing people from the physical limitations imposed by conventional libraries and enabling new work practices in virtual study and collaboration spaces.

It is clear that fundamental research issues in various disciplines need to be addressed before digital libraries can become a reality. These issues relate to areas as diverse as information retrieval, copyright regulations, databases, digital video, and user interfaces [Fox et al. 1995]. Just constructing the vast data repositories that will support knowledge-intensive activities poses problems of enormous dimensions. Two problems of particular importance are the digitization of existing data in various formats and media (such as books, magazines, microfiche and newspapers), and

the introduction of new data which is constantly generated and needed to keep the library's holdings complete and up-to-date. The use of *autonomous agents*, another emergent computer science area [Lesser and Gasser 1995; Sánchez 1997; Wooldridge et al. 1996], is a promising approach to addressing a number of these problems [Fox 1994].

This paper reports on the development of an agent-based environment that facilitates the task of entering new data into a digital library. This project has been undertaken in the context of floristic digital libraries, which support research activities conducted by botanists studying the flora of specific geographic areas, and also act as information centers for other scientists and the general public interested in biodiversity. The prototype is termed *Chrysalis* and takes advantage of the distinctive characteristics of floristic digital libraries and associated data collection and publishing processes to actively participate in the construction of a data repository. Chrysalis employs an agent-based style of human-computer interaction, in which semi-autonomous entities perform well-defined tasks on behalf of the user.

The paper is organized as follows. Section 2 provides a description of the floristic digital libraries context of the project and the specific problems found in their construction. Section 3 briefly discusses the notions of *user agent* and *task agent*, the agent categories into which Chrysalis can best be classified. Section 4 describes the conceptual design and prototypical implementation of Chrysalis. Finally, Section 5 discusses the project's accomplishments and the ongoing and future work.

2. Problem Area: Floristic Digital Libraries

The importance of plants for life on the planet cannot be overemphasized. Unfortunately, plant species are

disappearing today at a faster rate than scientists are able to study them using traditional means of collecting and sharing information. The use of information and communication technologies to support biodiversity activities (a field now known as *biological informatics*) has fostered the inception of *floristic digital libraries* (FDL), distributed virtual spaces comprising botanical data repositories and a variety of services offered to library patrons to facilitate the use and extension of existing knowledge about plants.

Flora of North America (FNA) and Flora of China (FOC) are specific instances of floristic digital libraries, currently under construction at the Missouri Botanical Garden. In a traditional sense, floras are printed inventories describing all the plant species occurring in a given geographical area. Electronic floras, or floristic digital libraries, represent an effort to make floras accessible to wider audiences and more useful to scientists by taking advantage of advances in areas such as computer networks, user interfaces, and automated information retrieval techniques. In fact, FNA is the first fully electronic floristic research project, and it is expected to produce an ever-expanding, continually refined digital library containing authoritative scientific information about the more than 20,000 species of vascular plants and bryophytes of North America north of Mexico [Schnase *et al*. 1997]. The FNA digital library will consist of a large collection of documents in a variety of media and formats, including text data, maps and illustrations, and will offer a wide range of services to patrons who will be able to use the library anywhere in the global network. FNA's main asset, scientific data about plants, is being collected, prepared and entered by more than 850 specialists working in different places throughout North America. FOC is a similar enterprise aiming at producing a digital library containing data about the more than 30,000 species of vascular plants of China.

2.1 Components of FDL Repositories

Four primary entities are maintained in a floristic digital library which are relevant to the work described in this paper: *taxonomic keys*, *distribution maps*, *illustrations*, and *treatments*. Taxonomic keys are hierarchical indices which are used by botanists as aids to identify specimens based on their inspectable characteristics. Keys exist for each taxonomic level (e.g. family, tribe, genus, species, etc.). Distribution maps graphically depict the exact geographical area where particular species are found, whereas illustrations are artistic renderings of typical specimens in a given taxonomic group. Treatments are the heart of a flora. They provide detailed *morphologic descriptions* of plants at the species and lower (infra-specific) levels. Treatments also refer to publications documenting plants' weedy or toxic nature, their endangered species status, or discussions about their classification and scientific names. Each morphologic description consists of series of descriptors referred to as *structures*, which are intermediate abstractions grouping a number of (often implicit) plant *characteristics* to which a value can be associated. Figure 1 shows an example of a morphologic description. Naturally, in a digital library environment, linking the available data to other information resources and providing means for their utilization greatly enhance the usefulness of a flora as the basis for discussion, academic collaboration and scholarly work.

2.2 FDL Construction Process

Making validated scientific data available (i.e. publishing) in a floristic digital library entails a complex edit-and-review process. The huge volume of information handled and the complexity of the interactions involved in the publication process requires careful activity planning and coordination (FNA is expected to be completed by year 2006). Researchers collect and study plants directly in the field, examine specimens at herbaria, and review previous related work before producing and submitting a treatment to an editorial committee. In the case of FNA, the project's editorial committee consists of 34 plant taxonomists distributed throughout the United States and Canada. Day-to-day project activities are coordinated at the FNA Organizational Center. Five distinct review processes (from scientific content to stylistic reviews) are carried out on each submitted manuscript. Usually, plant structures and characteristics making up the treatments' morphologic descriptions are organized into large spreadsheets to allow for editors to detect missing or inconsistent data. Of particular importance for users of FDLs is *parallelism* among treatments, which ensures that for any two given treatments in the same taxonomic group, values will be specified for the same set of characteristics, and these will appear in the same order. Visually checking manuscripts for parallelism and correcting them manually is a time-consuming and error-prone task.

Both treatment authors and reviewers would benefit from software environments that actively cooperate in the construction of the FDL's data repository. The work presented in this paper aims at assisting authors in the task of building morphologic descriptions, which make up a significant portion of treatments. In a typical scenario, the process of generating treatments and taxonomic keys can be regarded as consisting of two interleaving phases. In a "bottom-up" phase, the author writes treatments for individual instances of a given taxonomic level (e.g. species). Once a number of treatments are available, the author can "move up" and devise a taxonomic key, which summarizes the characteristics of the taxonomic group being described and identifies a higher taxonomic level (e.g. genus). In a "top-down" phase, any new treatment added to the lower taxonomic level will include all the characteristics of the taxonomic key just devised, plus any characteristics that make the taxonomic group distinct from others at the

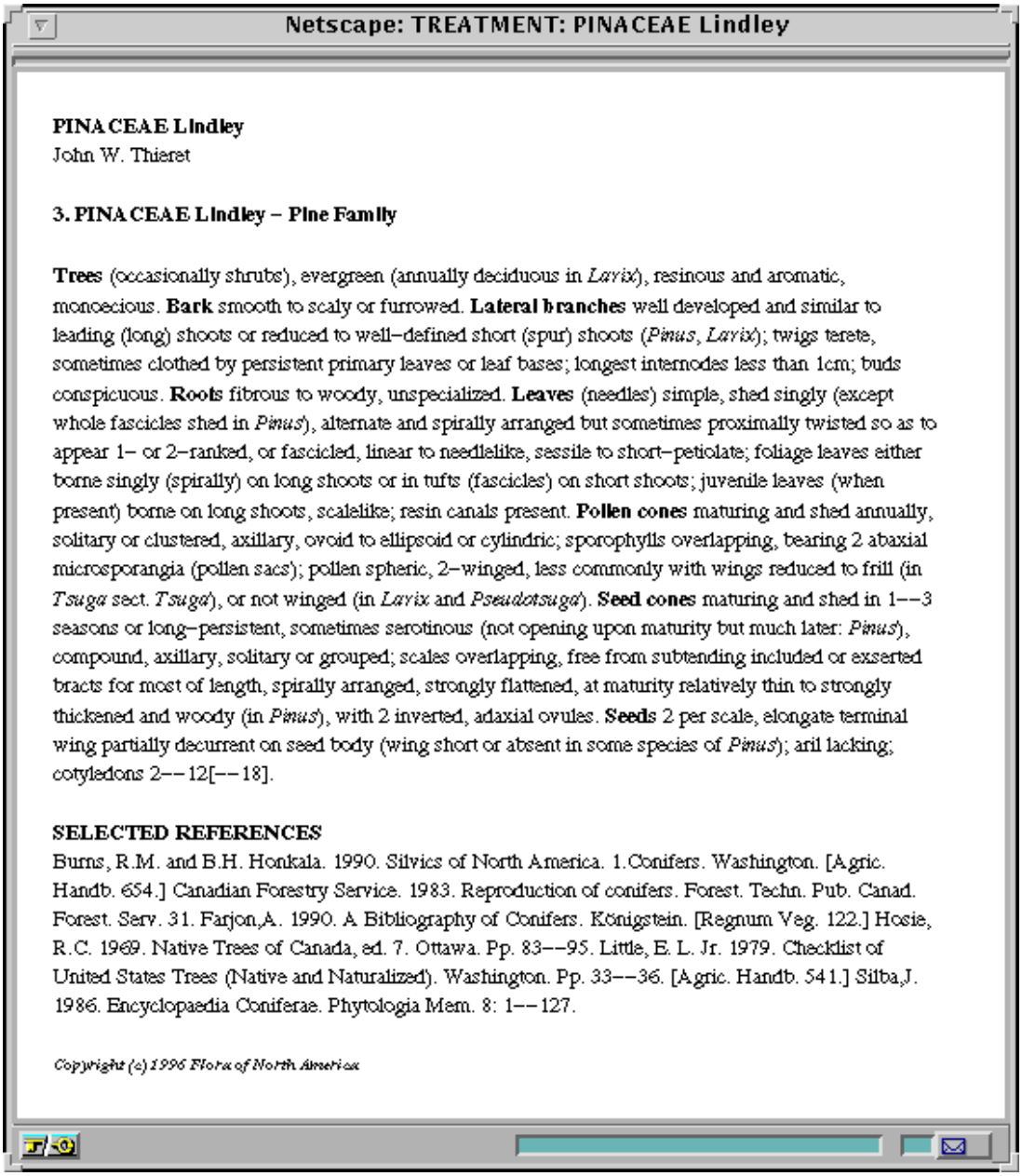

Figure 1. A sample (abridged) morphologic description for the family Pinaceae.

same level. Newly added treatments may result in changes to the taxonomic key and vice versa. In the traditional edit-and-review setting, including characteristics from higher level keys into lower level treatments results in very repetitive tasks for both authors and reviewers.

3. User Agents

One of the services users of digital libraries can have access to is that of *agents* to which they delegate tasks to be performed in a semi-autonomous fashion. This approach to presenting agents as digital library services was introduced by [Sánchez 1996] and [Sánchez and Leggett 1997]. The general notion of *agent* has been considered a highly useful abstraction in multiple areas of computer science. It has, in fact, motivated the formation of a diverse research community studying the potential of agents for a wide range of applications. We have concentrated on the development of and experimentation with *user agents* (also sometimes referred to as *interface agents*), a category of agents that can be directly perceived by end users as software entities

autonomously carrying out a delegated mission. Two other categories of agents, *programmer agents* and *network agents*, which refer to abstractions not necessarily accessible to end users, are discussed in the taxonomy of agents presented in [Sánchez 1997]. User agents and delegation are the basis for a promising style of human-computer interaction that is expected to supplement existing interface paradigms based primarily on direct manipulation and navigation. In the context of very large and dynamic information spaces (such as digital libraries), agents will play an important role in reducing complexity and information overload faced by end users.

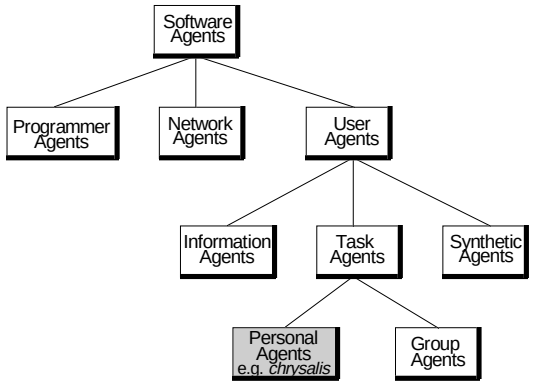

Figure 2. Chrysalis in the general agent taxonomy context.

Three subclasses of user agents can be identified [Sánchez 1997] . *Information agents* help users in dealing with information spaces that are typically unorganized and highly dynamic. A number of information agent prototypes have been developed for use in the WWW (see, for example, [Armstrong *et al.* 1995], [Balabanovic and Shoham 1997], [Lieberman 1995]). *Task agents* help users perform individual or group computer-supported tasks. These agents run concurrently with user applications, watch user activity and offer to automate certain actions. Examples of personal task agents include those documented in [Schlimmer and Hermens 1993] and [Selker 1994], whereas [Lakin 1994] and [Kautz *et al.* 1995] represent examples of group task agents. Finally, *synthetic agents* create engaging environments for users by introducing lifelike characters into the computer interface. Implemented mainly for entertainment purposes, examples of synthetic agents can be found in [Bates 1994], [Mauldin 1994] and [Maes 1995].

Chrysalis, the agent presented in the following section, assists individual authors in the construction of plant treatments to be integrated into a floristic digital library. Chrysalis can therefore be considered a personal task agent. Figure 2 summarizes the foregoing discussion by showing the context of personal task agents and Chrysalis in a general agent taxonomy.

4. Chrysalis: Agent-assisted Treatment Generation

As discussed in Section 2, creating and editing treatments amounts to a significant portion of the process of building a floristic digital library. Treatment authors engage in highly repetitive tasks while integrating plant data to be included in morphologic descriptions. Reviewers and editors must deal with multiple data formats and need to discover often implicit structures and characteristics in order to ensure treatment completeness and parallelism. This section describes Chrysalis, an agent-assisted treatment generation facility aimed at reducing these problems in treatment generation and expediting the entire process of library construction.

4.1 Conceptual Design

Figure 3 is a conceptual diagram of Chrysalis' main components, their inter-relationships, and the context in which they have been designed to operate. For simplicity, only those components of the digital library that are related to Chrysalis are shown in the diagram. The FDL repository is depicted as consisting mainly of data, a database management system (DBMS), and a common database schema on top of which a series of services are implemented. With the introduction of Chrysalis, the end users (treatment authors) will have at their disposal a set of tools which will allow them to remotely enter their treatment data directly into the FDL data repository and thus start the edit-and-review process. A task agent will attentively "watch" as the author enters plant characteristics and, if *sufficiently* similar treatments already exist in the library, the agent will volunteer the available structures and characteristics so the user can reuse or modify their values.

4.1.1 Web-based Editor

The treatment editing tool set includes a web-based treatment editor that automatically incorporates all characteristics appearing in a taxonomic key into the appropriate morphologic descriptions for a given treatment. Based on previously compiled lists of structures and characteristics, this editor facilitates the task of entering values for commonly occurring characteristics. The editor also allows the user to easily add new structures, characteristics and their corresponding values. Reviewers can efficiently check morphologic descriptions for

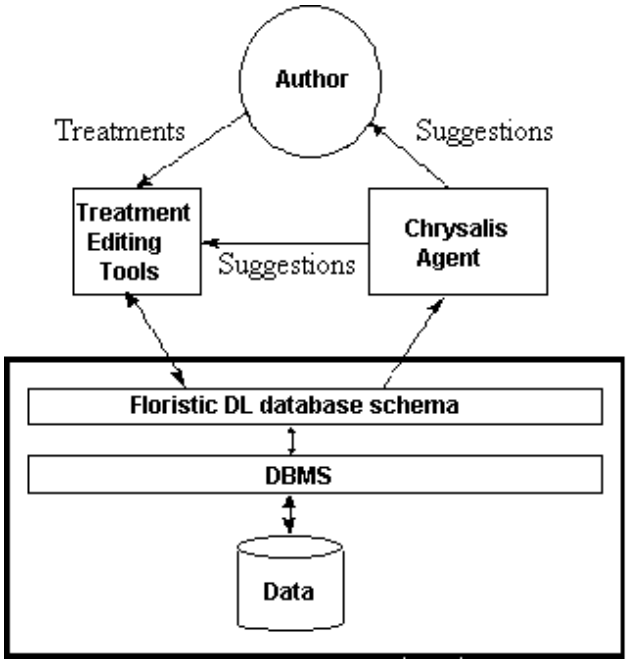

Figure 3. Agent-assisted treatment generation

consistency and style, whereas various mechanisms for measuring and enforcing parallelism can be implemented. By allowing the authors to enter treatment data directly into the FDL database and making a standard list of characteristics available for inclusion in treatments, time previously spent on reformatting or extracting data from manuscripts is saved and the potential for inconsistencies is considerably reduced. This is accomplished while still providing authors with great flexibility to generate dynamically completely new lists of structures and characteristics. The editor can function independently but its usefulness is enhanced when working in parallel with the Chrysalis task agent.

4.1.2 The Chrysalis Agent

Morphologic similarities are fundamental in defining taxonomic plant groups. Indeed, plants in the same genus or the same species share a number of characteristics and values. Some other characteristics may be present in plants within the same group but will have different values. The Chrysalis agent is a semi-autonomous process that takes advantage of these facts to produce suggestions for authors regarding which structures and characteristics to include in a treatment being constructed via the editor just described. As the author enters a morphologic description into the FDL repository, the agent compares the current description with existing treatments in the library. Based on a user-adjustable similarity measure, the agent determines whether an existing treatment might be retrieved from the library so the author can use it as the basis for the morphologic description being generated. If the user so instructs the agent, the characteristics of the suggested existing treatment are incorporated into the new treatment and the author may modify, add or remove characteristics, thus saving typing time and ensuring consistency.

Various mechanisms exist for measuring similarities among documents. Chrysalis' suggestions are based on the *vector space* technique discussed by Salton and McGill [1983]. Using this technique, morphologic descriptions in existing treatments are represented as vectors, each position associated with a characteristic and storing a value or weight. The characteristics being entered by the author are regarded as a *query vector* and the angle it forms with each of the existing treatments' vectors provides a measure of their similarity: the smaller the angle, the more similar the descriptions. Orthogonal vectors imply completely different morphologic descriptions, whereas zero-degree angles imply identical descriptions. A useful range for this similarity measure can be adjusted by the author or the agent as their interaction proceeds and according to the degree of success with which suggestions are produced.

4.2 Prototypical Implementation

A working prototype of Chrysalis has been developed in the context of the Flora of North America digital library. All major functions described in Section 3 have been implemented. Structures, characteristics and their values can be entered via HTML forms which invoke C programs using the CGI protocol. All accesses to data have been implemented using SQL queries via the C API and the "Web DataBlade" provided by Illustra [IUG 1995], the object-relational DBMS used to maintain FNA's repository. The agent's interface has been implemented using JavaScript.

Figure 4 illustrates a sample session with Chrysalis. Before actually entering a morphologic description, the author may select the taxonomic level at which the treatment is being generated. When a taxonomic level is picked, all characteristics and values associated with the corresponding taxonomic key are incorporated into the morphologic description in progress. The Chrysalis agent, represented here as a sunflower cartoon, records the author's choices and starts building a query vector with the initial set of characteristics. When new characteristics are added and a similar treatment is found in the library (see Figure 5), the agent smiles to indicate that a potentially useful morphologic description is available. The author may opt for viewing the suggested morphologic description and adding its data to the current treatment. Characteristics and values may be modified, added or removed according to the author's needs.

Figure 4. The Chrysalis agent watches as an author selects a taxonomic level.

5. Ongoing and Future Work

Agent-based user interfaces are a promising paradigm for human-computer interaction. Task agents such as Chrysalis can contribute to reduce the burden of the repetitive and error-prone tasks usually carried out during the construction of floristic digital libraries. The design and prototypical implementation of Chrysalis demonstrate the potential benefits of agent-assisted treatment generation: delegation of repetitive tasks to agents, reduced inconsistencies, dynamic generation of plant character lists, and increased parallelism.

Chrysalis has been developed in close cooperation with the staff of the FNA project and is currently undergoing a series of user tests in real treatment generation situations. Initial qualitative evaluations have been positive and user feedback is being used to refine the system's design. A full-fledged implementation of Chrysalis is expected to be released in the fall of 1997. Among the most important features to be implemented is user-adjustable similarity measures (in the current implementation, a fixed angle of 45 degrees is used as a threshold for determining whether an existing treatment is considered "potentially useful"). Future work includes adapting and deploying Chrysalis to the construction of the Flora of China digital library and other similar floristic projects.

Acknowledgments

This work is supported in part by grants from the Andrew W. Mellon Foundation and the National Science Foundation (DEB-9505383).

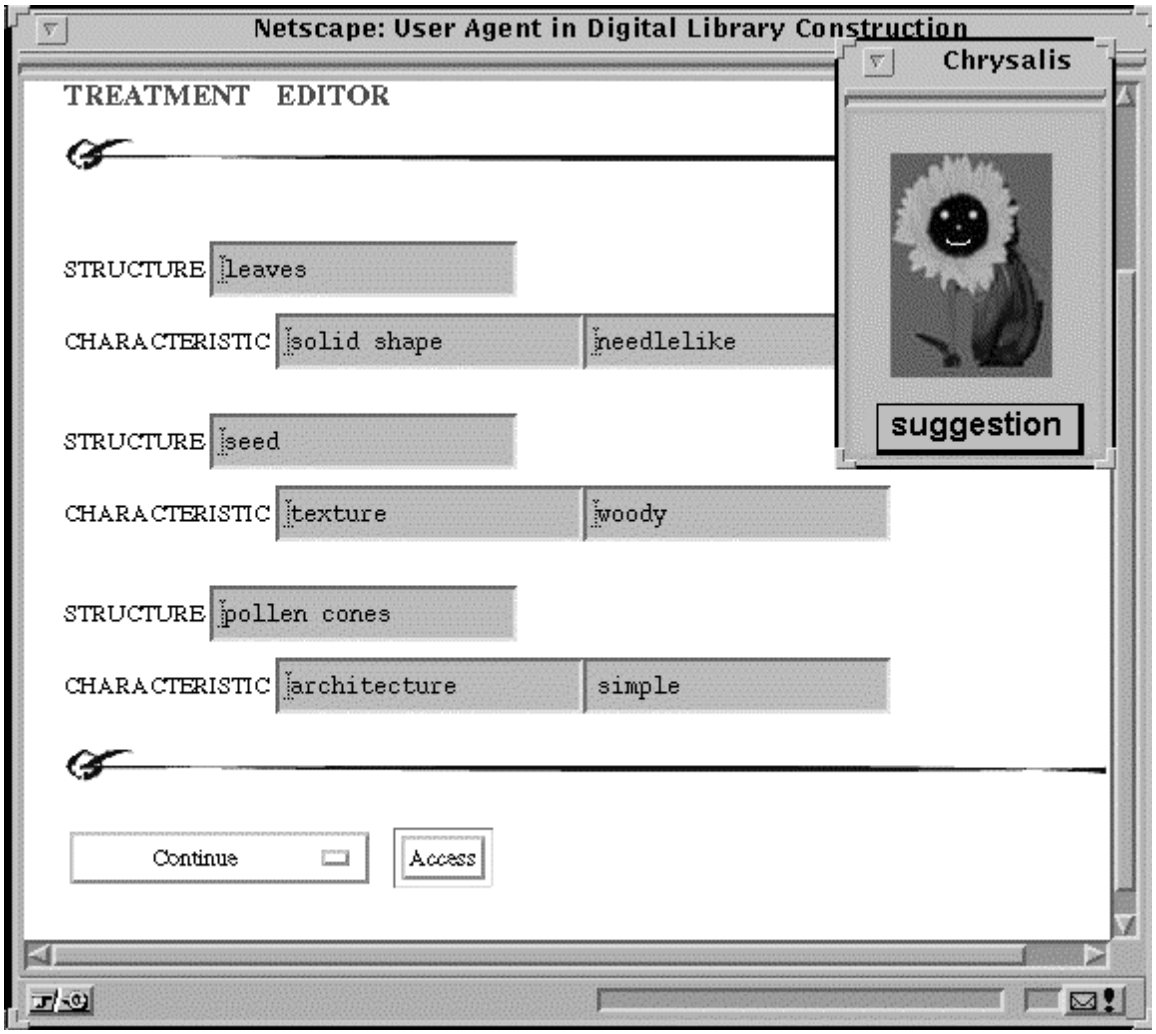

Figure 5. The Chrysalis agent suggests a morphologic description.

References

ACM. 1995. Special issue on digital libraries. *Commun. ACM 38*, 4 (April)

Armstrong, R., Freitag, D., Joachims, T., and Mitchell, T. 1995. WebWatcher: A learning apprentice for the World Wide Web. In *Information Gathering from Heterogeneous, Distributed Environments: Papers from the 1995 AAAI Spring Symposium* (Menlo Park, Calif., March), C. Knoblock and A. Levy, Eds. AAAI Press, Menlo Park, Calif., 6-12.

Bates, J. 1994. The role of emotion in believable agents. *Commun. ACM 37*, 7 (July), 122-125.

Balabanovic, M., and Shoham, Y. 1997. Fab: Content-based, collaborative recommendation *Commun. ACM 40*, 3 (March), 66-72.

Fox, E. 1994. How to make intelligent digital libraries. In *Methodologies for Intelligent Systems: Proceedings of the 8th International Symposium* (ISMIS '94) (Charlotte, N.C., Oct.). Springer-Verlag, New York, N.Y., 27-38.

Fox, E., Akscyn, R., Furuta, R., and Leggett, J. 1995. Digital libraries. *Commun. ACM 38*, 4 (April), 23-28.

Fox, E., and Marchionini, G. (Eds.). 1996. *Proceedings of the 1st ACM InternationalConference on Digital Libraries* (Bethesda, Md., March). ACM Press, New York, N.Y.

IUG. 1995. *Illustra User's Guide*. Release 3.2. Illustra Information Technologies, Inc., Oakland, Calif.

Kautz, H., Milewski, A., and Selman, B. 1995. Agent amplified communication. In *Information Gathering from Heterogeneous, Distributed Environments: Papers from the 1995 AAAI Spring Symposium* (Menlo Park, Calif., March), C. Knoblock and A. Levy, Eds. AAAI Press, Menlo Park, Calif., 78-84.

Lakin, F. 1994. A visual agent for performance graphics. In *Software Agents: Papers from the 1994 AAAI Spring Symposium* (Menlo Park, Calif., March). AAAI Press, Menlo Park, Calif., 103-106.

Lesser, V., and Gasser, L. (Eds.). 1995. *Proceedings of the First International Conference on Multi-Agent Systems* (ICMAS-95) (San Francisco, Calif., June). AAAI Press/The MIT Press, Menlo Park, Calif.

Lieberman, H. 1995. Letizia: An agent that assists web browsing. In *AI Applications in Knowledge Navigation and Retrieval: Papers from the 1995 AAAI Fall Symposium* (Menlo Park, Calif., Nov.), R. Burke, Ed. AAAI Press, Menlo Park, Calif., 97-102.

Maes, P. 1995. Artificial Life meets entertainment: Lifelike autonomous agents. *Commun. ACM 38*, 11 (Nov.), 108-114.

Mauldin, M. 1994. Chatterbots, tinymuds, and the Turing test: Entering the Loebner Prize competition. In *Proceedings of the Twelfth National Conference on Artificial Intelligence* (AAAI '94) (Seattle, Wash., August). AAAI Press, Menlo Park, Calif., 16-21.

Salton, G., and McGill, M. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, N.Y.

Sánchez, J. A. 1997. A taxonomy of agents. Tech. Rep. ICT-97-1. Laboratory of Interactive and Cooperative Technologies. Department of Computer Systems Engineering, Universidad de las Américas-Puebla, Cholula, Pue. 72820, México.

Sánchez, J. A. 1996. Agent services. Ph.D. Dissertation. Department of Computer Science, Texas A&M University, College Station, Tex., August.

Sánchez, J. A., and Leggett, J. J. 1997. Agent services for users of digital libraries. *Journal of Network and Computer Applications, 20*, 1 (Jan.), 45-58.

Schlimmer, J., and Hermens, L. 1993. Software agents: Completing patterns and constructing user interfaces. *Journal of Artificial Intelligence Research 1* (Nov.), 61-89.

Schnase, J. L., Kama, D. L., Tomlinson, K. L., Sánchez, J. A., Cunnius, E. L., y Morin, N. R. 1997. The Flora of North America digital library: A case study in biodiversity database publishing. *Journal of Networks and Computer Applications, 20*, 1, 87-103.

Schnase, J., Leggett, J., Furuta, R., and Metcalfe, T. (Eds.). 1994. *Proceedings of Digital Libraries '94* (College Station, Tex., June). Hypermedia Research Laboratory, Texas A&M University, College Station, Tex. (Also available from http://www.csdl.tamu.edu/DL94.)

Selker, T. 1994. Coach: A teaching agent that learns. *Commun. ACM 37*, 7 (July), 92-99.

Shipman, F., Furuta, R., and Levy, D. (Eds.). 1995. *Proceedings of Digital Libraries '95* (Austin, Tex., June). Hypermedia Research Laboratory, Texas A&M University, College Station, Tex. (Also available from http://www.csdl.tamu.edu/DL95.)

Wooldridge, M., Müller, J., and Tambe, M. (Eds.). 1996. *Intelligent Agents II*. Springer-Verlag, New York, N.Y.