CINDI: A System for Cataloguing, Searching, and Annotating Electronic Documents in Digital Libraries

Bipin C. Desai, Rajjan Shinghal, Nader Shyan, and Youquan Zhou

Department of Computer Science,
Concordia University,
1455 de Maisonneuve Blvd West,
Montréal,
CANADA H3G 1M8
Email contact: bcdesaics.concordia.ca
http://www.cs.concordia.ca/~faculty/bcdesai

Abstract. This paper describes a system called CINDI for cataloguing, searching, and annotating electronic documents in a digital library, the library being distributed over a computer communication network. A document is catalogued both on its syntactic and semantic content. This makes later searching for the document easier and more precise. On accessing a document, one can write annotations on the document, Such annotations can be read by people who access the document later. Graphical user interfaces are provided for each of cataloguing, searching, and annotating. The user of CINDI is helped by an expert system that mimics the expertise of professional librarians.

1 Introduction

The electronic documents distributed over a computer communication network in a digital library must be catalogued properly so *searchers* can easily locate them. Many systems (Koster, 1994; Mauldin, 1995; Pinkerton, 1994; Kahle, 1991) catalogue a document on words selected from it. A program (known in the literature as a *robot*, *worm*, *spider*, or *crawler*) traverses the network accessing the documents to be catalogued. Such a program, of course, adds to the traffic on the network. A number of systems use a document's semantics to catalogue it, for example, CORE (Cromwell, 1994), MARC (Petersen and Molholt, 1990) MLC (Horny, 1986), and TEI (Gaynor 1994). These systems can be used optimally only by professional cataloguers, who are usually expensive.

In this paper, we describe a system called CINDI (Concordia INdexing and DIscovery System), for cataloguing, searching and annotating electronic documents in a distributed digital library. A professional cataloguer is not needed, since the cataloguing is done by the *provider* of the document, as it is he (we are using *he* in a generic sense for both males and females) who knows his document best. A searcher can then locate and access the document. After accessing, the

searcher may annotate it for the convenience of those who may want to access the document later. CINDI provides a graphical user interface (GUI) for each of cataloguing, searching and annotating. The overall structure of CINDI is shown in Figure 1. The details of CINDY's cataloguing, searching, and annotating are given, respectively, in Sections 2, 3, and 4. Concluding remarks are in Section 5.

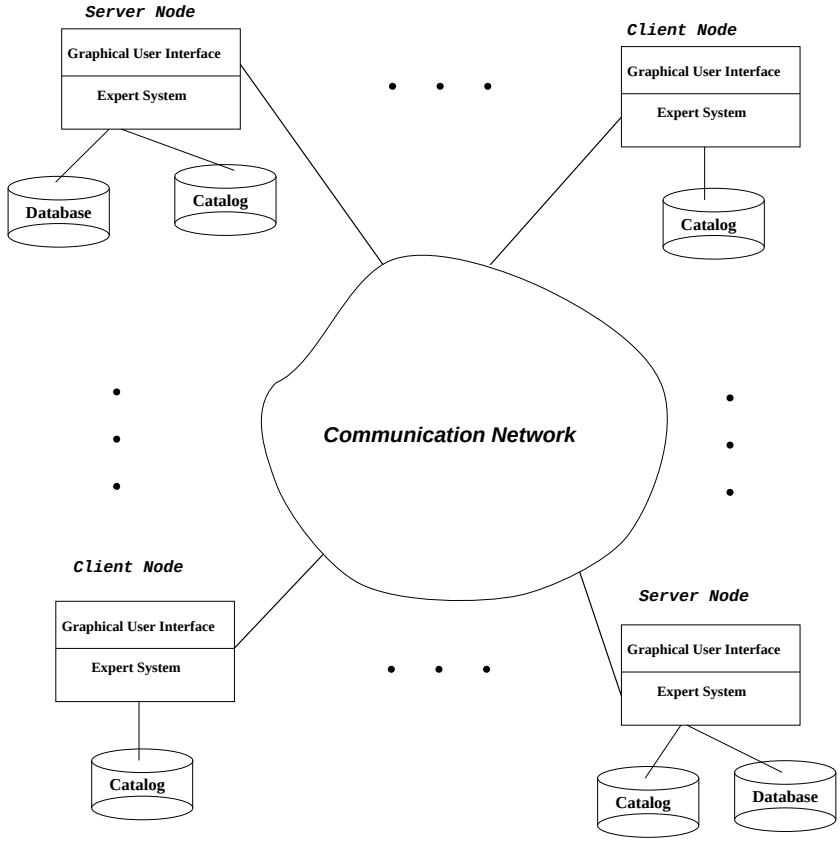

Fig. 1. Overall structure of CINDI.

2 Cataloguing

When a person puts a document on the net, he is the one who can best describe its syntactic and semantic contents. Accordingly, he fills in the slots of a meta data structure called the *semantic header* (Desai, 1997), part of whose GUI is shown in Figure 2, space restrictions preventing us from showing the full semantic header. Since the semantic header is designed to help searchers later

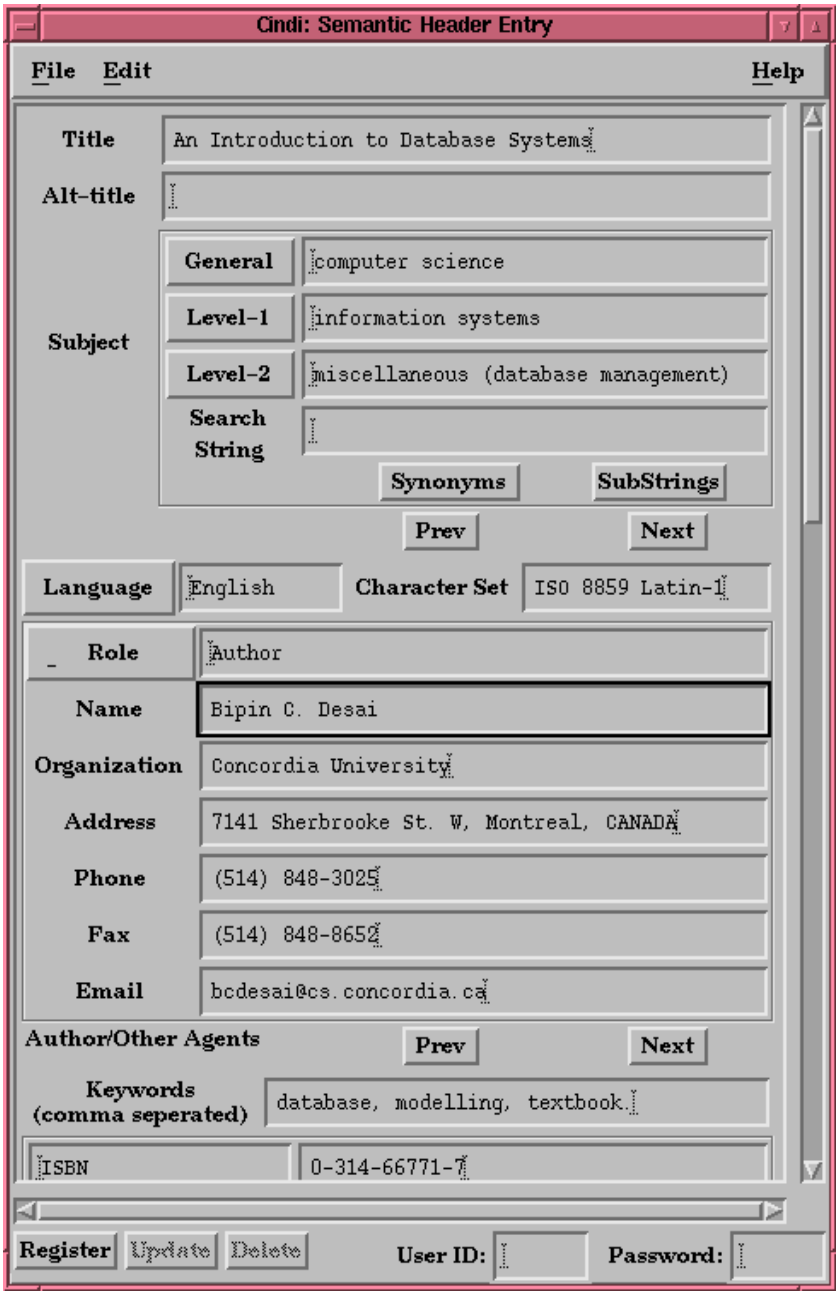

Fig. 2. Part of the semantic header GUI for cataloguing.

looking for the document, its slots contain those values that are most often used by searchers. An expert system—whose details are described by Chander, Desai, Shinghal, and Radhakrishnan (1997) and Desai and Shinghal (1996a)—mimics the expertise of a cataloguing librarian as it guides the provider in filling the slots of the semantic header. The semantic header has the following slots:

1. The **Title** of the document and its **Alt**ernative–**title**, if any.
2. The **Subject** area to which the document belongs. The subject areas are derived from the Library of Congress Subject Headings. CINDI maintains a database containing subject area hierarchies at three levels; for instance, game playing is a sub-area of artificial intelligence, which in turn is a sub-area of computer science. Thus computer science is at zero (or general) level, artificial intelligence at level-1, and game playing at level-2.
   When the provider wants to enter the subject area of the document being catalogued, CINDI offers him with a list of the possible general level subjects. He selects one subject, say computer science. The expert system then employs the stored subject area hierarchies to guide him in selecting first a level-1 subject and then a level-2 subject. If the provider changes the general level subject, then the old level-1 and level-2 subjects are erased, and he can select new level-1 and level-2 subjects. The provider is restricted to entering words from the control terminology standard in library practice. The provider can click on the Synonyms button to map his terminology, perhaps informal, to the control terminology.
   If the provider does not know the full name of the subject area, he can enter the partial name he knows and CINDI will help him find the full name. Suppose the provider enters 'com' and then clicks the SubStrings button in the GUI of the semantic header. A window displays the number of items that match 'com' at different subject levels. If the provider clicks on the General button, another window displays the list of general level subjects matching 'com'. The item selected from the list fills the slot for the general level subject, its level-1 and level-2 subjects being set to empty. Proceeding similarly, the provider fills the slots for the level-1 and level-2 subjects.
   A document may belong to more than one subject area. After entering one value in the subject slot, if the provider wants to enter another value, he clicks on the **Next** button. He then enters another value in the subject slot. If he want to modify the previous entry, he clicks on the **Prev** button and then enters the modified value. Corresponding Next and Prev buttons are similarly provided for all slots that can have multi-values.
3. The language of the document: English, French, Spanish, etc.
4. The characters used in the document: Latin, Cyrillic, Greek, etc.
5. The role of the person associated with the document, for instance, author, editor, and compiler. Then the provider enters the person's name, postal address, telephone number, fax number, and email address. Since this is a multi-valued slot (there can be more than one author), appropriate Prev and Next buttons are available.
6. One or more keywords for the document.

7. The URL (Universal Resource Locator) of the document. This is multi-valued slot (in case the document is at more than one place) and appropriate Prev and Next buttons are available.
8. The date on which the document was catalogued, and the date on which the document will expire. Moreover, there is information on the version number of the document and whether is supersedes any earlier version.
9. Does the document have a copyright? If the document has a copyright, a later searcher will not be able to down load the document without contacting the author and perhaps paying a fee.
10. The classification of the documents: usual type would be legal status of the document, its level of security or or dessemination (general, specialist, adult audience).
11. The audience at which the document is targeted.
12. The coverage indicates the targeted audience of the document or it may indicate cultural and temporal aspect of the document's content.
13. If the semantic header pertains to a collection of work, then its components and other related materail may be specified in this slot.
14. System Requirements: The hardware and software required to access the document.
15. The Source indicates the documents being referenced by the current document or which were required in its preparation. It could also be the main component for which the current document is an addendum or attachment. The identifier of the source would be URLs, ISBNs, etc.
16. An abstract of the document.
17. Annotations put in by readers of the document.
18. A Provider ID of at least six characters and a password of four to eight characters. More than one semantic header by the same provider can have the same ID and password. After the semantic header is completed, the provider registers it with CINDI by clicking on the **Register** button. CINDI then checks to see whether all the essential entries have been made. If any value is missing, CINDI asks the user to make the required entries. Once all the essential entries have been made, the semantic header is registered in the CINDI database.

    Later if the provider wants to update his semantic header, he needs to first supply his ID and password. Other users cannot make any changes in the entries made by the provider since they do not know the provider's ID and password. Other users, however, can make entries in the annotation slot since this slot gives users a forum for discussing the corresponding document. When updating his semantic header, the provider cannot change the contents of the annotation slot. By clicking on the **Delete** button in the GUI of his semantic header, a provider can delete his semantic header in case he wants to withdraw his document from the digital library.

The provider must fill at least the following information in the semantic header: the title, the general level subject, the name of the author, at least one keyword, the date the document is catalogued, and at least one identifier such as an URL,

an ISBN, or a FTP address to identify the document uniquely. Information in the remaining slots is optional. At least the name of one author is mandatory, or in the case the source is a corporate body, name of the organization is required.

3 Searching

A searcher looking for documents, first looks for the appropriate semantic headers in the CINDI registry. Once the appropriate semantic headers are located (Desai and Shinghal, 1996b), the actual documents are accessed. Since the registry for the semantic header is smaller than the collection of documents, searching becomes faster.

An expert system (Chander, Shinghal, Desai, and Radhakrishnan, 1997), mimicking the expertise of a reference librarian, helps the searcher formulate his query. For instance, if the searcher enters a general level subject, the expert system employs CINDI's stored subject area hierarchies to guide him in selecting level-1 and level-2 subjects. Overall, the searcher may formulate his query employing terms such as document title, author name, subject at different levels, keywords in the document, range of the dates when the document was put in the digital library, and language of the document. Moreover, these query terms can be nested in formulae using the logic operator AND. The graphical search interface is shown in Figure 3.

If the searcher makes a mistake in his query, CINDI issues an appropriate message asking the searcher to make corrections. Once the query is found to be correctly formulated, CINDI transforms it into a database query in reverse postfix. Then the CINDI client process at the searcher's workstation communicates with the nearest CINDI catalogue to determine the appropriate site where the required semantic headers can be found. Subsequently, the client process communicates with the site and retrieves one or more semantic headers.

The result of the search is returned to the searcher's site and displayed in her search GUI. Using the **Prev**and **Next** buttons, the searcher can scroll through the list of semantic headers displayed. To view the semantic header, the searcher clicks on its name in the list. By clicking on the **Access** button, the searcher invokes the Netscape browser to reach the document associated with any of the semantic headers. If the provider of the document has placed no restrictions (some providers, for instance, may ask for a fee to be paid) the searcher is able to access the document.

4 Annotation

The research community depends on peer review of documents submitted for publication. Such reviews are often not published. Nevertheless, comments to the journal editor made by readers of published papers are usually published and are accessible to the community. Many of the documents in a digital library will tend not to be reviewed. It would, however, be beneficial for the new reader of a document to see what past readers have said about it. CINDI allows searchers

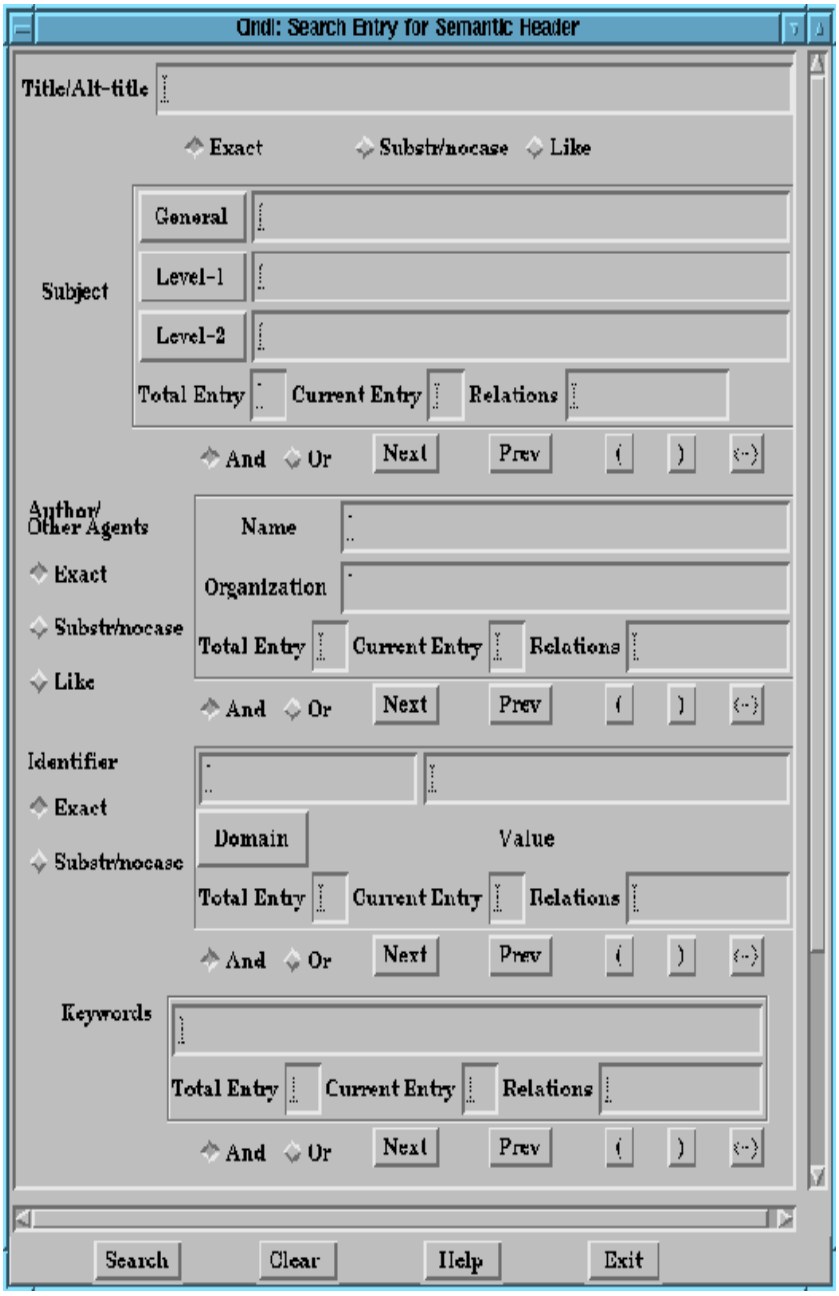

Fig. 3. Graphical interface for searching.

to add their annotations about a document in the **Annotations** slot of its semantic header. The annotations are stored together with the identity of the person who wrote the comments. The identity of the person includes his full name and email address. The identity must match his profile associated with his login name. Identifying the maker of the comments is aimed at preventing frivolous and libelous annotations.

5 Conclusion

In the CINDI system described in this paper, the provider of a document is the one who catalogues it. Such cataloguing is more reliable than one made by somebody else or the one obtained by scanning the document. By including the document's abstract in the cataloguing information, the provider is able to highlight the nature of the document. The provider uses a GUI to enter his cataloguing information. A GUI is similarly made available to the searcher of documents. Furthermore, a person who reads a document can make annotations available to later readers of the same document.

Acknowledgements

We express our gratitude to the following: Concordia University librarians Carol Coughlin and Lee Harris provided us with the expertise in cataloguing and searching documents. Their expertise was coded in the knowledge base of the expert system, earlier versions of which were developed by Dao Nguyen and Gokul Chander. Seagrams gave us a grant to fund this project.

References

1. Brody, H.: Internet@crossroad. *Technology Review*, 98(4), (1995) pp. 24–31, also at URL http://web.mit.edu/afs/athena/org/t/techreview/www/articles/may95/
2. Cromwell, W.: A New Bibliographic Standard, The Core Record. *Library Resources and Technical Services*, 38(4), (1994) pp. 415–424.
3. Chander, P. G., Desai, B. C., Shinghal, R., and Radhakrishnan, T.: An Expert System to Aid Cataloging and Searching Electronic Documents in Digital Libraries. *Expert Systems with Applications*, 12(4), (1997) pp. 405–416.
4. Desai, B. C.: Supporting Discovery in Virtual Libraries. *Journal of the American Society of Information Science*, 48(3), (1997) pp. 190–204.
5. Desai, B. C. and Shinghal, R.: Modeling Expert Search of Virtual/Digital Libraries. In *Poster Proceedings of the Ninth International Symposium on Methodologies for Intelligent Systems (ISMIS'96)* (Oak Ridge National Laboratory, UNC-Charlotte), Zakopane, Poland, (1996a) pp. 41–52.
6. Desai, B. C. and Shinghal, R.: Resource Discovery: Modelling, Cataloging, and Searching. In *Proceeding of the Seventh International Conference and Workshop on Database and Expert Systems Applications (DEXA'96)*(IEEE Press), Zurich, Switzerland, (1996b) pp. 70–75.

7. Gaynor, E.: Cataloging Electronic Texts: The University of Virginia Library Experience. *Library Resources and Technical Services*, 38(4), (1994) pp. 403–413.
8. Horny, K. L.: Minimal–level Cataloging: A Look at the Issues a Symposium. *Journal of Academic Librarianship*, 11, (1986) pp. 332–334.
9. Kahle, B.: *An Information System for Corporate Users: Wide Area Information Servers*, Thinking Machines Technical Report TMC-199, CA. (1991)
10. Koster, M.: Aliweb: Archie Like Indexing the Web.(1994)
URL *http://web.nexor.co.uk/aliweb/doc/aliweb.html.*
11. Mauldin, M. L.: Measuring the Web with Lycos. *Poster Proceeding of the Third International WWW Conference*, Darmstadt, Germany, (1995) pp. 26–29.
12. Petersen, T. and Molholt, P. (eds): *Beyond the Book: Extending MARC for Subject Access*, G. K. Hall, Boston, MA. (1990)
13. Pinkerton, B.: *WebCrawler*, (1994)
URL http://webcrawler.cs.washington.edu/WebCrawler/Home.html