# Extensible Networked Information Retrieval

Darren R. Hardy and Michael F. Schwartz
University of Colorado, Boulder
{hardy,schwartz}@cs.colorado.edu

IANET '93 Statement of Interest
May 1993

The growing diversity of Internet information demands flexible information services. We have implemented two prototypes that extend the flexibility of serving information through the Wide Area Information Servers (WAIS) interface. The first prototype, *Essence* [1], lets information providers tailor their indexing methods to support more varied data, improve precision, and reduce space in the index. The second prototype, *Dynamic WAIS* [2], lets them provide seamless gateways to other systems.

WAIS supports fine-grained information access by building full-text indexes, in which every keyword from a textual document appears in the index. To change the keyword extraction algorithm, WAIS must be modified and recompiled. Essence removes the responsibility for extracting keywords from WAIS, and provides an easy means for users to create arbitrarily sophisticated, file type-specific algorithms for extracting keywords. Essence uses *summarizers* to support many different keyword extraction algorithms. Summarizers use their knowledge of a file's semantics to carefully select keywords from a file. For example, a summarizer might extract the title, author, and abstract from a technical report. This approach has three main benefits. First, Essence can support more varied data, because summarizers may use any means to extract keywords. For example, a summarizer for a binary executable may extract keywords from related textual documents like UNIX manual pages, rather than from the binary file itself. Second, Essence improves the precision of the index, because summarizers carefully select keywords to reflect a file's contents. Finally, the index size is reduced (by an order of magnitude [1]), because Essence extracts fewer keywords on average than full-text.

To use WAIS, users first select a source of static, textual documents to search. In response, WAIS returns brief descriptions (''headlines'') of any documents matching the user-supplied keywords. Users then select among the headlines to retrieve entire documents. Dynamic WAIS extends this paradigm, allowing users to query remote information services, such as the Netfind Internet user location service, and the Archie Internet file location service. To support these gateways, Dynamic WAIS establishes mappings between the WAIS search-and-retrieve operations, and the underlying operations supported by remote information servers.

In the case of Netfind, for example, when the Dynamic WAIS user requests to search the *dynamic-netfind.src* WAIS source, the search is translated into a lookup in the Netfind seed database [3] to locate potential Internet domains to search. Dynamic WAIS returns the Netfind list of potential domains to search as WAIS headlines. When the user selects a headline, the result of a Netfind search in that domain is returned, in place of the usual WAIS static document. Similar mappings can be created to support gateways to other remote search services.

The Essence and Dynamic WAIS prototypes are available by anonymous FTP from ftp.cs.colorado.edu in /pub/cs/distribs.

## References

[1] D. R. Hardy and M. F. Schwartz, ''Essence: A Resource Discovery System Based on Semantic File Indexing,'' *Proc. Winter USENIX Technical Conf.*, San Diego, CA, January 1993, pp. 361-373.

[2] D. R. Hardy, ''Scalable Internet Resource Discovery Among Diverse Information,'' Tech. Rep. CU-CS-650-93, Dept. of Comp. Sci., Univ. of Colo., Boulder, CO, May 1993.

[3] M. F. Schwartz and P. G. Tsirigotis, ''Experience with a Semantically Cognizant Internet White Pages Directory Tool,'' *J. Internetworking: Research and Exp.*, 2(1), March 1991, pp. 23-50.

## Biography

Darren R. Hardy is a Ph.D. student in Computer Science at the University of Colorado, Boulder. His current research addresses developing scalable resource discovery techniques to cope with the Internet's explosive growth in information diversity and volume. He earned his B.S. from the same institution where he received the Steven Wozniak scholarship.

Michael F. Schwartz is an Assistant Professor of Computer Science at the University of Colorado, Boulder. His research focuses on issues raised by international networks and distributed systems, with particular focus on resource discovery and network measurement. Schwartz chairs an Internet Research Task Force Research Group on Resource Discovery and Directory Service, and is on the editorial boards for *IEEE/ACM Transactions on Networking* and for *Internet Society News*.