The Role of Resource Discovery in Support of a National Software Exchange

Michael F. Schwartz

Department of Computer Science
University of Colorado
Boulder, Colorado 80309
(303) 492-3902
schwartz@latour.colorado.edu

March 1991

The National Software Exchange (NSE) has been envisioned as a means to support collaborative research and development efforts needed to accomplish the goals of the High Performance Computing and Communications initiative. To build a system capable of supporting such collaboration, one must address the fundamental problem of how shared resources will be organized and searched by users of the system. My research group at the University of Colorado - Boulder has been investigating this *resource discovery* problem for the past three years. Our experiences indicate a number of issues relevant to the NSE.

Our research considers a broad formulation of the resource discovery problem, including resources such as network services, documents, retail products, current events, data, and people. We impose three key goals on our approaches. First, we consider very large environments, spanning national or international networks. Such environments place stringent scalability requirements on the algorithms that can be used. Second, we want to avoid imposing artificial constraints on the resource space organization. Traditional directory services (such as the CCITT X.500 standard [CCITT 1988]) rely on hierarchical organization to achieve good scalability. Unfortunately, the organization of a hierarchy becomes convoluted as an increasingly wide variety of resources is registered, and requires users to understand how the (increasingly deeply) nested components are arranged. Finally, we wish to minimize the need for global administrative agreement over protocols, information formats, and organizational structures. While standards are helpful, it is difficult to specify standards that are both globally adopted and technologically current. Moreover, standards based on a hierarchical organization require a high degree of agreement over the organization of at least the upper levels of the tree. As an increasingly diverse collection of institutions contribute to the global information infrastructure, smooth evolution will require the ability to support multiple organizational structures, and to interoperate with a heterogeneous set of protocols and information formats.

Our investigations cover a range of techniques. To date these efforts have included research into providing Internet "white pages" [Schwartz & Tsirigotis 1991a]; characterizing the organizational structure of distributed collaboration via electronic mail [Schwartz & Wood 1990]; devising probabilistic algorithms for supporting attribute-based ("yellow pages") searches [Schwartz 1989, Schwartz 1990]; supporting resource discovery among "anonymous" FTP sites [Schwartz et al. 1991a]; and supporting means of discovering and visualizing characteristics of large internets [Schwartz et al. 1991b]. The reader interested in more details about the various projects and prototypes is referred to [Schwartz 1991]. Our continuing efforts include research into supporting discovery of network services, software packages, devices, and other resources in a local internet environment; support for browsing/discovery among large scientific databases; a national collaborative experiment to support resource discovery on the TCP/IP Internet; routing and transport level network protocols to support large scale resource sharing; support for end users in discovering choices available to them in commercial internet environments; and ongoing measurements of wide area network performance and service reachability.

A number of observations may be drawn from our experiences [Schwartz & Tsirigotis 1991b]. Here we summarize the experiences most relevant to this workshop. First, organizing a resource space to be shared across administrative boundaries is not well supported by traditional database systems and directory services. Database systems typically focus on efficiently supporting queries against highly structured information stores, such as relational databases. Yet, reaching agreement on the database model, much less a particular schema, may be difficult or impossible in an environment that spans administrative

boundaries. Moreover, to date database research into scalability has focused on the problem of data size, rather than scope of distribution. Difficult problems arise when one attempts to build a database system that can support accesses by millions of users spanning thousands of sites across a wide area internet.

Distributed directory services, in contrast, typically support a very primitive means of organizing information, relying on hierarchical organization to achieve scalability. Yet, as noted above, it is difficult to search a large hierarchy effectively.

Our approach to this problem is to separate the process of organizing an information space from the process of searching for information. We have developed one set of techniques that support searches based on an understanding of the semantics of a particular resource discovery problem, allowing a user to search existing sources of relatively unstructured information. We have a second set of techniques that allow users to superimpose additional organization on a resource space in an incremental fashion. In this fashion, our systems can search heterogeneous repositories of resource information, but can also allow users to create new organizational structures, or recast existing organization structures into new *views*.

Another observation concerns how information about resources is generated. At one extreme, one can impose the restriction that all information be passed through a set of expert curators/moderators, in an effort to increase the quality of information. At the opposite extreme, one can allow any user to contribute information. While this approach could potentially reduce information quality and uniformity, it also has some potential benefits. First, the information space may be "populated" more quickly given the efforts of many different contributors. Moreover, using different contributors increases the breadth of perspective. We believe a hybrid solution (using both moderators and distributed contributions) could offer some advantages of each of these mechanisms. We are currently exploring this possibility.

Another observation is that a national system to support distributed collaboration will necessarily be used by many different people with many different interests. Therefore, it will be important for the system to support a fine-grained notion of shared interest groups, in support of collaborative work (for example, allowing users to build interest group specific views of accessible resources). Our electronic mail measurement study indicated that significantly more fine grained groups should be supported than, for example, the structure of current electronic bulletin boards and mailing lists. One idea we are exploring in this regard is to allow the system to construct dynamic notions of interest closeness, by monitoring the patterns with which people access various pools of information. Clearly, such a technique raises privacy issues, which we believe can be side-stepped by only using the technique when users are made explicitly aware that this is being done, and in situations not involving personal information (which may be many situations in the context of collaborative technical research).

Finally, we would like to offer a few suggestions about how a National Software Exchange should be implemented. The first suggestion concerns inducement to participate. We believe that the most important inducement comes from providing a needed service that is easy to use. Our most successful prototypes enjoyed real use because they were simple to use, and provided needed functionality (e.g., for discovering electronic mail addresses of Internet users). Real use, in turn, provides valuable insight into problems and potential improvements of tested techniques. In particular, we believe that it would be better to place effort into sophisticated search techniques than into sophisticated object specification languages. Most users will not use complex specifications, because of the overhead involved in learning to use the system.

While we advocate research into solving practical resource discovery problems, at the same time we believe it will be important to explore a number of competing strategies for supporting these functions. Some of our best insights into the problems of resource discovery have come from comparing our techniques for supporting a problem with those developed by other researchers.

Interested readers may obtain copies of many of the project related papers by anonymous FTP from latour.colorado.edu, in the directory pub/RD.Papers, or by contacting the author by electronic mail at schwartz@latour.colorado.edu.

Bibliography

[CCITT 1988]
CCITT. The Directory, Part 1: Overview of Concepts, Models and Services. ISO DIS 9594-1, CCITT, Gloucester, England, Dec. 1988. Draft Recommendation X.500.

[Schwartz 1989]
M. F. Schwartz. The Networked Resource Discovery Project. *Proc. IFIP XI World Congress*, pp. 827-832, San Francisco, CA, Aug. 1989.

[Schwartz & Wood 1990]
M. F. Schwartz and D. C. M. Wood. A Measurement Study of Organizational Properties in the Global Electronic Mail Community. Tech. Rep. CU-CS-482-90, Dept. Comput. Sci., Univ. Colorado, Boulder, CO, Aug. 1990. Submitted for publication.

[Schwartz 1990]
M. F. Schwartz. A Scalable, Non-Hierarchical Resource Discovery Mechanism Based on Probabilistic Protocols. Tech. Rep. CU-CS-474-90, Dept. Comput. Sci., Univ. Colorado, Boulder, CO, June 1990. Submitted for publication.

[Schwartz & Tsirigotis 1991a]
M. F. Schwartz and P. G. Tsirigotis. Experience with a Semantically Cognizant Internet White Pages Directory Tool. *J. Internetworking: Research and Experience*, 2(1), Mar. 1991.

[Schwartz et al. 1991a]
M. F. Schwartz, D. R. Hardy, W. K. Heinzman and G. Hirschowitz. Supporting Resource Discovery Among Public Internet Archives Using a Spectrum of Information Quality. To appear, *Proc. 11th IEEE Int. Conf. Distrib. Comput. Syst.*, May 1991.

[Schwartz et al. 1991b]
M. F. Schwartz, D. H. Goldstein, R. K. Neves and D. C. M. Wood. An Architecture for Discovering and Visualizing Characteristics of Large Internets. Tech. Rep. CU-CS-520-91, Dept. Comput. Sci., Univ. Colorado, Boulder, CO, Feb. 1991. Submitted for publication.

[Schwartz 1991]
M. F. Schwartz. Resource Discovery and Related Research at the University of Colorado. *To appear*, ConneXions - The Interoperability Report, *Interop, Inc., May 1991*.

[Schwartz & Tsirigotis 1991b]
M. F. Schwartz and P. G. Tsirigotis. Techniques for Supporting Wide Area Distributed Applications. Tech. Rep. CU-CS-519-91, Dept. Comput. Sci., Univ. Colorado, Boulder, CO, Feb. 1991. Submitted for publication.