Autonomy vs. Interdependence in the Networked Resource Discovery Project†

ACM SIGOPS European Workshop Position Paper

Michael F. Schwartz
Department of Computer Science
University of Colorado
Boulder, Colorado 80309-0430
July 29, 1988

As the world becomes increasingly inundated with information, value shifts from individual pieces of information to the structure placed on the information. While even single administrations can grow large enough that this becomes an issue (e.g., a large university library), the problem takes on new dimensions in a decentralized environment because of the privacy/security concerns of the participants, and because of the difficulty of reaching consensus on how the information should be organized.

As an instance of this general problem, the Networked Resource Discovery Project at the University of Colorado, Boulder is investigating mechanisms to allow users to discover the *existence* of resources in a large scale, administratively decentralized environment. This includes resources typically associated with computer systems (such as mailboxes and network services) as well as physical resources registered in various databases (such as retail products registered in corporate inventory databases, and people sharing particular interests, registered in special interest group membership lists). Hence, we would like to be able to support searches for a wide variety of resources, such as "a nearby laser printer", "the electronic mail names of graphics experts in New York City", "TCP implementations for IBM mainframes", "inexpensive lawn mowers", and "movies playing in town tonight".‡

Existing directory services do not by themselves solve this problem, largely because they are fragmented: one must access a variety of different mechanisms and databases to cover this wide variety of resources, including telephone directories, faculty rosters, bulletin boards, other people, etc. While it is in principle possible to build a database that registers all resources, doing so would cause difficult problems concerning consistency and transfer of authority in the large scale, rapidly evolving, and highly decentralized environment we are considering. Instead, we want to utilize resource information where it naturally resides, for example by accessing various specific corporate product line databases in an effort to find suppliers of a particular replacement part.

As of this writing, a prototype implementation effort has begun. Below, we consider the aspects of the design relevant to this workshop: privacy/security, and organizing the resource space.

Privacy/Security Issues in Sharing Resource Information

Since resource information is distributed among autonomous systems, sharing the information poses privacy and security problems. Our approach to this problem involves encapsulating each autonomous database by a *broker*, a program responsible for accessing the database and deciding exactly what information may be released to the outside world. Brokers are essentially abstract data types corresponding to human operators that currently allow the general public limited access to many existing databases. For example, telephone directory service operators are responsible for deciding to

† This material is based upon work supported in part by the National Science Foundation under Cooperative Agreement DCR-84200944.

‡ These examples are intended to demonstrate the general applicability of the problem. Whether we will be able to support searches as sophisticated as these is not yet clear.

refuse queries asking what person has a particular telephone number, but they will answer many other queries. Similarly, airline information operators will neither confirm nor deny whether a particular passenger boarded a particular airplane. In both examples, humans decide whether to allow a given query, based on privacy/security considerations. It would not be difficult to incorporate such considerations into brokers. Brokers can be built by information providers (or by a third party and inspected by the information providers) to ensure that they are trustworthy. Abstractly, this technique is similar to the "arms length" access control provided by "anonymous FTP" and electronic mail, where users outside an administrative domain are allowed access to a limited subset of the information in that domain, through a restricted interface.

A relatively large amount of resource information can be shared using this simple concept of a broker. For example, corporations would likely be willing to release product line information (other than sales figures, etc.), since that is essentially a form of advertising. Many individuals would also participate, just as they are currently registered in telephone directories and special interest group membership lists.

However, this model does not capture other interesting sharing relationships. A more sophisticated model would, for example, allow more information to flow within than across administrative boundaries. To support this, we are considering defining an *Information Sharing Domain (ISD)* as a grouping within which resource information is freely shared. An ISD may be smaller than an administrative domain. For example, within a company there may be some groups that do not freely share information with other groups. In this case, there should be ISDs surrounding each of the smaller groups within which information is freely shared, and a larger ISD surrounding them within which less information is freely shared. Beyond that is the outside world, in which even less (or no) information is shared. More generally, information may fall within multiple ISDs. For example, an employee in a software development group in a company could also be involved with a standards committee that crosses company boundaries. These separate roles each have their own ISD hierarchies. There could be some overlap, but in general different sharing restrictions apply to each ISD.

Extending the model this way clearly supports more general sharing behavior, but this generality comes at a fairly high price. First, the extended model requires an authentication mechanism, which is not required in the simpler model. Second, the extended model would involve more system complexity, e.g., a multiple inheritance hierarchy of objects, each of which performed "privacy filtering" for some ISD, plus a globally meaningful way to name and manipulate (add/delete, etc.) ISDs. It is not yet clear how seriously we will attempt to explore this more sophisticated model.

Organizing the Resource Space Without Centralized Administration

Organizing the resource space in a manner suitable to all participants is a difficult prospect. The obvious approach of constructing a single global hierarchy is problematic, since doing so would require centralized administration at some levels of the tree. More importantly, a strict hierarchy is inflexible. For example, in searching for persons having technical expertise in three dimensional graphics shading algorithms, one person might prefer to organize the world as "Computers.Graphics.3D.Experts", while another person might prefer "People.Interests.Technical.Graphics.3D". Moreover, as the world evolves, the resource space organization must change, and requiring centralized agreement for such changes would slow the process tremendously.

A simple hierarchy works reasonably well for name services because of their implicit assumption that people simply *know* names, usually by means outside the system (e.g., a telephone call to find out someone's electronic mail name). In contrast, the resource discovery problem is to find some specific information (e.g., an electronic mail name or a telephone number) about a collection of resources described in a less structured, attribute-based manner.

We want to allow the resource space to be organized according to a more general graph structure. Unfortunately, hierarchy is the only way that scalable computer systems have been built to date.

As a compromise, we will incorporate the notion of a hierarchy of *specialization subgraphs.* This is similar to the hierarchy of a simple tree, except that any resource could be a member of multiple graph structures (representing different organizational schema), and there could be back pointers and cycles in the graph (allowing shared information without the consistency problems of redundant copies). For example, one substructure could link databases containing information about automobile parts, while another substructure could link databases according to geographic boundaries. The automobile parts structure could, in turn, have one subgraph organized according to function (engine parts, tires, etc.), another subgraph organized according to manufacturer, etc.

This graph-based organization is essentially a special case of the network database model, where the network has more structure (i.e., specialization subgraphs). But there is an important additional issue. A primary concern in our project is avoiding the need for human administrators to take action when building or evolving the resource interconnection graph, which would slow the evolution process. Instead, we want to allow the resource space organization to evolve over time automatically, in accordance with what resources exist, and what types of queries users make. For example, when compact disk and other digital sound technologies entered the marketplace, the worlds of music equipment and information systems became closer, and a resource space reorganization became appropriate (a point demonstrated by how various music and computer trade magazines began advertising more products in each others' respective domains at that time). What we would like is that at any instant, the most popular organizational schema become the most efficient in which to search for resources, without restricting more specialized schema from coexisting. This approach is interesting because it is reminiscent of a *participatory democracy*, where individuals decide over time how the world is to be organized, rather than having an inflexible organization dictated by a centrally administered ruling body.

To accomplish this dynamically organized graph structure, we augment the database/broker conglomeration with a set of *agents* that dynamically construct links between the information repositories based on descriptive keywords, implementing the graph interconnection structure. Running an agent allows an organization to participate in the resource discovery network, at the cost of some amount of autonomy (since the agents must at least share some common protocols, and use some computational resources). This is analogous to the participation/loss of autonomy tradeoff of running UNIX† sendmail.

Establishing agent interconnections is a bootstrapping problem. Once they are established, we can use cache aging protocols to evolve the graph organization in accordance with common usage patterns (based on several dimensions of locality, including geographical, temporal, and functional). While our approach to the bootstrapping problem is outside the scope of this workshop, it is a major focus of the project, and our success in allowing the resource space to reorganize in a decentralized fashion without human intervention will depend heavily upon this mechanism. Hence, we now briefly discuss the approach.

In designing protocols to handle the bootstrapping problem, we have observed that, since failures may occur and various databases may be unavailable at any point in time, it is not possible to guarantee perfect service. This fact, in combination with the scalability problems presented by trying to ensure that the resource space can be exhaustively searched, led us to the decision that the system will only guarantee a reasonably high probability of success/exhaustiveness in resource searches. We plan to carry this decision one step further, utilizing a suite of probabilistic protocols to establish and search the graph. By doing so, we hope to gain a measure of scalability beyond what could be achieved by deterministic protocols, without a serious sacrifice in effectiveness. At this point in time, we are considering a technique involving *Sparse Diffusion Multicasts*, messages that reach only a small (e.g., logarithmic) proportion of the agent subgraph to which they are addressed. Through repeated use of this primative at multiple levels of the agent interconnection graph, combined with

---

† UNIX is a trademark of AT&T Bell Laboratories.

the aforementioned cache aging protocols to maintain reasonable storage and consistency profiles, we hope to construct a resource space organization that with high probability evolves to a useful configuration without undue overhead (as would be required by techniques requiring fully replicated information or full scale broadcast).

Summary

The goal of the Networked Resource Discovery Project is to explore a set of mechanisms that could provide an administratively decentralized means for users to navigate through an enormous resource space. A key problem is providing a system that does not impose a strict hierarchical structure on the resource space, and does not require manual human administration to maintain/evolve the resource space organization. We want to allow the resource space organization to develop and evolve dynamically in accordance with the set of existing resources, and with the types of queries users make.

Of relevance to this workshop are the privacy/security problems, and the resource space organization problems. Our approaches to these problems each strike useful and interesting compromises between autonomy and interdependence. In the case of privacy/security, we are considering two different models that would allow organizations to participate in the global resource discovery mechanism without giving up the autonomy to control exactly what information is released to the outside world. For resource space organization, participating in the agent network provides access to other organizations' resource information at the cost of some autonomy over the agent protocols, and over the computational resources consumed by the agents. Additionally, we allow the various autonomous participants to contribute to the resource space organization in a fully decentralized manner so that over time, the most popular organizational schemes will be the most efficient to use in resource searches.

More information about this project is contained in the following Technical Report: M. F. Schwartz. The Networked Resource Discovery Project: Goals, Design, and Research Efforts. Tech. Rep. CU-CS-387-88, Dept. Comput. Sci., Univ. Colorado, Boulder, CO, May 1988.