CNI White Paper on Networked Information Discovery and Retrieval

Clifford Lynch, University of California, Office of the President

Avra Michelson, MITRE Corporation

Cecilia Preston, Research Associate

Craig A. Summerhill, Coalition for Networked Information

This is the current working draft of the outline of the white paper on Networked Information Discovery and Retrieval being prepared for the Coalition for Networked Information. The chapters described in this outline will be placed on the Coalition's FTP server as they are available in draft during the course of Spring/Summer 1995; the final version of the white paper is targeted for Fall 1995. Comments are welcome, and should be sent to nidrcall@cni.org.

Table of Contents

Outline: Chapter 1

The Nature of the NIDR Challenge

- initial brief definition of terms: networked information discovery and retrieval; network resources (objects); metadata (including comments on etymology).
- scope of problem that is focus of this paper: how to improve ability of user to discover and access resources the current internet-based networked information resource environment. Extent of software mediation in the

NIDR process. How a mix of free and for-fee information will change the picture.

- Characterization of key features of the networked information environment relevant to NIDR problems:
- very large scale, rapid growth; dynamic addition and relocation of information resources
- extremely heterogeneous nature of resources
- wide variation in granularity of resources; hierarchical resource organization
- multiple generations of information resources and supporting access systems
- distributed and autonomously managed resources
- wide variation in quality of resource content and implementation
- growth of "self-publishing" models of information distribution
- combination of free & for-fee information resources
- combination of public and private information spaces
- no commitment by information providers to offer service registry within a central framework
- very heterogeneous user base; varying expertise and needs, varying access capabilities
- unrealistic (and poorly articulated) user expectations
- poorly defined user selection requirements
- information overload: too much overall information, and too much relevant information
- a closer look at the discovery process:
- discovery as an iterative research activity; different kinds of discovery.
- discovery as "catalog use"; performed by humans
- components of discovery as a process: selection, collocation, duplicate elimination, ranking/differentiation, browsing, determining "fitness for use".
- hierarchical searching and granularity; discovering systems/information spaces; knowing where to search
- the continued need for surrogates for objects in discovery on the net; arguments based on limited ability to fetch headers that are object components selectively, performance issues, economic and intellectual property issues (i.e. separate creation, control and distribution of surrogates and primary objects)
- automated support for discovery as a continuing process: SDI, personal agents, filters
- a closer look at the retrieval/access process:
- defined primarily by existing (simple) network retrieval protocols; these put an undue burden on discovery
- parameters of access processes, e.g. costs and formats (static vs. dynamic issues); poor accommodation by current protocols
- multistage, sequential nature of access/retrieval & subsequent use of network information objects.

- low levels of interoperability targeted (moving bits, or application-specific file formats)
- key problems with achieving current NIDR objectives:
- objects as viewed in the NIDR context are extremely simple
- classic information retrieval issues; heavy use of natural language
- lack of data sources (cataloging) upon which to base discovery
- networked information retrieval issues (extended retrieval)
- performance and architecture problems (technical issues) in large scale distributed systems
- incorporation of nontextual objects and their description
- nontechnical issues with major architectural implications: privacy, security, intellectual property, charging for information

Chapter One

The Nature of the NIDR Challenge

This is a revised draft of the first chapter of a white paper on Networked Information Discovery and Retrieval being prepared for the Coalition for Networked Information by Clifford Lynch (clifford.lynch@ucop.edu), Avra Michelson (avram@mitre.org), Craig Summerhill (craig@cni.org) and Cecilia Preston (cecilia@well.com). Subsequent chapters will be released in draft shortly, with a final version targeted for late 1995. Drafts will be available through the Coalition's FTP server (ftp.cni.org). Your comments on this draft are welcome, and should be sent to nidrcall@cni.org.

Draft of October 27, 1995

Scope and Focus of the White Paper

This paper explores the current state of the art in discovery and access for networked information resources, and ways in which the state of the art can be advanced. While the networked information environment can be interpreted broadly, we focus specifically upon the existing Internet as the host environment for networked information resources. The paper takes a perspective that is centered on the information user or consumers, rather than information creators, information providers or information managers; thus, our concern is not primarily with the management of information resources across time, for example, or with methods of publishing information in the network environment.

Specifically, we envision a network user who is seeking information via the network. As discussed later in this chapter, the specific information needs of network users vary widely, but in all cases he or she will locate a set of potentially relevant resources through various NIDR tools and services available on the network, make choices among these resources, and access or retrieve

one or more of these resources. It is these processes of location, selection and retrieval that form the primary focus of our analysis here.

The network user is a human being; his or her use of the network and the various information resources available through it are mediated and assisted by various software systems (usually called *clients* in current implementations). Today, the human being usually exercises very direct and close interactive control over these software systems; the client software primarily provides a graphical user interface to permit the user to interact directly with information services and resources (servers). Over time these software clients are expected to become increasingly autonomous and capable, and, at least in some situations, to require less detailed and continuous interactive control and direction by human network users. As this evolution occurs it will likely be less appropriate to refer to the user software simply as clients that work in conjunction with information servers and services accessible through the network; rather, the user software will be viewed as a complex suite of applications that includes and embeds such client functionality. In the future we may even be able to speak about the specific information requirements of these software applications -- perhaps now including groups of relatively autonomous network based collections of software "agents" -- in carrying out various activities which form subordinate steps in meeting the more broadly specified and longer term objectives of human network users.

We extrapolate in two major regards from the current Internet environment. Recognizing the increasing presence of commercial information providers on the network alongside organizations and individuals providing free access to information resources, we explicitly consider an environment where both free and for-fee information exists. We believe that this will soon characterize the information offerings available through the Internet to a much greater extent than it does today. The introduction of for-fee information will, we believe, significantly alter many of the existing mechanisms for discovery and retrieval of networked information.

Our second extrapolation of the present NIDR framework involves the consideration of a more extensive role for software proxies in the discovery and use of networked information resources. At present virtually all NIDR systems with which we are familiar involve a human being as an active, integral part of the NIDR process. Yet if one examines visions of future networked information environments, software "agents" (variously defined) play a large role in identifying, filtering, integrating, organizing and manipulating disparate networked information resources on behalf of human network users. There is a large gap between current practice and future vision, and part of our objective in this white paper is to examine the nature and origins of this gap and the barriers to bridging it. This is the gap between today's (human) network users directly querying a service like Archie or Yahoo on the one hand and tomorrow's network user interacting with his or her intelligent workstation to specify relatively broadly

defined information needs, to identify new information that has recently become available on the network, or to transact business.

Terminology and Definitions

Vocabulary to discuss the evolving world of networked information is far from standardized. In this section we provide preliminary definitions of a number of terms that will be used throughout this white paper. Later parts of the paper will explore these definitions in much greater depth.

Networked Information Resources are objects or interactive services that are available through the network in the broadest sense. These resources include files (which can have semantics as text, images, structured datasets, digital audio or video, or programs); interactive services such as data, audio or video feeds, interactive sessions using protocols such as Telnet; electronic mail based services such as reflectors and List Servers. They also include aggregations of information such as databases, file archives accessible through anonymous FTP, and archives from newsgroups or mailing lists; typically these aggregations support some type of browsing or searching once the aggregate object has been selected. Thus one can view networked information resources as having a wide range of granularities, and some networked information resources as being hierarchically organized.

Sometimes we will use the term *(networked information) object* to emphasize the subclass of networked information objects that excludes interactive services and to stress the idea of a collection of digital information that can be transferred from some host on the network to the user's machine for subsequent use.

In general a networked information resource is accessible through the use of some network protocol such as FTP, Telnet, Z39.50, HTTP or the like; the protocol may be used with a set of common conventions to provide a protocol-based service like anonymous FTP. In cases where subcomponents (subobjects) of a networked information resource are only selectable through some specialized interaction with a custom interface (for example, by interacting with a user interface via Telnet that is part of a database service) the individual subcomponents are typically *not* termed individual networked information resources.

A slightly disingenuous and circular definition of a networked information resource might be anything on the network that can be addressed by a Uniform Resource Locator (URL). The reason that this definition is somewhat circular is that while at present there is no URL which represents not only the establishment of a Telnet session (to take a specific case in point) but also a scripted interaction with a remote host once such a session has been established, there is no conceptual barrier to defining such a URL, only a sense that this is inconsistent

with the typical use of URLs. URLs are now being defined that allow the retrieval of specific records from a Z39.50 database, for example.

Discovery is a very broad term that is used to cover the entire process of identifying candidate network information resources that may be available to the user, and the management and navigation activities associated with such a set of candidate resources, which might include ranking or sorting, browsing, selection and similar activities. Typically, discovery involves the examination and manipulation of *surrogate* representations for actual networked information resources; these surrogates may be very simple (a URL that provides a path to the resource and perhaps some sort of name associated with the URL that provides some kind of description of the resource), or they may be quite complex (for example, a lengthy structured description of the resource).

Sometimes we will use the term *identification* to emphasize that part of the discovery process that focuses on the creation of a set of candidate resources (or their surrogates), and the term *selection* to emphasize the part of the process which focuses on making choices among these candidates.

Retrieval is a complementary process to discovery, and involves the actual fetching, access or invocation of networked information resources that have been found through the discovery process. To some extent, we view the actual *use* of a resource as taking place outside of and subsequent to the retrieval process; for example one might retrieve a complex numeric dataset that could subsequently be used as input to a simulation or visualization software package. Retrieval and use are of course interrelated: it is not very helpful to retrieval a network resource that one does not have the tools and computational/display capabilities to use, and the availability of the necessary tools and capabilities need to be considered as part of the discovery and retrieval processes. This is primarily a consideration for complex information objects; in other cases, such as simple text files or HTML pages, the boundary between retrieval and subsequent use (normally a simple display of the retrieved information) is so thin that the acts of retrieval and use are hard to distinguish.

The retrieval and discovery processes may be interleaved or iterated as a user retrieves one resource and returns to the discovery process informed by a better understanding of the nature of the resource. Ultimately, retrieval is likely to involve the evaluation of a URL, but it may involve much more (such as URN to URL resolution, or complex interactions with a host on the network that will result in URL evaluation). We will sometimes use the term *location* to emphasize the passage from an abstract description of (the contents of) a resource (such as a surrogate) which might be selected as part of a retrieval process to the identification of a particular instance of that resource somewhere on the network which could then be the subject of a retrieval operation.

The demarcation between discovery and retrieval is not clear-cut, in part because of the nature of the design of various network protocols and services that predate a model of network information access and use that distinguishes discovery and retrieval processes. Consider, for example, that the format of a textual document may well influence which document is selected (the discovery process); at the same time the format in which a document may ultimately be delivered to a user may be established as part of the retrieval process.

A great deal of the literature related to NIDR involves discussions of *metadata*. Metadata, literally, means "data about data"; our research has traced the earliest use of this term back to about 1976. (See the appendix to Chapter 3 for more details.) The origins of the term are murky; it seems to have been used to describe a range of concepts that evolved in the scientific data management, information management, archival, distributed/federated database, and perhaps even artificial intelligence research communities during the 1970s and 1980s. Our feeling is that at this point "metadata" as a descriptive term has become so debased by overuse (and means so many different things in different communities and contexts) that it is now virtually meaningless without extensive qualification; unfortunately, it has also become a very fashionable term. The very vagueness of the term metadata makes it all too easy to offer sophisticated-sounding proposals about using metadata in various ways that seem to be almost impossible to reduce to practice, or which are extremely pedestrian when actually implemented.

It is clear, for example, that the role of information as metadata is defined largely by the context of use; information can be data in one context and metadata in another. Indeed, as Michael Buckland has pointed out, the objectives and motivation of the network user may ultimately determine whether information is being viewed as data or metadata, not just the context of use. The dividing line between metadata and simply information that makes reference to other information (but that has an independent existence and perhaps independent status as intellectual property) -- for example abstracts, reviews, or descriptive cataloging -- is very poorly defined. At the same time, it is common to refer to parts of a specific networked information object (such as fields within a header) as metadata, even though these are sometimes inherently part of the object they describe or qualify.

In this paper we will try to minimize the use of the term "metadata" (except in describing the work of others that makes use of the term) and will prefer to speak of surrogates for networked information resources, and of specific types of data elements contained within these surrogates. We will also discuss data elements explicitly or implicitly contained within (and thus extractable or computable from) actual networked information resources; once extracted or computed these can be directly manipulated or can contribute to the construction of surrogates. Early NIDR systems operated primarily on very simple surrogates in the tradition of descriptive cataloging. The growing use of structured information representations

such as SGML and HTML markup or various types of headers associated with multimedia or structured datasets makes this latter class of extracted data elements an increasingly significant factor in the design of NIDR systems, and promises a much richer set of databases to support discovery and retrieval processes in the future. Similarly, the use of statistical textual analysis algorithms developed by the information retrieval research community to characterize textual objects has given new importance to the class of computed rather than simply extracted data elements. Chapter 3 of the paper explores these developments in detail.

We will try to avoid arguments about when to confer upon these data elements and surrogates some special, near-mystical status as metadata. Our intention is not to develop a specific working definition of the term "metadata" within the context of this paper.

While the term "metadata" is problematic, much of the thinking and writing about various types of metadata and the uses that can be made of such information is, in our view, both valuable and relevant to the NIDR framework, and we will examine parts of this literature at various points throughout the paper.

Characteristics of the Networked Information Environment relevant to the NIDR framework

In this section we review some of our assumptions about the current and near future Internet-based networked information environment which provide important contextual elements for examining the NIDR challenge. Many of these characteristics will be familiar to readers who have spent time using current networked information resources. In some cases we develop parallels or highlight distinctions between the networked information environment and the traditional print-based information environments, since experience with printed information has been a substantial influence in framing, understanding, and developing tools for the networked environment.

The Internet is now a very large scale distributed computing environment which is characterized by rapid growth and change. Virtually every host on the net can serve not only as an access point to network-based services and resources but also as a supplier of services or resources. New resources are being added at a rate that can no longer effectively be tracked by simply creating directories of new resources for human review. Further, resources are volatile; not only do new resources appear, but existing resources move from host to host and sometimes disappear, or become obsolete through lack of maintenance and support. A database or a collection of information may be created for, and funded as part of, a specific project, and may be an important, timely and comprehensive information source on some topic for a period of time; when the funding (or the interests of the developers) lapses, the information may stay on the network but become increasingly inaccurate or incomplete with no obvious indication to the

user. Some resources, such as newsgroups or information feeds, may by their nature change in focus and content as part of an evolution over time; beyond very broad topical characterizations, one can only describe their content relative to a fairly narrow window of time. This is in contrast to, for example, some file archives or databases which follow explicitly defined policies for content scope; while these policies may change occasionally, they tend to do so in a rather formal and well-announced fashion when compared to the casual introduction and exhaustion of topical threads within a newsgroup.

Networked information resources, as already indicated, are extremely heterogeneous in nature, volatility and coverage. They include a wide range of services and types of objects. This is part of what makes the NIDR challenge so difficult; the different types of networked information resources call for different types of descriptions and classification strategies and are appropriate for different types of information needs. Yet network users, at least in many cases, want to be able to view the collection of available information as a single universe rather than as a large number of collections organized by type of resource or methods of access. In this connection it is interesting to note that we are seeing the increased presentation of views of subsets of the networked information environment as relatively homogeneous information spaces (such as the Worldwide Web or Gopherspace) to the user community through NIDR systems, and it seems that the NIDR problem within these more constrained and homogeneous information spaces is considerably more tractable than the general problem. While this segmentation of the networked information environment into relatively homogeneous information spaces defined by access and navigational tools may facilitate the development of NIDR systems, we would argue that it is fundamentally at odds with the desire of network users to be able to discovery and retrieve resources based on content rather than the location of content within a specific information space such as the Web.

Networked information resources vary widely not only in character but in granularity and size. Describing a file archive containing tens of thousands of files or a database containing millions of documents is a very different problem than describing an individual file that contains a document or an image. In some cases it is difficult for a user to even determine the size of a given networked information resource. Yet again it seems that network users want to be able to search collections of resources at different levels of granularity as a unified whole. One can readily see the problems involved in this by imagining a search for networked information resources on a given topic resulting in the identification of two documents (files), a digital video clip (a file), a newsgroup in which the topic received major discussion from May to June of 1994, and also an indication that the Library of Congress, Yale university and the Nexis database may contain relevant information.

The Internet is host to an array of widely distributed and autonomously managed resources. There is a great cultural bias against centralized control and even

centralized registry of resources; this bias goes beyond merely technical issues, although technical issues (the "scaling" problem, in particular, and also reliability questions) are often raised in justifying this bias. Further, the world of Internet-based networked information resources has evolved in a piecemeal fashion over a lengthy period of time and over multiple generations of access and organizational technology; the NIDR framework must thus accommodate the characteristics of these multiple generations of technology.

A final important consequence of the autonomous and distributed nature of the network is the high degree of duplication of free information. Historically, it has been very common when assembling an collection of information on a given topic to simply make copies of relevant files that one discovered on the net. While this had the advantage that information tended to stay available even though one site might discontinue service or remove content in order to free up limited disk space resources for other, newer information, it had the very undesirable consequences that users would not only find many copies of the same objects when searching for resources but also would often be faced with multiple versions of the same object, since the "original" version of the object might be repeatedly updated but sites that had copied the object at some point in time would be unaware that it had been updated. This situation has improved somewhat with the broad deployment of technologies such as Gopher and the worldwide Web which permit a site to include an object at another site by "reference" (that is, by including a *pointer* to the object at the remote site rather than a copy of the object itself) but the proliferation of copies and persistence of obsolete versions of objects continues to be a problem for users and information providers alike.

There is a wide variation in the quality of resource content and also in the quality of implementation of these resources. Some resources are rigorously and professionally maintained by large, reasonably well-funded organizations as part of organizational missions; these resources are also supported by ample computational resources to provide reliable service with good response time. Other resources may be made available in a much more haphazard fashion, essentially as personal contributions to the pool of shared information on the network; these resources may be hosted on someone's personal workstation which is not always running and is often overloaded with other computational tasks. These resources may be maintained and updated only as long as their provider is interested in doing so. The growth of "self-publishing" models of information distribution in the Internet environment will continue to place stress on the variation in resource content and implementation quality.

Currently most information on the Internet can be used without fee. The more professionally maintained information resources are typically available because they are either part of some organization's mission -- a distribution site for government information, a university department distributing technical reports, a professional society offering preprints, a library offering an online catalog or access to digital images from a special collection -- or as part of a broader

commercial purpose -- for example a corporation making a catalog of products available for purchase through the network. We believe that in the near future the existing base of "free" information will increasingly be supplemented by a rich collection of information that one pays to use -- either by subscription or transactionally on some type of pay per use basis. The introduction of for-fee information will raise a large number of issues that are not currently well addressed in today's NIDR framework, including:

- Identifying what information is and is not free to use
- Determining the costs (and other restrictions) on using for-fee information resources.
- Describing the quality of for-fee information (both in terms of content and implementation), which is likely to be a much greater issue than it has been for free information, where many users will grudgingly agree that "you get what you pay for".
- Developing selection criteria that span free and for-fee information resources.
- Developing surrogates which facilitate the discovery of for-fee information resources without undermining the market for these for-fee resources on the one hand, and without leaving users feeling that they have been the victim of "false advertising" on the other.

Just as the Internet of the near future will increasingly combine free and for-fee information resources it will also, in our view, combine public (both free and commercially offered) and private information spaces. Not only will we see the increased development of personal information spaces housed on personal workstations, but also private spaces that support scholarly collaboration among closed communities of researchers and proprietary information spaces belonging to corporations and other organizations firewalled off or otherwise separated from the Internet and its public information spaces. At present many of the workgroup and organizational information spaces either use technologies that have not yet been scaled up to the public spaces of the Internet -- Lotus Notes being an excellent example -- or relatively low-technology internet tools such as private listservers or newsgroups. A few organizations are developing private versions of the internet to support organizational information resources -- the Mitre Information Infrastructure (MII) being perhaps the largest and most sophisticated such effort known to the authors.

The evolution of these private information spaces will emphasize the need for modularity and extensibility in the design of NIDR systems which can provide a user with an integrated view of the totality of the information resources the he or she has available -- personal, organizational and public. For the foreseeable future, it seems likely that due to considerations of scale, the public spaces of the internet will provide the most challenging arenas for application of NIDR technologies, however.

In comments on an earlier version of this chapter, Chris Weider and his colleagues suggested that for-fee information might be accommodated within a broader framework of use restrictions for networked information resources (such as resources which were for use only by specific closed communities). We believe that such limited access resources will be important -- for example, as components of private information spaces of various types -- and the need to support such resources highlights the need for an authentication and access control infrastructure that is built around much more than personal identities and can encompass questions of organizational or institutional affiliation of individuals; the development of such an infrastructure is a major unaddressed need in current work in security and authentication efforts. However, we do not believe that this is the most useful perspective with which to view for-fee resources, which are essentially being made available *publicly* within a context of commercial transactions and even advertised as available within this context, rather than being restricted from "view" as private resources are. The provider of a private resource will want to carefully control even the distribution of information about the existence of the resource (such as descriptive surrogates), while the provider of a for-fee resource will want to widely distribute managed surrogates indicating at least the general nature of the resource and the fact that it is available for use for a fee.

The Internet as an Information Retrieval "System"

We are seeing an increased diversity of users trying to locate and utilize the ever more diverse and complex set of resources accessible through the Internet. The expertise of the network user community -- both in terms of subject knowledge and also knowledge of (and patience with) NIDR tools and systems -- is now extremely varied; in particular, more and more users are viewing NIDR tools as a means to an end rather than an enjoyable pastime. The expectations of this user community -- and particularly the less network-sophisticated members of the community -- are extremely high and have been set largely by speculative portrayals of possible future NIDR technology through science fiction books and corporate advertising; they believe that a world of intelligent workstations, autonomous software agents and "knowbots' is close at hand. In terms of information quality, consistency and coherence, the expectations of these users have been conditioned by the well controlled, deliberately-designed stand-alone environments such as commercial database services and online library catalogs rather than the autonomous, distributed world of the Internet, which was never designed as an information retrieval environment.

In traditional information retrieval system design one attempts to define the set of "queries" that the system should be able to answer. There are several problems in extending this traditional methodology to the NIDR framework. There is a very wide range of queries, and many are ill-posed. Also, information seeking is often a protracted process, not the formulation of a single query. Indeed, user needs are frequently refined iteratively as the user obtains a better understanding of the

quantity and types of networked information resources that may be relevant to his or her needs in a particular context. It is also worth noting that while traditional information discovery and retrieval tools have over time established reasonably constrained sets of queries that they attempt to satisfy (or at least which they are optimized to be responsive to) user expectations in the NIDR environment are so high that the problem has not yet been well-constrained.

The increasingly broad network user community is also characterized by a more and more varied set of access capabilities. Some users have high speed network connections and very sophisticated, capable workstations; other users now access the network from dial-up lines at low speeds using low-end personal computers. This variation has implications both for the design of NIDR tools and also for the selection criteria that various classes of users will apply as part of the NIDR process.

A final observation goes beyond the NIDR framework and in our view will increasingly shape the evolution of the networked information marketplace, but it will also have a pervasive influence on the development of new NIDR systems. There is simply too much information, and more and more of it is on the network. Not only is there too much information overall, but increasingly users will find that there is too much *relevant* information on the network. Thus there will be a growing emphasis on precision in searching and on quality ranking and selection rather than simply identifying what is likely to be relevant information. The user's time will increasingly become the limiting factor in deciding what resources to examine.

There is a parallel here with the development of earlier information resources such as online public access library catalogs. When these databases were small, the system design goal was to try to ensure that the user was not turned away empty-handed (except in those few cases when there was really nothing relevant to his or her search criteria in the database); as the databases grew, huge search results became commonplace and the retrieval systems had to be extensively reengineered to support greater precision in specifying search criteria and to help the user to manage large retrieval results (by refining searches, for example, or by browsing or summarizing these result sets in various ways). Consideration was also given to supplementing traditional catalog databases with other types of resources such as bibliographies and pathfinders which offered users more concise responses to their searches. This redesign proved to be quite difficult and is in fact still ongoing, both as a research problem and an engineering effort. It is likely that the development of NIDR systems will follow this same evolutionary pattern.

The NIDR problem and the emergence of Digital Libraries

During the past 18 months digital libraries have emerged as a major research topic. While issues specific to the development of digital libraries are outside of

the scope of this white paper, the *relationship* among digital libraries, the broader networked information environment, and the development of NIDR technology is of central importance to the issues here.

For the purposes of this paper, we define a digital library (with some qualms about the appropriateness of this popular terminology) as simply an electronic information access system that offers the user a coherent view of an organized, selected, and managed body of information. In a real sense, digital libraries have existed since the 1970s: LEXIS, for example, certainly meets our definition of a digital library. Note that digital libraries need not be limited to "scholarly" content; we would expect to see digital libraries emerging to serve not only the scholarly community, but businesses in various areas, and all types of hobbyist and other amateur communities.

The networked information resources accessible through the Internet do *not* constitute a digital library; they do not represent an organized, selected or managed body of information. The information resources on the Internet are better compared to the output of the publishing industry, or perhaps more accurately, to the total output of the *printing* industry for a few years -- not only books and journals of possibly lasting importance, but business cards, menus, personal letters, announcements of events and the like. Internet information resources represent a part of the raw material from which digital library collections might be selected and organized (though there are admittedly some difficult questions about how the processes of selection, acquisition, and organization are to be accomplished with networked information resources; these are beyond the scope of this paper). A peculiarity of the Internet environment is that it provides access both to these raw materials and simultaneously to digital library information services that may include these raw materials within the context of deliberate collections.

We believe that NIDR technologies and approaches are relevant to digital libraries in at least two aspects. Certainly, the processes of discovery and retrieval need to be conducted *within* the context of digital library services; thus designers of digital libraries will build upon work that is being done in the NIDR area. In addition, as digital libraries proliferate over the next few years, we believe that one of the major uses of NIDR technologies will be to allow users to identify relevant digital library services that can satisfy specific information needs. Indeed, given the comments above about the overwhelming quantity of information on the network and the growing user demand to identify limited amounts of high-quality information in response to queries, we believe that it is likely the vast majority of users will want to limit their searching of the network to the identification of appropriate digital library services. NIDR technologies can be of service here: in fact, this is a much more constrained application than the "general" NIDR problem, in that digital libraries are resources at similar levels of aggregation when compared to one another rather than individual resources such as files. In addition, since digital libraries are typically managed resources, it

is not beyond belief to imagine that some systematic, relatively consistent method of describing these resources and their contents ("cataloging") might be established on an operational basis. And there will be orders of magnitude fewer digital libraries than there are files, Web pages, mailing lists, video and sensor feeds, and other individual networked information resources.

While the deployment of large numbers of digital libraries will, in our view, greatly diminish the long term importance of the "general" networked information discovery and retrieval problem for the average user's average query, the NIDR challenge will continue to be of great importance. Users of classic network-wide NIDR tools will be conducting research at the "fringes" of knowledge and information, beyond the boundaries of organized information in libraries: these users will likely include research scholars, financial analysts, intelligence and law enforcement analysts, detectives, and crisis managers. They may also include certain communities of information users who are not yet served by organized digital libraries.

A Closer Look at the Discovery Process

Goals and Tactics in the Resource Discovery Process

Resource discovery is not a simple, linear process of consulting some database or directory of networked information resources. Rather, we view resource discovery in this paper as an *human-centered* process. It has a great kinship to research, where the incremental acquisition of information and knowledge shape the ongoing process, and where biases, assumptions, experience and personal preferences, and indeed even chance inform decisions about what to do next. Resource discovery involves a number of tactical activities, many of which are supported by computer based tools, which are performed in an iterative fashion. The resource discovery process does not have a single, simple common structure or procession through stages; indeed, in many attempts to discover resources the process is shaped in a fundamental way by what resources the user is already familiar with from previous experience -- the goal is for the user to find *new* information on a topic. Resource discovery is also typically only a part of a broader information seeking activity that may span printed and broadcast resource, networked information, discussions with other people and perhaps even first-hand experimentation or other information gathering.

People approach the resource discovery process with many different goals. Typical examples include:

- Finding some *good* information about a topic; "good" here may mean one or more of many things, such as: recent, at an appropriate level of detail, assuming an appropriate level of prior knowledge of the topic, information that is quality-controlled or verified or authoritative in some fashion, or information that reflects a specific perspective desired by the user.

- Finding a known item (such as a document); in these situations the user may in fact think he or she knows what is wanted, but may have an ambiguous, incomplete or even incorrect description of the object.
- Finding everything that is available on a topic exhaustively, or at least determining how much information (and what kind of information) is available on a topic and how it is organized and structured, perhaps as a prelude to an exhaustive search for a more specific topic.
- Finding *new* information on a topic (that is, information that the user has not yet seen).

These are familiar goals that are in no way unique to the networked information environment; libraries have been helping people to achieve these goals many decades. Systems such as library catalogs are designed to support users with these goals, although to be sure they are far more effective in supporting some goals (for example, known-item or exhaustive searching) than others (finding a modest amount of "good" information).

Many tactics and methods have been developed over the years to support users in the pursuit of these goals. They include not only searching of various kinds of databases (or predecessors like printed indexes, card catalogs and bibliographies) to identify candidate resources, but also ways of organizing these candidate resources -- collocation or clustering of similar resources, elimination or consolidation of duplicates and differentiation of similar but distinct resources -- and the arrangement and presentation of candidate resources -- sorting on various criteria or ranking. It should be noted that catalogs -- both in print and computer-based -- in fact incorporate these organizational, presentation and arrangement features; in print, compilers were limited to static choices, while in computer based systems it is increasingly feasible, particularly as computational resources become larger and less costly, to tailor organization, presentation and arrangement dynamically to the needs of specific user inquiries.

If there are a large number of candidate resources various approaches may be used to provide abstracted views of the set, such a creating a listing of authors or of subject headings assigned to the resources in the candidate set. The user confronted by a large set of candidate resources may wish, more broadly, to explore how the structure of this set is related to the apparatus of classification that has been used to organize a body of information, such as a thesaurus or controlled vocabulary.

Users select among candidate resources thus organized and presented in a wide variety of ways: they browse or sample resources, they examine resource descriptions; they consider questions of availability and cost for obtaining access to the resources.

The examination of descriptive information is a particularly complex issue. Here the user brings knowledge that he or she has about what is being sought to

evaluate not only questions of relevance or quality but also "fitness for use". Fitness for use is a particularly valuable concept that we learned from the geospatial data community; it is one of the criteria that was used in defining the data elements that are part of the Federal Geospatial Metadata Standard. While fitness for use is not an unfamiliar concept even in the world of textual documents -- for example, if a person does not read German than documents in German are not likely to be particularly useful in most cases -- it takes on a very rich meaning in the networked information environment, where digital documents, images, audio or video resources may not be fit for use unless one has the necessary software, hardware and network bandwidth to exploit them, or where one is seeking remote sensing datasets at a specific minimum level or resolution, or structured data that includes specific data elements. Evaluation of fitness for use covers both semantic and syntactic considerations.

Hierarchy, Granularity and the Transversal of Information Spaces

One can view networked information resources as divided into two classes. There are actual objects -- documents, datasets, programs, and the like -- and there are information spaces which contain within them collections of objects. Each information space -- an interactive service, a database, a listserv or newsgroup, the WorldWide web -- comes with its own navigational and retrieval tools that operate within that space; further, objects within an information space are often organized, classified, and described according to specific schemes that mesh smoothly with the navigational and retrieval tools that define the information space. Information spaces may themselves be organized hierarchically; one information space may contain within it a number of subspaces, such as a system that houses many large databases but also contains navigational services which help the user to make selections among databases.

We have already discussed the difficulty of coherently presenting users with resources at different levels of granularity -- objects and information spaces -- and it should be clear that there is a dual problem in describing information spaces in a way that these descriptions are meaningful when intermixed with descriptions of individual objects. The description of information spaces must also be sufficiently flexible and detailed to reflect the evolution of their content over time (as in the case of a dynamically updated database or an active, wide-ranging newsgroup) if this description is to be helpful in directing users to the content of the information space.

The presence of information spaces among the range of networked information resources also contributes to the iterative nature of the discovery process. Discovery systems may help the user to identify candidate information spaces, but in order to explore and evaluate the contents of these information spaces, the user may need to not only employ specialized discovery and retrieval tools that are unique to each information space but also understand the rules and

conventions of information organization and description that are used within that information space. For a user in search of information or answers the retrieval of an information space is not a direct response but rather a suggestion about where to continue or focus the ongoing process of discovery.

Information spaces are not entirely disjoint; in fact, there has been an ongoing effort to make the contents of one information space visible to users of other information spaces through the user of gateways. These windows from one information space to another introduce distortions, and objects beyond the gateways often lack many of the descriptive attributes commonly attached to objects within the "home" information space. And, while a specific navigational or retrieval action may be able to reach through the gateway to another information space, the systems that build databases to support discovery processes may not be able to pass through gateways to inventory the contents of remote information spaces.

The situating of a user within a given information space also shapes the discovery process. Many users are now comfortable within a specific information space (and the tools used to navigate it) such as the Web. Resources that are difficult or awkward to describe effectively within the conventions of the Web (such as those behind gateways to other information spaces) are unlikely to be found as part of a discovery process; and, if discovered, the user may be reluctant to explore these resources because of the unfamiliarity of the navigational tools and information organization approaches within them. One might wish to start the discovery process at some base level where only information spaces were visible to the user, and where the first step in the discovery process was the selection of highest-level information spaces to explore further, but the nature of many of the existing information spaces, which are extremely large and which are defined on the basis of common tools and standards (such as the Web or Gopherspace) rather than along content lines (such as a database provider's offerings) suggests that this will be futile. The large technology oriented information spaces will be candidates for virtually every user's discovery process.

Granularity, hierarchy, and the boundaries of information spaces will likely be a continued problem in the discovery process. It is interesting to note that at present interactive database services (which can be viewed as information spaces) are frequently almost invisible to users of current NIDR tools. While these tools are beginning to successfully span multiple information spaces that contain relatively similar objects (for example, FTP archives, Gopherspace and the Worldwide web) it is much less clear how to usefully describe the unique, specialized information spaces represented by interactive database services or to present them alongside objects such as files and documents. And it seems to us that the number of these unique interactive information environments will grow rapidly in the near future: consider, for example, the efforts to develop collaborative information spaces to support research and learning, or the

transformation of traditional print newspapers into network-based information services.

Discovery as an ongoing process

While the discovery process as we have described it here can clearly be lengthy, it is bounded in the sense that the person seeking information is eventually satisfied (or at least sufficiently frustrated and exhausted to give up). This process of discovery may span a long period of time and involve many uses of various NIDR systems; it may be punctuated by extensive study of various resources that are retrieved during the process.

There is an additional type of discovery which needs to be supported by NIDR systems. Here the user has an *ongoing* need to be informed about newly available information on a given topic. The classical information retrieval literature often frames this as a problem of current awareness, selective dissemination of information, or filtering, rather than one of retrieval. There are architectural implications involved in supporting this type of discovery which we will explore in Chapter 2; essentially, the question is whether ongoing discovery is better supported by periodic searching or by examining new objects as they appear. In a highly dynamic and distributed environment the answers to this question revolve around both resource efficiency and user requirements for timely notification of the availability of new information.

Two issues involving ongoing discovery require highlighting here. The first is the much more extensive use of historical context in this class of discovery process (not only in identifying candidate resources but in ranking them). In ongoing discovery the NIDR system will need to consider what resources the user has already seen, and the extent to which he or she has found these resources useful. NIDR systems will be expected to develop measures of similarity between known useful objects and new objects and to use them in identifying and ranking new information. To a considerable extent, these issues are familiar from the classical IR environment (that is, a user interacting with a single database), although they take on new complexities because of the multiple and heterogeneous information sources in the networked information environment and the potential duplication of information among these sources.

What is fundamentally new and difficult in the networked information environment is the possible appearance of new information spaces as well as new objects as part of the results of an ongoing discovery process. These can be either entirely new resources (for example, a brand new database) or they can be existing resources that have just included relevant information as part of their content (for example, a new thread appearing on a newsgroup, or a document database that has added some new documents). A key question is the extent to which the NIDR system can reach inside an information space to retrieve information to the user, as opposed to the extent to which it can merely notify the user of the

existence of this newly relevant information space and invite the user to explore it directly, and perhaps use some information space specific NIDR tool to set up monitoring within the space for ongoing discovery purposes (if indeed such current awareness tools even exist for use within the information space). In the worst case -- where the high-level NIDR tool cannot reach inside the information space in question and no more specific tool or facility exists to monitor the availability of fresh information within the space -- the difficulty is how often to present the information space to the user as a new candidate resource. A network object may only change from time to time, and might be brought to the attention of the user anew when it has changed substantially (particularly if the user had previously found the object useful); an information space, particularly a large one, is likely to be constantly changing, and a NIDR tool that cannot reach inside it likely has no way of determining how often additional potentially relevant information has become available within the space, or how much new information of this type has been added recently. Thus it is unclear how often to tell the user who has initiated a process of ongoing discovery that such an information space should be re-examined for new relevant information.

The Role of Surrogates in Discovery

In the world of printed literature most discovery (with the exception of shelf browsing) operated purely with surrogate representations of the literature: cards in card catalogs, or entries in bibliographies and indexes. In the networked information environment where objects may be searched directly, the question is often raised as to whether the continued use of surrogates is useful, or whether discovery processes should operate directly on objects. We argue that surrogates will continue to play an essential role in networked information discovery and retrieval, although as discussed earlier it is important to recognize that the networked information environment offers new opportunities to derive (by extraction or computation) a much richer and more diverse set of surrogates from networked objects than the surrogates that were typically found in the print world. Chapter 3 will explore the scope and nature of the data elements that can contribute to surrogate construction in the networked information environment (and Chapter 4 will also explore an even more expanded role for such data elements). Our purpose here is to support the argument for the continued role of surrogates as a central architectural component. The justification includes the following points:

- Architectural: most retrieval protocols do not allow subcomponents of objects (such as data elements contained in headers) to be fetched separately.
- Performance: surrogates are often much smaller than the objects that they describe or represent; thus they require less resources to transmit, search, and store (including replication, which is important for scaling and reliability). There may well be cases when surrogates are larger than the base document they describe, particularly in cases where the surrogate is

computed from the base object, or perhaps the base object plus other objects linked to the base object.

- Scope: surrogates can represent or describe materials that are not necessarily immediately available in the networked environment; these might exist in some other form, such as print, or they may be housed in some form of tertiary storage (for example, large digital video or structured data files)
- Content: some data elements that are often found in surrogates are not actually part of the objects being described or represented, such as subject headings that might be independently assigned by human intellectual analysis, reviews, or linkages to related objects. Other data elements may require expensive computations (which perhaps may employ proprietary algorithms or software technology, or may include consideration not only of the primary object but also other databases such as authority files or dictionaries); storing the results of these computations as data elements in surrogates is much more practical then integrating the computations directly into the discovery process.
- Economics: In situations where there may be a fee for access to objects, surrogates are needed to permit users to make purchasing decisions. Similarly, some surrogates may themselves be marketed, independent of the economic framework controlling access to the objects that the surrogates represent or describe. The use of surrogates facilitates a market in networked information, and also a market in efforts to add value to the base of networked information by helping users to find the information that they need. Note that surrogates may be less expensive than the objects they represent (they may even be free) but equally they may be *more* expensive than the base objects they describe; we may well see services that offer reviews of publicly-available documents for a fee, for example.
- Intellectual Property: Just as surrogates permit much more flexibility in the economic framework, they also recognize the need to control access to intellectual property by some rightsholders while still disseminating awareness of the existence of this intellectual property. Similarly, some surrogates can represent intellectual property in their own right, independent of that which resides in the object being described or represented.
- Non-textual Resources. While it is possible to do considerable discovery on textual resources (documents) without the use of supplementary descriptions, particularly when these textual documents are structured through the use of a markup language, the state of the art in discovery and searching of non-textual information resources (interactive services, video clips, images, etc.) other than through descriptive text or other structured data elements that have been added to these information resources is extremely limited. To a great extent, we discover these resources through manipulation of supplementary textual surrogates.

A Closer Look at the Retrieval Process

In the Internet, retrieval systems were developed much earlier than resource discovery systems. The File Transfer Protocol has changed little in over a decade. In virtually all the existing information spaces -- anonymous FTP, Gopher and the Worldwide web are three excellent examples -- retrieval was the basic function of the system; discovery tools (even within the specific information space) were grafted on later, as an afterthought, rather than being an integral part of the architectural model and system design. In a very real sense, NIDR systems developed to permit people to find needed objects from among the vast number of objects available for retrieval through these protocols (or stored in these information spaces) only after the retrieval systems achieved very wide deployment and use.

Internet retrieval protocols have historically take very simple views of the objects that they retrieve. For example, the FTP protocol essentially understands binary objects and (ASCII) text objects. More recently designed protocols like HTTP and Gopher+ have more knowledge about types of objects, but the vocabulary is still relatively limited. And in most cases not much use is made of the information about object types as part of the retrieval process; a type designator is moved as part of the transfer operation so that the recipient can determine what local software should be invoked to process the received set of bits. There is no provision to transfer partial objects, except in cases where the designer of an object has segmented it into parts (for example, a series of linked HTML pages that form a single logical document).

In short, retrieval of a discovered resource in the Internet environment typically means either moving a collection of bits from some remote server to the local client, where this collection of bits is "cracked open" and processed by some viewer or other piece of local software (such as a spreadsheet processor), or invoking an interactive client to communicate with the remote resource. Most objects are copied and then used, rather than used across the net.

In contrast, most centralized, closed information retrieval systems have very minimal facilities to permit a user to fetch an object that they house; rather their design focus is searching these objects, and they offer a range of options for viewing or browsing search results in various formats, and perhaps some simple downloading facilities. The viewing and downloading functions are typically highly sensitive to the content and structure of the information objects being retrieved. In NIDR terminology, discovery was the emphasis, and the retrieval functions were designed in support of discovery.

The sparse functionality in retrieval protocols and lack of integration between discovery and retrieval functions in the networked information environment have, in our view, caused considerable confusion in the design of NIDR systems. Because retrieval protocols do not include provisions for negotiation between

clients and servers, and because servers are generally still functioning as simple suppliers of collections of bits, objects must be stored in multiple versions on servers to make the retrieval mechanisms work (rather than as the result of an engineering trade-off between computation on demand and storage of precomputed results).

Thus it is common to find documents stored in an ASCII format and also in multiple word processor formats on a server as separate files (often with no indication of which version of the document is the authoritative one), rather than simply having one object which can be delivered in multiple formats (via conversion at the server) along with some integrity and quality information indicating what the "native" format of the object is and some estimate (available as part of the retrieval process) of how much degradation is likely to occur as the result of a requested conversion. This proliferation of multiple formats makes the networked information environment even more confusing that it needs to be, in part by not representing objects at the proper level of abstraction due to limitations in the available retrieval mechanisms. It also weakens the integrity of network accessible information by not making content integrity part of the object's basic attributes.

Similarly, in some image applications one may actually find three separate stored copies of each image -- a thumbnail for browsing purposes, a screen-resolution image for viewing (precomputed based on some assumptions about screen resolution), and a high-resolution image for printing or for on-screen examination of details at high magnification (again precomputed based on some assumptions about the maximum resolution that is likely to be useful, as well as some consideration of the maximum resolution that the content provider is willing to offer). It would make more sense for the client to ask for what it needs based on the specifics of available hardware and usage, and for the existence of multiple versions of images to be hidden in the server if the server chooses to precompute them. Only the resolution of the highest-resolution image is a useful descriptive attribute of the image.

The cost to access an object is often proposed as a descriptive data element that belongs in an object surrogate. This seems completely inappropriate; while it is certainly useful to have some indication of whether access to an object is free in a surrogate, the cost of retrieving an object is going to depend on factors such as:

- The basic cost of the object, if any.
- Who is acquiring access to the object. An object may be free to certain communities, site licensed to certain communities, or discounted for access by certain communities.
- What format the object is being requested in. A thumbnail image may be free; a high resolution version of the same image may be very expensive.

- When the object is being accessed; access may be more costly during peak-use periods, and less expensive during off-peak hours.
- How long the resource is being accessed, for some types of interactive resources.

Similarly, other problematic but useful descriptive attributes of networked information objects, such as the size of the object (in bytes) can be more reasonably viewed as attributes of a retrieval transaction rather than as implicit in the object. A format-independent representation of some sense of object size (such as the number of words in a textual document) might then be more appropriate as an object attribute. It's unclear how useful a byte count for an image is outside of a retrieval transaction (where it might be represented by a labeled progress bar or some similar indicator); perhaps the dimensions of an image might ultimately provide a more reasonable sense of size once users obtain intuition about various levels of resolution.

A retrieval protocol that is better adapted to NIDR functions might well include the ability to negotiate costs for accessing an object, and also perhaps the ability to obtain a cost quote for access to an object without actually initiating a retrieval transaction. Note that determining and reporting the cost of access is a completely different matter from the mechanics of actually billing this charge back to the user; these issues of network-based electronic commerce are beyond the scope of this paper.

Closer integration between the discovery process and the retrieval protocols is needed for other reasons as well. If one examines NIDR systems today, they typically focus almost entirely on discovery and only offer access to retrieval tools for accessing discovered resources as an amenity. Some of the early discovery tools did not even go this far; they simply produced lists of resources that had to be passed explicitly to other retrieval tools, or they offered only minimal functionality in their implementations of retrieval protocols. In fact the integration between discovery and retrieval is complex and extensive, as illustrated by our description of the broader discovery process earlier. Users will want, for example, to browse *sets* of candidate resources that have been identified during discovery as a guide to what to do next; this means that a NIDR system should offer means of viewing groups of resources at various levels of abstraction (article summaries, image thumbnails, etc.) rather than forcing the user to examine them serially one at a time, moving back and forth between a list of candidate resources and a retrieval and viewing tool.

Key Problems In Achieving Current NIDR Objectives

This chapter has surveyed the networked information environment from a NIDR perspective and explored in some detail the discovery and retrieval processes from a user's point of view. It's clear that even limiting consideration to the current framework with an actively and continually engaged human information

seeker at its center the problems are extremely challenging. The following are some of the central problems:

- There seems to be a considerable mismatch between the complex iterative process of discovery and the more constrained operation of existing NIDR systems. Today's NIDR systems are handicapped by the limited amount of context and information about the user that they maintain.
- There needs to be a greater recognition of the implications of information spaces in the design of NIDR systems, and of the complexities of moving from one information space to another.
- Objects as viewed in the NIDR context (and particularly the perspective of retrieval protocols) are extremely simple; they are essentially collections of bits. This viewpoint, combined with the limitations of current retrieval mechanisms, creates a number of problems.
- There is a substantial reliance on classical information retrieval research concerning the retrieval of natural language documents. This is known to be a very difficult problem.
- There are a new set of issues involved in searching multiple information resources involving ranking, duplicate detection, and consolidation of information. These are not well explored.
- There are limited and inconsistent sources of descriptive metadata such as cataloging upon which to base discovery. This is particularly problematic with the growing among of non-textual information becoming available on the Internet.
- There are major technical issues involving appropriate architectures for very large scale distributed NIDR systems.
- Complementing these technical issues are a series of non-technical issues involving intellectual property, charging for information, privacy and control which will also influence NIDR architectures.

We will examine many of these problems in more depth in Chapters 2 and 3.

Yet, as we will discuss in detail in Chapter 4, the goals of the current NIDR efforts are in a real sense quite modest, and do not go far in accommodating ongoing discovery or increasingly autonomous discovery and retrieval agents or proxies. Nor do they facilitate the sharing and use of complex information resources other than to the extent that they can allow a user that understands such a resource to establish communication with it.

Outline: Chapter 2

Architectures and Technologies to Support the NIDR Process

Part I: The Machinery of Discovery

- targets for indexing; objects and information spaces
- the composition and subsumption of information spaces and the role of gateways among spaces
- defining collections and bounding searches
- the need for a modular indexing architecture (to allow introduction of new indexing methods)
- "push" vs. "pull" models: archie, veronica, webcrawlers, harvest.
- the central role of the "gatherer"; interactions with privacy, intellectual property, economics
- redistribution and aggregation of gathered indexing information -- index brokers and related proposals.
- protocol issues; interoperability, quality assurance, introduction of specialized extensions and new versions within an open architecture.
- the user side of discovery: "user history" databases. integration of multiple indexing systems. the mechanics of ranking and duplicate detection.

Part II: The Machinery of Retrieval

- characteristics and limits of current retrieval protocols
- "fetch and use" vs. "use across the net" models
- the URC framework for invoking retrieval
- the need to extend current retrieval protocols
- format conversion
- charging
- integrity issues
- browsing and sampling (thumbnails, derived data, sample data)
- component/subobject extraction and retrieval
- caching and replication; retrieving from managed distributed spaces

Outline: Chapter 3

Description and Metadata to Support Current NIDR Processes and Goals

Part I. Traditional forms of metadata information for NIDR

- self description by extraction and its limitations
- archie, veronica etc.
- HTML extraction (webcrawlers)
- SGML and DTDs; the use of the TEI header; linking semantics to DTDs.
- the descriptive cataloging tradition
- MARC practices; the 856 field and MARC as surrogate
- TOPNODE, GILS, OCLC Project

- issues around authority files, controlled vocabularies, thesauri, etc. problems of mixed controlled and uncontrolled vocabularies.

  simplified cataloging: RFC 1537, the "Dublin dozen" data elements

- combining multiple descriptive cataloging schemes
- attribute and data element mappings and hierarchies
- automatic indexing (IR)
- Essence project
- linkage with authority files (Gypsy); people and locations
- automatic type recognition of objects and use of heuristics
- limitations imposed by file systems without object typing
- nontextual media issues -- images, sound and video
- describing compound and aggregate objects and information spaces: how do you describe a database or other information space. Newsgroups as information spaces.

Part II: Enriching current NIDR with new types of metadata information

- usage data, citation data (links); self description by use.
- nondescriptive information: reviews, bibliographies, etc. from human beings. Pathfinders. Seals of Approval. Extent to which these are part of the "intellectual" infrastructure that supports NIDR processes as opposed to directly integrated with NIDR systems.
- the re-use of "management" metadata in retrieval processes. use in cache management based on access patterns. use in retrieval (expiration dates, regeneration schedules, classification and embargo). to what extent does pure management metadata exist?

Outline: Chapter 4

A Critique of Current NIDR Objectives: Are These Goals Too Limited?

- interoperation vs. access/retrieval as a goal
- limited semantics within current NIDR framework; simple vs. complex objects. objects that require extensive software mediation and interpretation.
- assumption of human discoverer and human user for resources
- alternative visions: object spaces, federated databases, megaprogramming, semantic level interoperability, exportable ontologies and metadata, agent economies and ecologies, ...
- programs (agents) as direct end NIDR service consumers
- redefining retrieval: use of complex objects within NIDR framework directly rather than after import by user

- open vs. closed information spaces.
- can the old and the new visions coexist?
- the formation of communities around common semantics and interchange standards.