Association of Research Libraries (ARL)
http://www.arl.org/resources/pubs/br/br217/br217mhp.shtml

ARL: A Bimonthly Report

ARL: A Bimonthly Report, no. 217 (August 2001)

Metadata Harvesting and the Open Archives Initiative

by Clifford A. Lynch, Executive Director, Coalition for Networked Information

Introduction

This article describes the Open Archives Metadata Harvesting Protocol, an important new infrastructure component for supporting distributed networked information services. The Metadata Harvesting Protocol—a mechanism that enables data providers to expose their metadata—is seeing very rapid deployment, and enables a fascinating array of new services and system architectures for a diverse set of communities. I will speculate about some of these services and discuss issues involved in their development. This article is not intended to be a definitive technical summary of the protocol; documents providing such a discussion can be found at http://www.openarchives.org/. Rather, the focus here is on the uses of the protocol and its strategic significance as an enabling technology.

There has been considerable confusion about the Open Archives Metadata Harvesting Protocol, mostly beginning with and stemming from its name. The protocol no longer has much to do with archiving or archives, other than in terms of its heritage. There is a broad movement now well established within the scholarly publishing world, championed by people like Stevan Harnad at the University of Southampton and Paul Ginsparg of Los Alamos National Labs,[1] to enhance public access to scholarly journal articles through the use of "e-print" servers. (These servers are sometimes referred to as "repositories" or as "archives"—for example, the Los Alamos Preprint Archive —although they are not necessarily archives in the technical sense that the digital preservation community speaks of repositories and archives.) The fundamental idea here is that authors would deposit preprints and/or copies of published versions of their articles into such servers, thus providing readers worldwide with a free way of obtaining access to these papers, without needing paid subscription access to the source electronic journals. The proponents of this movement argue that the refereed scholarly journal literature really belongs to the scholarly community and by extension to the world at large, and that such free access is better aligned with the interests of both authors and readers. The deposit of preprints would also speed up and democratize the frontiers of research and access to new knowledge; instead of a privileged circle of members of "invisible colleges" sharing preprints, these preprints would be available to everyone immediately, without the delays introduced by the journal refereeing and publication cycle. Proposals such as PubMed Central and the Public Library of Science build upon these ideas.[2] This movement is sometimes called "open archives" and addresses structural change in the system of scholarly communication.

The Open Archives Metadata Harvesting Protocol grew out of an effort (described in the next section) to solve some of the problems that were emerging as e-print servers became more widely deployed; it originated in the community concerned with advancing the development of e-print archives. However, as work on the protocol advanced it became clear that it provided a very general-purpose mechanism that could address a surprisingly

wide range of urgent needs and, in order to be useful in the widest variety of contexts, this mechanism had to be defined so that it was agnostic to assumptions about types of content, economic models, intellectual property constraints, or systems of information flow. For better or worse, by that time the protocol was widely known as the Open Archives Protocol, and the program to develop it was widely known as the Open Archives Initiative, so the decision was made to maintain the popularly known terminology.

The protocol is now often referred to as the Open Archives Metadata Harvesting Protocol in an attempt to reintroduce a bit more clarity. This Metadata Harvesting Protocol (MHP), as I will describe in more detail below, is simply an interface that a networked server (not necessarily an e-print server) can employ to make metadata describing objects housed at that server available to external applications that wish to collect this metadata. A server does not need to be part of an e-print program to use the protocol; indeed, it does not need to house journal papers at all. The server does not need to offer free access to the digital objects that it stores. The Metadata Harvesting Protocol can be implemented by servers housing commercial products available by subscription (like scientific journals) or available for transactional purchase, by libraries or museums offering access to digitized images, or by scientific data archives, to list only a few potential applications.

Many of the people who have been involved in the work on the MHP are also strong proponents of the broader program to open free public access to the scholarly journal literature, but they are usually careful to differentiate between the two; the MHP can certainly advance the deployment of e-print archives, but it is not limited to supporting that program. It is a much more general and more neutral technical-enabling mechanism. The term "Open Archives Initiative" is ambiguous; it refers both to a specific, organized, funded effort to advance the Metadata Harvesting Protocol, and to a separate, diffuse movement to promote the deployment and use of e-print archives involving a lot of the same people.

A Brief History of the Open Archives Metadata Harvesting Initiative

The Open Archives Metadata Harvesting Initiative had its roots in work done by Herbert Van de Sompel (then of the University of Ghent) working in collaboration with groups that included researchers and librarians at Los Alamos National Labs in the United States. In late 1999, a meeting was convened in Santa Fe, New Mexico to address a series of problems that he had been exploring in the e-print server world. Fundamentally, the problem was that as disciplinary e-print servers proliferated, often with overlapping disciplinary coverage and geographical affinity, there was a need to develop services that permitted searching across papers housed at multiple repositories. Repositories also needed capabilities to automatically identify and copy papers that had been deposited in other repositories. Further complicating matters, institutional e-print archives (such as the DSpace project being carried out at MIT in collaboration with Hewlett-Packard[3]) were beginning to emerge that would house the intellectual output of specific campus communities; it was undesirable to require scholars to deposit their work in multiple repositories, and hence it would be necessary for disciplinary archives to be able to identify and replicate material from these institutional archives and for federated search services to be able to search across both institutional and disciplinary archives.

The participants at the Santa Fe meeting explored many aspects of the issues involved in addressing these problems, including how to identify e-print archives of interest and how to codify acceptable use policies for material found in such archives. But the heart of their work was the definition of an interface to permit an e-print server to expose metadata for the papers that it held; this metadata could then be picked up by federated search services or by other repositories that wanted to identify papers for copying. The results of this

effort were documented in the Santa Fe Convention,[4] the precursor of the MHP. I will not go into the technical details of that work here, since changes were made as the Santa Fe Convention evolved into the Open Archives Metadata Harvesting Protocol, and these details are of primarily historical interest today.

The participants at Santa Fe made a key architectural decision that is worth highlighting here, however, because it has become a foundational principle for all subsequent work: they adopted a model that rejected distributed search (as might be done with the Z39.50 information retrieval protocol, for example) in favor of simply having servers provide metadata in bulk for harvesting services, subject only to some very simple scoping criteria, such as providing all metadata added or changed since a specified date, or all metadata pertaining to papers meeting matching gross subject partitions within an archive (such as physics, as opposed to mathematics).

This decision has generated some controversy, and has sometimes been misunderstood. Z39.50 is a mature, sophisticated, but unfortunately very complex protocol that allows a client to search a remote information server across a network. It can be used as a tool to build federated search systems; in such a system, a client sends a search in parallel to a number of information servers that comprise the federation, and then gathers the results, eliminates or clusters duplicates, sorts the resulting records and presents them to the user. It has proven very difficult to create high-quality federated search services across large numbers of autonomous information servers through Z39.50 for several reasons. Retrieval accuracy is a problem: different servers interpret Z39.50 queries differently, in part due to lack of specificity in the standard, leading to semantic inconsistencies as a search is processed at different servers. Also, there are scaling problems in the management of searches that are run at large numbers of servers; one has to worry about servers that are unavailable (and with enough servers, at least one always will be unavailable), and performance tends to be constrained by the performance of the slowest individual server participating in the federation of servers. And the user has to wait for a lot of record transfer and post-processing before seeing a result, making Z39.50-based federated search performance sensitive to participating server response time, result size, and network bandwidth.

For other metadata retrieval purposes that were also of interest to the Santa Fe group, such as guiding server replication, Z39.50 is an elegant and flexible way of retrieving metadata. But, unfortunately, it is far from trivial to implement either in clients or servers, although the availability of program libraries for the Z39.50 protocol has made this less of a problem than it once was.

The Santa Fe group wanted a very simple, low-barrier-to-entry interface, and to shift implementation complexity and operational processing load away from the repositories and to the developers of federated search services, repository redistribution services, and the like. They also wanted to minimize the interdependency between the quality of applications services as viewed by the user and the behavior of repositories that supplied data to the applications services. Thus, instead of using a distributed search model based on something like Z39.50, they chose to use an architecture that had been used in early networked information resource location systems like Archie (developed by Peter Deutsch and Alan Emtage in the early 1990s) and perhaps most elegantly articulated in the Harvest system developed by Mike Schwartz, Peter Danzig, Mic Bowman, and colleagues at the University of Colorado, Boulder in the mid-1990s (this is the system that gave us the verb "harvesting" for metadata). To a great extent, this is also the same architecture that had been validated by a long history of successful experience within the library community in building and operating very large-scale (centralized) union catalog databases. The Santa Fe participants recognized that every repository housed metadata, and so they devised a very simple way for repositories to export this metadata on

demand; service developers would then take the responsibility for actually collecting, or "harvesting," this metadata, and for processing (such as duplicate elimination or clustering) and normalizing this metadata to support functions such as federated searches. A user querying a federated search service would not interact with the repositories, but only with a database that the federated search service had already constructed from metadata harvested from participating repositories, for example. Hence the performance of the federated search service was largely independent of the performance or reliability of the participating repositories. The design goal was that a repository should be able to implement the Santa Fe Convention with a few days of programmer time, as opposed to months.

To clarify fully the relationship between Z39.50 and the Santa Fe Convention, let me make a few observations (which are equally applicable to the MHP discussed later). These two protocols are really meant for different purposes, with very different design parameters, although they can both be used as building blocks in the construction of similar services, such as federated searching. Neither is a substitute for the other. They make very different choices about the allocation of both developmental and operational complexity, cost, and effort among the components that participate in the delivery of a service, such as federated search, and about the characteristics of the resulting service (such as how quickly database updates are visible to the user of federated search). And we should not think about the world becoming partitioned between Z39.50-based resources and MHP-speaking resources, but rather about bridges and gateways. It is quite reasonable to think about a service that is constructed using the Santa Fe Convention/Open Archives Metadata Harvesting Protocol offering a Z39.50 interface to its user community, if such an interface is useful to that community. A Z39.50-speaking server can fairly easily be made MHP-compliant, and I would expect to see the development of gateway or broker services that make Z39.50 servers available for open archives metadata harvesting in cases where the individual server operators do not want to undertake this development work; this is not a major technical problem, assuming that there is common understanding about the metadata schemes to be supported.

Following the late-1999 Santa Fe meeting, there were several workshops held during 2000 at venues such as the ACM Digital Libraries meeting to share the thinking of the Santa Fe meeting with the broader networked information community. Out of these workshops a very surprising consensus emerged. Many other groups had very similar problems to those faced by the e-print community, including libraries, museums, commercial journal publishers, and communities of scholars who needed to share distributed data resources. The metadata that each community wanted to make available had unique features, but the fundamental mechanism of making metadata available for harvest subject to some very simple selection criteria seemed to be widely needed. Based on this recognition of common needs, the Coalition for Networked Information and the Digital Library Federation provided funding to establish an Open Archives Initiative (OAI) secretariat at Cornell University, managed by Herbert Van de Sompel (by then a visiting professor there) and Carl Lagoze (a research professor at Cornell widely known for his work in the development of advanced networked information access systems). An international steering committee was put in place to guide the effort, and a program was launched to generalize the Santa Fe Convention to support harvesting of all kinds of metadata, as well as to explore other infrastructure issues related to metadata harvesting (such as registries of sites available for harvesting, and interesting services that could be built given the availability of metadata to be harvested). The OAI convened a technical meeting at Cornell in September 2000 to rework the Santa Fe Convention and subsequently refined these specifications via e-mail review.

The revised specifications were made public in January 2001, with two day-long workshops (one in Washington, D.C. in January, and the other in Berlin in February) to

introduce them to potential implementers. The intention is that, with the exception of clarifications or correction of gross errors, these specifications will remain stable for at least a year while the community gains experience in using them. (In fact, there has already been one revision as of July 2001, because the protocol depends on a suite of XML-related standards and the Worldwide Web Consortium, which manages these standards, has made changes to them which necessitated corresponding changes in the MHP.) We will convene a meeting of technical experts again in very late 2001 or early 2002 to consider what revisions or extensions need to be made to the MHP specifications based on this year of experience. Following this process, we may submit the protocol to a formal standardization process through an organization like the National Information Standards Organization (NISO) or the Internet Engineering Task Force (IETF); at present it has no status as a formal standard endorsed by a formal standards body.

Meanwhile, implementation is moving ahead. A number of repositories already support harvesting according to the protocol (a list of these is available at the Open Archives website), and various services based on harvested metadata are under development. The Andrew W. Mellon Foundation has recently funded a number of proposals to help underwrite the development of experimental services that are built on metadata harvesting [see related article on p. 10].

The Metadata Harvesting Interface

Without going into technical detail that isn't relevant here, the Metadata Harvesting Protocol uses a very simple HTTP-based request-response transaction framework for communication between a harvester and a repository. A harvester can ask for metadata to be returned with optional restrictions based on when the metadata has been added or modified (in other words, it can obtain new or changed metadata since its last harvest interaction with a repository); it can also restrict metadata by server-defined "partitions" (think of these as gross subject-oriented subcollections housed on a server). The server returns a series of sets of metadata elements (in XML) plus identifiers (i.e., URLs) for the objects that the metadata describes.

Supporting this core harvesting transaction are a few ancillary transactions: for example, to permit a harvester to obtain a list of the names of partitions that a server has defined and which can thus be used as restriction criteria in harvesting requests.

Multiple metadata schemes are supported in the Open Archives Metadata Harvesting Protocol—this is really the key architectural change from the Santa Fe Convention. The protocol requires that all servers offer unqualified Dublin Core metadata (encoded in XML) as a lowest common denominator; however, each server is also free to offer metadata in one or more other schemes, and a harvester can request that metadata be provided in a scheme other than Dublin Core as part of the harvest request. There is also another auxiliary transaction that permits a harvester to obtain a list of the names of the metadata schemes that a given repository supports. The underlying idea here is that we will see communities of practice evolve that define metadata schemes that are richer and more precise than unqualified Dublin Core; for example, the e-print archives community is already working on one that encodes various important data elements for e-prints, such as author affiliations, bibliographic information if the paper has been published in a journal, and even the paper's cited references in a structured form. These community-specific schemes could be handled as qualified Dublin Core, or as de novo schemes; the only requirement is that they be transportable in XML. This is a very powerful method of enabling communities with common metadata to work together while still ensuring some minimal level of interoperability for very broad federated search or other services based on unqualified Dublin Core. It seems very likely that the MHP will drive development of community-specific metadata schemes. It will also be helpful in clarifying our

understanding of the value of unqualified Dublin Core for lowest common denominator cross-domain resource discovery.

Functionally, this is a reasonably complete description of the open archives metadata harvesting interface; I have ignored a number of technical details here, such as how requests for very large amounts of metadata are segmented across multiple harvest transactions. You can find these details in the technical specifications at the Open Archives website. But to really understand open archives metadata harvesting at a functional level, it is equally important to emphasize what is not within the scope of this interface.

Authentication and access management are not covered. A given repository can use a range of access-management mechanisms that are external to the protocol to decide which harvesters it will provide with metadata if it wishes to impose access control. Independent of access control decisions about providing metadata, the server only includes a pointer—such as a URL or URN—to each object described by metadata that it makes available for harvest; access controls on this base object may or may not exist. In addition, any authentication that the harvester wants to conduct on the repository is handled by external mechanisms.

The protocol does not address the very real issue of how harvesters will identify repositories that they wish to harvest, nor does it provide information to help determine when harvesting should occur, or how frequently. Questions about acceptable use of harvested metadata are not addressed by the protocol; these might be agreed upon explicitly as part of establishing a harvesting relationship with a server that is access-controlled, or they might be simply advertised as terms and conditions that any harvester automatically agrees to in the case of a publicly-accessible server, but in any case this is outside the scope of the harvesting protocol.

There will clearly be a need for some kind of registry of names for well-known community-specific metadata schemes. The MHP does not address this, though it clearly must be part of the broader infrastructure associated with the protocol. Along with the community-specific schemes, we will want agreements about how these schemes are downgraded to unqualified Dublin Core; documentation of these mappings is not part of the protocol but again may be part of the broader infrastructure.

Applications Enabled by the Metadata Harvesting Protocol

The most obvious applications that are enabled by the Metadata Harvesting Protocol are those that helped to motivate the work at the initial Santa Fe meeting: repository synchronization and federated search. For repository synchronization, one compares metadata from two or more repositories and decides what objects should be copied from one repository to another (along with the necessary metadata). The hard part here is in the application: deciding what repositories to examine, and determining the criteria for identifying what to copy. There is also a problem with the propagation of metadata from one repository to another; it's not clear (other than by using community standards) how to determine the most comprehensive metadata set describing an object so that all of the relevant metadata can be copied over.

Similarly, federated search using MHP is not hard in principle; one collects metadata from a number of sites, normalizes it, clusters it in some fashion to deal with duplicates as appropriate, and offers search services against the resulting database. In practice, all of the details are complex: what sites to harvest, how often to harvest them, how to normalize metadata (especially if one wants to do better than the lowest common denominator—unqualified Dublin Core—offered by each site), how to handle duplicate objects—these are all key design issues that need to be addressed. MHP provides a very powerful framework for building union-catalog-type databases (in the broad sense; not just union

bibliographic catalogs, but all sorts of union descriptive databases) for collections of resources by automating and standardizing the collection of contributions from the participating sites, which has traditionally been an operational headache in building and managing union catalogs. But there are many complex specifics that need to be coded into any actual implementation.

A set of applications closely related to federated search deal with the potential enhancement of web search engines in at least two distinct dimensions. One is providing a more efficient way for web search engines to crawl static HTML pages, and also to obtain metadata associated with these pages (there are other methods of getting the metadata, such as in-line META tags, but the MHP provides a much more flexible way of doing this). The second is being able to integrate various parts of what is sometimes called the "deep web" or the "invisible web" with the indexing of static web pages, including repositories of digital objects and databases that do not exist as retrievable and indexable static web pages, and also proprietary content, where the content owner may be willing to make metadata about the content available to facilitate finding it, but may be unwilling to permit arbitrary web-indexing programs to have direct access to the content in order to index it. The Metadata Harvesting Protocol allows a server to enumerate the objects it houses and to provide metadata associated with these objects, no matter what the nature of these objects and the access constraints that might apply to them.

Note that any application dealing with metadata created and stored on distributed sites faces issues about whether it can trust the metadata it is relying upon.[5] This means, for example, that it is unlikely that the public web search engines will use an OAI interface to harvest arbitrary sites anytime soon; however, federated searching or customized indexing of sites that are selected in some fashion—by being part of an organizational intranet (think of institutional portal sites in this context), or through some type of editorial policy that selects quality sites as part of a subject portal, for example—will likely make wide use of the protocol.

These are the most obvious applications (and the ones that will probably be available soonest), in that they directly extend or enhance existing practices. Interesting and novel applications are likely to emerge as well. For example, one can easily imagine the rise of intermediary services (reminiscent of the brokerage services in the original Harvest system) that collect raw metadata from sets of sites, consolidate and enhance it, and then redistribute it as a single feed or as custom-selected subsets to still other sites for reuse. Creative applications designers are at the very early stages of exploring the services that the MHP can enable, and I think we can expect some fascinating and unexpected developments in the next few years.

Open Questions and Future Directions for Open Archives Metadata Harvesting

While the Open Archives Metadata Harvesting Protocol solves one very important set of problems, it also focuses attention on a number of other issues that will have to be addressed as applications proliferate. Some of them will require progress in standards and/or other networked information infrastructure components; others are simply not well understood at this point and will require considerable research and experimentation to allow the development of a body of design knowledge and community practice. In this final section I will briefly sketch some of these issues.

Acceptable Use and Intellectual Property Issues for Metadata

Many sites make metadata about their holdings publicly available today in the sense that they offer publicly accessible search services operating against that metadata, which can then result in the display of individual records (possibly in a fully tagged format suitable for reuse by another computer application, or possibly not). This is the case with most library catalogs; for example, they offer searching to anyone who wants to search the

database, and some systems will even provide formatted displays of individual, full MARC records that are retrieved by such searches as an option. There is a great difference between this practice and simply making the entire catalog's contents, or major subsets, available for bulk downloading and re-incorporation into other databases and services. In a world of MHP-enabled network resources, metadata becomes migratory and recombinant.

Owners and operators of large databases and repositories will need to think through how comfortable they are in making their metadata generally available on this basis, and what limits, if any, they want to place on permissible reuse. In addition, this may focus new attention on metadata ownership; while nobody may much care about who reuses the odd individual bibliographic database record, copying an entire bibliographic database may be viewed quite differently. Finally, there is no real way to encode acceptable-use policies or restrictions on metadata harvesting within the MHP (though the earlier Santa Fe Convention did try to address this to some extent); while it would not be difficult to make reference to textual statements about such policies, encoding them (or even a modest range of common policies) in a machine-understandable way is a difficult problem that calls for collaboration between computer scientists, lawyers, and standards developers, among others. This is not an area where we have seen great progress to date. And infrastructure to permit sites to limit harvesting to a specific set of "partners" (perhaps those who have agreed to license agreements constraining the reuse and redistribution of the metadata that they will harvest) is still lacking; this is part of the general and very complex authentication and access management problem. Certainly there are readily available ad-hoc solutions for authenticating harvesters today, but they may not be as secure, as scalable, or as interoperable as one might wish.

Where to Harvest: Selection, Registries, and Trust Questions

Many applications will use manually-curated lists of sites to harvest, and, indeed, the editorial processes and selectivity that go into the development and maintenance of these lists will be part of the value of the application itself. As these services multiply, operators of new repositories will face the problem of bringing their site to the attention of the appropriate service operators so that their metadata can be harvested.

If we look at the operation of the public web search engines, which seek to be comprehensive, we find much more complex methods of identifying sites to crawl (the details of which vary from search engine to search engine, and are viewed as representing proprietary advantage to these search engines). They start with a base of websites that they have identified, or that site operators have submitted to the indexing services, but then they also dynamically discover new sites to index by analyzing links on sites that they visit. They also use a variety of techniques (both manual and automated) to determine how often to revisit sites looking for new material to index.

Some applications performing metadata harvesting will want to do a similar dynamic discovery of sites that are available to be harvested. This could be done in a number of ways. Certainly, for systems that harvest static HTML pages, it would be possible to program their crawlers to also attempt to do an OAI query against each site to see if it offered metadata for harvesting (though there would be a lot of overhead in this, since, presumably, for the foreseeable future only a very small percentage of sites would offer such metadata). It would be possible to establish a central registry of sites that offered metadata via the OAI secretariat; setting up a database that allowed sites to register themselves and interested parties to search the registry to identify harvestable sites is technically straightforward and not terribly resource intensive.

A prototype of this exists today on the Open Archives website; this is simply a list of MHP-compliant sites that have submitted their addresses to the OAI secretariat. The

secretariat does some essentially mechanical validation of the sites by running a program periodically that issues OAI queries; it checks that the site is still there and provides syntactically well-formed responses to these queries. This mechanical validation is very valuable in these early days of experimental implementations, though I think there are questions about how practical it will be to extend it to truly enormous numbers of sites.

But it is important to recognize the very limited value of such a registry. It does not address how one might classify sites (for example, by the nature of their content or their subject coverage) in order for applications to decide which sites should be harvested. Developing computer-based methods of doing this, at any level of generality, is a complex proposition (though there is always the expedient of having applications check out all newly registered sites a few times, and only continuing to harvest the ones that prove to have interesting material; this, of course, does not scale well.) Worse, it is not only complex, it is subjective and prone to debate. Someone, either the site operator or a third-party "cataloger" would need to describe each site, and the value of the registry would be closely tied to the quality and accuracy of this description. Finally, since we know so little about the spectrum of applications that are likely to emerge, any thinking about appropriate descriptive schemes for such a registry is almost pure speculation at this point in time. For all of these reasons, the OAI metadata harvesting initiative has decided that it is premature to do more than the simple prototype registry already described. But registries will clearly become an issue over time.

There is another dimension of the registry problem that is even more important and difficult. This has to do with rating the quality and accuracy of the metadata that the various sites offer for harvesting, which can range from impeccable to shoddy to actively deceptive (keep in mind that in some sense metadata is an unsubstantiated assertion about content that the site stores; the site may offer this content for inspection and use only under highly constrained conditions such as license). Many applications will likely be structured around the principle of offering in-depth access to high-quality information resources, which means that they will have to be concerned with the assessments of the quality of the metadata offered by the sites they choose to harvest. It's difficult to see how this can be handled at this point other than by evaluation and/or trust decisions by the individual applications. Scaling these decisions past manual editorial management of lists of sites to harvest takes us quickly to research areas such as reputation management systems, further complicated by the fact that both the harvesting services and their end users may want to be able to make choices about trust and reputation in their use of metadata.

Granularity

In developing MHP interfaces to databases or repositories, a key design decision has to do with the granularity of the objects for which the site exposes metadata. For the original e-print-archive applications this was straightforward; the objects of interest were individual article e-prints and thus they made available metadata for each article e-print. But as metadata harvesting is deployed in a wider range of contexts, the answers are less clear. For many repositories we have little or no metadata at the individual object level, and instead have various forms of collection-level description (such as the Electronic Archival Description, or EAD, files); in other cases we may have both collection- and item-level records, each of which has value in its own right, but it is not clear how to link the two and it also isn't always clear that it is useful to make huge numbers of item-level records available. In the extreme case, we have large databases, such as online catalogs or abstracting and indexing databanks, where it would be possible to expose truly vast numbers of individual records through the MHP (think here of RLG, PubMed Central, or OCLC's WorldCat), but where it is not clear which applications would actually find such large numbers of records useful. In some situations, we will want to be able to expose metadata for databases or services in the aggregate, so that users can identify and visit

services that may contain useful information; yet the effective description of such services or large databases in the aggregate is a research problem that has received only limited exploration to date, and this exploration has met with only limited success.

One very interesting OAI-enabled service that we may see, which will also perhaps re-invigorate research in how to effectively describe databases, is something I think of as a database summarizer. This service would use the MHP to pull metadata for all of the individual records in a database, and would then perform computations that allow it to provide some kind of summary metadata record for the entire database as output; these summary records could then be returned to the database owner or operator or could be made available to other resource discovery services. Prior to the availability of OAI metadata harvesting, any research group that wanted to explore this area had to actually go out and obtain copies of various databases to experiment with.

Metadata Schemes

The Metadata Harvesting Protocol is a means of making machine-processable metadata widely available for use. It will create tremendous pressure for standards programs that are now developing schemes to codify and represent such metadata, especially the Dublin Core Metadata Initiative. MHP specifies unqualified Dublin Core as the lowest-common-denominator mandatory metadata scheme for interoperability purposes. This will likely lead to the first truly large-scale test of the real utility and practicality of unqualified Dublin Core for resource discovery. It is certainly clear that almost any metadata scheme can be "downgraded" into unqualified Dublin Core (in a non-reversible fashion), but the actual usefulness of such a coarse metadata scheme has been the subject of considerable debate and speculation. OAI applications may provide large-scale empirical data on this, which will be invaluable to the future evolution of the Dublin Core work.

The Dublin Core Initiative is also codifying the notion of qualified Dublin Core—a series of extensions to the base Dublin Core elements that can provide much greater precision but that can be "interpreted down" to unqualified Dublin Core in a consistent fashion by applications that are not knowledgeable about specific extensions. The final versions of the specifications for qualified Dublin Core have been very slow in issuing, and because many communities of practice will likely want to build on Dublin Core, rather than creating de novo metadata schemes, OAI metadata harvesting will add to the pressure to advance these standards.

One of my personal hopes is that because MHP is designed to facilitate the interchange of computer-processable metadata it will lead to progress not only in descriptive metadata that is assigned by human intellectual activity, but also in the development and implementation of content-based computationally derived metadata—computed profiles of digital objects based on their content, such as word occurrence frequency modules for textual objects or spectral signatures for images—which will support distributed content-based retrieval as a complement to retrieval by intellectually-assigned metadata. To date, there has been little progress in this area, because content-based retrieval has always occurred within the context and boundaries of a specific system.

Finally, at a broader level, the idea of community-based metadata schemes is central to the vision of the Metadata Harvesting Protocol. There has been a great deal of work in this area over the past few years, in part driven by the promise of XML and the growing interest in the interchange of computer-processable information using the Internet. Fundamentally, the development of these standards is a social process within a community, though it calls upon a poorly-codified body of knowledge in data structuring, descriptive practices, classification, and other areas. It is still very slow, and very expensive, to develop these community standards, and it is hard to predict which ones will be successful in terms of broad adoption or in terms of effectively meeting

community needs over reasonably long periods of time. OAI metadata harvesting will increase the pressure to improve the speed and effectiveness of such standards development processes.

Conclusions

The Open Archives Metadata Harvesting Protocol opens many new possibilities which are yet to be explored. This means that it is difficult, and speculative, to establish strategies to exploit the new technology. But these opportunities are too important to be ignored.

For content suppliers, the way forward seems clear. They should prepare to offer metadata through the MHP interface. Yet they will need to think very carefully about what they are doing, both in terms of what metadata they want to expose and at what level of granularity, and in terms of the potential reuse of this metadata. This is particularly true for operators of online catalogs, though it is also a question for organizations mounting special collections of all kinds. Any organization offering access to a sophisticated networked information resource may find the MHP is a new way to make content available to a variety of innovative service providers.

For data-intensive scholarly communities in which data is widely distributed rather than centralized into a few key community databases, this interface may offer a new way to translate rather abstract investments in metadata standardization into tangible opportunities to contribute to operational systems for locating information resources. And it may have other far-reaching implications; for example, in communities where the resources to underwrite centralized databases haven't been available, or where the community practices emphasize local control of datasets by individual research groups, the base of available information may become much more visible to the community.

Finally, OAI metadata harvesting may offer a new bridge to bring innovation in networked information services and applications out of the research community more rapidly than has been the case in the past. Organizations that manage large databases and production information services are generally slow to innovate because their first priorities appropriately reflect the needs to exercise stewardship over the data and to provide reliable service to their user communities; most of their resources tend to be tied up in operations and maintenance. Researchers who want to explore new ways of organizing, presenting, or using these large data resources will now have a standardized way of extracting content without much disruption or cost to existing operational systems. This may be a powerful mechanism for enabling the development of new applications and services that have never before been possible.

Footnotes

1. It was announced in July 2001 that Ginsparg and the Los Alamos Preprint Archive will be moving to Cornell University.
2. Visit http://www.pubmedcentral.nih.gov/ and http://www.publiclibraryofscience.org/ for more information.
3. See http://web.mit.edu/dspace/.
4. See http://www.openarchives.org/meetings/SantaFe1999/sfc_entry.htm.
5. For more on this topic, see my article "When Documents Deceive: Trust and Provenance as New Factors for Information Retrieval in a Tangled Web" (Journal of the American Society for Information Science and Technology 52.1 [Jan. 2001]: 12-17), available at http://www-cse.ucsd.edu/~rik/others/lynch-trust-jasis00.pdf.

To cite this article

Clifford A. Lynch, "Metadata Harvesting and the Open Archives Initiative," *ARL: A Bimonthly Report*, no. 217 (August 2001): 1–9, http://www.arl.org/resources/pubs/br/br217/br217mhp.shtml.