Magnus Enger (11815)

The concept of 'overlay' in relation to the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)

Masters thesis in Documentation Science

Faculty of Humanities

University of Tromsø, Norway

Fall 2005

Version 1.0

Edit: #1056, 2005-11-18 14:50

Table of Contents

Chapter 1: Introduction

1.1 An ecology of documentation

The study of documentation in general (as defined by Lund 2001), and scholarly documentation in particular, might fruitfully be conducted within the metaphor of an ecology. In an earlier work I have have formulated a tentative definition of an ecology of documentation:

> The ecology of documentation is the study of documents in relation to the surroundings in which they are created and used. These surroundings are called the environment of the documents. This environment is made up of many different components, including other documents and their effects, and non-documentary (e.g. cultural, social, psychological and physical) factors. (Enger 2004, p. 5)

The species within such an ecology of scholarly documentary forms may be grouped into at least three genera:[1]

- *Primary documents* are e.g. books, articles or grey literature, created to aid in the dissemination of new findings or scholarly ideas.
- *Secondary documents* (or metadata) are created to describe (or document) the primary documents. These may be printed cards in a physical library catalogue, records in a database or in an XML document.
- *Systemic documents* are collections of secondary documents, arranged in such a way that they facilitate the discovery and/or locating of primary documents. Examples include library catalogues and reference databases.

Important environmental factors that influence how documentary forms develop and evolve are e.g.:

- Traditions in different fields that dictate what channels of distribution give the most status, or how one should relate to new technologies of distribution.
- Technological developments such as the invention of the Internet and the World Wide Web, and its increasing ubiquity in academia.
- The actions of individual or group actors within the ecosystem, such as influential scholars adopting or advocating a certain mode of distribution.

One of the predictions that can be drawn from the ecological way of looking at documentation

[1] "**genus** n. Taxonomic group of closely related species, similar and related genera being grouped into families." (Lawrence 1992, p. 203)

is that as one or more environmental factors change, so the documentary forms will change, to adapt to the changed environment.

Related to this is the concept of adaptive radiation, which can be defined as the "evolutionary process in which species descended from a common ancestor multiply and diverge to occupy different ecological niches" (Lawrence 1992, p. 9). Such processes can for example be observed when a group of islands is colonized by e.g. a new species of birds:

> Sometimes a founding population arrives on one of a group of islands and, as it colonises each one, each population is changed to suit conditions on that particular island. This results in a group of closely related species on the different islands. This process of speciation on a group of islands is a form of adaptive radiation. (Chapman and Reiss 1999, p. 247)

One of the hypothetical premises of the present work is that the Open Access movement and the creation of the Open Archives Initiative Protocol for Metadata Harvesting, which gives free access to metadata that describe openly available scholarly documentation, is creating a situation in which new forms of systemic documents will evolve. These new forms will initially be closely related to traditional systemic documents, from which they are "descended", but in time they will develop new features in order to adapt to the new environment. This process will then be one of adaptive radiation, as secondary and systemic documents adapt to the specific demands of the environment of different niches, such as specific scholarly fields.

1.2 The concept of "overlay"

There are at least two distinct sources for the concept of "overlay" in the sense in which it is used in the present work. The earliest occurrence of the concept in relation to scholarly documentation I have been able to find is in Ginsparg (1996):

> Any type of information could be overlayed on this raw archive and maintained by any third parties. (Ginsparg 1996, section 6)

He then goes on to elaborate further on this idea:

> One possibility is that some consortium of professional societies and institutional libraries will ultimately acquire the technical competence to provide umbrella sponsorship of the global raw research archive. Those societies that are as well non-profit publishers may continue to organize high-quality peer-reviewed overlays (though perhaps no longer as a means of generating income to subsidize other non-publishing ventures) [...] (Ginsparg 1996, section 8)

In the following year, referring back to a presentation of the same ideas held as early as 1993, John W. T. Smith launched the idea of the "deconstructed journal", by examining the roles and functions of journals as we currently know them, and by outlining how these roles and functions could be taken care of in a new, network-based and distributed model:

> As one might expect, at the core of this new model is a Web site/service [...]. This service contains links to relevant items of interest to its readers (subscribers). The New Scientific Journal (NSJ) is the visible replacement for the current Scientific Journal. Some of the important differences between this and the current paper-based and e-journals are:
>
> - The operators of this service do not own, or have any exclusive rights in, the items pointed to.
> - A major role of the service is to act as a 'filter' (as described in Part 2 above) between the contents of the net and the user or subscriber - not to be a repository of the said material.
> - The operators of this service (the NSJ) may, or may not, arrange the quality control (content) stage of the publishing process. (Smith 1997, section titled "The 'New Scientific Journal' - an Overview")

There is no reference to Ginsparg in this document, and the term "overlay" is not used, but the system described has striking similarities to that described by Ginsparg (1996).

By 2004 these two currents have met and merged into one:

> The name *overlay journal* comes (I believe) from a comment in Ginsparg (1996) where he discusses the possibility of information services provided as an 'overlay' on the Physics e-print archive. [...] An overlay journal (aka virtual journal) is basically a list of evaluated and commented links to full text articles held elsewhere. (Smith 2004, section titled "Overlay Journals". Emphasis in original.)

Some journals refer to themselves as "overlay journals". One example is *SIGMA*,[2] which says of itself:

> We are pleased to announce that SIGMA is an arXiv overlay journal. That the SIGMA is an overlay means that all published articles in the journal have been contributed or will be contributed to the arXiv. In addition the SIGMA web site has hyperlinks to the arXiv copies. [...] If an accepted for publication paper is

[2] <http://www.emis.de/journals/SIGMA/>

> already in the arXiv, the author should give to the Editors both the paper number and the password so that we can replace it with the typeset version.[3]

Front for the Mathematics ArXiv[4] defines overlay in the following terms:

> An overlay is any web site or collection of articles that refers to part of the arXiv. It can be as simple as a personal list of publications or as complicated as a full-fledged search engine.[5]

In the present work I will build on this last definition (but without limiting the view to only those services that are overlays to ArXiv.org), and apply the term to any service which "points to" primary scholarly documentation held in Open Access repositories (by harvesting metadata from these repositories), but which does not themselves host such documents. Since these are systemic documents that gather together secondary documentation in order to facilitate the discovery of primary documents, the precise name for my object of scrutiny should be *overlay systemic documents*.

1.3 The way forward

The goal of the present work is twofold: The first goal is to investigate how the emergence of a protocol for the exchange of structured metadata is facilitating the creation of new overlay systemic documents, and how they are adapting features of traditional systemic documents to the conditions of the new environment. This will be done through a general presentation of the Open Access "movement", which has prepared the way for the protocol, in chapter 2, as well as an introduction to some of the salient features of the Open Archives Initiative Protocol for Metadata Harvesting itself, in chapter 3. A short survey of the features of existing sites that fit the description of overlay systemic documents will then be presented in chapter 4, in order to uncover the status quo of the adaptive radiation of these forms of documentation.

The second goal is to report on an experimental overlay systemic document, which aims to implement some features not found in existing overlays, based on features from documentary forms that are native to the Web environment and that have already proved successful in this environment. The prototype from this experiment, as well as the experiences gained from it, will be presented in chapter 5.

Based on the survey from chapter 4 and the prototype from chapter 5, chapter 6 will discuss some of the possible ways overlay systemic documents might evolve, and how the nature of the OAI-PMH itself might influence this process.

[3] <http://www.emis.de/journals/SIGMA/about.html#overlay> (Accessed 2005-10-27)
[4] <http://front.math.ucdavis.edu/>
[5] <http://front.math.ucdavis.edu/overlays> (Accessed 2005-10-27)

Chapter 7 will draw some final conclusions and chapter 8 will make brief recommendations for future work that would explore further the ideas presented in this work.

Chapter 2: The changing face of scholarly documentation

There are several factors actively participating in changing scholarly documentation as we know it. On the one hand there are changes triggered by the move from the traditional regime of paper-based publishing of scholarly journals to the networked environment. On the other hand there is a growing dissatisfaction with how the distribution of scholarly documentation works, among its producers and consumers. Together these forces are likely to change the face of scholarly documentation as we have known it.

It should be noted that the present work will be concerned with what might generally be termed “articles”, or article-like documents. One key characteristic of these documents is that they do not result in a direct, economic compensation to the author when they are published. The classic examples are articles published in scholarly journals, which are given for free to the journal that publishes it, as opposed to books, for which the author gets a fee and/or royalties from sales. This is an important distinction in that it enables authors to choose the venues of publication that help give them the largest audience or impact in return for their efforts in creating the documents.

2.1 The remediation of scholarly documentation – from printed to networked

The first examples of scholarly, electronic journals distributed from computer to computer over a network go back a couple of decades, pre-dating the World Wide Web (WWW) by several years (Suber 2005b), but the number of such journals were initially low. With the advent of the World Wide Web, and the tremendous growth of users of on-line services in general, the number of journals available on-line have exploded. Large publishers like e.g. Elsevier have created on-line presences for their journals, while the printed journals are still being distributed in the traditional ways. Some journals have discontinued their printed editions and moved to a completely digital and networked mode of distribution. A lot of new, online-only, journals have also sprung up.

2.1.1 Recreating the printed journal in the networked environment

When commercial publishers have made their journals accessible in the networked environment, it is striking that much effort has been put into recreating the “look and feel” of printed journals. Portable Document Format (PDF) is often used to present the articles (although often accompanied by a HTML-version) in a way that closely resembles the look of the printed journal. The concepts of “volumes” and “issues” have also been retained. This obviously made a lot of sense when journals were printed, since lumping several articles together for printing and distribution is more cost-effective than handling each and every article on its own. In the networked environment however, it would be just as easy to make articles available as soon as the final version of the article is ready.

2.1.2 Adapting to the networked environment

As I have sketched out in Enger (2004), when a documentary form is recreated in a new medium, it is to be expected that it will initially retain a lot of its original features, but that after a while the form will adapt to the specific characteristics of the new environment.

There are signs that such changes are happening. As mentioned above publishers are providing articles in HTML-format, and some of these include "link-enabled cited references" (Jacsó 2004), which exploit the inherent hypertext-capabilities of the WWW. There is also a trend for articles to be made available electronically as soon as they have been peer-reviewed, and before they have appeared in the printed edition of a journal. The "Articles in Press"-feature of Elsevier's ScienceDirect is one example of this.[6]

There is, however, evidence of even more far-reaching changes, initiated not by traditional, commercial publishers, but by enterprising individuals or groups. Some of these involve the form of scholarly documentation, such as inclusion of video, datasets, interactive programs etc., see e.g. McKiernan (2002, 2001, 1999) for some examples. In the following I will be focusing on yet another aspect of scholarly documentation where changes are evident, that of the modes of distribution of such documentation. I will be focusing particularly on the Open Access "movement", and the phenomenon of so called Open archives or repositories.[7]

2.2 The Open Access "movement"

The Open Access "movement" is not a member organization with a board of directors and a clearly defined set of goals and motivations. Rather it is a confluence of different groups and individuals sharing more or less the same dissatisfactions with the status quo, striving towards goals that are more or less the same, using more or less the same methods. The main objective that everyone is working to achieve is the removal of access barriers (primarily understood as economic barriers such as subscriptions or pay-per-view systems) to scholarly journal articles, i.e. that this documentation should be available for free to anyone with access to the Internet.

The dissatisfaction with the system in its current form takes many shapes, and is variously voiced by different stakeholders:

[6] <http://www.sciencedirect.com/>

[7] "Open Archives" was the original name of this phenomenon and it is still retained in e.g. the name of the Open Archives Initiative <http://www.openarchives.org/>, but in the most recent literature there is a tendency to substitute the term *repositories* for *archives*, since archives are associated with a long tradition of curation and focus on longevity that is not necessarily evident in the Open Access movement. In the following I will use repositories as the preferred term.

Researchers are dissatisfied with the fact that restrictions on access to their published articles is hampering the impact of these articles, and thus the growth and development of their disciplines. Studies show that articles with Open Access have higher impact than those made available with access-barriers. Lawrence (2001) was the first to describe this phenomenon, Harnad and Brody (2004) give a summary of a more recent study that contrasts the impact of articles from the same journals, with and without access-barriers.

Librarians are dissatisfied with the increasing costs of journal subscriptions, and the fact that these costs makes it impossible to supply patrons with the full breadth of relevant materials in a timely manner.

The current system of scholarly publication is thought to introduce unnecessary delays in the availability of scholarly documentation, and this is seen as hampering the progress of the scientific endeavour.

An increasing number of writers are becoming aware of the fact that traditional publication in journals often entails signing the copyright in the published articles over to the publisher, which can strongly limit what the author can do with her article after it has been published.

There are also arguments in favour of Open Access that are of a more political kind:

Politicians and the public are dissatisfied with a situation where publicly funded research results in articles and other documentation that is given for free to commercial publishers, only to be bought back expensively by libraries that are also publicly funded.

As mentioned above, an objection to the current system is that even libraries at the largest and best funded institutions can not afford subscriptions to all the scholarly journals that might be relevant for its faculty. This situation is of course many times worse for libraries and researchers in developing countries, which are denied access to, and thereby the opportunity to build on, information that might ultimately be of importance for the development of a sustainable economy in their countries.

The dissatisfaction with the old regime of article-publication is not just spurred on by faults in the old system, there is also a recognition of some benefits that would come with Open Access to the research literature. For example there is the possibility of carrying out document-analyses such as that reported in Bollen et al (2005). Another interesting possibility is that of "open" citation analysis as described by Hitchcock (2002).

There are at least two complementary approaches in the Open Access movement. These could be labelled "Open Access journals" and "Open Access repositories":

2.3 Open Access journals

These are journals that have either evolved from traditional, paper-based journals, or new journals that have been created in the online medium.

2.3.1 Born-open journals

The first on-line, freely available scholarly journals started appearing in the 1980s, well before the invention of the WWW.

Gustafsson (2002) estimates that only 1.5% of the worlds scholarly journals are Open Access, and that 40-50% of the Open Access journals that existed in 1999 were discontinued by 2002. These numbers may reflect that early Open Access journals were created by individuals or small groups of enthusiasts, who are not able to keep up the energy required for long-term activities. There is however a tendency for better organized groups and established organizations to get involved in starting up new journals.

Some of the most high-profile efforts have been made by the Public Library of Science (PLoS), which is supported by several large grants:[8]

> [...] PLoS has initially published two journals - PLoS Biology and PLoS Medicine - that compete head-to-head with the leading existing publications in biology and medical research, publishing the best peer-reviewed original research articles, timely essays, and other features.[9]

2.3.2 Overlay journals

Some journals incorporate repositories in their infrastructure, while the role of the journal is primarily reduced to conducting the peer review-process and applying a seal of recognition to articles. This forms the basis of the journal-model known as the “distributed journal”, described by Smith (1997, 2004). See chapter 1.2 (p. 7) for details and examples.

2.3.3 Converted journals

Some journals that started out as traditional, printed, subscription-based journals have converted to the Open Access paradigm, and are now available on-line, free of charge. These journals may retain a printed version parallel to the electronic one, and this is of course not free.

2.3.4 Other journals of interest

Some journals that do not fit the description of Open Access journals are also interesting in this context:

[8] <http://www.plos.org/>

[9] <http://www.plos.org/journals/index.html>

2.3.4.1 Hybrid journals

One of the strategies for conversion from a “closed” to an “open” journal outlined by Crow and Goldstein (2003, p. 15-22) is that of the hybrid journal. This approach gives authors the choice of whether to provide Open Access to their articles or not. Authors who are not concerned about Open Access submit their articles in the normal way, but those who want to reap the benefits of Open Access can pay a fee that compensates the publisher for any loss of revenue from the Open articles. In this way Open and non-Open articles can co-exist in the same journal and even within the same issue of the journal. This is seen as an excellent way for journals to test the waters of Open Access, without committing to it completely.

2.3.4.2 Cooperating journals

A lot of traditional journals have not converted to Open Access or embraced the hybrid approach, and demand that authors sign over the copyright in the articles to the publisher, for the privilege of being published in its journals. By doing this authors are also relinquishing the rights to distribute the articles in any form, after they have been published in this particular journal.

But there are also some publishers that grant authors the right to make available versions of published articles on personal home pages and/or in institutional repositories. The policies of several thousand journals are charted by the SHERPA/RoMEO Publishers' Copyright Listings.[10]

This cooperation of a lot of the “traditional” journals have helped pave the way for the other branch of the Open Access movement, the Open Access repositories.

2.4 Open Access repositories

Repositories are web-based software systems that store and make available documentation as well as metadata describing that documentation. In the following I will only consider repositories that conform to, and make metadata harvestable through, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). The technicalities of this protocol will be discussed further in chapter 3 (p. 19), but first I will review some of the different roles repositories play.

Repositories can be divided into categories based on who is funding and maintaining them, and on who has the opportunity to supply content to them:

2.4.1 Institutional repositories

As the name implies, these are repositories run by institutions, and their goal is to capture the

[10] <http://www.sherpa.ac.uk/romeo.php>

research output of the local faculty. Theses and dissertations written by students might also be included.

2.4.2 Disciplinary repositories

Disciplinary repositories collect documentation that is relevant for a discipline, regardless of the institutional affiliation of its authors. ArXiv.org (described in McKiernan 2000), the oldest and largest of the repositories, is the prime example of such a repository.[11] These repositories are often run and maintained by a host organization, but funding can come from different sources, such as grants. Some vetting is often carried out to ensure the submitted materials are at least marginally relevant to the discipline, but this should in no way be confused with the quality control carried out by traditional journals.

2.4.3 Funder repositories

A third category of repositories are run by funding agencies. An emerging practices is for these bodies to require that any documentation that results from the funding they provide should be deposited in a repository run by that body. This does not stop authors from also making the documentation available in institutional and disciplinary repositories.

2.4.4 Personal repositories

Although the focus in the Open Access movement is on the three categories of repositories outlined above, there has also been at least one proposal for repositories that are meant to hold the output of one single individual, namely the Kepler framework presented in Maly, Zubair and Liu (2001), Liu, Maly and Zubair (2002), Liu (2002, chapter 7) and Maly, Liu and Zubair (2003). Repositories at the individual level raise a whole host of questions relating to reliability and the scalability of the OAI-PMH:

> The intention of OAI has been to support a contributing audience consisting of few data providers, each representing a digital library with a large holding (on the order of a hundred thousand to a million objects). In the Kepler service, the opposite is true: each data provider has only a few objects (e.g., an order of a hundred) but there may be, if the Kepler service is successful, tens of thousands [...] of such archivelets. (Liu 2002, p. 74)

Fong, Hui and Vu (2002) illustrate the difficulties in identifying scholarly publications presented on author home-pages, so if this is the alternative, personal repositories might still be preferable, at least from the point of view of the creators of overlay systemic documents, who want to gather metadata from all relevant sources.

[11] <http://arxiv.org/>

2.5 Forms of documentation

When it comes to documentary forms, repositories are usually able to store any file format. It is however interesting to note that the advent of repositories has created a convergence of the distribution of different documentary forms, in that documentation from different stages of the research process are now available through the same channel. It used to be that conference presentations were only available to those attending the actual conference or those that requested it from the author, drafts and pre-prints were only available to the author's circle of friends and colleagues, and the final, refereed versions of journal articles were only available to those who subscribed to the journal or were affiliated with a library that did.

Today documentation from all these different stages are all available through a single channel – repositories – and to anyone with an Internet connection. Distinctions between the different kinds of documents can be hard to draw, but some kinds stand out as particularly interesting.

2.5.1 Pre-prints

A pre-print is a draft of an article that has not yet been subjected to formal peer review, but which is intended for publication in a journal.

The Open Access movement can trace an important strand of its roots to the pre-print culture that existed in high energy physics, long before the advent of the WWW. It was customary for authors to circulate drafts of articles among a wide circle of colleagues before it was submitted to a journal. In this way new findings were made known as early as possible and it was possible for others to build on these findings, and to avoid repeating costly experiments that had already been carried out elsewhere. For years this exchange was paper-based, and it was taken care of by the authors themselves, or by enterprising individuals who established mailing lists of interested researchers (Kling and McKim, 2000 p. 1308; Kling 2004 p. 601-602). With the advent of the Internet and the WWW this informal communication was made ever more efficient, first through e-mail, under the auspices of Paul Ginsparg who established an electronic "bulletin board" (Taubes 1993), and then through the repository set up by Ginsparg at the Los Alamos National Labs (LANL) and today known as arXiv.org, the forebear of the Open Access repositories we know today.

When the question of quality control is raised in this context, one answer is that peer review is not really necessary in high energy physics, because the experiments that are needed are so costly that no-one is allowed to carry them out who is not thoroughly approved by those who fund the experiments in the first place. This field is also dominated by large, highly visible projects, which anyone who wants to be in the field needs to be aware of. So researchers are usually familiar with the researchers and institutions in their fields and can assess pre-prints based on this familiarity. Thus peer review and publishing in journals are an activity carried out after the fact, more to record the history of the discipline than to communicate and disseminate the newest findings.

One criticism often raised is that this way of disseminating scholarly documentation, with its roots in a narrow field of the sciences, will not transfer well to other fields, with different traditions. Kling and McKim (2000) discuss how these differences might impact on the transition from printed to hybrid or electronic forms of documentation.

2.5.2 Post-prints

Post-prints are articles that have been subjected to a formal process of peer review, and that have been accepted for publication in a journal. As mentioned above (2.3.4.2, p. 15), some journals allow articles from this stage to be posted on author home pages or in repositories. A further subdivision can be made between those that allow posting the "official" PDF of articles, containing the logo, formatting, layout and so on of the publishing journal, and those that allow posting just the plain text of articles, without the formatting provided by the journal. The latter kind can often be difficult to distinguish from a pre-print, but they will often contain a note stating what volume and issue of a journal it appeared in, and authors are encouraged to include such a note (Suber 2005a).

2.5.3 And everything else...

Post- and pre-prints (collectively known as e-prints) make up an important part of the content of repositories, but they are not alone. In fact, one of the most interesting aspects of the Open Access movement is the way it results in a convergence of documentation from all stages of the research process in one channel, namely repositories.

The most interesting thing about all these kinds of repositories, in the context of the present work, is that they can be built to comply with a protocol for metadata harvesting, which gives anyone who wants to construct systemic documents free access to the metadata describing the documents in the repositories. Some details of how this works are given in the next chapter.

Chapter 3: Anatomy of the OAI-PMH

The story of how the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) came into existence is described, by the authors of the protocol, in Lagoze and Van de Sompel (2003). Here I will only give a brief summary of the main features of the protocol, with some emphasis on those features that are of greatest interest for the construction of the prototype.

3.1 Data providers and service providers

The fundamental units in the infrastructure of the OAI-PMH are *data providers* and *service providers*.

3.1.1 Data providers

Data providers run the systems that store and make available primary documentation:

> A repository is a network accessible server that can process the 6 OAI-PMH requests in the manner described in [the specification]. A repository is managed by a data provider to expose metadata to harvesters. (Lagoze et al (eds) 2002a, chapter 2.2)

Initially these primary documents were Open Access scholarly articles, but there is nothing in the protocol that says this is the only kind of documents a data provider can hold. It is for example also possible to make available metadata about articles that are not Open Access, but where some sort of toll-gate is in place. The metadata could also be used to describe physical objects, such as artefacts in a museum.

Several Open Source software packages are available for running repositories that comply with the OAI-PMH. DSpace[12] and eprints.org[13] are among the best known and most widely deployed.[14] There is also a number of commercial services that offer to run and maintain such a repository on behalf of an organization for a fee, e.g. *Digital Commons@*[15] from UMI/ProQuest and *Open Repository*[16] from BioMed Central.

3.1.2 Service providers

Service providers collect metadata from data providers through an operation known as *harvesting*. The software that performs these operations is known as a "harvester":

> A harvester is a client application that issues OAI-PMH requests. A harvester is operated by a service provider as a means of collecting metadata from repositories.

[12] <http://www.dspace.org/>
[13] <http://www.eprints.org/>
[14] An extensive list of "OAI Tools" is available from <http://www.openarchives.org/tools/tools.html>.
[15] <http://www.umi.com/proquest/digitalcommons/>
[16] <http://www.openrepository.com/>

(Lagoze et al (eds) 2002b, chapter 2.1)

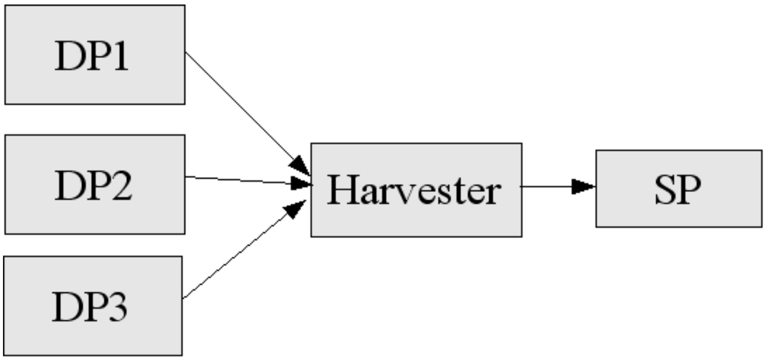


Illustration 3.1 The relationship between several data providers (DP1-3), one service provider (SP) and the harvester run by the service provider. Arrows indicate the movement of metadata.

A service provider can harvest metadata from one or more data providers. By utilizing some of the features of the OAI-PMH it is also possible to harvest sub-sets of the metadata that is available from a data provider.

The harvested metadata are used as the basis of different “overlay” services, such as searching across a collection of repositories, or the metadata can be enhanced in some way.

Some service providers also serve as data providers, by making the metadata harvested from data providers available to other service providers through the OAI-PMH.

It is important to note the difference between this approach of harvesting, and what is known as distributed searching. In a distributed search, a search-request is submitted to several services, records matching the search are retrieved from each service in real time and then combined in some way before they are presented to the user. With the harvesting approach, all metadata from the data providers are aggregated on a regular basis, and searches are run over the aggregated metadata, so there are no direct requests to the data providers when searching is conducted by a service provider.

Another important distinction should be made between harvesting of metadata, and the harvesting of the resources (or primary documents) that the metadata is about. In the OAI-PMH itself there is no mechanism for harvesting resources, only the records containing metadata *about* the resources. The URL of a resource (if it is network-accessible or -addressable at all) is usually included in the standard metadata, and it is possible to extract the URLs and automatically download the documents, but there is no provision for this in the OAI-PMH itself - as we will see below (chapter 3.5.6, p. 31) there is a verb in the OAI-PMH called GetRecord, but there is no verb called GetResource. See Van de Sompel et al (2004) for a discussion of this issue.

3.2 XML over HTTP

The OAI-PMH is built on top of HTTP, the protocol that is the basis of the World Wide Web, and inherits a lot of its characteristics from this basic protocol. One of these inherited characteristics is the notion of “request” and “response” between “client” and “server”. The relationship may be sketched like this:

	WWW	***OAI-PMH***
Protocol	HTTP	HTTP
Client	Browser, e.g. Firefox[17]	Harvester run by service provider, e.g. Celestial[18]
Server	Web server, e.g. Apache[19]	Data provider, e.g. DSpace
Request	HTTP POST, GET, PUT	HTTP POST or GET
Response	HTML or other document format	XML

Table 3.1: Comparison of HTTP and OAI-PMH.

The responses returned by data providers are always in XML format, and should conform to a publicly available XML Schema.[20]

3.3 Resources, items and records

The three central entities within the OAI-PMH are resources, items and records:

3.3.1 Resources

Resources are what the OAI-PMH is all about:

> A resource is the object or "stuff" that metadata is "about". The nature of a resource, whether it is physical or digital, or whether it is stored in the repository or is a constituent of another database, is outside the scope of the OAI-PMH. (Lagoze et al (eds) 2002a, chapter 2.2)

It is important to note the fact that the resources themselves are “outside the scope” of the protocol. This makes the protocol very flexible, since it does not tie implementations to any preconceived notions about what constitutes a resource. As we will see in chapter 3.6 (p. 32), the protocol is being used in applications well beyond that of sharing metadata about scholarly documentation.

3.3.2 Items

Items can be seen as the representations of resources in a repository.

[17] <http://www.mozilla.org/products/firefox/>
[18] <http://celestial.eprints.org/>
[19] <http://www.apache.org/>
[20] <http://www.w3.org/XML/Schema>

> An item is a constituent of a repository from which metadata about a resource can be disseminated. That metadata may be disseminated on-the-fly from the associated resource, cross-walked from some canonical form, actually stored in the repository, etc. (Lagoze et al (eds) 2002a, chapter 2.2)

For each resource (which can be outside or inside the repository) there is one item. This item can, on the other hand, be represented by one or more records.

3.3.3 Records

Records are the actual manifestations of metadata:

> A record is metadata in a specific metadata format. A record is returned as an XML-encoded byte stream in response to a protocol request to disseminate a specific metadata format from a constituent item. (Lagoze et al (eds) 2002a, chapter 2.2)
>
> A record is metadata expressed in a single format. A record is returned in an XML-encoded byte stream in response to an OAI-PMH request for metadata from an item. A record is identified unambiguously by the combination of the unique identifier of the item from which the record is available, the metadataPrefix identifying the metadata format of the record, and the datestamp of the record. (Lagoze et al (eds) 2002a, chapter 2.5)

It is important to note that a record is always in a specific format, and that it is always represented in XML. A record consists of three distinct parts:

3.3.3.1 Header

The header contains some higher-level information about the record in question:

> [It] contains the unique identifier of the item and properties necessary for selective harvesting (Lagoze et al (eds) 2002a, chapter 2.5)

The "properties necessary for selective harvesting" are identifiers of sets, which is described further below (chapter 3.5.3, p. 29). Unique identifiers are discussed in chapter 3.4 (p. 26).

3.3.3.2 Metadata

As stated above, the actual metadata that make up the "payload" of a record must be in a specific metadata format, and be encoded in XML. This manifestation is contained in the metadata-part of the actual record. In order to establish a basic level of interoperability among data providers, the OAI-PMH standard specifies that all complying repositories must be able to disseminate metadata about all its items in the unqualified Dublin Core format:

> At a minimum, repositories must be able to return records with metadata expressed in the Dublin Core format, without any qualification. Optionally, a repository may also disseminate other formats of metadata. (Lagoze et al (eds) 2002a, chapter 2.5)

The Dublin Core is a basic set of metadata elements that were initially described in 1996, which is widely used as a lowest common denominator for metadata in a lot of different contexts.

The Dublin Core consists of the following 15 elements. (Term names and definitions are taken from DCMI 2005, Section 2):

- **contributor** - An entity responsible for making contributions to the content of the resource.
- **coverage** - The extent or scope of the content of the resource.
- **creator** - An entity primarily responsible for making the content of the resource.
- **date** - A date associated with an event in the life cycle of the resource.
- **description** - An account of the content of the resource.
- **format** - The physical or digital manifestation of the resource.
- **identifier** - An unambiguous reference to the resource within a given context.
- **language** - A language of the intellectual content of the resource.
- **publisher** - An entity responsible for making the resource available
- **relation** - A reference to a related resource.
- **rights** - Information about rights held in and over the resource.
- **source** - A reference to a resource from which the present resource is derived.
- **subject** - The topic of the content of the resource.
- **title** - A name given to the resource.
- **type** - The nature or genre of the content of the resource.

The Dublin Core itself is independent of implementation, so it does not specify how the metadata elements should be represented. They can be plain text, XML or some other format. When Dublin Core is used in the context of the OAI-PMH, the metadata have to be encoded as XML, to comply with the protocol's demand that all metadata be represented in XML. See chapter 3.5.6 (p. 31) for a complete example of an OAI-PMH record with Dublin Core metadata encoded in XML.

Within the Dublin Core specification, all these elements are seen as optional and repeatable, which means that every element can be present zero or more times. Ward (2002, 2004) has shown that even this basic set of metadata are not utilized fully in repositories – while "title" was used by 98.8% of repositories, only 19.5% used "relation" (Ward 2004, p. 45).

One criticism that is often levelled at the Dublin Core set of metadata elements is that it does not specify how information should be represented. What format should e.g. the "date" element be in? This makes interpreting the data difficult for computers, and this has implications for what services can be built on top of the metadata, especially when metadata from different data providers are aggregated into a single service.

Even a cursory glance at the list of metadata elements available in the Dublin Core reveals that it is not ideal for dealing with metadata about many common forms of scholarly documentation. There is, for example, no good way to express information about the volume, issue and page-numbers of an article that has been published in a journal. This could perhaps be included in the "source" element, but the Dublin Core itself does not specify a standard way to do this, so parsing out the information for use in e.g. browsing or searching would be non-trivial.

To alleviate this a large number of richer metadata sets have been developed. Some pre-date the OAI-PMH while others have been developed specifically for this context.

To get an impression of the diversity of metadata formats, one can examine the page called "Distinct Metadata Schemas",[21] which is part of the "Experimental OAI Registry at UIUC".[22] This page lists the distinct URIs of XML Schemas that are used to define the syntax of the different metadata sets. Each Schema can have several "prefixes" associated with it, and these are also listed on the page. The list is ordered by number of occurrences, in descending order. The following table lists the 10 most used Metadata Schemas as of 2005-09-22:

Schema URI	Occurrences	Prefixes
http://www.openarchives.org/OAI/2.0/oai_dc.xsd	802	collexis, dare_didl, dc2, oai_dc, oai_dc2, oai_dcm, openURL
http://www.openarchives.org/OAI/1.1/rfc1807.xsd	145	oai_rfc1807, rfc1807
http://www.openarchives.org/OAI/1.1/dc.xsd	142	oai_dc, oai_dc_1.1

[21] <http://gita.grainger.uiuc.edu/registry/ListSchemas.asp>

[22] <http://gita.grainger.uiuc.edu/registry/searchform.asp>

Schema URI	Occurrences	Prefixes
http://www.openarchives.org/OAI/1.1/oai_marc.xsd	112	oai_marc
http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd	91	marc, marc21, marc21a, marc21b, marcxml
http://www.ndltd.org/standards/metadata/etdms/1.0/etdms.xsd	59	etd-ms, oai_etdms
http://www.openarchives.org/OAI/dc.xsd	42	oai_dc
http://www.language-archives.org/OLAC/1.0/olac.xsd	29	olac, olac_display
http://www.persistent-identifier.de/xepicur/version1.0/xepicur.xsd	18	epicur
http://ns.nsdl.org/schemas/nsdl_dc/nsdl_dc_v1.01.xsd	11	nsdl_dc

Table 3.2 The 10 most popular metadata formats as reported by the Experimental OAI Registry at UIUC

We see that different versions of the basic Dublin Core metadata set (with the standard prefix oai_dc) are the most popular, but that different needs are also being accommodated:

- Different formats related to MARC, usually associated with automated library systems, can be seen as indicators of a connection with legacy data from libraries, or an effort to establish interoperability with such data.
- The “electronic theses and dissertations” (ETD) formats is an example of metadata standards tailored to specific documentary forms.
- OLAC is a format that has been developed for the “Open Language Archives Community” - an example of a metadata standard developed for a particular scholarly community.

These are just three examples of specific needs that result in the development of new and more specific metadata standards to supplement the basic Dublin Core. As the OAI-PMH standard is used in new contexts and by new communities, we should expect to see an adaptive radiation of metadata formats.

3.3.3.3 About

> an optional and repeatable container to hold data about the metadata part of the record. The contents of an about container must conform to an XML Schema. Individual implementation communities may create XML Schema that define specific uses for the contents of about containers. (Lagoze et al (eds) 2002a, chapter 2.5)

Suggested uses for this section are information about the intellectual rights connected with the resource or the metadata, or information about the provenance of the metadata, e.g. if the metadata was originally harvested from another repository this could be recorded in this section. The OAI-PMH does not specify what form this information should take, other than that it should conform to a publicly available XML schema.[23]

3.4 Identifiers

One of the weakest points of the Internet and the World Wide Web as we know it, is the fact that it is dependent on URLs that point to the *physical* locations of documents. This is the source of so called "link rot", links that worked yesterday may result in a "404 Not Found"-message today. The underlying HTTP-protocol has made some basic allowances for this by providing marginally more informative status codes such as "301 Moved Permanently", "307 Temporary Redirect" and "410 Gone".[24] A more stable solution would be to introduce a level of indirection, i.e. the use of "logical", as opposed to "physical", identifiers – identifiers that would continue to identify the same document, even if that document was moved to another physical location. Several solutions along these lines have been proposed and implemented, e.g.:

- DOI – Digital Object Identifiers (see International DOI Foundation, 2004)
- URN – Uniform Resource Names (see Sollins and Masinter, 1994)

The OAI-PMH has learned a lesson from this situation, and provides for identifiers that are not directly related to physical locations, but instead uses a system of locally unique identifiers:

> A unique identifier unambiguously identifies an item within a repository; the unique identifier is used in OAI-PMH requests for extracting metadata from the item. Items may contain metadata in multiple formats. The unique identifier maps to the item, and all possible records available from a single item share the same unique identifier.
>
> The format of the unique identifier must correspond to that of the URI (Uniform Resource Identifier) syntax. [...] Repositories may implement the oai-identifier syntax described in the accompanying Implementation Guidelines document. (Lagoze et al (eds) 2002a, chapter 2.4)

[23] In early May 2005 a set of guidelines on "Conveying rights expressions about metadata in the OAI-PMH framework" (Lagoze et al (eds), 2005) was released.

[24] See <http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html> for a complete list of HTTP status codes, with explanations of their meanings.

The syntax of an identifier is formally defined in Lagoze et al (eds) (2002b, chapter 2.1) as:

```
oai-identifier = scheme ":" namespace-identifier ":" local-identifier
```

The “scheme” is always the literal string “oai”.

The “namespace-identifier” is usually related to the domain-name of the organization hosting the repository, in fact this is mandated by the standard:

> Organizations must choose namespace-identifier values which correspond to a domain-name that they have registered, and are committed to maintaining. [...] Domain name registration is used to avoid the need for any additional registration service for oai-identifiers. Domain name based identifiers guarantee global uniqueness without the need for OAI registration as required with the earlier, v1.0/1.1 specification. (Lagoze et al (eds) 2002b, chapter 2.2)

The page “Distinct Repository Identifiers” at the “Experimental OAI Registry at UIUC” indicates that not all repositories are complying with this demand: 43 repositories are listed with the repository identifier “GenericEPrints.OAI2” and 30 with “GenericEPrints”.[25] This is probably because users of the Eprints.org repository software have just accepted the default identifier, without customizing it to their own institution.

The “local-identifier” is some identifier which is unique in the context of this particular repository. An example of an identifier from ArXiv.org might look like this:

```
oai:arXiv.org:hep-th/9901001
```

Globally unique identifiers are useful in that they make it possible to trace a metadata record back to its source repository, and because they can be used to build rich and interlinked services based on metadata harvested from different repositories.

3.5 Six “verbs”

OAI-PMH requests made by harvesters to data providers can be any one of six types. These request-types are known as verbs:[26]

3.5.1 Identify

An Identify-request can be issued by a harvester in order to collect some basic information about the data provider, such as its name, what version of the OAI-PMH it supports, the e-

[25] <http://gita.grainger.uiuc.edu/registry/ListRepoIds.asp?self=1>

[26] The examples in this section are all taken from the D-LIST repository, located at <http://dlist.sir.arizona.edu/>. The actual XML responses have been obtained through the Repository Explorer, located at <http://re.cs.uct.ac.za/>. Some formatting has been applied to make the examples more readable. Responses to ListSets, ListIdentifiers and ListRecords have been abbreviated (using “[...]” to mark where deletions have been made) due to space constraints, and because these responses contain repeating patterns.

mail address of the administrator and any guidelines concerning the content and policies of the repository.

Example request:

```
http://dlist.sir.arizona.edu/perl/oai2?verb=Identify
```

Example response:

```
<?xml version="1.0" encoding="UTF-8" ?>

<OAI-PMH  xmlns="http://www.openarchives.org/OAI/2.0/"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
<responseDate >2005-04-15T10:17:52Z</responseDate>
<request  verb="Identify"
resumptionToken="">http://dlist.sir.arizona.edu/perl/oai2</request>
<Identify >
<repositoryName >DLIST</repositoryName>
<baseURL >http://dlist.sir.arizona.edu/perl/oai2</baseURL>
<protocolVersion >2.0</protocolVersion>
<adminEmail >mailto:paul@ahsl.arizona.edu</adminEmail>
<earliestDatestamp >0001-01-01</earliestDatestamp>
<deletedRecord >persistent</deletedRecord>
<granularity >YYYY-MM-DD</granularity>
<description >
<oai-identifier  xmlns="http://www.openarchives.org/OAI/2.0/oai-identifier"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai-identifier
http://www.openarchives.org/OAI/2.0/oai-identifier.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
<scheme >oai</scheme>
<repositoryIdentifier >DLIST.OAI2</repositoryIdentifier>
<delimiter >:</delimiter>
<sampleIdentifier >oai:DLIST.OAI2:23</sampleIdentifier>
</oai-identifier>
</description>
<description >
<eprints  xmlns="http://www.openarchives.org/OAI/1.1/eprints"
xsi:schemaLocation="http://www.openarchives.org/OAI/1.1/eprints
http://www.openarchives.org/OAI/1.1/eprints.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
<content >
<URL >http://dlist.sir.arizona.edu/information.html</URL></content>
<metadataPolicy ><URL >http://dlist.sir.arizona.edu/information.html
</URL></metadataPolicy>
<dataPolicy ><URL >http://dlist.sir.arizona.edu/information.html
</URL></dataPolicy>
<submissionPolicy ><URL
>http://dlist.sir.arizona.edu/information.html</URL></submissionPolicy>
        <comment >This system is running eprints server software (EPrints
2.2.1 (pepper)...</comment>
</eprints>
</description>
</Identify>
</OAI-PMH>
```

3.5.2 ListMetadataFormats

This verb is used to obtain a list of the metadata formats that the repository in question can disseminate. A harvester might be able to process some specialised metadata formats, but be forced to fall back on simple Dublin Core if the repository is unable to disseminate any of those formats.

Example request:

```
http://dlist.sir.arizona.edu/perl/oai2?verb=ListMetadataFormats
```

Example response:

```
<?xml version="1.0" encoding="UTF-8" ?>
<OAI-PMH  xmlns="http://www.openarchives.org/OAI/2.0/"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
<responseDate >2005-04-15T10:34:39Z</responseDate>
<request  verb="ListMetadataFormats"
resumptionToken="">http://dlist.sir.arizona.edu/perl/oai2</request>
<ListMetadataFormats >
 <metadataFormat >
  <metadataPrefix >oai_dc</metadataPrefix>
  <schema >http://www.openarchives.org/OAI/2.0/oai_dc.xsd</schema>
  <metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc/</metadataN
amespace>
</metadataFormat>
</ListMetadataFormats>
</OAI-PMH>
```

We see that this repository only supports one metadata format, oai_dc (bold text in the example above).

3.5.3 ListSets

The records in a repository can be divided into “sets” that can reflect some subject-based division or the structure of the parent organization, such as departments in a university. Each Record can belong to zero or more sets. By issuing the ListSets-verb a harvester can get a response that describes the structure of the sets in a repository. By passing the identifiers of sets along with the ListIdentifiers and ListRecords-verbs described below, a harvester can get back just those identifiers or records respectively, that belong to a given set.

Example request:

```
http://dlist.sir.arizona.edu/perl/oai2?verb=ListSets
```

Example response:

```
<?xml version="1.0" encoding="UTF-8" ?>

<OAI-PMH  xmlns="http://www.openarchives.org/OAI/2.0/"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <responseDate >2005-04-15T11:32:50Z</responseDate>
  <request  verb="ListSets"
resumptionToken="">http://dlist.sir.arizona.edu/perl/oai2</request>
  <ListSets >
  <set >
    <setSpec >7374617475733D707562</setSpec>
    <setName >Status = Published</setName></set>
  <set >
    <setSpec >7374617475733D756E707562</setSpec>
    <setName >Status = Unpublished</setName></set>
  <set >
    <setSpec >7374617475733D696E7072657373</setSpec>
    <setName >Status = In Press</setName></set>
[...]
</ListSets></OAI-PMH>
```

The list has been truncated to show just 3 sets. Each set is given a unique identifier in the form of a setSpec, and a human-readable setName.

3.5.4 ListIdentifiers

As well as specifying the verb, we have to include the metadataPrefix for the metadata format we are interested in. This has to be one of the formats described by the response to the ListMetadataFormats-verb.

Example request:

```
http://dlist.sir.arizona.edu/perl/oai2?verb=ListIdentifiers&metadataPrefix=
oai_dc
```

Example response:

```
<?xml version="1.0" encoding="UTF-8" ?>

<OAI-PMH  xmlns="http://www.openarchives.org/OAI/2.0/"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <responseDate >2005-04-15T11:15:34Z</responseDate>
  <request  verb="ListIdentifiers" metadataPrefix="oai_dc"
resumptionToken="">http://dlist.sir.arizona.edu/perl/oai2</request>
  <ListIdentifiers >
  <header >
      <identifier >oai:DLIST.OAI2:32</identifier>
      <datestamp >2002-07-17</datestamp>
      <setSpec >7374617475733D707562</setSpec>
      <setSpec >7375626A656374733D6C697365</setSpec></header>
  <header >
      <identifier >oai:DLIST.OAI2:33</identifier>
      <datestamp >2002-07-15</datestamp>
      <setSpec >7374617475733D707562</setSpec>
      <setSpec >7375626A656374733D646C6962</setSpec></header>
  <header >
      <identifier >oai:DLIST.OAI2:45</identifier>
      <datestamp >2002-07-15</datestamp>
      <setSpec >7374617475733D707562</setSpec>
      <setSpec >7375626A656374733D696E66737973</setSpec></header>
[...]
<resumptionToken >100/12503168</resumptionToken></ListIdentifiers></OAI-
PMH>
```

This example has been truncated to show just three identifier. Along with the identifiers themselves datestamps that show the last modification date for the record is shown, along with setSpecs that point to the sets returned by the ListSets-verb.

3.5.5 ListRecords

This verb can be used to retrieve all the records from a repository with a given metadataPrefix, but it is also possible to limit the response to records from a given set or records that have been added or updated within a given period of time.

Example of a simple request that will retrieve all records:

```
http://dlist.sir.arizona.edu/perl/oai2?verb=ListRecords&metadataPrefix=oai_
dc
```

Example response:

```
<?xml version="1.0" encoding="UTF-8" ?>
<?xml-stylesheet type='text/xsl' href='/oai2.xsl' ?>

<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2005-04-22T09:42:31Z</responseDate>
  <request verb="ListRecords" metadataPrefix="oai_dc"
resumptionToken="">http://genie.sir.arizona.edu/perl/oai2</request>
  <ListRecords>
  <record>[...]</record>
  <record>[...]</record>
[...]
  <resumptionToken>archive/100/13042613/oai_dc</resumptionToken>
  </ListRecords>
</OAI-PMH>
```

This example has been truncated to show the containers of just two records. See the example response to the GetRecord-verb below for an example of a complete record.

3.5.6 GetRecord

The GetRecord-verb requires a metadataPrefix and the locally unique identifier of the item that we want to retrieve metadata about.

This is an example request for oai_dc-metadata from the item with identifier “oai:DLIST.OAI2:32”:

```
http://dlist.sir.arizona.edu/perl/oai2?verb=GetRecord&metadataPrefix=oai_dc
&identifier=oai%3ADLIST.OAI2%3A32
```

Example response:

```
<?xml version="1.0" encoding="UTF-8" ?>
<?xml-stylesheet type='text/xsl' href='/oai2.xsl' ?>

<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2005-04-22T09:50:52Z</responseDate>
  <request verb="GetRecord" identifier="oai:DLIST.OAI2:32"
metadataPrefix="oai_dc"
resumptionToken="">http://genie.sir.arizona.edu/perl/oai2</request>
  <GetRecord>
  <record>
    <header>
      <identifier>oai:DLIST.OAI2:32</identifier>
      <datestamp>2002-07-17</datestamp>
      <setSpec>7374617475733D707562</setSpec>
      <setSpec>7375626A656374733D6C697365</setSpec></header>
    <metadata>
      <oai_dc:dc xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd"
xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:dc="http://purl.org/dc/elements/1.1/">
        <dc:title>Interdisciplinarity: The Road Ahead for Education in
Digital Libraries</dc:title>
        <dc:creator>Coleman, Anita</dc:creator>
        <dc:subject>Library and Information Science Education</dc:subject>
        <dc:description>This article reviews the state of education in
digital libraries and curriculum planning documents from professional
associations in two areas: Library and Information Science; and Computing.
It examines suggestions for integration and interdisciplinarity in
education for digital libraries curricula using definitions of a
discipline, interdisciplinarity, and the transdisciplinary structure of a
university in order to discover how such integration may be successfully
accomplished. A plan to use learning communities and develop an
interdisciplinary curriculum for Knowledge Organization is briefly
discussed.
</dc:description>
        <dc:date>2002-07-01</dc:date>
        <dc:type>Journal Article (On-line/Unpaginated)</dc:type>
        <dc:identifier>http://genie.sir.arizona.edu/archive/00000032/</dc:i
dentifier>
        <dc:format>html
http://genie.sir.arizona.edu/archive/00000032/02/07coleman.html</dc:format>
</oai_dc:dc></metadata></record></GetRecord></OAI-PMH>
```

3.6 Extending the OAI-PMH

The OAI-PMH in its current incarnation is very general, and it provides a very basic set of operations related to harvesting of metadata records. Several projects have been undertaken that extend the functionality of the OAI-PMH, by utilizing it for novel purposes, by adding new verbs, or by adding new arguments to the existing verbs. Some examples of these approaches:

- Van de Sompel, Young and Hickey (2003) describe novel uses for the OAI-PMH, e.g. as a framework for distributing a thesaurus as well as usage logs from a digital library
- The ODL framework described in Suleman (2002) extends the OAI-PMH by adding completely new verbs, as well as new arguments to existing verbs, in order to use the OAI-PMH as a basis for services such as peer review, annotations and recommendations. This set of extensions are sometimes referred to as XOAI-PMH.
- The Kepler framework (see e.g. Liu 2002, chapter 7) proposes an extension of the protocol to accommodate large numbers of small data providers.

The OAI-PMH has proven to be a viable tool for lightweight metadata harvesting. There is no reason why it should not be extended to facilitate more advanced services than those available today, but how this will be done is as yet unclear. We might get a lot of locally developed extensions that take the base protocol and “embrace and extend” it, such as we see in the ODL and Kepler projects, or we might see a modularization of the protocol, which provides mechanisms for extensions that still ensure interoperability at a basic level. A third possibility, which is much easier to accomplish than the other two, is to use the protocol in novel ways by putting new and innovative forms of metadata inside the protocol as it exists today. Examples of this can bee seen in the first example mentioned above.

Chapter 4: The diversity of existing OAI-PMH Service Providers

4.1 A survey of Service Providers

The OAI provides a list of "Registered Service Providers".[27] All the service providers listed on this page as of 2005-09-01 were visited during the first two weeks of September, along with the services reviewed by Brogan (2003) and McKiernan (2003a, 2003b, 2004). The features of each service was surveyed, based on an inspection of the sites themselves, as well as a perusal of any documentation about the service found at the site or in the literature.

4.2 Summary of features

The following table summarizes the features found in the surveyed service providers:

	Search	*Browse*	*Outgoing links*	*Equation search*	*Citation analysis*	*Annotations/comments*	*Collaborative filtering*	*"Shopping cart"*	*Search history*	*Alerting*	*Virtual collections*	*Rating*	*Reviewing*
AmericanSouth.org	✓	✓											
ARC	✓	✓											
Archon	✓		✓	✓		✓							
BASE	✓												
citebaseSearch	✓				✓								
DP9	✓		✓										
ePrints UK	✓												
ETD OAI	✓	✓											
Grainger	✓							✓	✓				
MeIND	✓	✓						✓	✓	✓			
METALIS	✓		✓										
NCSTRL	✓	✓											
OAIster	✓	✓											
OLAC	✓												
Perseus	✓	✓											
Public Knowledge Harvester	✓	✓	✓										
SAIL-Eprints	✓	✓											

[27] <http://www.openarchives.org/service/listproviders.html>

	Search	***Browse***	***Outgoing links***	***Equation search***	***Citation analysis***	***Annotations/comments***	***Collaborative filtering***	***“Shopping cart”***	***Search history***	***Alerting***	***Virtual collections***	***Rating***	***Reviewing***
Scirus	✓												
Sheet Music	✓	✓									✓		
SuUB Bremen	✓							✓					
TORII	✓	✓	✓			✓	?	✓		✓		✓	✓
UIUC Digital Gateway	✓	✓						✓	✓				

Table 4.1: Summary of features found in 22 service providers

4.2.1 Search

Some form of search functionality is implemented by all the surveyed service providers. Searching can be done across all the data harvested by the service, and in some service providers it can also be limited to metadata from a given data provider. Some services also allow searching to be limited by field, such as author or title.

4.2.2 Browse

Browsing is the act of “leafing through” records in some order defined by the system. Browsing by archive seems to be the most common function, but some services also allow browsing by certain metadata fields, such as author. One service, MeIND, allows browsing by the Dewey Decimal Classification system (DDC), but there is no information on the site as to how records are assigned to the classes of the DDC.

4.2.3 Outgoing links

Some of the services provide links to other services that can provide information relevant to a specific record, which is not available in the first service. These links can be of three different kinds:

- Links that conform to the OpenURL standard (Sompel and Beit-Arie 2001) can point to a resolver which provides further links to relevant services. These can be e.g. links to a search in a library catalogue or links to articles in publisher's databases or collections of electronic services. METALIS gives an example of this, where the service provides a default resolver, but users can also provide the URL of their own preferred resolver.
- Links that contain the OAI-PMH identifier can connect users with information relevant to a

particular record at other sites. TORII provides links based on identifiers to a service called iCite, but following these links results in a “Server not found” message.

- Other links may be based on elements of the harvested metadata, such as links to web search services that contain titles or names of authors.

4.2.4 Equation search

Archon, a service provider aimed at the physics community, is unique in enabling users to search for equations:

> In Archon, many metadata records contain equations in LaTeX and other formats. These equations are harvested as text format and not easy for users to browse and view. It is a value-added service to search equations by traditional text query but present it in a user-friendly way (e.g GIF file). By this method we build virtual metadata (images) over the original flat text metadata. (Maly et al 2002, p. 3)

In order to make the equations searchable, they are extracted from the source documents and stored in a database, along with a visual representation in the form of a GIF-file. When searching for equations the search has to be specified in the same format as the original document:

> To realize this function, LaTeX strings that are used to express equations are extracted from the metadata records. The extracted LaTeX strings are filtered and cleaned to eliminate errors and illegal symbols. Then the clean LaTeX strings are converted into GIF images.

Equations can also be browsed.[28]

4.2.5 Citation analysis

Citebase Search is the only surveyed service provider which provides citation analysis. This is done by harvesting records from data providers, locating the URLs of full text documents in the metadata, downloading those documents and identifying the bibliographies or lists of references that they contain. This information is then used to identify which documents are being cited by other documents, in order to build a service that can provide some of the same information that is available in the ISI Web of Science.[29] The service is, however, keen on stressing that it is by no means complete or ready to compete with the ISI:

> Citebase is currently only an experimental demonstration. Users are cautioned not to use it for academic evaluation yet. Citation coverage and analysis is incomplete

[28] <http://archon.cs.odu.edu:8066/archon/eqnresult.jsp?formname=subject&subjectList=ALL>

[29] <http://isiknowledge.com/wos>, only available to subscribers or users affiliated with subscribing institutions.

> and hit coverage and analysis is both incomplete and noisy.[30]

Another similar service is CiteSeer.IST[31] (Giles, Bollacker and Lawrence 1998). It should however be noted that this service does not rely on the OAI-PMH to locate full text documents, but relies on Web search services:

> Currently, CiteSeer uses Web search engines (e.g. AltaVista, HotBot, Excite) and heuristics to locate papers (e.g. CiteSeer can search for pages which contain the words "publications", "papers", "postscript", etc.). CiteSeer locates and downloads Postscript files identified by ".ps", ".ps.Z", or ".ps.gz" extensions.

The citation data gathered by CiteSeer.IST can be accessed through the OAI-PMH, so it acts as a data provider, but falls outside the scope of the present survey since it is not a service provider (i.e. it is not based on harvesting metadata from other data providers).

4.2.6 Annotations/comments

A couple of services allow users to attach arbitrary text to harvested records, in the form of annotations or comments. In Archon anyone can provide public comments by entering their name and e-mail address along with the comment. In TORII registered users can choose to make comments public (viewable by anyone) or private (viewable only by the user making the comment). Comments may be a way to facilitate discussions around records, or they may be used to direct fellow users to other records that are related to a given record, such as e.g. updated versions, critiques or reviews.

4.2.7 Collaborative filtering

Kohrs and Merialdo define collaborative filtering in the following terms:

> In collaborative filtering objects are selected for a particular user, which are relevant to similar users. Generally, in collaborative filtering the content of the objects is ignored and only other users' opinions on the considered objects are relevant. (2001, p. 696)

On its front page, TORII states that it is

> [...] a web environment that allows unified access to open archives of the scientific community and provides useful services on them like full-text search, cognitive and collaborative filtering, storing in a personal folder, and autonomous citation extraction.[32]

[30] <http://citebase.eprints.org/cgi-bin/search> Accessed on 2005-10-28.

[31] <http://citeseer.ist.psu.edu/>

[32] <http://torii.sissa.it/torii/index.jsp>

Further down on the same page it says that

> You will also have the possibility of defining profiles of interests; the system will then screen your daily browsing by ordering documents by their relevance according to your profiles.

This explanation does not agree with the definition of collaborative filtering above. There seems to be no evidence that recommendations in TORII is based on “other users' opinions”, and so it would seem that what this service provides is not actually collaborative filtering as defined by Kohrs and Merialdo (2001).

4.2.8 “Shopping cart”

“Shopping carts” originated in online shopping services, and make it possible for a user to collect together several items, before proceeding to buy them. In the context of OAI-PMH service providers, “shopping carts” provide a way to collect records during a session. These carts can be temporary, and expire when a session is terminated, or they can be persistent, so that they are still available when the user logs on at a later time. An example of this is found in MeIND, which allows registered users to create named carts. These carts can also be made publicly available.

What the contents of the cart can be used for will also vary between services. Some do not provide any special functions, so carts are only useful for collecting records during a session and then e.g. printing each one at the end of the session. Services that could be provided are functions for exporting references or e-mailing them to others.

4.2.9 Search history

Some services record the searches that a user makes during a session, so as to make it easier for the user to re-run a previous search. This is typically useful for users who try a large number of searches in order to assess what search strategies or criteria yield the most relevant or highest number of hits. A search history makes it easy to go back to the successful searches, as well as keep track of which approaches have already been tried.

4.2.10 Alerting

Keeping an eye on new documents that are available on e.g. journal homepages, commercial databases, and OAI-PMH service providers is an ever more daunting task. This is partially due to the inherent “pull” nature of the WWW, where users have to request information actively. This means that users have to visit all the sites that are relevant for that user regularly, to see what is new. Solutions to this provide “push” mechanisms, where information is pushed from services to users, and collected in one, more or less useful, place. Traditionally this has been done through e-mail, letting users sign up for newsletters or other dispatches of news. Some of the surveyed service providers, e.g. MeIND, have a functionality whereby users can specify a search, and then have new records that satisfy the criteria of that search e-mailed to them as they become available.

4.2.11 Virtual collections

The Sheet Music Consortium provides a feature called virtual collections. This allows end users to create collections of records, either as a set of recommendations, or in order to make explicit some useful grouping of records. These collections are given a name and a list of them are displayed on the site, along with the username of the "owner".[33]

4.2.12 Rating and reviewing

TORII is the only service which provides functionality that goes some way towards realizing the idea of overlays as evaluative tools. As stated above, registered users can rate and review records. It is, however, not possible to get an impression of to what extent these features are actually used.

4.2.13 Conclusion

Federated search and some sort of browsing based on the inherent structure of the harvested metadata seems to be a common denominator for the service providers surveyed. There are few signs of any human intervention in the presentation of metadata.

[33] <http://digital.library.ucla.edu/sheetmusic/librarian?GETVCLIST=0>

Chapter 5: The prototype – colLib.info

The purpose of the prototype is to explore the possibilities of constructing a service that acts as an overlay to documentation made available in subject, institutional or funder repositories, based on metadata gathered through the use of the OAI-PMH. The prototype should chart new land by implementing features not found in existing service providers. Based on the findings in the survey above, it is evident that the service implemented needs to go beyond federated searching, although some sort of search facility will probably be an basic feature for any service provider.

5.1 Web-specific forms as inspiration

When coming up with a set of interesting features for the prototype, I have looked to, and drawn inspiration from *tagging*, *wikis*, and *content syndication*. These are all documentary forms that are native to the Web, in the sense that their existence is reliant on the basic features of the Web, and could not be conceived outside the environment of the Web. The motivation for doing this has been the hope that this inheritance would be useful, and that users should respond positively to it.

5.1.1 Folksonomies/tagging/social bookmarks

On the Web there is a growing popularity of services based on metadata supplied by authors or end-users, in the form of so called “tags”. This can be seen in a number of popular services:

- Flickr allows its users to upload digital images to its server, and supply tags that describe the pictures. Images from different users that use the same tag are co-located on a page, to make browsing easy.[34]
- del.icio.us is a Web-based tool for creating and maintaining “bookmarks” (URLs to Web-pages and other Web-accessible documents).[35] The system makes it possible to see a list of bookmarks that have been marked with the same tag by one user or by all users, or to see all the tags that all the users have assigned to any given URL. (Golder and Huberman 2005)

An important difference between these two services is that while Flickr only allow authors (i.e. photographers) to add tags to items, de.licio.us allows anyone to do it, not just the author of a Web page. Hammond et al (2005) gives a general overview of several similar services.

“Tags” can be proper names or nouns, or any combination of signs that are meaningful to the person creating the tag, or to a closed circle of initiates.

There is no consensus on a name for this phenomenon of “organizing and grouping by tags”.

[34] <http://flickr.com/>
[35] <http://del.icio.us/>

One novel term that has been suggested is “folksonomy”, a combination of “folk” (because this is an activity carried out by “ordinary people”, not specially educated librarians or subject experts) and “taxonomy”. The use of this word has been criticised as confusing the rigorous tradition of taxonomies with the anarchic, free-for-all organization that it is trying to describe. “Social bookmarks” is a term that is also used (e.g. in Hammond et al 2005 and Lund et al 2005), especially for services like de.licio.us that actually deal with bookmarks and URLs, but this seems inappropriate for a service like Flickr, which deals with images. In the following I will use “tagging” to denote activities like those made possible by Flickr and del.icio.us.

One service that has been especially important as a source of inspiration for the prototype is CiteULike.[36] The site sums it self up in the following words:

> CiteULike is a free service to help academics to share, store, and organise the academic papers they are reading. When you see a paper on the web that interests you, you can click one button and have it added to your personal library. CiteULike automatically extracts the citation details, so there's no need to type them in yourself.[37]

The social nature of this service is also noted:

> Because your library is stored on the server, you can access it from any computer. You can share you library with others, and find out who is reading the same papers as you. In turn, this can help you discover literature which is relevant to your field but you may not have known about.

CiteULike uses a system of plug-ins to extract metadata from sites, but it is also possible to enter any web-based document into the system, by manually entering the appropriate metadata:

> Only links to the papers are stored, the papers themselves stay in archives like CiteSeer or PubMed. [...] The system currently supports: AIP Scitation, Amazon, American Geophysical Union, American Meteorological Society, Anthrosource, Association for Computing Machinery (ACM) portal, BMJ, CiteSeer, IEEE Xplore, IngentaConnect, JSTOR, MathSciNet, MetaPress, NASA Astrophysics Data System, Nature, PLoS Biology, PubMed, PubMed Central, Science, ScienceDirect, SpringerLink, Usenix, Wiley InterScience, arXiv.org e-Print archive, but you can post any other article from any non-supported site on the web

[36] <http://www.citeulike.org/>

[37] This and the following quotes describing CiteULike are taken from <http://www.citeulike.org/faq/all.adp>, accessed on 2005-09-04.

> - you'll just have to type the citation details in yourself.

CiteULike uses tags in much the same way as del.icio.us. When users add a reference to their "library" they can also associate tags to the reference. It is then possible to browse ones own collection of references with these tags, or one can browse all the references with the tags used by all users.

CiteULike also has an awareness-function which allows a user to "watch" both tags and other users, so that any references added to a watched tag or the "library" of a watched user is displayed on a single page.

The popularity of the concept of tagging in relation to scholarly documentation can be seen in the sheer number of services similar to CiteULike, e.g. Get cited[38] and Connotea[39] (described by Lund et al (2005)).

5.1.2 Wikis

Leuf and Cunningham give the following definition of a wiki:

> The WikiWikiWeb server concept, most often called simply " a wiki", originated with Ward Cunningham. A wiki is a freely expandable collection of interlinked Web "pages", a hypertext system for storing and modifying information – a database, where each page is easily editable by any user with a forms-capable Web browser client. (Leuf and Cunningham 2001, p. 14.)

They elaborate on this in the following manner:

> At the functional level, which is what the user sees, the essence of Wiki can be summarized by these statements.
>
> - A wiki invites all users to edit any page or to create new pages within the wiki Web site, using only a plain-vanilla Web browser without any extra add-ons.
> - Wiki promotes meaningful topic associations between different pages by making page link creation almost intuitively easy and by showing whether an intended target page exists or not.
> - A wiki is not a carefully crafted site for casual visitors. Instead, it seeks to involve the visitor in an ongoing process of creation and collaboration that constantly changes the Web site landscape. (Leuf and Cunningham 2001, p. 16)

For the purposes of the present project, the features of wikis that are most interesting are the

[38] <http://www.getcited.org/>

[39] <http://www.connotea.org/>

following:

Wikis are editable by anyone, through a web-browser. Although this was the starting-point of the wiki movement, it is also possible to have a wiki where there is a limited group of editors that have permission to write to the wiki, while ordinary users only have permission to read. Wikis may also be used on intranets, where all users have write-permissions, but the group of users is limited, for example by institutional affiliation. Write-permissions can also be limited on a per-page basis in an otherwise open wiki, so that e.g. pages that are deemed of special importance are not writeable by all users.

Wikis are written in a simple markup-language (which is converted to HTML by the wiki software when a page is requested for ordinary viewing). Since rich interlinking is a feature of wikis, the creation of links are made especially easy. In the earliest wiki implementations a link to another page in the same wiki was simply denoted by giving the name of the other page in so called “CamelCase”, that is a word with internal capital letters. There has been some dissatisfaction with this linking mode, and several wiki engines now use special characters to denote a link, such as this: [[name of page]]. If such a link is created that does not point to an existing page, the resulting link does not point to an empty page or an error-message, but rather to a form which allows for the easy creation of a new page. This makes the creation of new pages very easy.

Probably the best known example of a wiki today is Wikipedia (see McKiernan 2005 for an introduction), a freely available encyclopaedia written by volunteers.[40] There is quite a lot of controversy about the quality and trustworthiness of this resource, a discussion of which is outside the scope of the present work. It should however be noted that Wikipedia has attracted a lot of users, and these users have demonstrated an impressive willingness to invest time and energy in the development of Wikipedia. As of 2005-09-04 the Wikipedia's own statistics-page reports that the English version alone has 416,325 registered users, 715,852 articles and that 22,844,518 edits have been made since July 2002.[41] Based on this, and the fact that there are a lot of other software-packages that implement wiki features, and a lot of projects that use this software to create websites, it should be safe to say that the wiki form has been successful in attracting users.

5.1.3 Content syndication, RSS/Atom

An XML-based document format called RSS was launched by Netscape in 1999, as a convenient way for a website to export it's headlines in a machine readable way, so that

[40] <http://en.wikipedia.org/wiki/Main_Page>

[41] <http://en.wikipedia.org/wiki/Special:Statistics>

services could be created that aggregate these “feeds” and present headlines from different sources in a convenient way (see Hammersley, 2003).

CiteULike uses RSS extensively for its awareness-services. It is possible to subscribe to RSS feeds for tags and users, thus making it easy to discover new references that have been added to a tag or by a user that someone deems interesting.

5.2 Features of the prototype

The basic feature of the prototype is the connection between an OAI-PMH service provider and a wiki. This is realized by connecting one page in the wiki to each of the harvested records. Records are displayed in the standard manner of the service provider, but on the same page as the metadata, the contents of the corresponding wiki page are also displayed. This opens up for at least two applications:

- Browsing will be made possible through inclusion of “tags” (wiki-links) on a page, and all the records that have been associated with a given tag will be displayed on a special page for that tag.
- Any information can be added to the wiki-page connected with a record, so this can be used for all kinds of annotations, e.g. links to newer versions of articles or information about related records.

RSS feeds are provided for the newest records as well as separate feeds for each of the “tags” that are created.

5.3 Implementing the prototype

When choosing the tools with which to implement the prototype, I had the following premises in mind:

- The learning-curve should be as small as possible, since the limited amount of time available for this project should preferably be devoted to actual development, not to learning to use new tools.
- A search should be made for existing software packages that could be included in and/or modified for the purposes of the present project. The source code of any such programs should preferably be available under the GNU Public License (GPL)[42] or a similar licence approved by the Open Source Initiative (OSI).[43]

[42] <http://www.gnu.org/copyleft/gpl.html>

[43] <http://www.opensource.org/>

5.3.1 Selecting the platform: “LAMP”

“LAMP” is used as an acronym for a combination of tools that is often used in Open Source, Web-related projects. The acronym is derived from the first letters of Linux[44], Apache[45], MySQL[46] and Perl[47], PHP[48] or Python.[49] All of these components are Open Source and available free of charge.

A Web server can bee seen both as a machine that serves Web-pages, and as the specialised program that takes care of the actual serving, often known as a HTTP daemon. In the context of LAMP, Linux is used as the operating system on the server, and Apache is the actual Web server program.

To handle the metadata that the prototype needs to store and make searchable, a relational database is often the best solution. MySQL is a popular database management system. Although it does not have all the advanced features of enterprise solutions like Oracle or IBM's DB2, it is often considered well suited to Web based projects.

There are of course a large number of programming languages available that are suited to Web based projects. The present author has some knowledge of both Perl and PHP, and these are generally well suited to Web projects. A lot of existing projects are implemented in one of these languages, and this comes in handy when one is looking for source code that can be modified for a new project.

5.3.2 OAI-PMH tools considered

The list of “OAI Tools” provided by the OAI was taken as a starting point in the search for a suitable metadata harvesting system.[50] Some of the tools that satisfied the criteria given above were installed and tested on a development server. Two different approaches to harvesting were considered before the tools were selected:

“Just in case” harvesting is the normal mode of harvesting within the OAI-PMH framework. Harvesters collect all new and updated records (optionally limited to one or more sets) from repositories on a regular basis, and make them available to users.

The main benefits of this approach is that the reliability of the service provider is minimally affected by the availability of the data providers. If a harvest is attempted, but fails to be completed for some reason, the harvester can return to the repository at a later time and make

[44] <http://www.linux.org/>
[45] <http://www.apache.org/>
[46] <http://www.mysql.com/>
[47] <http://www.perl.org/>
[48] <http://www.php.net/>
[49] <http://www.python.org/>
[50] <http://www.openarchives.org/tools/tools.html>

a new attempt at harvesting the new and updated records.

If one is going to build a subject-specific service provider, it can be difficult to decide a priori which data providers are relevant to the intended audience, especially if the subject is one with a lot of related subjects that might be relevant. One solution to this might be to harvest all available data providers and extract only those records that are relevant, but this quickly becomes a daunting task when the total number of records are in the millions. This would also create unnecessary overhead in terms of bandwidth usage and load on data providers.

To avoid the problems associated with just-in-case harvesting, one could envision a system based on “just-in-time” harvesting, where the discovery of relevant metadata records is done through services external to the service providers. These might be the sites of data providers themselves, or other service providers such as Arc or METALIS. A mechanism would be needed for the new service provider to be told the identifier of the record that someone deemed relevant, this could then be resolved to the base URL of the originating data provider, so that a request for the metadata of the record could be made through the OAI-PMH, and the record added to the database of the service provider.

The advantage of this approach would be that metadata could be harvested from any repository, without having to harvest the whole of that repository. This would save bandwidth and load both for the data and service provider. On the other hand, if a repository is temporarily down when a request for metadata is being made to it, users will experience this as problem of the service provider, and the perceived quality of the service will suffer.

Some effort was expended in creating a just-in-time harvester based on the Net::OAI::Harvester mentioned below, but it quickly became apparent that the task of obtaining OAI identifiers from external services and resolving them to the base URLs of originating repositories was too cumbersome, so it was decided to base the prototype on a regular just-in-case harvester.

The experiences from testing four of the available systems are recounted below:

5.3.2.1 ODL framework

- Web site: <http://oai.dlib.vt.edu/odl/>
- Programming language: Perl
- Database: MySQL
- Licence: Unknown

This is a set of modules developed as a part of the *Open Digital Libraries* project at the

Digital Library Research Laboratory of Virginia Polytechnic Institute and State University, described in Suleman (2002). The modules each implement some functionality often found in digital libraries, such as searching, browsing, annotations and recommendations. The system uses the OAI-PMH to harvest metadata from data providers, but it also uses the same protocol (with some custom extensions) for communication between the modules.

The ODL framework was rejected because the separation of the components seems to introduce some unnecessary, and potentially troublesome, duplication of the harvested metadata.

5.3.2.2 Celestial

- Web site: <http://celestial.eprints.org/> and <http://oai-perl.sourceforge.net/>
- Programming language: Perl
- Database: MySQL
- Licence: GPL (according to the file “README.TXT”, which is part of the downloadable source code)

Celestial is developed as a cache for metadata:

> Celestial is software that harvests metadata from OAI-compliant repositories and re-exposes that metadata to other services - in effect an OAI cache.[51]

This can be useful in a couple of different ways. The metadata can be normalised in some way, so that other harvesters will not have to face issues such as malformed XML or faulty character encodings. Celestial could also be used as an intermediary for small repositories such as those envisioned by the Kepler project (described in chapter 2.4.4, p. 16). In this context a service based on Celestial could reduce the problem of unreliable repositories, as well as give other service providers the benefit of only having to harvest one data provider. Because Celestial is a fully fledged OAI-PMH metadata harvester, as well as a service provider, it could of course be used as the basis for a regular overlay service. The features that make it a data provider as well could then be turned off, or seen as a handy complement to the service provider features, for making any local enhancements to the metadata available to other service providers.

Celestial was rejected as a basis for the current project because it would never report a successful harvest, even when all the records from a repository were in fact harvested.

[51] <http://celestial.eprints.org/>

5.3.2.3 Net::OAI::Harvester

- Web site: <http://search.cpan.org/dist/OAI-Harvester/>
- Programming language: Perl
- Database: None
- Licence: “may be distributed under the same terms as Perl itself”,[52] i.e. under the terms of the GPL or the “Artistic Licence”.[53]

As explained in Summers (2004), Net::OAI::Harvester is a Perl module designed to make it easy to build OAI-PMH compliant harvesters. The module provides methods for harvesting metadata, and extracting the actual data from the XML formatted records, but it does not provide any facilities for storing or searching the metadata.

Net::OAI::Harvester was rejected as a basis for the current project, because building all the features needed to complement the actual harvester would be too time-consuming.

5.3.2.4 PKP OAI Harvester

- Web site: <http://pkp.sfu.ca/pkp-harvester/>
- Programming language: PHP
- Database: MySQL
- Licence: GPL

The PKP OAI Harvester is a fully fledged service provider, which is easy to install and has a Web-based administration interface for adding to and maintaining the list of harvested data providers. It also has an end-user interface which includes searching and browsing. A working demo with 176.809 metadata records harvested from 182 data providers (as of 2005-09-28) is available at the PKP site.[54]

The PKP OAI Harvester was chosen to serve as the basis for the prototype, on the grounds that it provides a basic, working and modifiable service provider.

5.3.3 Selecting the Wiki

When it was decided that the PKP OAI Harvester should be used as the basis for the prototype, it was time to select a Wiki for it to interact with. Since the harvester is written in PHP, it would make sense to find a Wiki that is implemented in the same language, to make the integration of the two easier.

[52] <http://search.cpan.org/src/ESUMMERS/OAI-Harvester-1.0/README.txt>
[53] <http://dev.perl.org/licenses/artistic.html>
[54] <http://pkp.sfu.ca/harvester/>

The first idea pursued was to find a complete content management system (CMS) that also included a Wiki, so that the CMS could take care of issues such as user registration and authentication, so that this feature did not have to be implemented from scratch. Two open source CMS-solutions were tested:

- PostNuke has a Wiki module, but the overall architecture of the framework was found to be too complicated to achieve easy integration with the harvester.[55]
- Mambo did not have a wiki-module at the time the tests were conducted, although several have since emerged.[56]

After these false starts a search was made for a pure Wiki tool that could also keep track of users. MediaWiki, the software that Wikipedia runs on, was tested and found to be adequate for the task at hand. It has a quite advanced wiki syntax, with user registration and a mechanism of so called "special pages" which facilitates extending the system with custom functionality. MediaWiki has also got the concept of a "watchlist", a feature that makes it easy for users to remember and keep track of changes to pages in the wiki that they find particularly interesting.

5.3.4 Installing the tools

The tools mentioned above were tested on a non-public development website. When the PKP OAI Harvester and the MediaWiki had been decided upon as the main building blocks, a new server-space, with the domain collib.info,[57] was rented from a commercial Web space provider. The server runs Apache on Linux, provides a connection to a MySQL database and gives access to the PHP programming language. MediaWiki and the harvester were then installed, and the metadata harvested during testing was moved to the new server.

5.3.5 Layout of the prototype service

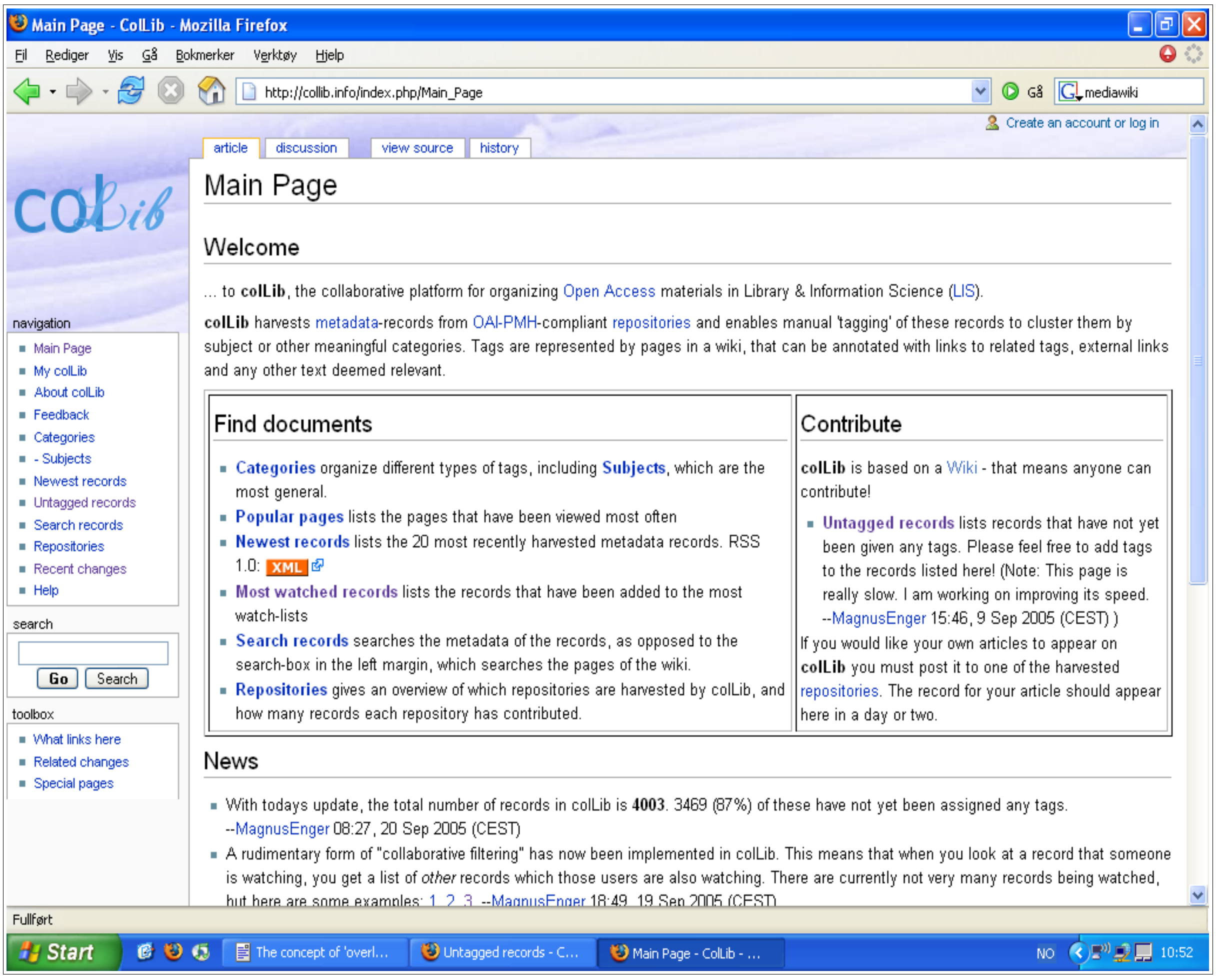
5.3.5.1 Design

The main layout of the prototype follows the "monobook" skin of the MediaWiki software. This is the same layout found on Wikipedia at the time of writing. The logo in the upper left corner was change to reflect the name of the prototype, and the colours of the page background was slightly altered to set it somewhat apart from Wikipedia. The layout was otherwise retained, so that any users familiar with Wikipedia should feel "at home". Below is a screen shot showing the front page of the prototype as of 2005-09-27:

[55] <http://www.postnuke.com/>

[56] <http://www.mamboserver.com/>

[57] <http://collib.info/>. The name was chosen to be a combination of the words "collaboration" and "library".



Main Page

Welcome

... to **colLib**, the collaborative platform for organizing Open Access materials in Library & Information Science (LIS).

colLib harvests metadata-records from OAI-PMH-compliant repositories and enables manual 'tagging' of these records to cluster them by subject or other meaningful categories. Tags are represented by pages in a wiki, that can be annotated with links to related tags, external links and any other text deemed relevant.

Find documents

- **Categories** organize different types of tags, including **Subjects**, which are the most general.
- **Popular pages** lists the pages that have been viewed most often
- **Newest records** lists the 20 most recently harvested metadata records. RSS 1.0: XML
- **Most watched records** lists the records that have been added to the most watch-lists
- **Search records** searches the metadata of the records, as opposed to the search-box in the left margin, which searches the pages of the wiki.
- **Repositories** gives an overview of which repositories are harvested by colLib, and how many records each repository has contributed.

Contribute

colLib is based on a Wiki - that means anyone can contribute!

- **Untagged records** lists records that have not yet been given any tags. Please feel free to add tags to the records listed here! (Note: This page is really slow. I am working on improving its speed. --MagnusEnger 15:46, 9 Sep 2005 (CEST))

If you would like your own articles to appear on **colLib** you must post it to one of the harvested repositories. The record for your article should appear here in a day or two.

News

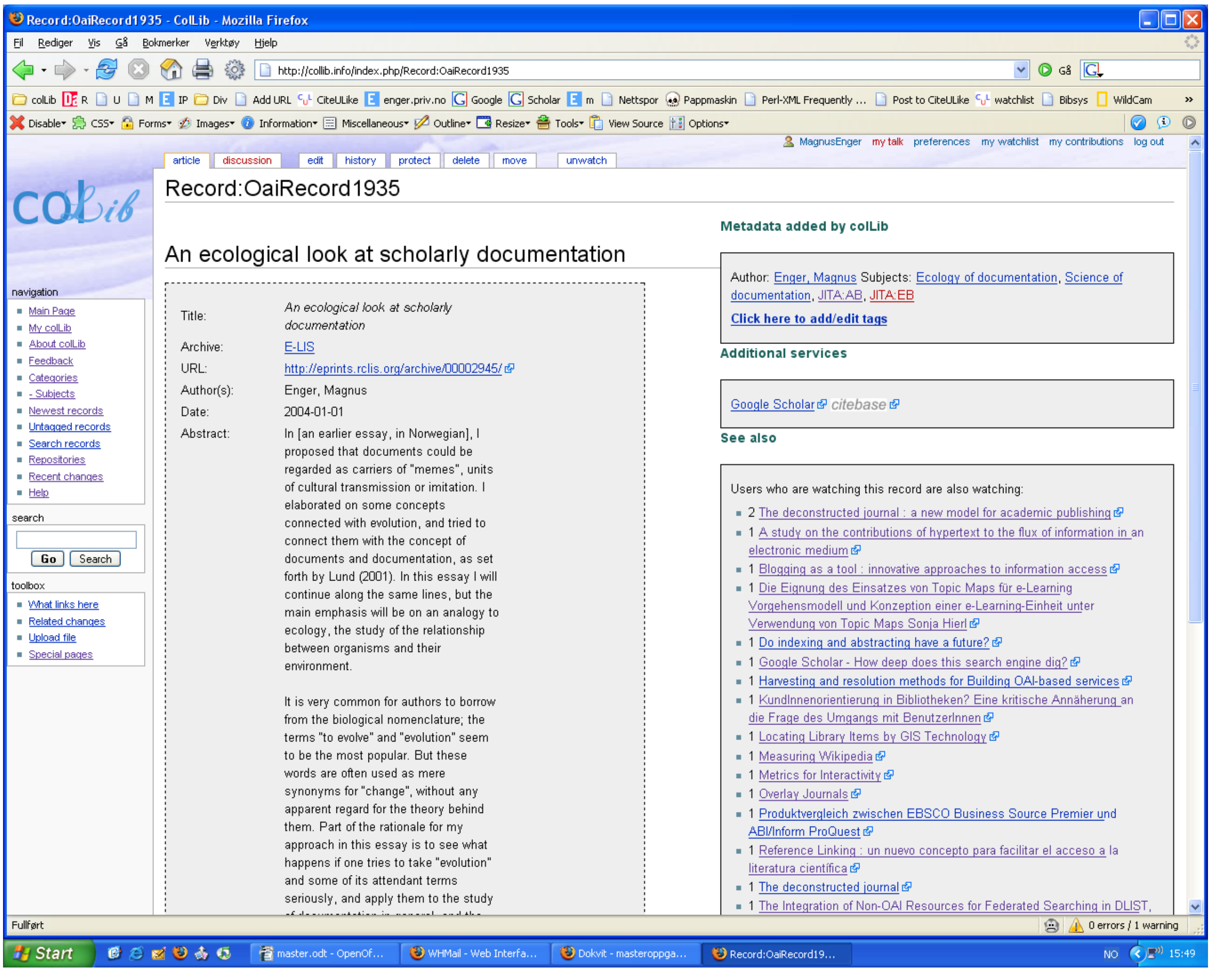
- With todays update, the total number of records in colLib is **4003**. 3469 (87%) of these have not yet been assigned any tags. --MagnusEnger 08:27, 20 Sep 2005 (CEST)
- A rudimentary form of "collaborative filtering" has now been implemented in colLib. This means that when you look at a record that someone is watching, you get a list of *other* records which those users are also watching. There are currently not very many records being watched, but here are some examples: 1, 2, 3. --MagnusEnger 18:49, 19 Sep 2005 (CEST)

Illustration 5.1: Front page of prototype as seen in the browser Firefox 1.0.1 on 2005-09-27.

5.3.5.2 Display of records

The harvested metadata of a record is displayed on a page with a URL like e.g. <http://collib.info/index.php/Record:OaiRecord4014>. The number at the end corresponds to the running numbers that the harvester assigns to records as they are harvested. This looks like the normal URL for a page in MediaWiki. The source code of the system has, however, been altered so that pages with a name containing the literal string "Record:OaiRecord" are treated in a special way. The information on the page is divided into three or four parts, depending on whether anyone has added the page in question to their watchlist or not:

The four sections are as follows:



Record:OaiRecord1935

An ecological look at scholarly documentation

Title:	*An ecological look at scholarly documentation*
Archive:	E-LIS
URL:	http://eprints.rclis.org/archive/00002945/
Author(s):	Enger, Magnus
Date:	2004-01-01
Abstract:	In [an earlier essay, in Norwegian], I proposed that documents could be regarded as carriers of "memes", units of cultural transmission or imitation. I elaborated on some concepts connected with evolution, and tried to connect them with the concept of documents and documentation, as set forth by Lund (2001). In this essay I will continue along the same lines, but the main emphasis will be on an analogy to ecology, the study of the relationship between organisms and their environment. It is very common for authors to borrow from the biological nomenclature; the terms "to evolve" and "evolution" seem to be the most popular. But these words are often used as mere synonyms for "change", without any apparent regard for the theory behind them. Part of the rationale for my approach in this essay is to see what happens if one tries to take "evolution" and some of its attendant terms seriously, and apply them to the study

Metadata added by colLib

Author: Enger, Magnus Subjects: Ecology of documentation, Science of documentation, JITA:AB, JITA:EB

Click here to add/edit tags

Additional services

Google Scholar citebase

See also

Users who are watching this record are also watching:

- 2 The deconstructed journal : a new model for academic publishing
- 1 A study on the contributions of hypertext to the flux of information in an electronic medium
- 1 Blogging as a tool : innovative approaches to information access
- 1 Die Eignung des Einsatzes von Topic Maps für e-Learning Vorgehensmodell und Konzeption einer e-Learning-Einheit unter Verwendung von Topic Maps Sonja Hierl
- 1 Do indexing and abstracting have a future?
- 1 Google Scholar - How deep does this search engine dig?
- 1 Harvesting and resolution methods for Building OAI-based services
- 1 KundInnenorientierung in Bibliotheken? Eine kritische Annäherung an die Frage des Umgangs mit BenutzerInnen
- 1 Locating Library Items by GIS Technology
- 1 Measuring Wikipedia
- 1 Metrics for Interactivity
- 1 Overlay Journals
- 1 Produktvergleich zwischen EBSCO Business Source Premier und ABI/Inform ProQuest
- 1 Reference Linking : un nuevo concepto para facilitar el acceso a la literatura científica
- 1 The deconstructed journal
- 1 The Integration of Non-OAI Resources for Federated Searching in DLIST,

Illustration 5.2: The display of harvested and added metadata.
<*http://collib.info/index.php/Record:OaiRecord1935*>

Original record details

These are the actual metadata that were harvested via the OAI-PMH. For the sake of making programming the page easier, the metadata are fetched through an HTTP request to to the harvester, and then incorporated into the page.

Metadata added by colLib

This is the actual content of the page in the Wiki that has the name Record:OaiRecord2168. This information can be edited by registered users by clicking on the link that reads “Click here to add/edit information”.

Additional services

These are links to other services that may be relevant to the record that is being displayed.

A link to a search for the title of the record in Google Scholar is always provided.[58] This can help locate alternative locations for the resource the metadata is about, and it also gives access to the citation data and links that might be available in Google Scholar.

[58] <http://scholar.google.com/>

The records from the E-LIS data provider are indexed and citations analysed in Citebase, so for records that are from E-LIS a link to this information in Citebase is provided, incorporating the OAI identifier of the record in question (e.g. <http://www.citebase.org/cgi-bin/citations?id=oai:eprints.rclis.org:3351>).[59]

See also

If the record being displayed is on the watch-list of one or more users, the other records that these users are also watching are displayed here, on the premise that if you find this record interesting, you might benefit from seeing what others who also find it interesting are watching. This is a simple form of collaborative filtering.

If a record is not being watched by anyone, this section is not displayed at all.

5.3.5.3 Special pages

The MediaWiki software makes it possible to implement custom functionality through so called "special pages". This facility was used to integrate the functionality from the PKP OAI harvester into the prototype. The following special pages were created:

- Newest records <http://collib.info/index.php/Special:OaiRecent>
  Displays the 20 most recently harvested records, newest records at the top. This list is also available as an RSS feed: <http://collib.info/xml.php>.

- Untagged records <http://collib.info/index.php/Special:OaiUncollib>
  This page displays 10 random records that have not yet been given any tags. A new selection is displayed each time the page is visited or reloaded.

- Search records <http://collib.info/index.php/Special:OaiSearch>
  This page allows the harvested metadata to be searched, as opposed to the search box in the left hand menu, which only searches the contents of the pages in the Wiki.

- Repositories <http://collib.info/index.php/Special:OaiRepos>
  Displays a list of the repositories that are being harvested by the prototype, along with the number of records for each repository. Clicking on the name of a repository displays the information about the repository that was obtained through the Identify-verb (see chapter 3.5.1, p. 27).

- My colLib <http://collib.info/index.php/Special:MyCollib>
  This page displays the pages that have been added to a users watchlist. These are grouped into "Categories", "Tags" and "Records". For records, the title and first author are

[59] <http://citebase.eprints.org/>

displayed.

- Most watched records <http://collib.info/index.php/Special:OaiMostWatched>
  Displays the title and first author of those records that are watched by the largest number of registered users. The list is sorted by how many are watching each record, in descending order.

- OAI Statistics <http://collib.info/index.php/Special:OaiStats>
  Displays some statistics about the harvested records. As of 2005-09-27 the following information is displayed:

  - Total number of records

  - The number and percentage of records that have not had corresponding pages created in the wiki yet.

  - The distribution of language-codes in the harvested metadata, sorted by frequency and displayed in descending order.

5.3.5.4 Extensions

Tags are represented by pages in the Wiki. In order to display the records that are connected with a tag on the page for that tag, some custom mechanism is needed. This could have been accomplished by altering the source code of the system, in a similar way to how it was done for the display of records. This would have made creating pages for new tags easier, but it would have involved some compromise of flexibility in how records are presented. It would also still mean that some content would have to be entered on the page, otherwise it would simply not exist. This would be counter-intuitive, in that a page for a tag should be able to display just the records, without any further content.

The mechanism that was eventually chosen was to write an "extension".[60] An extension in the context of the MediaWiki system works in the following way: A pair of opening and closing tags surrounding some string of characters are entered on a page, e.g.:

```
<xyz> abc </xyz>
```

The string that is enclosed in the tags, in this case "abc", is then passed as an argument to a custom script, xyz. The script can then process that argument in any way it wants, and pass some output back. This output will then replace the tags and the argument in the displayed page.

In order to display records associated with a tag, an extension called "oairecords" was created, so that including the following in a page:

[60] <http://meta.wikimedia.org/wiki/Write_your_own_MediaWiki_extension>

```
<oairecords>Some page</oairecords>
```

results in a search being made through all the pages associated with records (i.e. with a name that begins with "Records:OaiRecord"), and a list returned of those that contain the string "Some page" enclosed in double brackets (i.e. the search is in fact looking for the literal string "[[Some page]]").

5.4 Data providers harvested

There is no standard procedure for discovering data providers that are relevant to any given subject or scholarly field. One way to identify repositories that may be relevant is to browse the (very long and not very informative) list of "Registered Data Providers", provided by the OAI.[61] Another is to have a look at which repositories general service providers or service providers in related fields are harvesting metadata from, since this information is usually available at the site of the services (cf. the field "Sources" in the survey of service providers). If the creators of the service provider are practitioners in the field that the service will cover, they are likely to become aware of interesting data providers through e.g. e-mail lists, presentations at conferences and so on. A last possibility is that maintainers of service providers are contacted directly by maintainers or advocates of data providers, with suggestions for inclusion of their repository in the service.

After the prototype under discussion here was launched, the present author was contacted by three people with suggestions for data providers that might be harvested by the prototype. One of these turned out not to be relevant, and the other two had technical problems that made harvesting impossible.

"Library and information science" was chosen as the field of the prototype, because this is a field the author is somewhat familiar with, and will be able to "tag" in a meaningful way. Based on this choice, and the "methods" outlined above, the following data providers were harvested by the prototype service provider (unless otherwise noted, the descriptions in italics are taken from the homepage of the service):

5.4.1 ALIA e-prints

URL	http://e-prints.alia.org.au/
Base URL	http://e-prints.alia.org.au/perl/oai2
Description	*This seems to be an installation of the Eprints software which has not been sufficiently customised, so the only information displayed is the default information provided by the software.*
Metadata	oai_dc

[61] <http://www.openarchives.org/Register/BrowseSites>

formats[62]	
Sets	See Appendix A

Table 5.1: Details for ALIA e-prints

5.4.2 @rchiveSIC - Sciences de l'Information et de la Communication

URL	http://archivesic.ccsd.cnrs.fr/
Base URL	http://archivesic.ccsd.cnrs.fr/perl/oai20
Description	@rchiveSIC is an author self-archiving server for articles and working papers in the field of the Communication and Information Sciences (SIC). The majority of the articles are in french.
Metadata formats	ccsd_tel, oai_dc
Sets	See Appendix A

Table 5.2: Details for ArchiveSIC

5.4.3 CaltechLib - Caltech Library System Papers and Publications

URL	http://caltechlib.library.caltech.edu/
Base URL	http://caltechlib.library.caltech.edu/perl/oai2
Description	Caltech Library System Papers and Publications is an archive of the papers and publications of the professional staff at the Caltech Library System. We have organized the papers according to the JITA classification scheme developed by the LIS service provider Research in Computing, Library and Information Science (RCLIS).
Metadata formats	oai_dc
Sets	See Appendix A

Table 5.3: Details for CaltechLib

5.4.4 DLIST

URL	http://dlist.sir.arizona.edu/
Base URL	http://dlist.sir.arizona.edu/perl/oai2
Description	DLIST is the Digital Library of Information Science and Technology, an Open Access, cross-institutional repository of full-text electronic resources in the domains of Library and Information Science (LIS) and Information Technology (IT).

[62] The available metadata formats were determined by issuing the ListMetadataFormats-verb to the data providers through a custom Perl-script based on the Net::OAI::Harvester module described in chapter 5.3.2.3 (p. 48).

Metadata formats	oai_dc
Sets	See Appendix A

Table 5.4: Details for DLIST

5.4.5 E-LIS

URL	http://eprints.rclis.org/
Base URL	http://eprints.rclis.org/perl/oai2
Description	E-LIS is an open access archive for scientific or technical documents, published or unpublished, on Librarianship, Information Science and Technology, and related areas. [...] We serve LIS researchers by facilitating their self-archiving, ensuring the long-term preservation of their documents and by providing word-wide easy access to their papers.
Literature	Kumar and Kalyane (2004), Medeiros (2004), De Robbio and Coll (2005)
Metadata formats	cnr_eprints, oai_dc
Sets	Based on the JITA scheme[63], see Appendix A for details.

Table 5.5: Details for E-LIS

5.5 Marketing

To attract users to the prototype in order to gain some feedback and be able to observe how it might be utilized by actual users, some sort of marketing was necessary. This was accomplished in two stages. The URL of the prototype was initially announced on the private web page of the author, before all the features were in place. This led to its being mentioned in several places, and being indexed by Google.[64]

On 2005-09-08, information about the prototype was posted to the following e-mail lists:

- <biblioteknorge@nb.no>
- <oss4lib-discuss>
- <boai-forum@ecs.soton.ac.uk>
- <SCHOLCOMM@ala.org>
- <web4lib@webjunction.org>
- <BIBLIST@SEGATE.SUNET.SE>
- <diglib@infoserv.inist.fr>

[63] <http://eprints.rclis.org/jita.html>. Accessed 2005-11-14.

[64] <http://www.google.com/>

- <oai-eprints@lists.openlib.org>
- <ifla-l@infoserv.inist.fr>
- <oss4lib-discuss@lists.sourceforge.net>

Recipients on these lists have reported that they have forwarded the messages to Spanish, Catalan, and Latin-American mailing lists as well.

Information about the service was included on the OAI list of service providers from 2005-09-09.[65]

In the middle of November, a short description of the prototype was published in *D-Lib Magazine* (Enger 2005).

5.6 Experiences from running the prototype

5.6.1 Harvesting

The PKP OAI Harvester comes with a command-line interface for running the harvester. This can be run interactively by logging in to the server that the harvester runs on and running the script from the command-line. The script can also be scheduled to run at set intervals, through e.g. the "cron" service available on Linux-based servers. During the time that the prototype has been running, the harvester has been run interactively to monitor directly (rather than through a log file) any problems associated with the harvesting.

Very few problems have occurred. The only incidents have been a temporary failure to harvest from one repository, but re-running the harvester immediately has returned no errors. This would not have been a problem if the harvester was running on a schedule, since the potentially new records that were not harvested would be harvested the next time the harvester was run.

5.6.2 The harvested metadata

There have been some cases of illegible records due to the use of non-western character sets, but some of these are also badly rendered in the web-pages of the data-provider, so this is not necessarily a problem of the software behind the prototype. Records that are unreadable have been marked with the tag "needsfixing", as suggested by one of the users of the prototype.[66]

5.6.2.1 Number of records per data provider

A total number of 4109 records had been harvested by 2005-10-30. The distribution of records among the data providers was as follows:

[65] <http://www.openarchives.org/service/listproviders.html>

[66] <http://collib.info/index.php/Needsfixing>

Repository	*Records*
E-LIS	2959
ArchiveSIC	613
DLIST	480
CaltechLib	35
ALIA	22
Sum:	4109

Table 5.6: Number of records harvested from repositories

5.6.2.2 Document types

As of 2005-11-08 the following document types were present in the database, according to the descriptions in the harvested metadata.

Description	*Number*
Journal Article (Print/Paginated)	933
Journal Article (On-line/Unpaginated)	860
Conference Paper	655
Text	613
Presentation	338
Journal Article (Paginated)	153
Preprint	110
Report	82
Thesis	70
Other	69
Conference Proceedings	48
Book Chapter	44
Technical Report	38
Conference Poster	31
Monograph	24
Book	21
Newspaper/Magazine Article	18
Guide	13
Tutorial	11
Project/Business Plan	8
Bibliography	8
Guide/Manual	8
Library Instructional Material	7
Conference or Workshop Item	7

Description	*Number*
Journal (On-line/Unpaginated)	6
Journal (Paginated)	4
Dataset	2
Article	2
Departmental Technical Report	1
Departmental Report	1
Interactive Material	1
In Collection	1

Table 5.7: Number of distinct document types as described in the harvested metadata.

5.6.2.3 Language of harvested records

The page <http://collib.info/index.php/Special:OaiStats> gives details about the language codes that are used in the harvested metadata. As of 2005-11-15, the number of records with a two-letter language-code were 3509 (84%). The distribution of records among language codes is as follows:[67]

Language code	*Language*	*Number of records*
ES	Spanish (Catalan)	1200
EN	English	821
FR	French	595
IT	Italian	416
CA	Catalan	112
SR	Serbian	98
DE	German	79
PT	Portuguese	65
HR	Croatian	63
ZH	Chinese	29
TR	Turkish	6
EL	Greek	5
ID	Indonesian	5
RU	Russian	5
PL	Polish	4
CS	Czech	2
RO	Romanian	2
SH	?	1

[67] The language codes are presumed to follow the ISO 639 standard, see <http://en.wikipedia.org/wiki/List_of_ISO_639_codes> for a list of these codes. This list has also served as the source for the language names associated with two-letter codes.

Language code	Language	Number of records
NE	Nepali	1

Table 5.8: Number of records associated with language codes.

It is of some interest to note that English is not the language with the highest number of records, and it is also heavily outnumbered by the other languages taken together.

5.6.2.4 Alternative metadata formats

Two of the harvested data providers provided metadata formats other than the basic oai_dc. These formats were ccsd_tel, used by @rchiveSIC and cnr_eprints used by E-LIS.

Schema definition for metadata format ccsd_tel:[68]

```
<schema targetNamespace="http://tel.ccsd.cnrs.fr/OAI/2.0/ccsd_tel/"
        xmlns:ccsd_tel="http://tel.ccsd.cnrs.fr/OAI/2.0/ccsd_tel/"
        xmlns="http://www.w3.org/2001/XMLSchema"
        elementFormDefault="qualified" attributeFormDefault="unqualified">

<annotation>
<documentation>
  Schema for CCSD theses-en-ligne metadata format, 2002.
  Schema validated at http://www.w3.org/2001/03/webdata/xsv on 05-09-2001
  Server theses-en-ligne is available at http://theses-en-ligne.in2p3.fr/
</documentation>
</annotation>

<element name="tel" type="tel:telType"/>

<complexType name="telType">
<choice minOccurs="0" maxOccurs="unbounded">
<element name="language"  minOccurs="0" maxOccurs="unbounded"
type="string"/>
<element name="datesubmit"  minOccurs="0" maxOccurs="unbounded"
type="string"/>
<element name="formats"  minOccurs="0" maxOccurs="unbounded"
type="string"/>
<element name="abtract"  minOccurs="0" maxOccurs="unbounded"
type="string"/>
<element name="defencedate"  minOccurs="0" maxOccurs="unbounded"
type="string"/>
<element name="keywords"  minOccurs="0" maxOccurs="unbounded"
type="string"/>
<element name="type"  minOccurs="0" maxOccurs="unbounded" type="string"/>
<element name="depositby"  minOccurs="0" maxOccurs="unbounded"
type="string"/>
<element name="advisor"  minOccurs="0" maxOccurs="unbounded"
type="string"/>
<element name="title" minOccurs="0" maxOccurs="unbounded" type="string"/>
<element name="abtracten"  minOccurs="0" maxOccurs="unbounded"
type="string"/>
<element name="thesistype"  minOccurs="0" maxOccurs="unbounded"
type="string"/>
<element name="subject"  minOccurs="0" maxOccurs="unbounded"
type="string"/>
<element name="timesubmit"  minOccurs="0" maxOccurs="unbounded"
type="string"/>
<element name="institution"  minOccurs="0" maxOccurs="unbounded"
type="string"/>
<element name="subjectfr"  minOccurs="0" maxOccurs="unbounded"
type="string"/>
```

[68] <http://tel.ccsd.cnrs.fr/OAI/2.0/ccsd_tel.xsd> Accessed 2005-11-17.

```
<element name="author"  minOccurs="0" maxOccurs="unbounded" type="string"/>
<element name="urlpage"  minOccurs="0" maxOccurs="unbounded"
type="string"/>
<element name="urldocpdf"  minOccurs="0" maxOccurs="unbounded"
type="string"/>
<element name="msc2000"  minOccurs="0" maxOccurs="unbounded"
type="string"/>
<element name="acmccs"  minOccurs="0" maxOccurs="unbounded" type="string"/>
<element name="altloc"  minOccurs="0" maxOccurs="unbounded" type="string"/>
<element name="department"  minOccurs="0" maxOccurs="unbounded"
type="string"/>
<element name="comments"  minOccurs="0" maxOccurs="unbounded"
type="string"/>
</choice>
</complexType>
</schema>
```

Schema definition for metadata format cnr_eprints:[69]

```
<schema xmlns="http://www.w3.org/2001/XMLSchema"
        xmlns:cnr="http://eprints.bo.cnr.it/cnr_eprints/"
        targetNamespace="http://eprints.bo.cnr.it/cnr_eprints/"
        elementFormDefault="qualified" attributeFormDefault="unqualified">

<element name="cnr_eprints">
<complexType>
<choice minOccurs="0" maxOccurs="unbounded">
<element name="title" type="string" minOccurs="0" maxOccurs="unbounded"/>
<element name="creator" type="string" minOccurs="0" maxOccurs="unbounded"/>
<element name="subject" type="string" minOccurs="0" maxOccurs="unbounded"/>
<element name="description" type="string" minOccurs="0"
maxOccurs="unbounded"/>
<element name="publisher" type="string" minOccurs="0"
maxOccurs="unbounded"/>
<element name="date" type="date" minOccurs="0" maxOccurs="unbounded"/>
<element name="type" type="string" minOccurs="0" maxOccurs="unbounded"/>
<element name="format" type="string" minOccurs="0" maxOccurs="unbounded"/>
<element name="identifier" type="string" minOccurs="0"
maxOccurs="unbounded"/>
<element name="language" type="string" minOccurs="0"
maxOccurs="unbounded"/>
<element name="rights" type="string" minOccurs="0" maxOccurs="unbounded"/>
<element name="journal" type="string" minOccurs="0" maxOccurs="unbounded"/>
<element name="volume" type="string" minOccurs="0" maxOccurs="unbounded"/>
<element name="issue" type="string" minOccurs="0" maxOccurs="unbounded"/>
<element name="pages" type="string" minOccurs="0" maxOccurs="unbounded">
<attribute name="pgstart" type="string"/>
<attribute name="pgend" type="string" />
</element>
</choice>
</complexType>
</element>
</schema>
```

5.6.3 Spam

Two incidents of spamming were detected in the middle of October. Both were directed at the page <http://collib.info/index.php/Help:Contents> and consisted in the addition of links to “suspect” Websites, in such a way that the links were not visible to ordinary users, but would presumably be picked up by search engines. The edits were made by users who were not logged in, and from several different IP-addresses. One interpretation of this behaviour might

[69] <http://eprints.bo.cnr.it/xsd/cnr_eprints.xsd> Accessed 2005-11-17.

be that these were just tests to see how the site would react to spamming, and that it might have been followed by a wave of similar activity, if the attempts had not been removed. To counter this, adding and editing pages was turned off for non-logged in users on 2005-10-20.

On 2005-11-12 a new incident of spamming was detected, this time made by a registered user, but otherwise similar to the incidents above. The page was reverted to its original state and the user was "blocked" for 100 days, which means that this user will not be able to make changes to this or other pages for that period of time.

5.6.4 Use statistics

Some statistics about the use of the prototype are available at the Web site.[70] The following graph gives an overview of the usage of the site:[71]

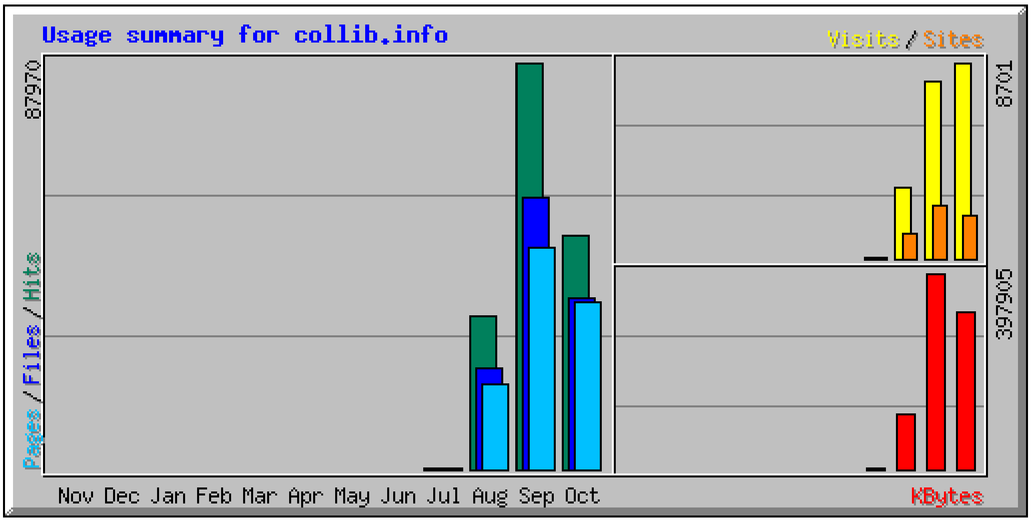


Illustration 5.3: Usage of collib.info as of 2005-10-30.

The effects of announcing the service to international mail-lists in September can be seen in the sharp increase of usage from August to September. In October the initial interest seems to have died down. It is interesting to note that even though the number of page views, files requested and hits (on the left) has dropped, the number of visits has actually increased from September to October. This becomes clear if we look at the actual numbers:

Month	***Visits***	***Pages***	***Files***	***Hits***	***Pages per visit***
Aug 2005	3164	18620	21729	33252	5.9
Sep 2005	7835	48088	58733	87970	6.1
Oct 2005	8701	36007	36945	50383	4.1

Table 5.9: Usage statistics for collib.info, monthly totals.

This may reflect a difference between first-time visitors, who view a higher number of pages

[70] <http://collib.info/statistikk/>

[71] The terms used in the graph are defined here: <http://www.mrunix.net/webalizer/webalizer_help.html>

to get an impression of the site, compared to returning visitors who e.g. visits the pages of specific topics to see if there are any new records added since the last time they visited.

It should however be noted that these numbers include spiders run by e.g. Web search engines, that these can account for a large proportion of the numbers, and that this can skew the impression one gets from looking at them. Thus it might be more interesting to look at how actual features of the site have been used.

5.6.5 Use of site features

The MediaWiki software provides some features that reveal details of the usage of the site. All the numbers given below are current as of 2005-10-30.

5.6.5.1 Popular pages

The ten most popular pages, or pages that have been viewed the most times, are as follows:[72]

1. Main_Page (4259 views)
2. Open_Access (378 views)
3. OAI-PMH (273 views)
4. FRBR (153 views)
5. Metadata (148 views)
6. Record:OaiRecord1935 (146 views)
7. Overlay (144 views)
8. Repositories (140 views)
9. Information_literacy (126 views)
10. Bibliometrics (117 views)

5.6.5.2 Added metadata and pages in the Wiki

721 of the 4109 records harvested have been edited/tagged, leaving 3388 records untouched.

19 users have been active in adding and editing pages. A count of who has done the current edits to pages in the Wiki yield the following results:

User	***Current edits***
User A	919
User B	140
User C	88

[72] Based on <http://collib.info/index.php/Special:Popularpages>.

User	*Current edits*
User D	21
User E	18
User F	15
User G	12
User H	9
User I	6
User J	4
User K	4
User L	3
User M	3
User N	3
User O	2
User P	2
User Q	1
User R	1

Table 5.10: Number of current edits per user. ("User A" is identical with the author).

Users in this list are a combination of the names of registered users and the IP addresses of unregistered users or registered users who are not logged in. This means that some users may be represented by two or more rows in the table above.

5.6.5.3 Categories

Pages that are created can be assigned to categories. The following categories have been created:

Category	*Number of pages*
Computer languages	2
Document formats	5
Document types	10
Geography	41
Libraries	5
Natural languages	7
Organization	8
Persons	39
Projects	27
Protocols	1
Services	27
Software	4

Category	*Number of pages*
Sources	5
Standards	10
Subjects	139
Total:	330

Table 5.11: Number of pages per category

5.6.5.4 Registered users and features available to them

There are 49 registered users.[73] Registered users are able to "watch" pages in the Wiki, and have them displayed in "My colLib". 12 of the registered users have chosen to watch pages, and they are watching the following numbers of pages:

User	*Pages watched*
User A	42
User B	20
User C	7
User D	7
User E	4
User F	3
User G	3
User H	2
User I	1
User J	1
User K	1
User L	1

Table 5.12: Number of pages watched by registered users. ("User B" is identical with the author.)

5.6.5.5 RSS feeds

From the usage statistics it is possible to extract the following usage for the RSS feeds provided on the site:

Month	*Hits*
Aug 2005	227
Sep 2005	4822
Oct 2005	5305

Table 5.13: Number of hits on the RSS feeds

Interpreting these numbers is non-trivial, since such feeds can often be polled (downloaded) several times a day to check for new additions, resulting in very high numbers from just a few

[73] <http://collib.info/index.php/Special:Listusers>

users. On the other hand, online services such as Bloglines[74] can download feeds once and provide updates to several users of their service.

5.7 Conclusion

All the statistics shown above indicate a very low usage of the “advanced” features of the prototype. One conclusion to be drawn from this is that browsing records harvested from Open Access repositories in Library and Information Science may be welcome, but very few people are willing to expend time and effort on contributing to the organization of the records.

On the other hand, quite a lot of records have had metadata added, so it might not take that many users to maintain a useful service. The site will live on at least until the summer of 2007, to see what developments, if any, will occur.

[74] <http://www.bloglines.com/>

Chapter 6: Discussion

6.1 Adaptive radiation of overlay systemic documents

As was seen in table 4.1 (p. 35), there is quite a lot of variation between different overlay systemic documents. What is striking about the features in this table is that almost all of them are of a general nature, i.e. features that could be useful in any service provider, independently of the subject or scholarly field being served. It should be expected that as the services mature, they will adopt more of these general features, so that the differences between services actually become smaller.

For a feature to be adaptive, it should have some functionality that is tailored to the requirements of a scholarly discipline (or a narrow group of disciplines), such that services displaying that feature should have a higher chance of becoming successful in the niche they are catering to.

The only feature found in the survey that fits this description is the equation search in Archon, which is tailored specifically to the physics community. This is clearly an example of a feature that is built to serve a special need in its intended audience. A similar feature would probably be relevant for mathematics too, but for fields that do not use equations extensively or at all, it would be irrelevant. On the other hand, a similar feature for molecular formulas could be of interest to e.g. chemical, biological and medical service providers.

As can be seen in the survey of OAI-PMH-based service providers, there are already complementary services that could be said to compete for the attention of users. One kind of competition is between the general service providers, whose aim seems to be harvesting metadata from as many data providers as possible. Another kind of competition is between service providers that cater to users in more or less the same scholarly field, such as physics. The general services could also be seen as competing with the specialised services, since a general service can contain exactly the same metadata records as the specialised ones.

One of the factors that will decide who wins this competition for attention is the features that the services can offer. Specialised services can attract users from the general services by offering features that cater specifically to the needs of the users in its chosen niche. An example of this is the equation search found in Archon.

Whatever features are implemented in an overlay service provider, they have to rely on some kind of data. These can be the metadata harvested from data providers, or they can be added to the harvested records by the overlay service itself:

6.2 Creating features based on harvested data

6.2.1 Metadata

As the survey shows, the one feature implemented by all service providers is federated search, or search across the data from all the data providers harvested. Further refinements to this would be fielded search, where users can choose to search e.g. authors name, titles or descriptions, or a search by source, where the search would be limited to metadata from a subset of the data providers harvested.

One problem with this is the potential multitude of languages present in the harvested metadata. As shown in chapter 5.6.2.3 (p. 59), there were 19 different language codes represented in the 5 data providers harvested by the prototype, and only 84% of the records had an easily interpreted language code. This means that searching for a subject in one language will result in too few hits, since records about that subject in other languages will not show up in the results. One solution to this is the use of a (mono- or multilingual) controlled vocabulary, which names the relevant subjects in a consistent manner and can be applied across all the languages present in the service provider. On the other hand any one user will be able to read a limited set of languages, so hits describing documents in a language the user can not read will be of little use. Specifying which languages are of interest as one of the search criteria could alleviate this, but records without a valid language code would still be a problem.

This is definitely an issue with the prototype, where English “tags” have been applied across all the languages.

It is interesting to note that metadata formats are undergoing an adaptive radiation of their own. In chapter 5.6.2.4 (p. 60) we saw two examples of metadata formats that are created to be able to describe specific forms of documentation better than the basic oai_dc format. ccsd_tel seems to be geared towards describing theses and dissertations, with elements such as *defencedate*, *advisor* and *thesistype*, while cnr_eprints is well suited to describing journal articles, with elements such as *journal*, *volume*, *issue* and *pages*. This creates a new challenge for service providers, who need to be able to parse and store all the different formats that are available, in order to exploit them for creating richer services for end users.

6.2.2 Full text

As stated earlier, the OAI-PMH does not provide mechanisms for harvesting the resources described by metadata (full text documents in the case of most scholarly repositories). On the other hand, URLs pointing to the locations of the primary documents can be included in the metadata that is harvested through the protocol, and the documents themselves can be downloaded through some means external to the OAI-PMH.

Once a service provider is able to obtain the primary documents, opportunities for creating new services arise, such as the equation search found in Archon. Other applications might be full text searching or extraction and analysis of references.

To facilitate the creation of services based on the full text of primary documents a simple metadata format containing not much more than the URLs of primary documents could be created. If this were widely adopted by data providers, it would facilitate the creation of services based on full text.

6.2.3 Sets

A third type of data that can be utilized as a substrate for services is the set-structure made available by data providers, depending on how they are organized. Sets that are created in order to reflect the structure of an organization such as a university may not be all that interesting in this context, but sets created in order to reflect the topical divisions of a field would.

The sets provided by the 5 repositories harvested by the prototype are listed in Appendix A. We see immediately that there is quite a bit of variation in how sets are realized, making it difficult to use them to provide e.g. an integrated browsing-experience across all the records. On the other hand there are similarities, for example both CaltechLib and E-LIS base their sets on the JITA classification scheme (CaltechLib uses just the main classes, while E-LIS uses the whole scheme), and all the repositories have sets that are at least related to the same subject-matter. Based on these similarities, it might be possible to create a mapping between sets, creating a consistent, but quite high-level subject hierarchy that could be used for browsing by subject.

6.3 Creating features based on adding data

Table 5.7 (p. 59) highlights some of the problems connected with basing features on the metadata harvested from multiple data providers. Different terms are used for what is presumably the same category, e.g. “Journal Article (Print/Paginated)”, “Journal Article (Paginated)” and “Journal (Paginated)”. One solution to this might be that the data providers in a given field such as library and information science agree on a standard set of terms for describing e.g. document formats. Another solution which does not require collaboration on the part of data providers would be to map the different categories into one vocabulary, so that all the terms quoted above were mapped to one common term like “Journal Article

(Print/Paginated)"

If the metadata available in records or sets is not consistent or detailed enough to serve as the basis for e.g. browsing records, then adding the necessary metadata is an option, although a labour-intensive one. This is the approach chosen in the prototype, where the addition of "tags" to the wiki-pages connected with harvested records, makes it possible to browse the records by subject, author, document type or other criteria. As table 5.10 (p. 64) shows, very few users have been willing to invest time and effort in adding metadata to records.

6.4 Selective inclusion of records

One important factor in the competition between service providers is how they create features based on the metadata, full text and sets that they have access to. Another factor, which seems not to be utilized at all in the surveyed service providers, is how they select records for inclusion in the service. For the general services this may be quite easy, they will probably want to get their hands on as many metadata records as possible, to be able to provide for the broadest possible range of subjects. For specialised services the opposite might be true, they will want to have a collection of metadata that is as relevant as possible for the intended audience, so that e.g. searching returns as few falls hits ("noise") as possible. There are several ways to obtain a focused collection of metadata.

6.4.1 Man v.s. machine

One of the main distinctions that can be made among different overlays in this respect is whether the selection and inclusion of records is carried out by human intervention or done automatically. The OAI-PMH is in its basic modus operandi mechanical. Records are harvested on a regular basis and included in the database of the service provider. Through the mechanism of sets, it is possible to harvest only a part of the records that a data provider makes available, but service providers are completely reliant on the data provider here, since they can only choose among the sets that the data provider has seen fit to provide.

Some mechanical selection could be introduced, so that records are filtered on a set of given criteria after they are harvested, but before they are included in the database. This could run the whole gamut from a simple search for keywords in titles and descriptions to complex operations based on artificial intelligence and machine learning. Filtering could also be based on structured metadata such as classification codes that are included in the metadata made available by service providers.

In this way a service provider could harvest records from a large number of repositories, and create an overlay that is especially relevant to some narrow field of interest, by excluding records that are not relevant to this field, before the records are presented to the users of the service. This approach would be particularly effective in a field with an agreed upon classification scheme that was used by all relevant data providers. In such a case, the data providers could also create sets based on the classification, so that harvesting could be limited to relevant sets, rendering post-harvest filtering obsolete.

However, the reality of most service providers is probably that they harvest metadata from a number of disparate data providers, without the benefits of any common classification. Filtering records based on searching for keywords will also probably be problematic for most, since it would be difficult, if not impossible, to construct a set of keywords that effectively define the perimeters of the field, resulting in relevant records being excluded and irrelevant records being included.

A low-tech, but labour-intensive, solution to the problem of selecting records for inclusion in a choosy service provider would be to assign human editors the task of choosing which records should be included. As new records were harvested, they would be placed in a queue, and only those records that gained the approval of an editor would be included and made available as part of the service.

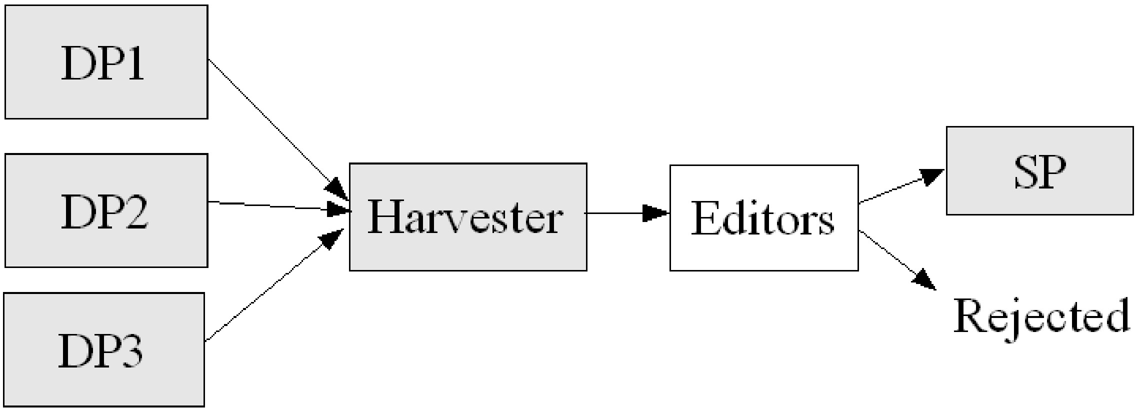


Illustration 6.1: Outline of an overlay with manual intervention. DP = Data provider, SP = Service provider. Arrows indicate the flow of metadata. Grey boxes represent the standard elements of the OAI-PMH framework.

The prototype fits this description of a manual overlay only partially. Records are automatically imported into the database, and are made available for searching, but to be included in the browsing interface, human intervention is required.

6.4.2 Editor v.s. audience

The proportion of editors to the number of people in the audience may vary between overlays, resulting in quite different services. Here are some of the possible combinations:

Editor(s)	***Audience***
One	One
One	Many
Few	Many
Many	Many

Table 6.1: Possible relations between editors and audience

On one end of the spectrum is an overlay where the editor and the audience is the same person. The OAI-PMH is not designed with an eye to individuals running harvester software, but rather that each harvester should be able to provide services to several people. On the other hand there is nothing stopping individuals from running their own harvesters. The more acceptable way to do this (from the point of view of the data providers) would be to run a service where individuals could register and create their own overlays, in the form of bibliographies or lists of references. The "My colLib" feature of the prototype (described briefly in chapter 5.3.5.3, p.52) is an implementation of this idea, limited to one list of records per user.

The next step is for individuals to be able to share the bibliographies they create with others. An example of this can be seen in the "Virtual collections" feature of the Sheet Music Consortium.

If more people are added on the editorial side of the equation, we get a situation that more closely resembles that of the traditional journal, where there may be one editor in charge, and several people doing the actual work of selection and quality control.

At the opposite end of the spectrum to the personal overlay is the group-edited overlay, where anyone can choose to be an editor. The main browsing feature of the prototype is an example of this, where anyone can register and influence what records should be made browsable, and what keywords they should be associated with. In the prototype this feature is built on top of a Wiki, but it could also be implemented in more traditional way, users could e.g. get to select keywords that should be connected with a metadata record from a pre-defined list, or they could be able to contribute new terms to this list themselves.

6.5 Different roles

The material found in Open Access repositories will typically be a combination of articles that have been published in a journal (or is intended for such publication), and a host of other documents. Service providers may choose to relate to these in different ways.

6.5.1 Overlays to published documentation

Virtual Journals in Science and Technology provides a collection of five journals: *Applications of Superconductivity*, *Biological Physics Research*, *Nanoscale Science & Technology*, *Quantum Information* and *Ultrafast Science*.[75] The journals are described in the following manner:

> This series of "virtual" journals in the physical sciences has been jointly developed by the American Institute of Physics (AIP) and the American Physical Society (APS). Each of the virtual journals presents an online collection of relevant papers from a broad range of "source" journals in the physical sciences. Participating source journals include all journals published by APS and AIP, journals from participating publishers on AIP's Scitation, Science, and Nature.[76]

Current Cites (Tennant 2005) operates along similar lines:[77]

> A team of librarians and library staff monitors information technology literature in both print and digital forms, each month selecting only the best items to annotate for this free publication. The resulting issue of 8-12 annotated citations of current literature is emailed to a mailing list and is redistributed on other electronic fora.[78]

Both these journals provide pointers to documents that are available elsewhere, and I would therefore label them as overlays. They provide a service of filtering the vast streams of new documents available in their fields, narrowing it down to a trickle that even the most overworked researcher or practitioner has a chance to keep up with. By doing so they are also adding a seal of approval, by focusing on what is perceived to be the “best” or “relevant” new documents.

6.5.2 Overlays to published documentation in repositories

One could also envision an overlay service which did not aim to filter published articles, but merely to organize and make browsable metadata for articles available in repositories, that have already been published in journals. This service could be organized by publishing journal, so that it would be possible to view a list of journals, and then to see what articles that were published in that journal are also available through Open Access from some repository.

colLib could easily be used to this end, and in fact there is the embryo of such a structure in the “Sources”-category.[79]

[75] <http://www.virtualjournals.org/>
[76] <http://www.virtualjournals.org/vjs/aboutvj.jsp> Accessed 2005-10-20.
[77] <http://lists.webjunction.org/currentcites/>
[78] From the homepage, accessed 2005-10-20.
[79] <http://collib.info/index.php/Category:Sources>

This raises some interesting questions about the suitability of the OAI-PMH in this scenario. The first stumbling block is the lack of information about publishing journal in the basic oai_dc metadata format. This could however be surmounted by utilizing richer metadata sets that are designed to convey this information. Other, more problematic issues are related to trust. Can we trust a piece of metadata that says the document it describes was published in a particular journal? Another important question concerns versioning – is the representation of the published article that is available from the data provider an early draft, a PDF downloaded from the site of the journal or a post-print with corrections and additions? How should this information be conveyed – in the metadata or in the primary document? And again, if the information is furnished by the author – how do we know if we can trust it? One function of an overlay service may actually be to assess and provide information on what version of an article is available from a given repository.

6.5.3 Overlays to unpublished documentation in repositories

The two scenarios sketched above both rely on actors external to the overlay to provide the rigorous quality control that is often expected in conjunction with scholarly publishing. This means that they can co-exist both with traditional "closed" journals, and with the new breed of Open Access journals (see chapter 2.3, p.14). The "Distributed Journal" model described by Smith (2004) goes beyond this, however, and posits that we can dispense completely with journals as we know them, and distribute the functions they perform today among alternative actors. In this model, repositories will take care of the physical storage of scholarly documentation, while quality control is carried out by independent bodies, known as Certification Agents.

Some journals, e.g. in mathematics, have begun to make open repositories a part of their infrastructure. *SIGMA* describes its procedures for submitting an article like this:

> Submitting a paper to the journal can be performed in two easy ways:
>
> - submit the paper to the arXiv.org and send the archive number to sigma@imath.kiev.ua or
> - send zipped paper in TeX/LaTeX format directly to sigma@imath.kiev.ua (please include PDF or PostScript file as well).[80]

This means that instead of just sending drafts of articles directly to the editor of the journal, authors can upload them to arXiv, making them immediately available to anyone with an Internet connection. If and when the article is accepted for publication by the journal, with or

[80] <http://www.emis.de/journals/SIGMA/about.html#submit> (Accessed 2005-10-27)

without editing of the contents, the original pre-print can presumably be replaced with the final version, including any formatting provided by the journal. The journal does not, however, point to the finished articles in arXiv, but provides the articles for download from their own site.

The *Front for the Mathematics ArXiv* elaborates on this procedure in their advice on how to "convert journals to overlays":

> You can either ask authors to contribute new papers to the arXiv themselves and ask them to give you the arXiv numbers, or you can contribute articles to the arXiv for them by proxy, or ideally both. If your overlay-to-be is a journal with its own typesetting, you will need a proxy submitter; in addition, if a submitted paper is already in the arXiv, the author must give you both the paper number and the password so that you can replace it with the typeset version.[81]

6.5.4 Registering v.s. evaluating overlays

There is a spectrum of different stances that an overlay can take to the harvested metadata. In the case of something calling itself a journal it would be expected that the overlay should maintain the same standards of quality as is found in other journals.

One could also envision an overlay that consists of review-articles that do not only include or exclude references based on their quality, but that discuss some issue based on documents that are available in Open Access repositories. References in the running text could then be links to the full bibliographic description, with a link to the full text at the originating repository. Such articles could be written by one individual, by a group of individuals who each write articles on their own topic of expertise, or they could be written collaboratively, e.g. in the form of a Wiki. The prototype would be well suited to this form of collaboration, since links to records are easily embedded in running text, and collaboration is the *raison d'être* of wikis.

At the other end of the spectrum one could find overlays that do not concern themselves with quality at all, just with relevance to some scholarly field or subject. This could be likened to a bibliography which aims to include all the literature relevant to a given subject.

The OAI-PMH in it self seems not to be an obstacle in this context, the services described above could all be created on top of the protocol.

6.6 Some problems with the present model

I think this discussion has uncovered some problems with the present implementation of the OAI-PMH:

[81] <http://front.math.ucdavis.edu/overlays#convert> (Accessed 2005-10-27)

One of the questions that has to find a solution if the “distributed” journal is ever going to gain any serious ground is how to tell if an item in a repository has been given a “seal of approval” by an overlay service. This information could be incorporated in the primary document itself, as we saw in the procedure described by the *Front for the Mathematics ArXiv* above, where pre-prints should be replaced by the formatted post-print once it was accepted by the journal. This relies on a one-to-one relationship between articles and journals, and would not work in a situation where multiple overlays may “approve” of an article. The author may want to convey this information in some way, because the inclusion in a high status overlay may confer some status on the article and the author. This could perhaps be expressed in the metadata instead.

Another serious issue is the fact that there are no absolute guarantees that documents in repositories are not changed or removed. If a document is approved by an overlay service, users of that service need to be sure that the document has not been altered after the approval was given. One solution to this might be for the overlay service to download and make available the original documents in the state they were at the time the approval was given, another might include the creation of digital fingerprints, as described in Appendix A of Smith (2004).

Chapter 7: Conclusion

Based on the ecological metaphor, one can predict that the new opportunities offered by large amounts of freely available metadata, easily harvested by means of the OAI-PMH, will result in an adaptive radiation of a wide range of new overlay systemic documents, as enterprising individuals and groups adapt tools and techniques to the specific demands of new niches. These services will then compete for important resources such as the attention of users, and those that are best suited to the demands of the environment will prosper.

The survey reported in chapter 4 showed that there is already quite a lot of variation between overlay systemic documents created on the basis of metadata harvested through the OAI-PMH, but this variation appears more random than adaptive, since very few of the features seem to be created on the basis of the specific needs of researchers and practitioners in a given field.

From the experiences drawn from creating and running the prototype as reported in chapter 5, it seems that the Open Archives Initiative Protocol for Metadata Harvesting is ready and able to provide the infrastructure on which to build a lot of different overlays. The present exercise has shown that the real challenge is not in constructing the actual service, but to come up with ideas for services that are designed to serve the needs in a field, and to attract users to these services.

Some mention has been made of the opportunities that exist for extending the OAI-PMH as it exists today. My opinion is that one should not forget the possibilities offered by creating new metadata formats. On the other hand, the plethora of formats that are already available can be problematic for service providers, who need to be able to parse and utilize the data available in different formats. This will be a challenge for those who are developing the software of service providers, in order to create flexible frameworks for these services.

In the end, only time will tell what forms of overlay systemic documents the OAI-PMH will lay the foundations for, and how they will evolve. In these times of quick technical developments, let us not forget the librarians adage that everything we do should be “to the benefit of the user”.

Chapter 8: Suggestions for future work

The scope of the present work has meant that I have only had the opportunity to scratch the surface of some very interesting phenomena that could be worthy of further research in and of themselves. I think these can be grouped into at least three distinct families of questions.

8.1 Open Access

The Open Access movement is only now beginning to produce results, e.g. in the form of creating and filling repositories with scholarly documentation. This raises some interesting questions:

- Who is making what available where, and when?
- What kinds of documents are being made available? At what stage of the publication or research process?
- How (and why) does practices differ among scholarly fields?
- How are metadata formats evolving, who are creating new formats and how are they being adopted by data as well as service providers?

8.2 Web-specific forms

The new, Web-specific forms of documentation briefly described in these pages should provide a fertile ground for some interesting research into how new forms evolve and adapt to changing circumstances. Some of the forms that could be given attention are:

- Wikis.
- Syndication and aggregation through channels such as RSS and Atom.
- Sites organizing documents based on end-user tagging.

8.3 The ecology of documentation

I think the metaphor of an ecology of documentation has served well as a framework for the present work, and deserves some more attention. This could take at least two forms:

- Theoretical investigations to find new approaches to the study of documentation based on the ecological metaphor.
- Practical research that utilizes the metaphor as a framework, in order to test the suitability of this framework in different contexts.

Appendix A: Survey of service providers

Some of the service providers listed in the sources mentioned above were excluded from the survey:

- **Freescience** (From "digitAlexandria", formerly known as "Biblioteca d'Alessandria") <http://www.bdaweb.net/>. Listed by the OAI, but is not a web-based service.
- **Callima** <http://www.callima.com/>. Server not found.
- **CILEA Open Archives Platform** <http://www.cilea.it/>. Not yet available, but the planned features are described by McKiernan (2004).
- **CYCLADES** <http://www.ercim.org/cyclades>. The front page contains a link labelled "CYCLADES (beta version) available", but clicking on this results in a time out. Described by Brogan (2003, chapter 6.2.1).
- **iCite** <http://icite.sissa.it/ >. This service is seemingly no longer available. The URL redirects to a description of SISSA/ISAS (Scuola Internazionale Superiore di Studi Avanzati/International School for Advanced Studies).
- **my.OAI** <http://www.myoai.com/>. Server not found. Described by McKiernan (2003a).
- **OASIC** <http://oasic.ccsd.cnrs.fr/ >. "403 Forbidden" status-message received. Listed by the OAI.
- **Repository Explorer** <http://purl.org/net/oai_explorer>. The Repository Explorer does not store metadata harvested through the OAI-PMH, and hence it falls outside the scope of this survey. It is, however, on the OAI list of registered Service providers.

The results of the survey is given below. (Descriptions in italics are taken from the home- or help-pages of the surveyed service, unless otherwise noted.)

AmericanSouth.org

URL	http://americansouth.org/
Description	*AmericanSouth.Org is an on-going project undertaken at Emory University in collaboration with a large number of Southern research libraries that seeks to improve access to digital resources. Put simply, AmericanSouth harvests metadata, or information about information, from an amalgam of library and museum archives, pulling this metadata into a central location for aggregation, indexing, search, and discovery. The resulting digital library collection enables a researcher to conduct a combined search of the materials held by many*

	different institutions.
Subjects	*The cultures and histories of the American South.*
Sources	http://americansouth.org/archives.php
Features	• Federated search • Browse by data provider.
Literature	Brogan (2003, chapter 6.4.3), McKiernan (2003b)

Table A.1: Survey of AmericanSouth.org

ARC

URL	http://arc.cs.odu.edu/
Description	*Arc is an experimental research service of Digital Library Research group at Old Dominion University. Arc is used to investigate issues in harvesting OAI compliant repositories and making them accessible through a unified search interface. It is not a production service and may be subject to unscheduled service interruptions and anomalies.*
Subjects	General
Sources	http://arc.cs.odu.edu:8080/oai/results.jsp
Features	• Federated search (simple and advanced). • Browse by data provider.
Literature	Brogan (2003, chapter 6.2.1), McKiernan (2003a), Liu et al (2001)

Table A.2: Survey of ARC

Archon

URL	http://archon.cs.odu.edu/
Description	*This is a collaborative project between Old Dominion University, American Physical Society and Los Alamos National Laboratory. This project is building an Open Archives Initiative compliant federated digital library with an emphasis on physics for the The National Science Digital Library. [...] This physics digital library will federate holdings from the physics e-print server arXiv, Physical Review D from the American Physical Society. We are also working with CERN to federate their collection. Other holdings are being imported from the Arc project. [...] We are supporting high-level services such as cross-reference linking, which is based on OpenURL and leverages the Citebase research at Southampton.*
Subjects	Physics
Sources	APS, CERN, Emilio, NTRS-Physics, arXiv

Features	• Federated search (simple and advanced) • Equation search (Maly et al 2002, p.3) • Reference linking (Maly et al 2002, p. 5) • Annotations • Display of references
Literature	Brogan (2003, chapter 6.2.3), Maly et al (2002)

Table A.3: Survey of Archon

BASE: Bielefeld Academic Search Engine

URL	http://www.base-search.net/index.php?l=en
Description	*BASE integrates scientific OAI-resources as one information type among others into the local digital library environment, together with catalogues, article databases, digitised collections. The search interface features many characteristics of internet search engines, thus offers a new type of search interface for a local digital library. BASE uses the search technology of FAST Search & Transfer.* [From the OAI -list]
Subjects	General
Sources	http://base.ub.uni-bielefeld.de/about_sources_english.html
Features	• Federated search (basic and advanced)
Literature	Summann and Lossau (2004)

Table A.4: Survey of BASE: Bielefeld Academic Search Engine

citebaseSearch

URL	http://citebase.eprints.org/, http://www.citebase.org/
Description	*Citebase Search is a semi-autonomous citation index for the free, online research literature. It harvests pre- and post- prints (most author self-archived) from OAI-PMH compliant archives, parses and links their references and indexes the metadata in a search engine.*
Subjects	*Citebase contains articles from physics, maths, information science, and (published only) biomedical papers.*
Sources	http://citebase.org/help/info_press.php
Features	• Federated search (by “Metadata”, “Citation” or “OAI Identifier”) • Citation analysis. • Statistics of full text downloads.
Literature	Brogan (2003, chapter 6.2.3), McKiernan (2003a)

Table A.5: Survey of citebaseSearch

DP9

URL	http://arc.cs.odu.edu:8080/dp9/index.jsp
Description	*DP9 is a gateway service that enables indexing of an OAI data provider by an Internet search engine. DP9 does this by providing a persistent URL for repository records, and converting this to an OAI query against the appropriate repository when the URL is requested. This allows search engines that do not support the OAI protocol to index the "deep web" contained within OAI compliant repositories.*
Subjects	General
Sources	http://arc.cs.odu.edu:8080/dp9/index.jsp
Features	• DP9 is mainly meant to be used by automated web crawlers used by e.g. search engines to index the Web. The following additional services are made available: • Persistent and bookmarkable URL for OAI record • OAI Identifier Resolver • Service Linking • Support parallel metadata Set
Literature	Brogan (2003, chapter 6.5.2)

Table A.6: Survey of DP9

ePrints UK

URL	http://eprints-uk.rdn.ac.uk/
Description	*ePrints-UK harvests metadata from approximately 30 institutional repositories on a daily basis. A demonstration of a simple search service based on the harvested metadata has been made available now. In future the project intends to enhance metadata available for searching by means of name authority file checks and automatic subject classification.*
Subjects	General, but limited to institutional repositories in the UK.
Sources	About 30 UK-based institutional repositories.
Features	• Federated search. • More features are being developed.
Literature	McKiernan (2004)

Table A.7: Survey of ePrints UK

ETD OAI Union Catalog

URL	http://rocky.dlib.vt.edu/~etdunion/cgi-bin/index.pl
Description	*This is a service built by harvesting metadata from Open Archives of electronic theses and dissertations.The underlying technology is based on layered Open Archives with data being harvested from source archives and then stored in a Union Catalog. This Union Catalog is then front-ended with a search engine for demonstration purposes, but the data is just as easily accessible to other service providers, both local and remote.*
Subjects	General, but limited to electronic theses and dissertations
Sources	14 data providers are listed on the search-page
Features	• Federated search • Browse by data provider (institution)
Literature	McKiernan (2004), Brogan (2003, chapter 6.2.2)

Table A.8: Survey of ETD OAI Union Catalog

Grainger Engineering Library at UIUC

URL	http://g118.grainger.uiuc.edu/engroai/
Description	*This site predominantly provides access to scientific e-prints, technical reports, theses and dissertations, and e-journals collections. At present this service is intended primarily for local institutional use. As a result, it does not provide any context or documentation about its mission, scope of operation, or collection policy.* (Brogan 2003, chapter 6.2.3)
Subjects	Engineering, Computer Science, and Physics
Sources	http://g118.grainger.uiuc.edu/engroai/LastHarvest.asp
Features	• Federated search • “Book bag”, a feature for “remembering” records. • Search history
Literature	Brogan (2003, chapter 6.2.3)

Table A.9: Survey of Grainger Engineering Library at UIUC

MeIND

URL	http://www.meind.de/ , English language: http://www.meind.de/?ln=en
Description	*The Service-Provider covers all subjects form different data-providers in Germany. The Pub-Types include doctoral theses, diploma, articles, magazines and digitised materials.* [From the OAI list.]
Subjects	General

Sources	http://www.meind.de/info/oai-template.html
Features	• Federated search • Browse by document type or by subject (based on the Dewey Decimal Classification) • Search history • “Shopping basket” • Alerting (based on “saved searches”)

Table A.10: Survey of MeIND

METALIS

URL	http://metalis.cilea.it/
Description	*METALIS is a Service Provider for the Library and Information Science field. We collect (harvest) metadata from institutions that offer full-text papers and documents about Library and Information Science.*
Subjects	*Library and Information Science*
Sources	http://metalis.cilea.it/
Features	• Federated search (simple and advanced) • Outgoing OpenURL-compliant links. The service provides a default resolver, but users can also provide the URL of a preferred resolver.
Literature	Tajoli (2005)

Table A.11: Survey of METALIS

NCSTRL

URL	http://www.ncstrl.org/
Description	*NCSTRL provides unified access to technical reports and eprints from computer science departments, institutes and laboratories. This is an OAI-based implementation of the NCSTRL project.* [From the OAI list]
Subjects	Computer science
Sources	http://www.ncstrl.org:8900/ncstrl/body.html
Features	• Federated search (simple and advanced) • Browse by harvested repository/institution

Table A.12: Survey of NCSTRL

OAIster

URL	http://oaister.umdl.umich.edu/o/oaister/

Description	*OAIster is a project of the University of Michigan Digital Library Production Service. Our goal is to create a collection of freely available, previously difficult-to-access, academically-oriented digital resources [...] that are easily searchable by anyone.*
Subjects	General
Sources	http://oaister.umdl.umich.edu/o/oaister/viewcolls.html
Features	• Federated search • Browse by institution (data provider) • More features are planned: http://oaister.umdl.umich.edu/o/oaister/future.html
Literature	McKiernan (2004), Brogan (2003, chapter 6.2.1), Hagedorn (2001)

Table A.13: Survey of OAIster

Open Language Archives Community (OLAC)

URL	http://www.language-archives.org/ Search: http://linguistlist.org/olac/
Description	*OLAC, the Open Language Archives Community, is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources.*
Subjects	Linguistics
Sources	http://www.language-archives.org/archives.php4
Features	• Federated search (simple and advanced)
Literature	Brogan (2003, chapter 6.2.2), McKiernan (2003b)

Table A.14: Survey of Open Language Archives Community (OLAC)

Perseus

URL	http://www.perseus.tufts.edu/, search: http://www.perseus.tufts.edu/cgi-bin/vor
Description	*Perseus is an evolving digital library, engineering interactions through time, space, and language. Our primary goal is to bring a wide range of source materials to as large an audience as possible* This digital library comprises a combination of original records and records harvested from external data providers.

Subjects	History
Sources	The "American Memory" project, amongst others.
Features	• Search • Browse
Literature	Brogan (2003, chapter 6.3.2)

Table A.15: Survey of Perseus

Public Knowledge Harvester

URL	http://www.pkp.ubc.ca/harvester
Description	*The Public Knowledge Project has developed a number of discipline-specific Research Support Tools (RST), which accompany individual research studies indexed from e-journal and conference paper websites covering a wide range of disciplines. The RST utilizes the study's metadata to search relevant open-access databases for related studies, theory, news, policies, and other resources, as well as offering access to the study's metadata and citation, to a personal portfolio, and to email and comment options.*
Subjects	General
Sources	http://pkp.sfu.ca/harvester/archives.php
Features	• Federated search (Simple and advanced) • Browse by archive • When a single record is displayed, a "Research Support Tool" is available which contains links to several off-site services. Which links are available depends on which discipline the record is associated with.
Literature	McKiernan (2004)

Table A.16: Survey of Public Knowledge Harvester

SAIL-Eprints

URL	http://eprints.bo.cnr.it/
Description	*SAIL-eprints (Search, Alert, Impact and Link) is an electronic open access service provider for finding scientific or technical documents, published or unpublished, in Chemistry, Physics, Engineering, Materials Sciences, Nanotechnologies, Microelectronics, Computer Sciences, Astronomy, Astrophysics, Earth Sciences, Meteorology, Oceanography, Agricolture, and related application activities. SAIL-eprints has been designed primarily to collect information on scientific documents (metadata) authored by CNR researchers and deposited as preprints or postprints in CNR institutional open*

	access archives. Also, SAIL-eprints collects metadata from other data repositories all over the world publishing materials in the same scientific fields.
Subjects	The sciences, see the description above.
Sources	http://eprints.bo.cnr.it/cgi-bin/info.pl
Features	• Federated search (Simple and advanced) • Lists of recently harvested records • Browse by author and deposit-date
Literature	McKiernan (2003a)

Table A.17: Survey of SAIL-Eprints

Scirus

URL	http://www.scirus.com/
Descriptio n	*Scirus is the most comprehensive science-specific search engine on the Internet. Driven by the latest search engine technology, Scirus searches over 200 million science-specific Web pages [...]. Search engines are all different in the Web sites they cover, and the way they classify these Web sites. Scirus, the search engine for science, focuses only on Web pages containing scientific content. [...] Scirus returns results from the whole Web, including access-controlled sites that other search engines don't index.*
Subjects	General
Sources	http://www.scirus.com/srsapp/aboutus/#sources
Features	• Federated search (simple and advanced). • Advanced search gives the possibility to limit a search to “Preferred Web sources”, these are the actual OAI-PMH Data providers. • Save, mail or export references • Refine search using keywords found in the results
Literature	Scirus (2004), Brogan (2003, chapter 6.5.1)

Table A.18: Survey of Scirus

Sheet Music Consortium

URL	http://digital.library.ucla.edu/sheetmusic/
Descriptio n	*The Sheet Music Consortium is a group of libraries working toward the goal of building an open collection of digitized sheet music using the Open Archives Initiative:Protocol for Metadata Harvesting (OAI:PMH). Harvested metadata about sheet music in participating collections is hosted by UCLA Digital*

	Library Program, which provides an access service via this metadata to sheet music records at the host libraries.
Subjects	Music
Sources	Members of the consortium are listed on the front page, a drop-down list of available collections can be found in the advanced search form.
Features	• Federated search • Browse by title, year or collection • Virtual collections, where users can collect records and make these collections available to other users. These collections can be password-protected, or open for anyone to edit.
Literature	Brogan (2003, chapter 6.2.2), McKiernan (2003b)

Table A.19: Survey of Sheet Music Consortium

SuUB Bremen

URL	http://suche.suub.uni-bremen.de/
Descriptio n	*The State and University Library Bremen - Germany offers a library catalogue called SuUB Bremen, where selected OAI-data providers are included. The search engine delivers record about localized library ressources of print and electronic materials (books, journals, doctoral thesis, articles, magazines, reports, etc.).* [From the OAI list]
Subjects	General
Sources	http://suche.suub.uni-bremen.de/oai-archives.html
Features	• Federated search ("Standard" and "Erweitert") • Shopping cart

Table A.20: Survey of SuUB Bremen

TORII

This site requires the Microsoft Internet Explorer browser in order to display properly.

URL	http://torii.sissa.it/
Descriptio n	*[...] a web environment that allows unified access to open archives of the scientific community and provides useful services on them like full-text search, cognitive and collaborative filtering, storing in a personal folder, and autonomous citation extraction.*
Subjects	Physics, computer science
Sources	http://torii.sissa.it/html/torii_service_provider.html

Features	• Federated search • Browse by repository • The following features are available to users who have registered with the site: • "Shopping cart" that persists between sessions. • Filtering based on specifying areas of interest or creating "cognitive profiles", where users enter keywords, text or identifiers of documents that are known to be relevant, and the system recommends new records based on this information. • The possibility to add private and public comments to records, and view the public comments of others. • The possibility to rate (from zero to five stars) or review (e.g. assigning values to the criteria "Research quality", "Presentation" and/or "Intended audience") records.
Literature	Bertocco (2001)

Table A.20: Survey of TORII

UIUC Digital Gateway to Cultural Heritage Materials

URL	http://nergal.grainger.uiuc.edu/cgi/b/bib/bib-idx/
Description	*The OAI Metadata Harvesting Project at Illinois focused on creating a deep, domain-specific portal designed to search metadata describing selected manuscript archives and digitized cultural heritage information resources.*
Subjects	Cultural heritage
Sources	http://oai.grainger.uiuc.edu/AboutCollections.htm
Features	• Federated search (simple and advanced) • Browse by type ("Images and video", "Museum and Archives Collection" or "Text, Sheet Music, and Websites) • Search history • "Bookbag"
Literature	Brogan (2003, chapter 6.2.3), Cole (2003), McKiernan (2003b)

Table A.21: Survey of UIUC Digital Gateway to Cultural Heritage Materials

Appendix B: Sets

A small Perl-script based on the Net::OAI..Harvester module described in chapter 5.3.2.3 (p. 48) was created to extract information about the sets available from the data providers harvested by the prototype. The findings are listed below.

Sets from ALIA e-prints

Status = In Press
Status = Published
Status = Unpublished
Subject = Aboriginal culture
Subject = Aborigines
Subject = Academic libraries
Subject = Acquisitions (libraries)
Subject = Arts
Subject = Australian Library and Information Association
Subject = Books
Subject = Bookselling
Subject = Childrens libraries
Subject = Childrens literature
Subject = Colleges of technical and further education
Subject = Communications
Subject = Computer networks
Subject = Computers
Subject = Continuing professional education
Subject = Copyright
Subject = Culture
Subject = Educational institutions
Subject = Employment
Subject = Evaluation
Subject = Information
Subject = Information economy
Subject = Information literacy
Subject = Information management
Subject = Information technology
Subject = Inservice training
Subject = Internet
Subject = Internet and censorship
Subject = Internet and copyright
Subject = Internet and libraries
Subject = Internet and publishing
Subject = Intellectual property
Subject = Libraries
Subject = Library circulation
Subject = Library cooperation
Subject = Librarianship
Subject = Museums
Subject = National libraries
Subject = Periodicals
Subject = Public libraries
Subject = Publishing
Subject = School libraries
Subject = Special libraries
Subject = State and territory libraries
Subject = Statistics
Subject = Vocational education and training

Sets from @rchiveSIC

Others
Bibliometry, scientometry
Cinema, art, esthetics
Local authorities
Scientific communication and information
Conflicts, information strategy, intelligence
Information retrieval
Information/communication law
Economy, cultural industry
Electronic publishing
Education, e-learning, training
Public Sphere
Geopolitics
Knowledge management
History of information/communication
Hypertext, hypermedia
Information system engineering
Mass media
Museology
Organisation and communication
Sociology of information and communication
Theory of information/communication

Sets from CaltechLib

Status = In Press
Status = Published

Status = Submitted
Status = Unpublished
Subject = All Records
Subject = All Records: Housing technologies
Subject = All Records: Industry, profession and education
Subject = All Records: Information sources, supports, channels
Subject = All Records: Information treatment for information services (Information functions and techniques)
Subject = All Records: Information technology and library technology
Subject = All Records: Libraries as physical collections
Subject = All Records: Management
Subject = All Records: Publishing and legal issues
Subject = All Records: Technical services in libraries, archives and museums
Subject = All Records: Users, literacy and reading
Subject = All Records: Information use and sociology of information
Subject = All Records: Policy Documents
Subject = All Records: Theoretical and general aspects of libraries and information

Sets from DLIST

Status = In Press
Status = Published
Status = Unpublished
Subject = Academic Libraries
Subject = Artificial Intelligence
Subject = Anthropology
Subject = Archaeology
Subject = Archives
Subject = Bibliometrics
Subject = Cataloging
Subject = Citation Analysis
Subject = Classification
Subject = Co-citation Analysis
Subject = Cognitive Science
Subject = Communications
Subject = Computational Linguistics
Subject = Computer Science
Subject = Databases
Subject = Database Searching Instructions
Subject = Digital Libraries
Subject = Data Mining
Subject = Distributed Learning
Subject = Economics
Subject = Economics of Information
Subject = Epistemology
Subject = Electronic Publishing
Subject = Evaluation
Subject = Geographic Digital Libraries
Subject = Geography
Subject = Geographic Information Science
Subject = Human Computer Interaction
Subject = Hypertext and Hypermedia
Subject = Information Analysis
Subject = Information Extraction
Subject = Information Literacy
Subject = Indexing
Subject = Internet
Subject = Informetrics
Subject = Information Architecture
Subject = Information Ethics
Subject = Information Systems
Subject = Interdisciplinarity
Subject = Information Seeking Behaviors
Subject = Information Science
Subject = Journalism
Subject = Knowledge Management
Subject = Knowledge Organization
Subject = Knowledge Representation
Subject = Knowledge Structures
Subject = Libraries
Subject = Library Instruction
Subject = Library Statistics
Subject = Library Systems
Subject = Linguistics
Subject = Library and Information Science Education
Subject = Learning Science
Subject = Library Science
Subject = Map Librarianship
Subject = Media Studies
Subject = Metadata
Subject = Medical Libraries
Subject = Management
Subject = Management Information Systems
Subject = Museums
Subject = Neuroscience

Subject = Natural Language Processing
Subject = Ontology
Subject = Philosophy
Subject = Psychology
Subject = Qualitative Research
Subject = Quantitative Research
Subject = Reference Services
Subject = Research Methods
Subject = Scholarly Communication
Subject = Social Epistemology
Subject = Social Informatics
Subject = Sociology
Subject = Software
Subject = Standards
Subject = Science Technology Studies
Subject = Training
Subject = User Studies
Subject = Virtual Communities
Subject = Web Metrics
Subject = Web Mining
Subject = Wireless Technologies
Subject = World Wide Web
Subject = XML
Subject = Z39.50

Sets from E-LIS

Status = In Press
Status = Published
Status = Unpublished
Subject = A. Theoretical and general aspects of libraries and information.
Subject = A. Theoretical and general aspects of libraries and information. : AA. Library and information science as a field.
Subject = A. Theoretical and general aspects of libraries and information. : AB. Information theory and library theory.
Subject = A. Theoretical and general aspects of libraries and information. : AC. Relationship of LIS with other fields .
Subject = A. Theoretical and general aspects of libraries and information. : AZ. No one of these, but in this section.
Subject = B. Information use and sociology of information.
Subject = B. Information use and sociology of information.: BA. Use and impact of information.
Subject = B. Information use and sociology of information.: BB. Bibliometric methods.
Subject = B. Information use and sociology of information.: BC. Information in society.
Subject = B. Information use and sociology of information.: BD. Information society.
Subject = B. Information use and sociology of information.: BE. Information economics.
Subject = B. Information use and sociology of information.: BF. Information policy
Subject = B. Information use and sociology of information.: BG. Information dissemination and diffusion.
Subject = B. Information use and sociology of information.: BH. Information needs and information requirements analysis.
Subject = B. Information use and sociology of information.: BI. User interfaces, usability.
Subject = B. Information use and sociology of information.: BZ. No one of these, but in this section.
Subject = C. Users, literacy and reading.
Subject = C. Users, literacy and reading.: CA. Use studies.
Subject = C. Users, literacy and reading.: CB. User studies.
Subject = C. Users, literacy and reading.: CC. User categories: children, young people, social groups.
Subject = C. Users, literacy and reading.: CD. User training, promotion, activities, education.
Subject = C. Users, literacy and reading.: CE. Literacy.
Subject = C. Users, literacy and reading.: CF. Reading and story telling.
Subject = C. Users, literacy and reading.: CZ. No one of these, but in this section.
Subject = D. Libraries as physical collections.
Subject = D. Libraries as physical collections.: DA. World libraries.

Subject = D. Libraries as physical collections.: DB. National libraries.
Subject = D. Libraries as physical collections.: DC. Public libraries.
Subject = D. Libraries as physical collections.: DD. Academic libraries.
Subject = D. Libraries as physical collections.: DE. School libraries.
Subject = D. Libraries as physical collections.: DF. Government libraries.
Subject = D. Libraries as physical collections.: DG. Private libraries.
Subject = D. Libraries as physical collections.: DH. Special libraries.
Subject = D. Libraries as physical collections.: DI. Science libraries.
Subject = D. Libraries as physical collections.: DJ. Technical libraries.
Subject = D. Libraries as physical collections.: DK. Health libraries, Medical libraries.
Subject = D. Libraries as physical collections.: DL. Archives.
Subject = D. Libraries as physical collections.: DM. Museums.
Subject = D. Libraries as physical collections.: DZ. No one of these, but in this section.
Subject = E. Publishing and legal issues.
Subject = E. Publishing and legal issues.: EA. Mass media.
Subject = E. Publishing and legal issues.: EB. Printing, electronic publishing, broadcasting.
Subject = E. Publishing and legal issues.: EC. Book selling.
Subject = E. Publishing and legal issues.: ED. Intellectual property: author's rights, ownership, copyright, copyleft.
Subject = E. Publishing and legal issues.: EE. Intellectual freedom.
Subject = E. Publishing and legal issues.: EF. Censorship.
Subject = E. Publishing and legal issues.: EZ. No one of these, but in this section.
Subject = F. Management.
Subject = F. Management.: FA. Co-operation.
Subject = F. Management.: FB. Marketing.
Subject = F. Management.: FC. Finance.
Subject = F. Management.: FD. Public relations.
Subject = F. Management.: FE. Personnel management.
Subject = F. Management.: FF. Funding.
Subject = F. Management.: FG. Local government.
Subject = F. Management.: FH. Reorganization.
Subject = F. Management.: FI. Unitary authorities.
Subject = F. Management.: FZ. No one of these, but in this section.
Subject = G. Industry, profession and education.
Subject = G. Industry, profession and education.: GA. Information industry.
Subject = G. Industry, profession and education.: GB. Software industry.
Subject = G. Industry, profession and education.: GC. Computer and telecommunication industry.
Subject = G. Industry, profession and education.: GD. Organizations.
Subject = G. Industry, profession and education.: GE. Staff.
Subject = G. Industry, profession and education.: GF. Biographies.
Subject = G. Industry, profession and education.: GG. Curricula aspects.
Subject = G. Industry, profession and education.: GH. Education.
Subject = G. Industry, profession and education.: GI. Training.
Subject = G. Industry, profession and education.: GZ. No one of these, but in this section.
Subject = H. Information sources, supports, channels.
Subject = H. Information sources, supports, channels.: HA. Periodicals, Newspapers.
Subject = H. Information sources, supports, channels.: HB. Gray literature.
Subject = H. Information sources, supports, channels.: HC. Archival materials.
Subject = H. Information sources, supports, channels.: HD. Rare books and manuscripts.
Subject = H. Information sources, supports, channels.: HE. Print materials.
Subject = H. Information sources, supports, channels.: HF. Microforms.
Subject = H. Information sources, supports, channels.: HG. Non-print materials.

Subject = H. Information sources, supports, channels.: HH. Audio-visual, Multimedia.
Subject = H. Information sources, supports, channels.: HI. Electronic Media.
Subject = H. Information sources, supports, channels.: HJ. CD-ROM.
Subject = H. Information sources, supports, channels.: HK. Online hosts.
Subject = H. Information sources, supports, channels.: HL. Databases and database Networking.
Subject = H. Information sources, supports, channels.: HM. OPACs.
Subject = H. Information sources, supports, channels.: HN. e-journals.
Subject = H. Information sources, supports, channels.: HO. e-books.
Subject = H. Information sources, supports, channels.: HP. e-resources.
Subject = H. Information sources, supports, channels.: HQ. Web pages.
Subject = H. Information sources, supports, channels.: HR. Portals.
Subject = H. Information sources, supports, channels.: HS. Repositories.
Subject = H. Information sources, supports, channels.: HZ. No one of these, but in this section.
Subject = I. Information treatment for information services
Subject = I. Information treatment for information services: IA. Cataloging, bibliographic control.
Subject = I. Information treatment for information services: IB. Content analysis (A and I, class.)
Subject = I. Information treatment for information services: IC. Index languages, processes and schemes.
Subject = I. Information treatment for information services: ID. Knowledge representation.
Subject = I. Information treatment for information services: IE. Data and metadata structures.
Subject = I. Information treatment for information services: IF. Information transfer: protocols, formats, tecniques.
Subject = I. Information treatment for information services: IG. Information presentation: hypertext, hypermedia.
Subject = I. Information treatment for information services: IH. Image systems.
Subject = I. Information treatment for information services: II. Filtering.
Subject = I. Information treatment for information services: IJ. Reference work.
Subject = I. Information treatment for information services: IK. Design, development, implementation and maintenance
Subject = I. Information treatment for information services: IZ. No one of these, but in this section.
Subject = J. Technical services in libraries, archives, museum.
Subject = J. Technical services in libraries, archives, museum.: JA. Aquisitions.
Subject = J. Technical services in libraries, archives, museum.: JB. Serials management.
Subject = J. Technical services in libraries, archives, museum.: JC. Withdrawals.
Subject = J. Technical services in libraries, archives, museum.: JD. Stock taking.
Subject = J. Technical services in libraries, archives, museum.: JE. Record keeping.
Subject = J. Technical services in libraries, archives, museum.: JF. Paper preservation.
Subject = J. Technical services in libraries, archives, museum.: JG. Digitization.
Subject = J. Technical services in libraries, archives, museum.: JH. Digital preservation.
Subject = J. Technical services in libraries, archives, museum.: JI. Circulation.
Subject = J. Technical services in libraries, archives, museum.: JJ. Document delivery.
Subject = J. Technical services in libraries, archives, museum.: JK. Interlibrary loans.
Subject = J. Technical services in libraries, archives, museum.: JZ. No one of these, but in this section.
Subject = K. Housing technologies.
Subject = K. Housing technologies.: KA. Resources centers.
Subject = K. Housing technologies.: KB. Library, archive and museum buildings.
Subject = K. Housing technologies.: KC. Furniture.

Subject = K. Housing technologies.: KD. Vehicles.
Subject = K. Housing technologies.: KE. Architecture.
Subject = K. Housing technologies.: KF. Planning, Design, Removal.
Subject = K. Housing technologies.: KG. Safety.
Subject = K. Housing technologies.: KH. Disaster planning.
Subject = K. Housing technologies.: KZ. No one of these, but in this section.
Subject = L. Information technology and library technology.
Subject = L. Information technology and library technology.: LA. Telecommunications.
Subject = L. Information technology and library technology.: LB. Computer networking.
Subject = L. Information technology and library technology.: LC. Internet, including WWW.
Subject = L. Information technology and library technology.: LD. Computers.
Subject = L. Information technology and library technology.: LE. Scanners.
Subject = L. Information technology and library technology.: LF. Digital cameras.
Subject = L. Information technology and library technology.: LG. Photocopiers.
Subject = L. Information technology and library technology.: LH. Computer and network security.
Subject = L. Information technology and library technology.: LI. Authentication, and access control.
Subject = L. Information technology and library technology.: LJ. Software.
Subject = L. Information technology and library technology.: LK. Software mehtodolgies and engigneering.
Subject = L. Information technology and library technology.: LL. Automated language processing.
Subject = L. Information technology and library technology.: LM. Automatic text retrieval.
Subject = L. Information technology and library technology.: LN. Data base management systems.
Subject = L. Information technology and library technology.: LO. Object-oriented DBMS.
Subject = L. Information technology and library technology.: LP. Intelligent agents.
Subject = L. Information technology and library technology.: LQ. Library automation systems.
Subject = L. Information technology and library technology.: LR. OPAC systems.
Subject = L. Information technology and library technology.: LS. Search engines.
Subject = L. Information technology and library technology.: LZ. No one of these, but in this section.

References

Bertocco, Sara (2001). "Torii, an Open Portal over Open Archives" (online). *High Energy Physics Libraries Webzine*, Iss. 4. URL: <http://library.cern.ch/HEPLW/4/papers/4/>. Accessed: 2002-12-29.

Bollen, Johan et al (2005). "Trend Analysis of the Digital Library Community" (online). *D-Lib Magazine*, Vol. 11, Iss. 1. URL: <http://www.dlib.org/dlib/january05/bollen/01bollen.html>. Accessed: 2005-01-25.

Brogan, Martha L. (2003). "A Survey of Digital Library Aggregation Services" (online). URL: <http://www.diglib.org/pubs/brogan/>. Accessed: 2004-01-19.

Chapman, J.L. and Reiss, M.J. (1999). *Ecology : principles and applications*. Cambridge : Cambridge University Press. 330 s. ISBN 0-521-58802-2.

Cole, Timothy W. et al (2003). "Implementation of a Scholarly Information Portal Using the Open Archives Initiative Protocol for Metadata Harvesting" (online). URL: <http://oai.grainger.uiuc.edu/FinalReport/Mellon_FinalReport.doc>. Accessed: 2005-09-15.

Crow, Raym and Goldstein, Howard (2003). "Guide to Business Planning for Converting a Subscription-based Journal to Open Access" (online). URL: <http://www.soros.org/openaccess/oajguides/business_converting.pdf>. Accessed: 2003-01-31.

DCMI (2005). "DCMI Metadata Terms" (online). URL: <http://dublincore.org/documents/dcmi-terms/>. Accessed: 2005-09-21.

De Robbio, Antonella and Coll, Imma Subirats (2005). "E-LIS : an International Open Archive Towards Building Open Digital Libraries" (online). *High Energy Physics Libraries Webzine*, Iss. 11. URL: <http://library.cern.ch/HEPLW/11/papers/1/>. Accessed: 2005-09-22.

Enger, Magnus (2004). "An ecological look at scholarly documentation" (online). URL: <http://eprints.rclis.org/archive/00002945/01/spes-1.1.pdf>. Accessed: 2005-10-16.

Enger, Magnus (2005). "colLib.info : OAI-PMH meets Wiki" (online). *D-Lib Magazine*, Vol. 11, Iss. 11. URL: <http://www.dlib.org/dlib/november05/11inbrief.html#ENGER>. Accessed: 2005-11-16.

Fong, A. C. M.; Hui, S. C. and Vu, H. L. (2002). "Effective techniques for automatic extraction of Web publications". *Online Information Review*, Vol. 26, Iss. 1, p. 4-18.

Giles, C. Lee; Bollacker, Kurt D. and Lawrence, Steve (1998). "CiteSeer: An Automatic Citation Indexing System" (online). URL: <http://www.neci.nec.com/homepages/lawrence/papers/cs-dl98/latex.html>. Accessed: 2002-12-02.

Ginsparg, P. (1996). "Winners and Losers in the Global Research Village" (online). URL: <http://xxx.lanl.gov/blurb/pg96unesco.html>. Accessed: 2002-12-04.

Golder, Scott A. and Huberman, Bernardo A. (2005). "The Structure of Collaborative Tagging Systems" (online). URL: <http://www.hpl.hp.com/research/idl/papers/tags/tags.pdf>. Accessed: 2005-08-26.

Gustafsson, Thomas (2002). "Open Access - En empirisk undersökning om fritt tillgängliga vetenskapliga journaler på Internet" (online). URL: <http://oacs.shh.fi/publications/GraduGustafsson.pdf>. Accessed: 2004-09-24.

Hagedorn, Kat (2001). "OAIster : a "no dead ends" OAI service provider". *Library Hi Tech*, Vol. 21, Iss. 2, p. 170-181.

Hammersley, Ben (2003). *Content Syndication with RSS*. Beijing : O\'Reilly. ISBN 0-596-00383-8.

Hammond, Tony et al (2005). "Social Bookmarking Tools (I) : A General Review" (online). *D-Lib Magazine*, Vol. 11, Iss. 4. URL: <http://www.dlib.org/dlib/april05/hammond/04hammond.html>. Accessed: 2005-04-21.

Harnad, Stevan and Brody, Tim (2004). "Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals" (online). *D-Lib Magazine*, Vol. 10, Iss. 6. URL: <http://www.dlib.org/dlib/june04/harnad/06harnad.html>. Accessed: 2004-06-23.

Hitchcock, Steve (2002). "Open Citation Linking - The Way Forward" (online). *D-Lib Magazine*, Vol. 8, Iss. 10. URL: <http://www.dlib.org/dlib/october02/hitchcock/10hitchcock.html>. Accessed: 2002-12-04.

International DOI Foundation (2004). "Overview of DOI" (online). URL: <http://dx.doi.org/10.1000/202>. Accessed: 2005-01-25.

Jacsó, Péter (2004). "Link-enabled cited references". *Online Information Review*, Vol. 28, Iss. 4, p. 306-311.

Kling, R. and McKim, G. (2000). "Not just a matter of time : field differences in the shaping of electronic media in supporting scientific communication". *Journal of the American Society for Information Science*, Vol. 51, Iss. 14, p. 1306-20.

Kling, Rob (2004). "The Internet and unrefereed scholarly publishing". *Annual Review of Information Science and Technology*, Vol. 38., p. 591-631.

Kohrs, Arnd and Merialdo, Bernard (2001). "Creating user-adapted Websites by the use of collaborative filtering". *Interacting with Computers*, Vol. 13., p. 695-716.

Kumar, Anil and Kalyane, Venkatrao Lakshmanrao (2004). "Bibliographics for the 983 eprints in the live archives of E-LIS : trends and status report up to 7th July 2004, based on author-self-archiving metadata" (online). URL: <http://eprints.rclis.org/archive/00001927/>. Accessed: 2004-12-27.

Lagoze, Carl et al (eds) (2002a). "The Open Archives Initiative Protocol for Metadata Harvesting" (online). URL: <http://www.openarchives.org/OAI/openarchivesprotocol.html>. Accessed: 2005-01-20.

Lagoze, Carl et al (eds) (2002b). "Specification and XML Schema for the OAI Identifier Format" (online). URL: <http://www.openarchives.org/OAI/2.0/guidelines-oai-identifier.htm>. Accessed: 2005-01-20.

Lagoze, Carl et al (eds) (2005). "Conveying rights expressions about metadata in the OAI-PMH framework" (online). URL: <http://www.openarchives.org/OAI/2.0/guidelines-rights.htm>. Accessed: 2005-05-06.

Lagoze, Carl and Van de Sompel, Herbert (2003). "The making of the Open Archives Initiative Protocol for Metadata Harvesting". *Library Hi Tech*, Vol. 21, Iss. 2, p. 118-128.

Lawrence, Eleanor (1992). *Henderson's Dictionary of Biological Terms*. 10th ed.. Harlow : Longman Scientific & Technical. 645 s. ISBN 0-582-06433-3.

Lawrence, Steve (2001). "Online or Invisible?" (online). URL: <http://external.nj.nec.com/~lawrence/papers/online-nature01/>. Accessed: 2002-11-27.

Leuf, Bo and Cunningham, Ward (2001). *The Wiki way : quick collaboration on the Web*. Boston : Addison-Wesley. 435 s. ISBN 0-201-71499-x.

Liu, X.; Maly, K. and Zubair, M. (2002). "Enhanced Kepler Framework for Self-Archiving" (online). URL: <http://kepler.cs.odu.edu:8080/kepler/publications/kepler.pdf>.

Accessed: 2003-01-14.

Liu, Xiaoming (2002). "Federating heterogeneous digital libraries by metadata harvesting" (online). URL: <http://www.cs.odu.edu/~liu_x/paper/thesis/thesis.pdf>. Accessed: 2005-03-08.

Liu, Xiaoming et al (2001). "Arc - An OAI Service Provider for Digital Library Federation" (online). *D-Lib Magazine*, Vol. 7, Iss. 4. URL: <http://www.dlib.org/dlib/april01/liu/04liu.html>. Accessed: 2003-02-12.

Lund, Ben et al (2005). "Social Bookmarking Tools (II) : A Case Study - Connotea" (online). *D-Lib Magazine*, Vol. 11, Iss. 4. URL: <http://www.dlib.org/dlib/april05/lund/04lund.html>. Accessed: 2005-04-21.

Lund, Niels Windfeld (2001). "Omrids af en dokumentationsvidenskab : anno 2003". *Norsk tidsskrift for bibliotekforskning*, Iss. 16, p. 92-127.

Maly, K. et al (2002). "Archon : A Digital Library that Federates Physics Collections" (online). URL: <http://www.cs.odu.edu/~liu_x/paper/archon/archon.pdf>. Accessed: 2005-03-08.

Maly, K.; Liu, X. and Zubair, M. (2003). "Kepler Proposal and Design Document" (online). URL: <http://kepler.cs.odu.edu:8080/kepler/publications/finaldes.doc>. Accessed: 2003-01-14.

Maly, Kurt; Zubair, Mohammad and Liu, Xiaoming (2001). "Kepler - An OAI Data/Service Provider for the Individual" (online). *D-Lib Magazine*, Vol. 7, Iss. 4. URL: <http://www.dlib.org/dlib/april01/maly/04maly.html>. Accessed: 2003-01-14.

McKiernan, Gerry (1999). "Embedded multimedia in electronic journals". *Multimedia Information and Technology*, Vol. 25, Iss. 4, p. 338-343.

McKiernan, Gerry (2000). "arXiv.org : the Los Alamos National Laboratory e-print server". *International Journal on Grey Literature*, Vol. 1, Iss. 3, p. 127-138.

McKiernan, Gerry (2001). "New Age E-Journals, Indexes, and Services" (online). URL: <http://www.public.iastate.edu/~gerrymck/EJIS.pdf>. Accessed: 2004-05-13.

McKiernan, Gerry (2002). "E is for Everything: The Extra-Ordinary, Evolutionary [E-] Journal" (online). URL: <http://www.public.iastate.edu/~gerrymck/SLv41n3-4.pdf>. Accessed: 2002-12-20.

McKiernan, Gerry (2003a). "Open Archives Initiative Service Providers : Part I: Science and

Technology" (online). URL: <http://www.public.iastate.edu/~gerrymck/OAI-SP-I.pdf>. Accessed: 2004-01-19.

McKiernan, Gerry (2003b). "Open Archives Initiative Service Providers : Part II: Social Sciences and Humanities" (online). URL: <http://www.public.iastate.edu/~gerrymck/OAI-SP-II.pdf>. Accessed: 2004-01-19.

McKiernan, Gerry (2004). "Open Archives Initiative Service Providers : Part III: general". *Library Hi Tech News*, Iss. 1, p. 38-46.

McKiernan, Gerry (2005). "WikimediaWorlds : Part I: Wikipedia" (online). URL: <http://www.public.iastate.edu/~gerrymck/WMW-I.pdf>. Accessed: 2005-11-07.

Medeiros, Norm (2004). "A repository of our own : the E-LIS e-prints archive". *OCLC Systems & Services*, Vol. 20, Iss. 2, p. 58-60.

Roy Tennant (2005). "Current Cites : What a Long, Strange Trip It's Been" (online). *D-Lib Magazine*, Vol. 11, Iss. 9. URL: <http://www.dlib.org/dlib/september05/09inbrief.html#TENNANT>. Accessed: 2005-10-17.

Scirus (2004). "Scirus White Paper" (online). URL: <http://www.scirus.com/press/pdf/WhitePaper_Scirus.pdf>. Accessed: 2005-09-14.

Smith, John W. T. (1997). "The Deconstructed Journal" (online). URL: <http://library.kent.ac.uk/iccc/1997/papers/deconjnl.htm>. Accessed: 2005-01-25.

Smith, John W. T. (2004). "The Deconstructed (or Distributed) Journal : an emerging model?" (online). URL: <http://library.kent.ac.uk/library/papers/jwts/OI04.html>. Accessed: 2005-01-25.

Sollins, K. and Masinter, L. (1994). "Functional Requirements for Uniform Resource Names" (online). URL: <http://www.ietf.org/rfc/rfc1737.txt>. Accessed: 2005-11-11.

Sompel, Herbert Van de and Beit-Arie, Oren (2001). "Open Linking in the Scholarly Information Environment Using the OpenURL Framework" (online). *D-Lib Magazine*, Vol. 7, Iss. 3. URL: <http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html>. Accessed: 2003-01-05.

Suber, Peter (2005a). "Archived postprints should identify themselves" (online). *SPARC Open Access Newsletter*, Iss. 85. URL: <http://www.earlham.edu/~peters/fos/newsletter/05-

02-05.htm#brand>. Accessed: 2005-06-14.

Suber, Peter (2005b). "Timeline of the Open Access Movement" (online). URL: <http://www.earlham.edu/~peters/fos/timeline.htm>. Accessed: 2005-01-25.

Suleman, Hussein (2002). "Open Digital Libraries" (online). URL: <http://scholar.lib.vt.edu/theses/available/etd-11222002-155624/>. Accessed: 2005-01-06.

Summann, Friedrich and Lossau, Norbert (2004). "Search Engine Technology and Digital Libraries - Moving from Theory to Practice" (online). *D-Lib Magazine*, Vol. 10, Iss. 9. URL: <http://www.dlib.org/dlib/september04/lossau/09lossau.html>. Accessed: 2004-09-10.

Summers, Ed (2004). "Building OAI-PMH Harvesters with Net::OAI::Harvester" (online). *Ariadne*, Iss. 38. URL: <http://www.ariadne.ac.uk/issue38/summers/>. Accessed: 2004-02-03.

Tajoli, Zeno (2005). "METALIS, an OAI Service Provider" (online). URL: <http://eprints.rclis.org/archive/00003612/01/art_creta1.pdf>. Accessed: 2005-09-13.

Taubes, Gary (1993). "Publication by electronic mail takes physics by storm". *Science*, Vol. 259., p. 1246-1248.

Van de Sompel, Herbert et al (2004). "Resource Harvesting within the OAI-PMH Framework" (online). *D-Lib Magazine*, Vol. 10, Iss. 12. URL: <http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html>. Accessed: 2004-12-16.

Van de Sompel, Herbert; Young, Jeffrey A. and Hickey, Thomas B. (2003). "Using the OAI-PMH ... Differently" (online). URL: <http://www.dlib.org/dlib/july03/young/07young.html>. Accessed: 2003-08-22.

Ward, Jewel Hope (2002). "A quantitative analysis of Dublin Core metadata element set (DCMES) usage in data providers registered with the Open archives initiative (OAI)" (online). URL: <http://foar.net/research/mp/Jewel_Ward-MPaper-November2002.pdf>. Accessed: 2004-06-09.

Ward, Jewel Hope (2004). "Unqualified Dublin Core usage in OAI-PMH data providers". *OCLC Systems & Services*, Vol. 20, Iss. 1, p. 40-47.

Acknowledgements

My most sincere thanks are due to

- **Per Bäckström** for being the supervisor on this project.
- **Niels Windfeld Lund** for inventing the “Science of Documentation” at the University of Tromsø, Norway.
- **Øyvind Bjørkås, Per Olav Bøyum and Jørn Hjørungnes** for reading and commenting on drafts at different stages of completion.
- **Elin** for endless support, understanding, encouragement, draft-reading and waffles.
- **Rosa & Benoni** for being friendly and furry, and insisting on a routine of daily mid-day walks.
- The developers and maintainers of the **OAI PKP Harvester** and the **MediaWiki**, for releasing their software under a licence that allowed me to build on their work.
- The developers and maintainers of **Linux**, **Apache**, **MySQL** and **PHP**, for creating tools that are open and useful.
- The developers and maintainers of the **OpenOffice.org** office-suite, with which this thesis was created.

Without these outstanding individuals the present work would never have been what it is today. The responsibility for all errors and short-comings rests, of course, solely with the author.