The Consortium of European Research Libraries

Searching Facility for Manuscripts & Hand-Press Book Catalogues

March 2003

Radcliffe Interactive
The Old Malt House
Paradise Street
Oxford, OX1 1LD

Tel: +44 (0) 1865 790 446
Fax: +44 (0) 1865 201 370
email:
peter.ashby@radcliffe-interactive.com

CONTENTS

1 Introduction
2 Terms of reference
3 Definition of the problem
4 Current online projects
5 Conceptual framework
6 Encoding standards
    MARC
    EAD
    TEI
    MASTER and EAMMS
7 Search engine architecture
    Open Archives Initiative
    RLG
    DFAS
    LOCKSS
    The European Library
8 Communication protocols
    Z39.50/ ZING
    Open URL
9 Proposed CERL system
10 Outstanding issues
    User requirements
    Institutional perspectives
    Common Access Points
    Management and logistics
11 Next steps
    Final investigations
    Pilot/proof of concept
    Implementation
12 Indicative schedule

13 Supplement to the March 2003 Report – Potential Technical Partners with Costs

Appendix A – Sources
Appendix B – Articles and papers online
Appendix C – Weblinks
Appendix D – Distributed searching diagrams
Appendix E – Example of XML Encoding
Appendix F – Illustration of minimal vs. maximal encoding from MASTER
Appendix G – Proposed CERL manuscript access system diagram
Appendix H – Appendix H : Article by Mary Jackson on the Advent of Portals

Glossary

CERL Catalogue Searching Facility for Manuscripts & Hand-Press Book Catalogues

1. Introduction

The Consortium of European Research Libraries has for some years recognised that its experience with printed materials of the hand-press period (that is, from 1455 to c. 1830) might be extended to manuscript materials. The Hand-Press Book database, which at present gives unified access to some million and a half records contributed by seventeen major libraries, has proved its value by being available to a wide range of users who indeed make use of these materials for a great variety of purposes.

At the end of 1999 the Consortium's members decided to explore whether similar unified access to manuscript materials would be desirable and feasible, thus aiming to provide access in a single system to as much of the written heritage of the Western world, manuscript and print, as is at present available through a variety of databases.

Successive surveys and reports led to the decision to commission a technical report from Radcliffe Interactive, in order to make the best possible use of what rapidly developing technology can offer.

The present report provides a general survey, and then selects several options as suitable for what the Consortium, its members and an ever widening circle of users may come to expect of such a new research tool. The report is therefore not only intended to set out the options for the Consortium's management and its members, but also for wider dissemination as a discussion document for potential end-users, especially in the academic world.

Before embarking on technical detail it may therefore be useful to relate briefly the several clear directions we have already received in the course of our explorations:

a. Not to impose a date limit on the material to be included in order to parallel the practice of the largest libraries providing automated access.

b. To give access to records in a variety of cataloguing formats, and at a variety of levels.

c. To give access to both manuscript and printed material, if requested by the user.

d. To aim to restrict to a minimum the efforts required from contributing institutions, and certainly not to be prescriptive in any way.

e. Access to be provided through common access points, to be agreed, which are to include the standard 'bibliographical' information as well as provenance information, bindings where applicable, and selected data relevant to manuscript material.

The benefit of such an indexing system that gives immediate access to any further detailed information contained in the contributing databases is, in one word, contextualization. Any one object (or selected set of objects) whether searched by author, text, place of origin, date, producer (whether a scriptorium, an artist, a printer or a publisher) or the history of ownership, can be placed within a range of other dispersed over the world. One of the important functions of combining manuscript and print may in due course lie in a large-scale instrument for the virtual reconstruction of collections in earlier periods. An indispensable support in making these data accessible, in spite of being recorded in a variety of traditions and languages, will be the CERL Thesaurus system that has been successfully developed to meet such requirements.

The proposed system is a new concept, in that, by providing direct access to sophisticated data, it is more than a finding list, a simple index or even a short-title catalogue. In the twentieth century, the contextualization made possible by the existence of, for example, P. O. Kristeller's *Iter Italicum*, or the short-title catalogues of printed books, first published in book-form and later, dramatically more useful, in automated form, changed the methodology of scholarship. Whatever use scholars of the future will make of the new instrument – and there should be a great variety of uses – we can be confident that it will have a profound effect on the way the written heritage is approached and interpreted.

Lotte Hellinga

2. Terms of reference

Considerable work has been carried out in Europe and the USA to provide access to manuscript information online, including the development of union catalogues (e.g. in France, Germany. Italy and Portugal), and the creation of common standards and technical routes for cross-collection searching (e.g. MASTER, MALVINE).

CERL is exploring the feasibility of creating a technical solution which would give researchers a single point of access to these online resources, and has commissioned this initial report in order to define the technological issues and to begin to map out the best way forward, in terms of timing, cost and future scalability.

Logically, two options present themselves: a union catalogue and a distributed system.

Work carried out by CERL to date indicates that the creation of a single, central bibliographic database in the manner undertaken for the HPB database is not desirable because of the disparity of cataloguing standards and the very considerable demand on resources that would be required. A distributed search system was therefore mooted, which would carry out searches across the various online resources rather than through a central database. There are technical challenges posed by this type of solution, which are considered fully in this report, but in any case while CERL expects to sponsor and manage a searching system it does not wish to become a technology provider – something for which it has neither the expertise nor the resources. CERL will however incorporate The Hand Press Book database (HPB) which covers printed material from the mid-fifteenth to the mid-nineteenth century, as well as the CERL Thesaurus (CT), in the planned developments set out here. The proposed integration of access between manuscripts from all periods with printed materials will be of great scholarly importance.

This report was therefore required to: consider CERL's initial contributions by way of the HPB and CT as part of the major databases that are to be cross - searched; consider the pros and cons of existing search systems; explore any likely solution(s); consider implementation issues; provide a preliminary costing.

3. Definition of the problem

Manuscript material falls into two categories:

(1) Automated catalogue records and other finding aids giving access to collections

(2) Digital images made available by institutions as substitutes or as aids for identification of particular items , either within an institution or compiled in a project across institutions (e.g. Digital Scriptorium , IRHT in Paris)

Scholars are primarily interested in access to the manuscripts either to originals by using on-line finding aids (automated indexes, automated catalogue records), and/or by obtaining online access to digital substitutes. An image can provide the scholar with more detailed information than any catalogue description, but, in order to be accessible it still requires systematic metadata on the lines of catalogue records that have to be integrated into an index system.

Substitutes in digital form are not available for the vast majority of manuscripts, and will not be in any short to medium term timescale. Large finding aids are already available online, and the process of making more available is currently a priority in many major research libraries and institutions. CERL therefore recommends a concentration on available finding aids both of catalogue records and of metadata giving access to digital images. This approach is set out in Lotte Hellinga's reports of October 2001 and November 2002, which discuss the need for Common Access Points to existing databases.

The issue therefore becomes one of defining the best technical domain capable of delivering a search across bibliographic metadata extruded through Common Access Points possibly in combination with, and alongside more extensive automated finding aids and manuscript catalogues. Given the above, the approach adopted in this report has been as follows:

- to survey the "landscape" in order to define the major technical issues affecting the implementation of the facility required by CERL;
- to define an overall conceptual framework for the desired solution;
- to consider both previous projects and current, largely Internet based developments in order to define the likely nature of a solution;
- to examine potentially relevant technologies and methodologies;
- to discuss this solution with several key experts as an initial "reality check"

4. Current online projects

Gordon Dunsire's survey on behalf of CERL of existing manuscript projects clearly shows that there are currently no common standards for the storing and retrieval of bibliographic data relating to manuscripts. For example: EAMMS uses specialized versions of MARC and SGML/XML under the TEI aegis, with recognition to MASTER; the Bodleian Library uses an XML encoded EAD format; the Bibliothèque Sainte-Geneviève in Paris uses a unique cataloguing system; and Bibliothèque Nationale de France is running a project that is creating HTML catalogue pages. (See the Glossary for an explanation of the acronyms). Even where common "standards" such as MARC are used they may vary significantly from one implementation to another.
Gordon Dunsire lists a number of initiatives designed to make sense of this diversity. Some are cross-collection "gateways" or portals, while others (such as MASTER) attempt to establish a common standard for representing the bibliographic data. Many of the gateways are experimental or short-term projects. None have aimed at being comprehensive and there is no reason to believe that they will become so.

As a concrete example, the Manuscript Studies Portal website currently under development at the University of London Library aims to provide an online research environment. One element will be some form of cross-catalogue access (see http://www.palaeography.ac.uk/portalplan11.doc). However, work during 2003 will focus on the creation of the web environment and assessment of a candidate search engine following which further funding will be required. Moreover the plan talks of providing "links to library catalogues" which is far from being international searching of the kind envisaged by CERL. In fact this and other gateways would benefit from being able to link in to the CERL search facility.

On a wider scale, there are several major collaborative projects in various European countries and North America which have resulted in union catalogues, together with large individual electronic cataloguing projects in major national collections such as the British Library. Some of these catalogues are searchable online. Yet there is no harmonisation between them – in each case the institutions have developed systems which remain under their own control, are designed to meet the perceived needs of their own academic community and are adapted to the particular nature of the collections available to the institutions.

5. Conceptual framework

In order to take best advantage of the many manuscript projects already taken forward by institutions, it was thought best **not** to attempt to develop an all-encompassing union catalogue. A union catalogue has many advantages but it is prescriptive and is very costly in terms of both finance and resources. CERL has highly relevant experience of this with the creation of the HPB, and creating a similar facility for manuscripts would seem to be even harder.

A more realistic approach points to only two possible types of model for consideration: distributed searching or centralised index. These are shown diagrammatically in Appendix B and explained below.

In **distributed searching** the data is located on multiple databases and is searched by clients using standard protocols such as Z39.50, ZING/SRW or XML Query. There is no centralisation of the data.

A central search engine sends a user's search request to every participating online database, having first translated it so that it can be understood by each database. The remote databases carry out their searches locally and return the results to the central search engine. The central engine waits for all the results to come in then it checks the quality of the incoming responses, groups apparent duplicate hits and then sorts and displays the results.

Clifford Lynch of the Coalition for Networked Information, in a review of the Open Archives Initiative (see Appendix A) notes that "there are scaling problems in the management of searches that are run at large numbers of servers; one has to worry about servers that are unavailable (and with enough servers, at least one always will be unavailable), and performance tends to be constrained by the performance of the slowest individual server participating in the federation of servers. In these circumstances the user has to wait for a lot of record transfer and post-processing before seeing a result, making… federated search performance sensitive to participating server response time, result size, and network bandwidth." Other experts voice similar views. These problems are ameliorated but not removed by the adoption of alternative protocols to Z39.50.

There is the further problem that some smaller catalogues do not support online searching – some are not even in database format.

Distributed search works best among a limited number of large, powerful databases. It is not a suitable solution for a cluster of highly heterogeneous systems such as we find with manuscript collections and the HPB.

The creation of a **centralised index** has much to recommend it. It is emphatically *not* the same as a union catalogue in that it provides a lowest common denominator view of the information available in the remote systems. Catalogue data from each remote system is "harvested", then converted into a single standard format and incorporated into a single database which provides Access Points back into the original catalogues.

This is the methodology chosen by the Open Archives Initiative, albeit for providing central access to repositories of archival records rather than catalogues. Lynch states that the Santa Fe Convention – which was the starting point for the OAI – "recognized that every repository housed metadata, and so they devised a very simple way for repositories to export this metadata on demand; service developers would then take the responsibility for actually collecting, or "harvesting," this metadata… A user querying a federated search service would not interact with the repositories, but only with a database that the federated search service had already constructed from metadata harvested from participating repositories, for example. Hence the performance of the federated search service was largely independent of the performance or reliability of the participating repositories. The design goal was that a repository should be able to implement the Santa Fe Convention with a few days of programmer time, as opposed to months."

Elizabeth Shaw, speaking from her experience in indexing data of diverse origin (multi-institutional database and encoded prose) for the Digital Scriptorium, comments that OAI provides a model, although probably not a set of technologies, for the preparation of a central index by CERL. But she believes that the only way that a central index can be effective in this idiosyncratic and often chaotic arena is if it provides **an acceptable minimum of Access Points but no more.**

Clearly an index needs to be an index into *something*. If the catalogue entries are then also digitised through re-keying or data captured as formatted text and cross-linked to the index, then a user could experience much of the functionality of a full union catalogue.

The example in Appendix E, of a bibliographic entry with an average level of mark-up may help to assess the problem. The key elements at this level indicate the sort of Access Point structure needed by CERL, and when linked to the manuscript description itself probably represents a reasonable functionality. The gains to be made by a higher level of mark-up would be far outweighed by the enormous problems of achieving consistency across many different catalogues.

This low-impact approach seems to fit CERL's criteria closely. Further, if the harvesting process is extended, through re-keying of data, to include the capture of currently non-electronic or highly idiosyncratic material, it has the potential to be more inclusive than distributed searching.

Janifer Gatenby's paper (Appendix B) discusses research by The European Library project (TEL) into the canonical forms of searching outlined in this section. TEL has also identified the "virtual union catalogue", which is discussed in the paper by Clifford Lynch presented to the Caslin Conference in 1999. These works provide a good background to the arguments presented here.

The DFAS project which is considered later, examined a "replicated search" model, where a central index is locally replicated among participating institutions who can use it alongside their own catalogue search engines. The DFAS team felt that "replication has significant value where the institution that has created finding aids is not able to mount an effective system for retrieval, or where consortia interests are strong”. This clearly has relevance for CERL.

Subject to further discussion later in this report, we therefore take as the overall conceptual framework for the CERL solution that it should be based around searching of a central index created – in some manner to be determined – from data gathered from the individual catalogues rather than a distributed search system or union catalogue. **CERL needs to consider the best mechanism for making copies of the index available to participating institutions.**

6. Encoding standards

Even starting from the most agnostic basis as to the technical architecture of the eventual solution, it is clear that if it is to be based around a central index it must involve normalisation of metadata.

Normalisation means the rendering of heterogeneous data into a common encoding standard. Since, as discussed in section 3, there is no conformity between the collections, any cross-collection searching must involve mappings between the various finding aids and whatever common standard the CERL searching facility implements. That is, the relevant information from each catalogue that is to be used in creating the Access Points has to be extracted and transformed into a single, common format. There has been considerable work in this area over many years and it is possible to be more certain how it can be done in the case of the CERL facility. Defining the manner in which searching should then be carried out is less clear because there are several possible technical architectures with much to recommend each of them. There are relatively few standards that are directly relevant to the question.

a. MARC

Most electronic library catalogues in Europe and America use MARC encoding for printed matter, and MARC bibliographic records are widely accessible through local and national databases. Libraries with MARC-based cataloguing systems can be expected to maintain them for the foreseeable future. However, MARC was not designed for cataloguing manuscripts from all periods and the fields have to be redefined or expanded to allow accurate encoding. For the development to AMREMM see below p. 11. MARC is extremely cumbersome from the point of view of creating a specialised cross-collection searching system. While the system must of course cope with the fact that many records are in MARC it should not use MARC in its underlying data structures. Records will have to be translated out of MARC and into whatever format the CERL system expects.

b. EAD (Encoded Archival Description)

The EAD Document Type Definition (DTD) is a standard for encoding archival finding aids using SGML and XML. EAD was devised primarily to describe the organization of series of documents, at the collection level rather than the individual manuscript that is a book. It could be said therefore that it is primarily aimed at collections of modern manuscripts. It is a non-proprietary encoding standard for machine-readable finding aids such as catalogues, inventories, registers, indexes, and other documents created by archives, libraries, museums, and manuscript repositories to support the use of their holdings. EAD elements include, for example:

- Archival description
- Description of subordinate elements
- Adjunct descriptive data (bibliography, indexes, etc.)
- Controlled access points (names, title, subject, etc.)
- Pointer, reference & linking elements (URL, etc.)

Some of the EAD elements will be of value to the CERL index, and EAD is very widely used. However, the DFAS project (see below) identified a lack of standardization in the way EAD is applied as a possible source of difficulties in developing a cross-collection searching facility, especially as EAD is designed to support a very dense description of items, sometimes nested many levels deep. **The findings of the DFAS project therefore stress specifically the importance of reaching consensus amongst the participating institutions as to how the EAD elements are applied.**

c. TEI (Text Encoding Initiative)

The TEI is an international SGML/XML standard that helps libraries, publishers, museums and scholars encode text for digitising literary and linguistic texts. TEI guidelines provide a means of tagging electronic text to mark its structure and other features of interest.

While it may sound as if this is only of value where the documents themselves have been digitised, TEI is also designed to support document metadata. The TEI Header provides a basic structure for bibliographic information but it is general purpose and does not, of itself, provide a bibliographic structure that is suitable for use in the CERL searching facility.

d. MASTER (Manuscript Access through Standards for Electronic Records) and EAMMS (Electronic Access to Medieval Manuscripts)

EAMMS was a project to develop guidelines for encoding and storing catalogue descriptions of medieval and Renaissance manuscripts in electronic form – it provides a template for encoding descriptions of manuscripts, and suggestions for composing electronic finding aids for manuscript repositories. EAMMS is linked to the Digital Scriptorium and MASTER in its SGML/XML implementation; its MARC implementation, termed AMREMM, is also discussed further on.

MASTER was funded by the EU from 1999-2001 in order to provide to provide a singe international standard for manuscript records, based on the TEI. MASTER concludes with a small pilot project and report yet to be evaluated.

MASTER is being used in several large European projects, primarily as a normalised structure for data transfer between different database systems. However the principal manager of MASTER Peter Robinson, notes that each institution is customising the MASTER data structure for its own purposes. He suggests that a (customised) subset of the MASTER DTD is probably suitable for CERL's normalisation requirements.

There was a considerable degree of cross-fertilisation between MASTER and the TEI Workgroup on manuscript descriptions (whose members include Consuelo Dutschke and Lou Burnard), resulting in a joint DTD (http://www.tei-c.org.uk/Master/Reference). The state of play now is that the TEI Council is calling for a comparison between the TEI Working Group's dtd with that of the MASTER dtd to see how/if the two should be harmonized.

Dr. Dutschke points out "EAMMS is committed to devising standards for the cataloguing (essentially intended as on-line) of medieval and renaissance manuscripts across a broad spectrum of national traditions. Implementation of the cataloguing is two-fold in method: MARC and SGML/XML under the TEI aegis. The MARC work was carried out by Dr. Gregory Pass, Director, Vatican Film Library, St. Louis University, St. Louis Missouri, under the heading "AMREMM," which is an anagram for Ancient, Medieval, Renaissance and Early Modern Manuscripts. AMREMM has received full approval from the American Library Association and is being printed as a supplement to AACR2. The second prong of the EAMMS project is to develop for the same purpose a dtd originally in SGML, but by now in XML; to this end, EAMMS sponsored a TEI Working Group, co-chaired by Dr. Dutschke and Ambrogio Piazzoni (Vatican Library). The TEI Working Group functioned in cooperation with the MASTER project, as long as that group continued to develop its dtd. Presently, the TEI Working Group is finishing its dtd and current TEI plans call for an examination of the TEI dtd and its subset, the MASTER dtd, to note points of difference, in order to eventually support one dtd."

In summary, legacy computer-based systems have implemented MARC and will continue to use it, but tremendous energy is going into SGML and XML implementations. EAD provides many valuable elements but is not always applied in a standardised manner. Elizabeth Shaw argues that TEI and MASTER contain a massive superset of the necessary elements. She also notes that there has been some recent work by the Digital Scriptorium which may be of relevance and should be investigated further. TEI, EAMMS and MASTER are XML-based and indicate the emergence of a common viewpoint on encoding standards for manuscripts and manuscript metadata. Given also that XML is platform-neutral and is the emerging standard for data transfer – and will be for the foreseeable future – the CERL facility should look to the work of TEI and MASTER for an approach to its data structures.

One point cannot be stressed too strongly: there can be no question of CERL prescribing a common bibliographic standard for institutions to adopt, as most are already committed to their particular way of doing things. And in the light of the HPB it should be clear that **any normalisation solution must make as small an impact on institutional resources as possible**.

7. Search engine architecture

Defining a suitable architecture for the CERL system is considerably more difficult than defining an encoding standard. Each of the possible models has something to recommend it and the final decision will be based as much on management criteria such as cost, logistics, resourcing and control as on technical criteria and extensibility. At this stage therefore we will only look at key initiatives that help to define the main issues.

a. OAI (Open Archives Initiative)

The OAI was introduced earlier in this report. It is now becoming relatively widely used. It must be understood that the OAI is about access to repositories of primary materials which have a certain degree of homogeneity, not to extremely heterogeneous collections of finding aids. On the other hand, the OAI states that:

> "The roots of the OAI lie in the E-Print community, which promotes and maintains web-accessible archives of scholarly papers as a means of increasing access to scholarly research. Initial work in the OAI was motivated by a desire to develop interoperability frameworks for federating E-Print archives. It soon became evident, however, that the concepts in the OAI interoperability framework - exposing multiple forms of metadata through a harvesting protocol - had applications beyond the E-Print community. Therefore, the OAI has adopted a mission statement with broader application: opening up access to a range of digital materials."

Although OAI uses Dublin Core (see Glossary) it applies equally well to other XML schemas. Robust OAI software is in use in many institutions. OAI has thus created a large base of experience on harvesting, normalising and combining data. **CERL can benefit from this experience.**

b. RLG (Research Libraries Group, Inc.)

The Archival Resources database developed by RLG adopts the model of a union catalogue of finding aids. It combines summary collection descriptions and finding aids together with a single search and display interface that attempts to accommodate all the various local practices. Writing in D-Lib Magazine about the DFAS project, January 2000, MacKenzie Smith of Harvard University Library points out the limitations of the RLG Model:

> "The Archival Resources database developed by RLG had begun exploring the union catalog approach by combining summary collection descriptions and finding aids together with a single search and display interface that attempted to accommodate all the various practices. While the RLG system works very well, we felt the model had inherent limitations for accommodating local practice in a scalable way. In order to accommodate institutional variation, the system is forced to handle each new case for both indexing and display programs. Beyond a few dozen institutions, this could become quite difficult to accomplish and maintain over time."

Given the need to include information from many more than a dozen institutions, CERL's varied membership, its relatively limited resources, and the overall experience of the HPB, scaleability is clearly an important aspect which CERL must bear in mind. At this stage of investigation it appears that the RLG system is probably not suitable. Nevertheless a more considered analysis taking soundings from institutions already participating in the RLG system, needs to be made once the details of the proposed CERL system have been agreed.

c. DFAS – the Distributed Finding Aid Server.

During 1998-99, the Digital Library Federation underwrote a project to develop an automated system for distributed online searching of EAD-encoded finding aids. The participating institutions were Harvard University, the University of Michigan, Columbia University, Indiana University and Oxford University. The Distributed Finding Aid Server project (DFAS), was completed in July 1999.

DFAS adopted a combination of the distributed searching and normalisation models by bringing together online finding aid catalogues developed at multiple institutions and tailored to each institution's finding aid structure and mark-up practices, while displaying search results from the different systems using mappings mutually agreed between the institutions. DFAS had great difficulty in creating these mappings because of the considerable diversity of practice between institutions.

DFAS looked at both distributed searching of remote sites and local search of replicated remote catalogues. The DFAS team concluded that true distribution of searching is preferable from a methodological point of view, but that from an architectural viewpoint, replication has many advantages.

DFAS investigated many areas of considerable importance to the CERL project including:

- techniques for tackling the diversity of practice in mark-up of finding aids;
- CAP implementation;
- methodological and technological difficulties in implementing distributed versus locally replicated index searching;
- management of extremely large finding aids and persistent result sets;
- navigation of finding aids;
- implementation of standardised “middleware” software;
- “web crawler” synchronization of replicated indexes;
- cost factors

d. LOCKSS (Lots of Copies Keep Stuff Safe)

The LOCKSS model originated at Stanford University. LOCKSS (“Lots of Copies Keeps Stuff Safe”) creates digital “caches” of e-journal content housed locally at participating institutions. Accuracy and completeness of LOCKSS caches is assured through a robust and secure polling system – basically, the databases talk to each other to synchronise themselves.

There is relevance to CERL's requirements, particularly as regards distribution, local autonomy and ease of updating. We originally considered the LOCKSS model as a front runner but have come to realise that the update frequency of the CERL index will be probably not be so high that a complex technology to synchronise replicated indexes will be necessary (this also applies to OAI and RLG). It should be borne in mind however, in case more detailed study shows that updating and local control are significant issues and render any solution based on this approach unsuitable.

e. TEL (The European Library)

The European Library is an EU-funded 'virtual library' which will allow users to search for, and access, digital and other collections from all the participating libraries. The TEL WP3 and WP4 teams have investigated metadata issues and search architectures.

There are strong similarities with the work of both OAI and DFAS – metadata is harvested and normalised into an extensible format based on Dublin Core, and searching can be either distributed or through a central index. The TEL test portal (see link in Appendix C) allows for searching of a central index. It is thus the closest example to an existing system that matches our conceptual model for CERL's system. Note that this is only a test system - it is not meant for end users but for testing purposes and it therefore shows the raw data fields rather than a neat and tidy display.

In discussions with Theo van Veen of TEL it became clear that TEL is unfortunately a long way from being able to provide a system that meets CERL's requirements. One issue of significant concern is that the TEL metadata structure is not only general purpose but also flat. It therefore does not appear at this stage to support the particular requirements for access points outlined by Lotte Hellinga. TEL is currently geared up for large, frequently updated collections rather than specialist needs.

If TEL succeeds in developing a pan-European framework for library access it will certainly provide a framework within which CERL's system might sit in a few years time. CERL's activities in developing its search facility should therefore be informed by the ongoing work at TEL.

Summary: The experience of DFAS and TEL points to the RLG solution as being inappropriate to CERL's requirements because it take the approach of a centralized union catalogue of finding aids and is unlikely to be scalable for the diversity of institutions and material CERL needs to consider.

A peer-to-peer system similar to LOCKSS is probably unnecessary but is worth bearing in mind. OAI, DFAS and LOCKSS represent a valuable corpus of reference on how to implement a peer-to-peer access system and between them have tackled most of the important issues.

8. Communication protocols

It would be premature at this stage to specify the communication protocols that the CERL system will eventually use. However, two protocols crop up frequently in discussion and need to be mentioned.

Z39.50 / ZING

Z39.50 is very widely supported but is now twenty years old and is not coping well with the Internet. ZING – "Z39.50-International: Next Generation" – refers to a number of initiatives by the Z39.50 community to build on the existing base of experience in order to make Z39.50 more attractive by lowering the barriers to implementation.

ZING includes protocols such as SRW and SRU which allow Z39.50 to work more easily across the Web.

The CERL system should probably include a ZING interface of some sort so that it could be linked in to participating institutions' searching systems. The system should also have a Web-based search screen. The TEL portal combines both of these – the Web screen translates queries into SRU, which is the format required by the search engine.

OpenURL

OpenURL is an important technology in the development of shared, open, reference linking services. It extends "traditional" URLs by adding extended information about a bibliographic reference that can provide a user with access to an online record, or to various extended services.

For example, if a user of the CERL system sees that a search "hit" mentions the existence of an online transcription, will the user be allowed access to that resource? An ordinary URL will simply point at the resource, which may be unavailable to unauthorised users. The OpenURL on the other hand can carry extended information that will help in determining whether the CERL user should be given access, so perhaps when the user clicks on the OpenURL link he will trigger a process of electronic negotiation between the CERL system and the resource-holding institution which will then grant him access.

OpenURL is still in its infancy but it should be investigated carefully as a potential component of the CERL system.

9. Proposed CERL system

In light of all the aforesaid a technical solution which could give researchers a single point of access to manuscript material begins to take shape:

- CERL should implement a searching and access facility based around a central index harvested from the catalogues of participating institutions, including the HPB database. The data is located on distributed databases but there is a centralised index. Internet search engines follow this model. They retrieve data from servers, index the data then discard it, retaining URL pointers to the data.

- For each catalogue that is to be encompassed by the system a mapping is prepared (in the manner envisaged by DFAS) to extract the Common Access Points into an XML structure based on TEI/MASTER. A CERL index is therefore created combining a simplified view of entries from each catalogue – not so much a "union catalogue" as a "union index".

- Wherever possible the entries appearing in this union index could also be supported by machine readable versions of the full manuscript descriptions or catalogue entries. These machine readable descriptions could be searched by a free text search engine.

- The system should make full use of the CERL Thesaurus to give access to the HPB database in combination with manuscript material, thereby delivering a "critical mass" of content from the outset.

- References in the original catalogue entries to online substitute resources such as facsimiles and images should be preserved so that researchers will be able to link through from the results list, as more and more of these become available in digital format.

- The harvesting process probably does not need to be automatic, but should be frequent enough to reflect changes being made by participating institutions. We envisage an initial "load" of the system followed by, periodic updates. Maintenance would be needed if collections were moved, merged or broken up, but this will be unusual.

- Given the limited number of access points needed by CERL and the idiosyncratic procedures adopted in catalogues, it may be cost effective to prepare much of the index through digital data capture methods such as offshore mark up and keying.

- Users will search the index using a search engine provided by CERL and may in some cases then be offered "Open" links to richer bibliographic data or other digital materials such as facsimiles or images, held at participating institutions.

- The CERL search engine should have a machine-friendly interface, probably ZING-based, in addition to a user-friendly web interface. In both cases duplicate records must be accommodated and links between them created.

- CERL should consider offering copies of the index to each participating institution for use within their own library systems.

This proposal is shown diagrammatically at Appendix G.

10. Outstanding issues

User requirements

At this stage no formal work has been undertaken to determine the exact functionality required by scholars. We assume that this will include a minimal set of Common Access Points and links to online resources where available, but the Access Points themselves have not yet been fully specified. It is important that a sounding is taken from a wide and representative array of potential users in the final stage of this investigation about the kind of functionality they would expect from such a resource.

Institutional perspectives

It has been assumed throughout this report that institutions will be prepared to make their catalogues available for the creation of the central index. Clearly, CERL will need to confirm that this is the case, and be ready to discuss the levels of support and any restrictions that may accompany a conditional willingness to participate.

Institutions will have concerns about security, copyright and access. It may therefore be necessary to incorporate a mechanism which allows institutions to alter or even withdraw their data from the index. This can be done by adding appropriate control codes and identifiers to each entry.

As in the case of potential users it will also be important to begin to take soundings about all these issues from potential participating institutions in the final stage of this investigation. Additionally it will be important to obtain some representative sample catalogue entries and manuscript descriptions and to determine the extent to which the information is already available in digital format. An effective way to achieve this will be build on the survey of Manuscript Online Projects already undertaken by Lotte Hellinga in 2001, through a combination of a written questionnaire and telephone interviews.

Common Access Points

DFAS identified CAPs as perhaps the most important issue to be tackled in the creation of a search system:

> “We examined the existing finding aid catalogs of the participating institutions and immediately discovered that there were no indexes common across all five participants: some had chosen more structural indexes… while others had chosen more traditional bibliographic indexes… The structural CAPs might not be useful if researchers don't understand the structure of finding aids ahead of time and if finding aids aren't constructed consistently (which they often aren't across institutions). But the semantic CAPs could also be confusing if encoders applied these more subjective tags inconsistently (which they often did)...”

CERL requires a small number of CAPs, which ought in theory to be derivable with a reasonable level of consistency from the participating institutions’ catalogues. However we consider that **it would be appropriate for CERL to carry out a survey as part of a pilot study to make sure that this is in fact the case**.

Outstanding issues

Management and logistics

As commented above, the initial loading of this system from each institution will be labour intensive and may involve re-keying. Many institutions are only part way through the preparation of electronic catalogues of manuscript holdings so the system must therefore be able to cope with regular updates for several years after initial implementation. Management structures will be required which will ensure the necessary long term stability.

Most institutions will not be able to afford any significant level of resource to work on this project. Moreover, technical staff may be on contract, so CERL should not rely on their continued availability at each institution.

11. Next steps

We can distinguish three further phases for this project: final investigations; pilot /proof of concept studies; implementation.

a. Final investigations

- To survey and narrow the field of potential technology-supply partners.
- To further investigate the feasibility of OpenURL linking to online resources.
- To begin to identify possible participating institutions and draft a survey to be carried out during the pilot phase of the project.
- Further research into the possible list of CAP's with a view to finalizing this during the pilot phase of the project.
- To begin to formulate a tentative budget and project plan and begin to scope out the elements that would go into a proof of concept project, drawing upon the experience of similar projects such as MALVINE, and that of the German Union Catalogue of medieval manuscripts (Manuscripta Mediaevalia). Central to all costs are those of setting up and maintaining the Central Index.

b. Proof of concept definition/Technology Pilot

- To survey user requirements to explore and define functionality, including multilingual requirements and associated cost and operational implications. Notwithstanding the usefulness of the CT in this respect; this must be seen as a major component of the project.
- To carry out a full survey of possible participants to determine the general level of cooperation which could be expected. The survey will identify the interplay of "push" (feel obliged to participate) and "pull" (want to use/participate) institutions. There will be a wish to be part of the community and enthusiasm, but will this be enough? An indicator will be the ability to obtain a wide array of samples of material which will be needed in trials, and in parallel, the beginning and sharing of the mapping, timing and logistics of work.
- Equally, it will be important to give some consideration to the issue of control of access to the system.
- Determine the extent of any material which would need to be re-keyed and the potential cost of organising and executing the effort.

- During the pilot phase a small-scale “harvesting” test will need to be carried out.
- To develop a technology demonstration similar to the TEL test portal.
- To draft a full system specification.
- To identify possible sources of funding for the project

c. Implementation

- It is anticipated that full implementation would include the development of contracts with technology partners and agreements with participating institutions, research libraries and collaborating projects.
- Software development.
- Initial harvesting and normalisation of data.
- Full development, testing and roll out of the system.

12. Indicative schedule

- **Final investigations**

  Timeframe Spring 2003 (in progress)

- **Pilot / proof of concept**

  Definition Spring 2003

  Cost subject to negotiation with participating third parties

  Begin implementation Summer 2003

  Review Pilot Implementation at November 2003 AGM

- **Implementation**

  November 2003 AGM - Review anticipated costs and identify possible sources of funding.

  Full specification of functionality Spring 2004

  Agreements with technology providers by Autumn 2004

  Begin deployment mid 2005

13. Supplement to the CERL March 2003 Paper – Potential Technical Partners

This is a supplement to the March 2003 paper on A Searching Facility for Manuscripts and Hand Press Book Catalogues, prepared for the LIBER conference.

Continuing investigations since October 2002 indicate two possible technology partners, both library automation companies of those that have been identified as possible participants in the development of CERL's proposed searching facility. They are Fretwell Downing http://www.fdgroup.com and MuseGlobal http://www.museglobal.com. Two further options from the library arena emerged in the course of these investigations: the Kalliope open information system developed by the Staatsbibliothek zu Berlin http://kalliope.staatsbibliothek-berlin.de and the possible adaptation of the system being developed by the Association of Research Libraries Scholars Portal Project http://www.arl.org/arl/pr/scholars_portal.html There has not yet been sufficient time for a closer study of the two library resources but this should not detract from the possible relevance of these options, on the contrary a fuller investigation is recommended.

More importantly in the course of these investigations it has become increasingly evident that perhaps a more suitable way of describing the required system is as a **"library portal**" rather than a search engine, starting with the central premise of the requirement for a single point of access not just to manuscript material, but along with that integrated access to the the HPB and CERL Thesaurus

There is a lot of information on the internet about library portals, and a good starting point is with the Library of Congress Portals Applications Issues Group (LCPAIG) http://www.loc.gov/catdir/lcpaig/paig.html whose primary goal is to search for portal products that would best meet the reference and research needs of Library of Congress. Another purpose of this group is to promote best practice not least through the Bibliographic Control of Web Resources: A Library of Congress Action Plan, that calls for development and enhancement of portal functionality for the benefit of the library community in general. It was last revised in November 2002. Indeed the principal objectives of the plan (set out here) could be considered a useful set of background references for the current work proposed by CERL:

1 increased availability of standard records for Web resources;

2 enhanced record display and access across multiple systems

3 collaboration among metadata standards communities for better bibliographic control of Web resources;

4 development of automated tools for harvesting and maintaining metadata;

5 provision of appropriate training for the Web environment; and

6 support of research and development to enhance bibliographic control of Web resources.

A somewhat less daunting but extremely well informed and useful overview of the reasoning behind the advent of library portals was presented by Mary E. Jackson, senior program officer for the Association of Research Libraries, in the September 2002 issue of Library Journal (the article is attached in its entirety) http://libraryjournal.reviewsnews.com/index.asp?layout=articleArchive&articleid=CA242296%20 This paper makes many points worth considering at this stage of the CERL project:

- *"Librarians in institutions of all types and sizes want to provide users with a single point of access to their high-quality resources. Increasingly they see portals as the key to integrating access to the growing range of information resources in a vast number of formats.*

- *The core feature of any portal will be integrated, cross-database searching of a local catalog, other library catalogs, selected web sites, locally licensed full-text and abstracting/indexing databases, and public domain or publicly accessible abstracting and indexing services*

- *Integrated searching is a key feature of a portal. It distinguishes it from a web site. ..Library web sites usually do not permit users to conduct a single search of multiple resources, nor do web sites deliver integrated results. Users usually have to integrate the results from their separate searches as another step*.

- *Both the multiplicity of standards and the lack of standards are challenges in developing integrated, cross-database searching. Many online catalogs can be accessed by the international standard for search and retrieval, Z39.50, but additional search techniques are required for such resources as XML datasets or web resources using different metadata schemes such as MARC, Dublin Core, Computer Interchange of Museum Information (CIMI), and Encoded Archival Description (EAD).*

- *Research libraries, like academic, public, and special libraries, view library portals as an enabling tool to support the mission of their institutions. Librarians have high expectations for how portals will facilitate access to a wide range of high-quality content. The number of companies marketing portal products is expanding rapidly, and the number of libraries offering portals with integrated searching and multiple supporting services will continue to grow. The challenge and the fun ahead is to design portals with enough flexibility to respond continually to user preferences in the discovery, presentation, and use of high-quality information resources*."

Three overarching conclusions emerge from this article:

1. Portals cannot be developed in isolation, they need to take into account the whole continuum of information use and the retrieval process; database searching is just the beginning, integration with other library resources and processes should also be taken into consideration.

2. The technologies are developing very quickly and the definition of portals and function are still in flux. Hence any portal must be based on open standards and architecture which will ensure it will be extensible into the future.

3. Finally, although not stated overtly in the article, it follows that that the choice, the role and relationship with technology providers is crucial in the development of any such cross searching portal application. Mary Jackson identified 13 possible providers of portals applications in her article, including Fretwell Downing and MuseGlobal.

It is against this background that we will present the two possible technology partners investigated for the CERL project to date: Fretwell-Downing and MuseGlobal.

Fretwell-Downing, http://www.fdgroup.com/fdi/company/home.html, is a group of companies, with operations in Sheffield, England, Kansas City, USA, and Victoria, Australia. FD have been established for ten years, they are global suppliers of information management systems, specializing in library automation and networked information services including:

- Portals
- Remote authentication services
- ILL and document delivery
- Meta–data servers
- e–Document Rights Management

FD provides an impressive suite of software products based on open standards designed to integrate with each other, and with other components from third parties. This gives a great deal of flexibility towards achieving the company's vision of "a library without walls" – the mixing of Internet, Intranet, digital and printed resources, from search to delivery to the desktop. Main customers include university, college and public libraries, state and regional library consortia, research organizations, corporate & special libraries. The company's website gives a full range of FD's range of Products. Special attention is drawn here to:

ZPORTAL Common Information Portal (ZPORTAL) which allows access to disparate information from a single interface supporting the whole process from searching through to delivery. ZPORTAL can identify resources from libraries, museums, archives, and the Web, irrespective of how and where they are stored. These resources can be seamlessly integrated over the whole information finding process, from 'discovery to delivery'. Andrew Cox, LITC calls it "one of the most advanced solutions available" for "resolving technical interoperability issues raised by cross searching of multiple sources, managing authentication for users and supporting choices between document delivery options" (LTWorld Wednesday, 21 Nov 2001 www.sbu.ac.uk/litc/lt/2001/news2213.html)

CPORTAL is a community information network solution that allows online library catalogs and other Web sites to integrate with pre-existing internal databases, giving access to previously inaccessible information. Initially designed for government, state agencies and other organizations to connect their existing databases and services and integrate them with Web-based information resources. CPORTAL complies with internationally recognized e-government standards.

Z'MBOL provides an open meta-data server platform for publishing data sets as Z39.50 servers. This allows integration of arbitrary meta-data and data sets into the digital library environment.

Z2WEB a generic toolkit for making proprietary interfaces accessible as Z39.50 targets, allowing them to be integrated into Z39.50 portals. Z2WEB opens the way for librarians to create a search interface that simultaneously searches traditional library catalogue information and the Web.

Open Linking tools for dynamic content linking which automatically locate the appropriate resource for a user for a given citation while providing management and authentication services. It is integrated with other products from the FDSuite such as ZPORTAL, VDX and eDRM.

FD seems well established with a substantial number of very important libraries and consortia as customers including the Ohio Public Library Network, State Library of

Colorado, Ontario Council of University Libraries, the Minnesota Higher Education Service Office. A good overview of customers exists on the Fretwell-Downing website.

More importantly and of direct relevance to the current project, in 2001 Fretwell Downing was selected after a thorough review by the Association of Research Libraries' Scholars Portal Working Group from a long list of possible suppliers to participate in the ARL Scholars Portal project http://www.arl.org/arl/pr/scholars_portal.html. Mary Jackson made the point in her article that FD were the only supplier who offered high proportion of the functionality requirements identified by the group, in fact 80% and that they were willing to collaborate with the members of the working group to develop the other 20%.

There is much information about this project available on the internet, on the www.arl.org site, but a very good summary is provided by Barbara Quint in Information Today www.infotoday.com/newsbreaks/nb020513-2.htm. She writes:

"The Scholars Portal Project will provide software tools that allow an academic library to supply a community of users with a single point of Web access that can reach a full array of diverse, high-quality information resources and deliver material directly to the user's desktop. Initially, it will use Fretwell-Downing's ZPORTAL and several related products as a base. Once deployed, ZPORTAL will offer cross-domain searching of licensed and open Web content in a range of subject fields from multiple institutions. The portal will then aggregate and integrate search results. In time, the designers plan to add other improvements, such as the integration of searching within local online learning and course environments, links to 24/7 digital reference services for immediate consultation with reference librarians, transfer of orders to document delivery outlets, etc. ..

Current plans anticipate that, in most academic environments, the tools developed through the Scholars Portal Project will function as a library channel within a university-wide portal. The project grew out of an ARL Scholars Portal Working Group that was set up in 2000 to explore how best to establish a collaborative research library presence on the Web. The Scholars Portal Project will demonstrate the viability of that vision with one vendor's products, although plans don't limit future software development to any single vendor. ARL planners hope the project will encourage other vendors to enter the marketplace with competitive tools to advance portal functionality. ARL will continue to monitor available software tools that can "meet the needs of the 21st-century academic Web user."

"Seven of ARL's major member libraries—the University of Southern California (USC), University of California–San Diego, Dartmouth College, University of Arizona, Arizona State University, Iowa State University, and the University of Utah—will collaborate on the initial release of the Scholars Portal Project. Over the course of the 3-year undertaking, ARL plans to expand the number of participating libraries. Mary E. Jackson, ARL's senior program officer for access services, notes that several other ARL members have expressed serious interest in joining the project. "We're now welcoming all other ARL members. Not all will be interested, but the project has the ability to handle any number. It's anyone's guess what the eventual number will be."

Jerry Campbell, chief information officer and dean of University Libraries at USC, chaired the ARL Scholars Portal Working Group. In the announcement, he said: "Fretwell-Downing, Inc. is not the only portal game in town. We selected them to work with us in this project for two reasons. The first reason is that we believe FD's existing ZPORTAL product suite will work together to take us significantly down the road toward achieving our initial project goals. And secondly, the leadership of the company shares our vision of a portal and has *committed to bearing some of the costs of developmental work that will lead to enhancements that both we and FD see as important."*

The full report of the ARL Scholars Portal Working Group Report," May, 2001 is at www.arl.org/access/scholarsportal/may01rept.html

Naturally the ARL working group is a good starting point for further soundings about this company, particularly Mary Jackson, the project manager. But note that this is planned but has not yet been completed.

The CERL Searching Facility Report was read (with a confidentiality agreement in place) by Fretwell Downing in mid-February with the aim of obtaining feedback and identifying costs for a possible pilot project. Neil Smith of FD states that he followed the logic and endorses the main recommendations. He makes these important points:

1. **The growing importance of access to digital surrogates of manuscripts** , touched on in section 2 of the report", led him to point out that "whilst it is stated that the majority of items in the index will not have digital resources associates with them initially, it is important not to underestimate the financial resources being made available for digitisation projects. The searching facility should be designed to allow access to digital resources from the outset as there is a role here for **the use of linking technologies based on OpenURL in the search facility**. The goal would be to allow linking to an appropriate copy of a digital resource associated with the metadata record in the central index. There are 2 main aspects to this:
    a. Parsing of the metadata to point to the URI of the resource. Whilst this would not need to conform to the OpenURL standard, it would need **the metadata to be sufficiently rich to identify the item along with a definable algorithm for determining the URL for an item from the metadata within each institution**.
    b. It would also be advantageous **to be able to identify the type of resource being linked to as different permissions might apply** to, say, a digital image of a cover page compared to a PDF surrogate of the whole resource or an HTML file containing the extracted text."

2. In respect of architecture, the report's review of distributed search technologies is endorsed. He also points out the study by Sebastian Hammer of Index Data (http://www.indexdata.dk/paraz/parallel_search.html) which, whilst refuting Z39.50's reputation as a 'heavyweight' protocol recommends a pragmatic approach based on a combination of union lists and parallel searches.

3. He endorses the report's proposal for a central index to be created whilst allowing for searching of other resources (e.g. the HPB database) in parallel, and goes on to point out that physically distributing the central index to institutions, should not be technically necessary.

4. In respect of encoding standards he concurs that "an adaptation of the OAI model would be the most appropriate starting point - and that there may well be valuable pointers in the DFAS work on how to achieve this.

5. He cautions against putting too much emphasis on ZING/ SRW at the moment believing that there are many legacy Z39.50 resources still around. "I would design

a system which allows access to 'legacy' Z39.50 resources (of which there are many) as well as allowing other resources to be integrated (e.g. resources searchable by http interfaces only - see FDI's Z2Web gateway(http://www.fdusa.com/products/z2web.html)

6. Authentication is an important element that the report has so far not detailed -- determining whether the user is authorised to access a particular copy of a digital resource. "It is this issue, which would also apply to any centralised web portal, which is perhaps worthy of expanded consideration in the report. FDI's experience in Scholar's Portal and in State-wide resource sharing systems in the USA has shown that a distributed authentication and authorisation model is technically feasible. Further information would be required about the authentication systems in place at each institution to fully assess the practicality of this as an approach."

Based on the initial findings of the report Fretwell Downing concurs that a 'technology pilot' should include the following four areas:

- Harvesting of metadata from at least 2 institutions and combining in a single index

- Provision of a web based search interface onto the index which
  - provides a 'simple search' function (similar to Z39.50 'any')

  - allows all common access points to be searched

  - allows Boolean operands for advanced searching

  - provides parallel searching with at least 1 other resource (e.g. HPB database)

- External search interface to the central index using both Z39.50 and ZING/SRW

- Proof of concept in the following areas:

  - distributed authentication/ authorisation

  - 'open' linking to digital resources (including access control)

The pilot project should run for 6 months to be ready for review by November 2003. The pilot would be based on existing software (both FDI and Open Source as appropriate) and be hosted on FDI's Managed Service in the UK and made globally accessible via http and other protocols if appropriate. FD indicate they would not charge a licence fee for the duration of the pilot. Also provided would be licensed software as contribution 'in kind' to the project. FDI offer a 20% discount on their normal daily labour rate.

Before moving on to another possible technology partner it is worth quoting David Dorman, columnist for the periodical American Libraries in his regular column *Technically Speaking :*"A vigorous debate seems to be raging within and among many state library agencies over whether it is preferable to create one consolidated statewide catalog, or to rely on a virtual catalog brought into being by broadcast searching of multiple catalogs, the results of which are deduplicated "on the fly" by computer algorithms. The debate is far from over, but if I were a betting man, I would put my money on a consolidated approach, because achieving bibliographic coherence out of the current situation requires more than relying on computer processing and voluntary standards to herd our library cats."

The same issue of Library Journal contains an item about the next potential technology provider, MuseGlobal. "MuseGlobal—with the 83 branch libraries of the New York Public Library, for MuseSearch and MuseBridge. MuseBridge provides a single interface to multiple information resources and MuseSearch manages broadcast searching among those resources. In recognition of the severe budget cuts sustained by NYPL after 9/11, the company offered the library a two-year subscription to these services free of charge."

MuseSearch™ -- MuseGlobal, Inc. http://www.museglobal.com
Provides library technology which makes possible unlimited numbers and types of information sources to be searched simultaneously with a single user query whether Web, Z39.50, SQL, telnet, proprietary, intranet, full-text database, image database, or any other data source; the company describe this technology like an adaptor that can be built for any platform. Further information from the company's website: "MuseSearch is a solution that allows unlimited numbers and types of information sources to be searched simultaneously. Each search is translated into the native language and protocol of each source, so results are of the highest quality. MuseGlobal work directly with the manufacturers of the information databases used by their clients to ensure continued access at the highest level even when information supplier interfaces change. And because MuseGlobal specialise in providing interfaces to information sources they can even interface any custom written or legacy databases…"

The company is based in the US with offices in Utah and New Mexico, and has two representatives in the UK with a track record of some major automation projects, including recently with the Statewide California Library Consortium. A good overview of projects to date can be found in pdf at http://www.libit.de/download/lospeng.pdf.

MuseGlobal also signed a confidentility agreement and read the March version of the Report. They observe that while they could see the merit of a central index approach the process of deciding on Common Access Points and the encoding schema could take a while to get in place. They point out that their technology MuseSearch could be built and plugged in immediately, while an index was being compiled, providing access immediately before the central index is fully in operation.

In fact they indicate that they would be willing to build two pilot implementations, to commence installation in the Summer of 2003. The pilots could be ready for review by November, 2003. Both proposed implementations are given in full below:

Proposal #1: Searching and Linking with a Central Index Structure

> This goes to serve the recommendations set out in Appendix G of the CERL/HPB report. It is not possible to be very detailed here as CERL still needs to identify and secure participating libraries, outline functionality, delineate desirable mappings for

data, define and publish the Common Access Points, and obtain agreements on authentication and access.

Bearing in mind those issues CERL has yet to determine about the project's scope and functionality, MuseGlobal proposes to create and tune connections for up to 20 databases at between 10 and 15 CERL institutions to provide search/retrieval/linking around a central index built from the selected sources and institutions. This proposal includes working on a suitable thesaurus for the index.

Proposal #2: Distributed Searching and Linking

This proposal allows CERL to test a distributed search where MuseGlobal provides database connectors and search translation and authentication to each relevant database.

Bearing in mind those issues CERL has yet to determine about the project's scope and functionality, MuseGlobal proposes to create and tune connections for up to 20 databases at between 10 and 15 CERL institutions to provide search/retrieval/linking in a common interface. Additionally, the search interface will provide tools and options for sorting, ranking, filtering, and de-duplicating search results.

The value of this pilot is to establish the viability of such a model as either a fast start or support for the centralized index model as it is under construction. It seems clear that it will take significant effort and time to create a suitable centralized index and acceptable thesaurus. The point of this pilot is to provide a distributed searching functionality for the databases right away. It could be used in place of the centralized index until that index is complete. At that point it will be easy enough to "gather" in the distributed searches and point the search tool at the central index.

It will always provide support for the central index model to cover data that cannot be included in the central index for some reason and to provide a fast and straight forward way to bring in new databases and institutions without waiting for the central index to be updated. It may prove to be an efficient enough model to postpone or reduce dependence on a centralized index for the project.

It is worth observing that the cost of the FD pilot including the proof of concept elements roughly converts to the cost for the two combined pilots from FD. Perhaps there is room here to consider a combination of working with both technology providers in developing a portal for CERL, taking the best features of both. Further investigation and technical assessment of whether this might work still needs to be undertaken.

Another possibility which emerges for further investigation is the adaptation of the ARL Scholars Portal, and/or working with Fretwell Downing to achieve this. Kalliope has been mentioned at the beginning of this paper but there has not yet been time to give any consideration in detail at this resource.

Nevertheless some clear recommendations emerge so far:

Staged Development:

The current proposed development of a cross searching resource for CERL should be discussed in the context of a staged development of a library portal, with continuing, ongoing, future enhancements, rather than a one off searching tool.

Open Protocols:

Any technology used should be based on open protocols rather than proprietary platforms in order to ensure future extensibility.

Ability to accommodate high rate of change in Technology:

The project will be challenging as technology in this area is developing and changing rapidly. Therefore it will be important to select technology partners who are reliable, responsive and capable of working within the parameters of a short chain of communication avoiding any hint of convoluted internal politics.

Appendix A – Sources

Third parties consulted in the preparation of this report were:

Peter Robinson (MASTER)
Elizabeth Shaw (TEI/DS)
Theo van Veen (Koninklijke Bibliotheek & TEL)
Lawrence Mielniczuk (Bodleian Library & DFAS)
Jutta Weber (Staatsbibliothek, Berlin)
Consuelo Dutschke (Columbia University)

Appendix B – Articles and papers, online

DFAS – 'The Distributed Finding Aid Search System'
by MacKenzie Smith, Office for Information Systems, Harvard University Library
http://www.dlib.org/dlib/january00/01smith.html

LOCKSS, 'Lots of Copies Keep Stuff Safe'
- A Cooperative Archiving Solution for E-Journals
by Victoria A. Reich, Director, LOCKSS Program, Stanford University
http://www.istl.org/02-fall/article1.html

MASTER – 'Ref Manual for the MASTER Document Type Definition' (Discussion Draft)
edited by Lou Burnard for the MASTER Work Group
http://www.tei-c.org.uk/Master/Reference/

OAI - 'Metadata Harvesting and the Open Archives Initiative'
by Clifford A. Lynch, Exec Director, Coalition for Networked Information
http://www.arl.org/newsltr/217/mhp.html

Open URL - Working Documents
http://library.caltech.edu/openurl/Working_Documents.htm

'ParaCite' – An Overview
by Michael Jewell
http://paracite.eprints.org/docs/overview.html

'Aiming at Quality & Coverage Combined
– Blending Physical & Virtual Union Catalogues'
by Janifer Gatenby, Consultant ITC, OCLC PICA (NL)
TEL Milestone Conference, 29-30 April 2002
http://www.europeanlibrary.org/doc/tel_milconf_presentation_gatenby.doc

'Building the Infrastructure of Resource Sharing –
Union Catalogs, Distributed Search, and Cross-Database Linkage'
by Clifford A. Lynch, Exec Director, Coalition for Networked Information
http://www.caslin.cz:7777/caslin99/a3.htm

'Metadata' – serving services
by Theo van Veen, Project Leader, R & D Dept, Koninklijke Bibliotheek (NL)
TEL Milestone Conference, 29-30 April 2002
http://www.europeanlibrary.org/doc/tel_milconf_presentation_vanveen.doc

Appendix C – Weblinks

CERL Thesaurus –

http://www.cerl.org/Thesaurus/thesaurus.htm

DECOMATE II – Developing the European digital library for economics

http://www.bib.uab.es/decomate2

D-Lib Magazine

http://www.dlib.org/

DLPS – The Digital Library Production Service (of the University of Michigan)

http://www.umdl.umich.edu/

EAD – Encoded Archival Description

http://www.loc.gov/ead/

EAMMS – Electronic Access to Medieval Manuscripts

http://www.hmml.org/eamms/

The Hand Press Book Database –

http://www.cerl.org/HPB/hpb.htm

iPort – Pica Internet Information Portal service (based on Decomate II software)

http://www.pica.nl/en/news/iport.shtml

MALVINE

http://www.malvine.org/

MASTER – Manuscript Access through Standards for Electronic Records

http://www.cta.dmu.ac.uk/projects/master/index.html

OAI – Open Archives Initiative

http://www.openarchives.org/

OpenURL – Development of an OpenURL Standard

www.niso.org/committees

PiCarta – part of OCLC PICA

http://oclcpica.org/?id=102&ln=uk

TEI – Text Encoding Initiative

http://www.tei-c.org/

TEL – The European Library

http://www.europeanlibrary.org/

TEL test portal

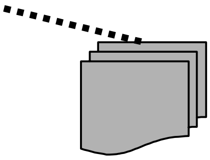
http://krait.kb.nl/coop/tel

ZING – Z39.50 International: Next Generation (also: SRW and CQL)

http://www.loc.gov/z3950/agency/zing/

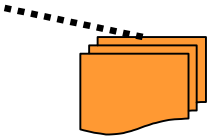
Appendix D – Diagrams of searching

D1



Paper-based catalogue – **not searchable**

D2



Paper-based
catalogue –
potentially
searchable

Appendix E – Example of XML EAD encoding

XML EAD encoded record of Bodleian MS. Fr.e.32. Courtesy of the Bodleian Library.

```
<c01 langmaterial="French">
  <did>
<unitid type="shelfmark">MS. Fr. e. 32</unitid>
<unitid type="SCN">Not in SC (late accession): no description
available</unitid>
<unittitle>
<title><emph render="italic">La Chevalerie Vivien</emph>, and
  <emph render="italic">Aliscans</emph>.</title>
<geogname>French, North-east</geogname>
<unitdate>12th century, late or <emph render="italic">c</emph>.
  1200</unitdate> </unittitle>
<physdesc><physfacet type="Material">parchment</physfacet></physdesc>

  </did>
  <scopecontent>
<head>Decoration</head>
<p>Simple red and decorated initials.</p>
  </scopecontent>
  <odd>
<head>Images</head>
<daogrp>
<daoloc href="images/aaq0404.gif" title="Vol. 2 Part 2, p. 1045"
role="SC">
</daoloc>
</daogrp>
<daogrp>
<daoloc href="images/bar052203.jpg" title="1r" role="MS">
  <daodesc>
<p>Whole page with initial O[mnibus] and border.</p>
  </daodesc>
</daoloc>
</daogrp>
  </odd>
  <admininfo>
<custodhist>
<p>Thomas Arnold, St. Augustine's, Canterbury ('Liber fratris T.
  Arnold' de libr' sancti Augustini Cantuariensis'; see A. B. Emden,
  <emph render="italic">Donors of Books to St. Augustine's Abbey
  Canterbury</emph>, Oxford 1968, 5), with 15th-cent. letter-identifier '.Cum.
  .H.' (M. R. James, <emph render="italic">The Ancient Libraries of Canterbury
  and Dover</emph>, Cambridge 1903, 374, no. 1533); Savile, sale, Sotheby's 6
  Feb. 1861, lot 16, bought by Powis for £150; Sir Thomas Phillipps (1792-1872),
  MS. 25074; Sotheby's, 30 Nov. 1971, lot 495, pl. 7 (fol. 28r).</p>
</custodhist>
  </admininfo>
  <add>
<head>Bibliography</head>
<bibliography>
<bibref>Duncan McMillan, 'La <emph render="italic">Chevalierie
  Vivien</emph> dans le MS. dit "de Savile": notes prolégoméniques', in
  <emph render="italic">Études de langue et de littérature du Moyen Age: offertes
  à Félix Lecoy par ses collègues, ses élèves et ses amis</emph> (Paris, 1973),
  pp. 357-75.</bibref>
<bibref>Ian Short, 'An early French epic manuscript: Oxford,
  Bodleian Library, French e. 32', in <emph render="italic">The medieval
  Alexander legend and romance epic: essays in honour of David J.A. Ross</emph>
  ed. Peter Noble, Lucie Polak, and Claire Isoz (Millwood, NY, 1982), pp.
  173-91</bibref>
<bibref>D. McMillan, 'Un manuscrit hors série: le cas du Manuscrit
  <emph render="italic">S</emph> de la <emph render="italic">Chevalerie
  Vivien-Aliscans</emph> (Bodléienne, French e. 32)', in <emph
  render="italic">Symposium in honorem prof. M. de Riquier</emph> (Barcelona,
  1986), pp. 161-207.</bibref>
</bibliography>
  </add>
</c01>
```

Appendix F – Illustration of minimal vs. maximal encoding from MASTER

Example MASTER encoding of bibliographic data (taken from http://www.cta.dmu.ac.uk/projects/master/gentintr.html)

Original record:

Oxford, Corpus Christi College, MS 198
Geoffrey Chaucer The Canterbury Tales. c. 1400
Folios 1r-266v. The Canterbury Tales. A274-I290. Defective at beginning and end.
Parchment, trimmed. 33.5 x 22.5 cm. Quires [14, 15, and 28] were disordered in the previous binding. They have been reordered and refoliated, with the old foliation being the uppermost. Two consecutive folios are numbered '64a' and '64'
Written by the scribe identified by Doyle and Parkes as 'Hand d'
Dated c. 1400 (personal communication, Malcolm Parkes). On fol. 146r is the name 'Burle' in drypoint, in the margin next to E1396. Cp came to the College as a bequest of William Fulman, according to a note on fol. 1r : 'Liber C.C.C.Oxon Ex dono Gulielmi Fulman A.M. hujus Collegii quondam socius.'

Minimal encoding:

```
<msDescription>
   <msIdentifier n="1">
      <country reg="GB">Great Britain</country>
      <settlement>Oxford</settlement>
      <repository>Corpus Christi College</repository>
      <idNo>MS 198</idNo>
   </msIdentifier>
   <msHeading>
      <title>The Canterbury Tales</title>
      <author>Geoffrey Chaucer</author>
      <origPlace>?</origPlace>
      <origDate notBefore="1395" notAfter="1420">c. 1400</origDate>
      <textLang langKey="ENM">Middle English</textLang>
   </msHeading>
</msDescription>
```

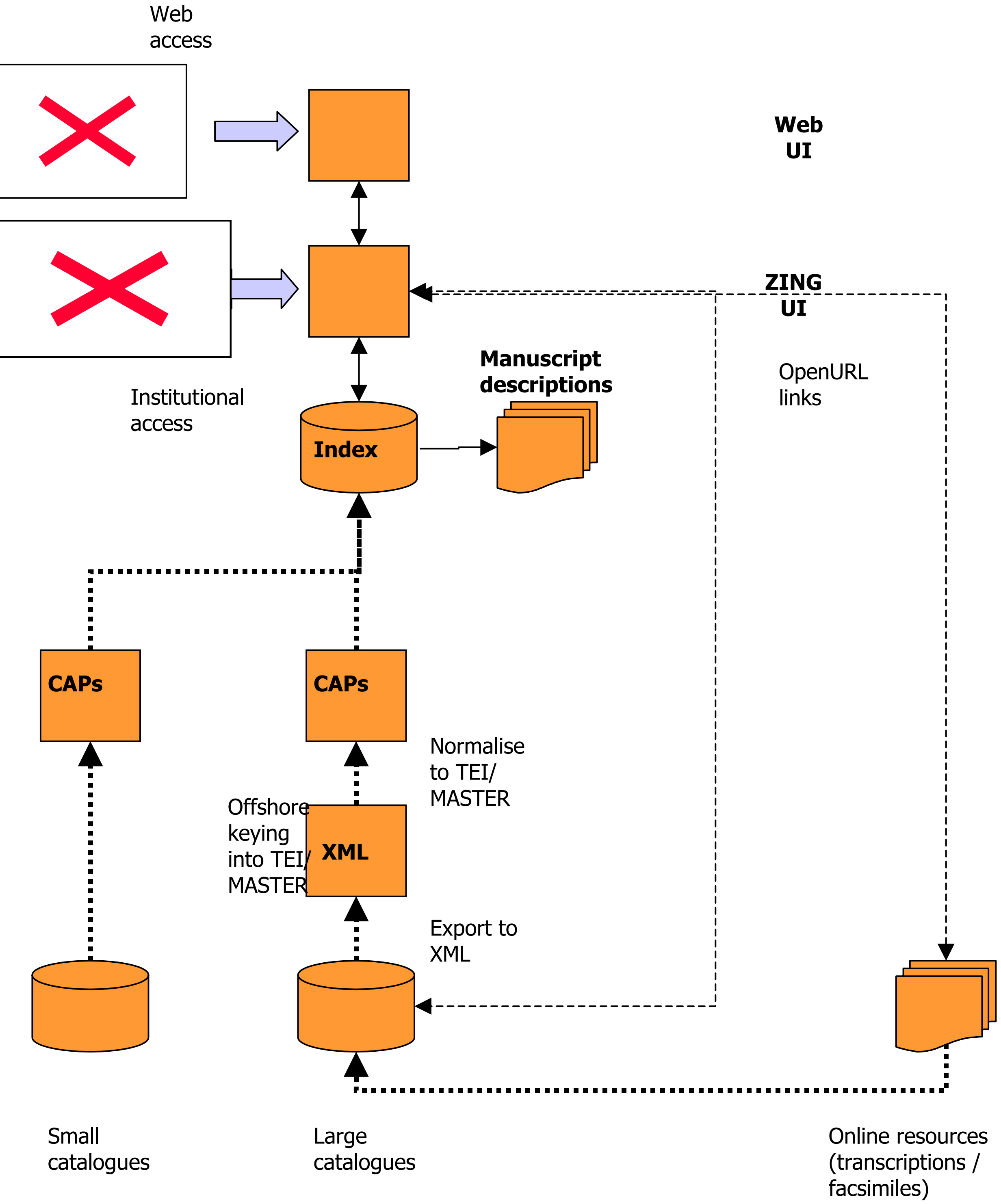
Maximal encoding:

```
<msDescription>
   <msIdentifier n="1">
      <country reg="GB">Great Britain</country>
      <settlement>Oxford</settlement>
      <repository>Corpus Christi College</repository>
      <idNo>MS 198</idNo>
   </msIdentifier>
   <msHeading>
```

```
    <title>The Canterbury Tales</title>
    <author>Geoffrey Chaucer</author>
    <origPlace>?</origPlace>
    <origDate notBefore="1395" notAfter="1420">c. 1400</origDate>
    <textLang langKey="ENM">Middle English</textLang>
  </msHeading>
  <msContents>
    <msItem n="1" defective="yes">
      <locus from="1" to="266">Folios 1r-266v</locus>
      <title type="uniform">The Canterbury Tales</title>
      <bibl>
        <biblScope>A274-I290</biblScope>
      </bibl>
      <note>Defective at beginning and end</note>
    </msItem>
  </msContents>
  <physDesc>
    <form>
      <p>Codex.</p>
    </form>
    <support>
      <p>Parchment, trimmed.</p>
    </support>
    <extent>266.<dimensions type="leaf" scope="all">
        <height>33.5</height>
        <width>22.5</width>
      </dimensions>
    </extent>
    <collation>
      <p>Quires [14, 15, and 28] were disordered in the previous binding. They
have been reordered and refoliated with the old foliation being the uppermost. Two
consecutive folios are numbered '64a' and '64'</p>
    </collation>
    <msWriting hands="1">
      <handDesc scribe="Hand D (Doyle/Parkes)" script="Anglicana"
medium="ink" scope="sole">
        <p>Written by the scribe identified by Doyle and Parkes as '<name
type="person" role="scribe" key="DPhandD">Hand d</name>'</p>
      </handDesc>
    </msWriting>
  </physDesc>
  <history>
    <origin notBefore="1395" notAfter="1420" certainty="high"
evidence="conjecture">
      <p>Dated c. <origDate>1400</origDate> (personal communication,
Malcolm Parkes).</p>
    </origin>
    <provenance>
      <p>On fol. 146r is the name 'Burle' in drypoint, in the margin next to
E1396. </p>
    </provenance>
    <acquisition>
```

```
            <p>Cp came to the College as a bequest of William Fulman, according to a
note on fol. 1r : <q>'Liber C.C.C. Oxon Ex dono Gulielmi Fulman A.M. hujus Collegii
quondam socius.'</q>
            </p>
        </acquisition>
    </history>
</msDescription>
```

Appendix G – Proposed CERL access system



Thursday, September 12, 2002
Library Journal

The Advent of Portals

By Mary E. Jackson

The internet, digital electronic resources, and database technology have transformed the way people search for information.

Today many people rely on keyword searches in single-step search engines--e.g., Google or AltaVista--that retrieve information from unspecified slices of the web. This is eroding the use of traditional library reference and information services and could ultimately result in a set of services and resources that are less authoritative but more convenient.

Users frequently cite ease and convenience as the main reasons they prefer commercial search engines over gaining access to electronic sources through a library's web site. Sources such as online journals, online public domain materials, or locally developed databases are often passed by because of these convenient options. Libraries must gear up to provide a competing level of convenience while retaining the authority and quality of information delivery for which they have been traditionally known.

For a host of queries, of course, the Google or AltaVista search may be all a user needs. Librarians must reconcile themselves to this fact and refocus the mission of library information services and resources to the deeper, more complex information needs of users engaged in searches that require higher levels of authority and more comprehensive scope.

Imagine one web site that can combine the powerful searching of web resources with the searching of local catalogs, online journals, or locally digitized resources. Add to this the ability to initiate a reference question, submit an interlibrary loan (ILL) request, and transfer into course management systems a citation or portion of a journal article, all without leaving that web site.

The dream portal

Librarians in institutions of all types and sizes want to provide users with a single point of access to their high-quality resources. Increasingly they see portals as the key to integrating access to the growing range of information resources in a vast number of formats.

While the library and IT communities have not agreed on a single definition of a portal, there is growing consensus on the essential features and functions of one.

Michael Looney, cofounder of the portal company GoCampus, Inc., and Peter Lyman, at UC-Berkeley's School of Information Management and Systems, offer one definition: "Portals gather a variety of useful information resources into a single, 'one-stop' web page...[that] allow users to customize their information sources by selecting and viewing information they find personally useful."

Like Looney and Lyman, Andrew Cox and Robin Yeates, both associate directors of the UK's LITC, expect portals "to integrate the diverse licensed and owned electronic holdings of libraries for users, through the whole process of discovery and searching to final delivery, regardless of the content's format, the metadata standard in use, publisher interface, or authentication mechanism."

Sarah Michalak, director of the University of Utah Libraries, has defined a "dream portal" as a super discovery tool that specializes in high-quality content. The dream portal is fast and powerful. It searches across formats and resources and returns results that are deduped and relevancy ranked. It is more than a discovery tool because it delivers full text or information objects whenever available. The dream portal integrates appropriate applications such as course management software. Finally, the dream portal supports authentication and permits customization and personalization, e.g., alerts, saved hits or searches, and custom views of resources. This dream portal will help users overcome "infoglut." It is "Google with good content supported by a range of library services." A portal combines powerful searching with the diverse resources and services that patrons find when they use a library. Portals should provide library experience of that quality without requiring people to come to the library.

Portals are more than enhanced web pages, although some have evolved from library web sites. Librarians are just beginning to define the requirements for portal products. The portals and librarian expectations of their functionality will become more refined and sophisticated as libraries adopt them. Users may "discover" the library portal from a Google search, a visit to the library, or in any of the ways users now discover new resources in libraries. The core feature of any portal will be integrated, cross-database searching of a local catalog, other library catalogs, selected web sites, locally licensed full-text and abstracting/indexing databases, and public domain or publicly accessible abstracting and indexing services.

Cross-database searching

Integrated searching is a key feature of a portal. It distinguishes it from a web site. Many library web sites provide access to the online catalog, licensed resources, vetted web sites, and links to one or more commercial search engines. However, access to these disparate resources is most frequently accomplished by searching one source at a time. Library web sites usually do not permit users to conduct a single search of multiple resources, nor do web sites deliver integrated results. Users usually have to integrate the results from their separate searches as another step.

Both the multiplicity of standards and the lack of standards are challenges in developing integrated, cross-database searching. Many online catalogs can be accessed by the international standard for search and retrieval, Z39.50, but additional search techniques are required for such resources as XML datasets or web resources using different metadata schemes such as MARC, Dublin Core, Computer Interchange of Museum Information (CIMI), and Encoded Archival Description (EAD).

Keyword searching is common for web sites. Licensed resources may have proprietary search strategies. As a result, portals must support various search standards and protocols (Z39.50 and http), and they must integrate the results. Portals also must support a variety of controlled vocabulary or thesauri. Library users accustomed to searching Medline, for example, will expect comparable results if they search Medline through a portal interface.

Commercial search engines like Google present results ranked by relevance. Users, however, often prefer one kind of relevancy over another. One user may want full-text resources displayed first. A second may opt for results from a specific journal or resource. A third user might choose all materials held locally. Portals must be able to rank search results differently to meet the needs of different users. Users should be given choices in how those results are ranked or listed. Some users, for example, may want only items from one source while others will want to see only a listing by date of publication.

Not just searching

While discovering an information citation or resource is obviously a core function of a portal, it is not enough for portals to be just search engines. Users want to use the information they discover and that means portals must provide for that use. The portal's supporting services must supply the ability to capture, integrate, manipulate, and distribute the information and offer ways to consult others and collaborate with them in the process.

At the minimum, users will expect to capture the information resource and bring it to their desktops. Most people prefer to get the full-text or full-image object rather than simply a citation. If the electronic document is not available, then capturing the citation is necessary. Portals will increasingly offer the functionality to read an OpenURL, transfer bibliographic or descriptive metadata, and check that the user has suitable permissions to access each relevant resource and then enable links to applicable resources. The OpenURL standard also enables dynamic linking to local copies of electronic journal articles, library catalog records, and remote commercial article services. It also helps maintain static URLs, seamlessly directing users to the most appropriate copy.

Another new open standard will permit users to access their circulation or ILL records from the portal interface. NISO members approved the NISO Circulation Interchange Protocol (NCIP) standard in July 2002. When NCIP is incorporated into portals, users will be able to place holds, recall items, or check on the status of ILL requests.

Portals can also offer the ability to transfer captured citations into ILL requests, commercial document delivery requests, or requests for the library to purchase the item. Portals that do this eliminate the current need to search in one tool and enter ILL requests using another. It is the seamless flow from discovery to requesting that will make for a successful portal.

After capturing the information, users expect to be able to integrate some or all of it into a variety of related applications. For example, users may wish to put images from the Library of Congress's American Memory Project into class assignments created within Blackboard or WebCT, perhaps including an audio clip relating to the image. Users expect to manipulate the captured information objects or citations by excerpting text, annotating citations, creating bibliographies, and manipulating images. Integration with local e-mail systems, calendars, and campus schedules must be part of new portals.

The center of education

Students in many universities now expect to learn in a collaborative environment, either physical ones such as the Information Commons at the University of Arizona Libraries or Emory University Libraries or electronic ones such as chat rooms or shared workspaces. In a portal environment, students will expect to share information resources with their classmates, and faculty will expect to distribute class assignments, engage in chat-room discussions, and provide pointers to information resources. Portals that seamlessly link to learning management environments will bring library-vetted resources to students and faculty rather than forcing them to use the library web site to search for materials.

Links to virtual reference services will permit users to seek the help of reference librarians when they need it and without physically going to the library. Portals will be viewed as virtual libraries, with a range of services equal to those currently provided in the library.

Integration, not isolation

Library portals cannot be developed in isolation. They must be interactive with many other systems, including university portals, content and course management systems, and document management systems. For academic institutions, the library portal must link to the learning management systems (LMS) or course management software. So far there has been little interaction between these systems. Learning management systems such as Blackboard and WebCT have been developed with minimal input from librarians. These systems support a variety of functions, including access to faculty-developed course materials, required and optional readings, chat rooms, and grades.

One area of potential integration is for LMSs to access electronic content licensed by the library rather than having the faculty or department license identical resources from a commercial supplier. Transfer of data between a portal and an LMS should result in seamless interaction between the two applications. For example, when students are searching the library portal, they should be able to transfer citations as well as electronic resources from the online catalog or other resources directly into the LMS. Likewise, students should be able to search the online catalog and other library resources from within the LMS and find references to resources or the electronic resources themselves.

Many models

Many academic institutions are developing university or college portals. Some view the library portal as simply a channel on the university portal, while others view the library portal as parallel to the university portal. Both alternatives will evolve, and regardless of which option predominates, access to information resources will be available to information seekers.

Research libraries, like academic, public, and special libraries, view library portals as an enabling tool to support the mission of their institutions. Librarians have high expectations for how portals will facilitate access to a wide range of high-quality content. The number of companies marketing portal products is expanding rapidly, and the number of libraries offering portals with integrated searching and multiple supporting services will continue to grow. The challenge and the fun ahead is to design portals with enough flexibility to respond continually to user preferences in the discovery, presentation, and use of high-quality information resources.

The ARL Scholars Portal Initiative Mary E. Jackson To advance the concept of a collective research library presence on the web, the Association of Research Libraries (ARL) established the Scholars Portal Working Group in early 2000.

In a May 2000 white paper, Jerry Campbell, chief information officer and dean of university libraries, University of Southern California (USC), developed ideas from the 1999 ARL-OCLC Strategic Issues Forum. He asked ARL members to pursue seriously the feasibility of developing a "library.org" web presence, argued for a collaborative partnership approach, and asserted that research librarians are better qualified to create a Scholars Portal than anyone else. The first to articulate some of the portal's key features and functions, Campbell stated that it would be "the place to start for anyone seeking academically sound information."

The working group, confirming that the ultimate goal was to establish a suite of scholarly productivity tools and services, said it was essential to define an initial step. That step was the development of what Brian Schottlaender (UC-San Diego) termed a "super discovery tool" that operates across both licensed and openly available content in a broad range of fields and delivers high-quality resources. A list of required and desired features of this "super discovery tool" is included in the final report of the Working Group.

After an environmental scan in spring 2001, the working group found that technology exists to facilitate cross-database searching and related supporting services. A number of nonprofit and commercial agencies were found to be engaged in efforts driven by the same or a very similar vision. Some of these initiatives would accomplish the integration of metadata across different electronic resources and databases only within a relatively narrow scope (e.g., within a particular proprietary environment or subject area), or the scope was biased by commercial objectives. The working group narrowed its search and eventually selected one vendor.

The ARL board specified that the working group and the vendor should not describe the resulting project as a commitment on the part of ARL libraries to use the product, professing that the particular product would not carry an "ARL imprimatur" or endorsement. The board asked ARL staff to continue to monitor other vendor developments and library applications of search engines and resource integration software tools and to develop a set of "best practices" generated by libraries that implement the portal. That list is still in progress.

The Scholars Portal Project (SPP), a collaboration between several ARL member libraries and Fretwell-Downing, Inc. (FD) was launched in May 2002. At the October 2001 ARL Membership Meeting, Campbell had noted that a handful of vendors offered some reasonable percentage of the requirements, but it was the working group's belief that FD had more available, about 80 percent. In addition, FD was willing to collaborate with the working group members to build the other 20 percent. The SPP is expected to enable academic desktops to be connected to the web more effectively by presenting academic-quality collections and library expertise in a way that more closely matches the searching style and expectations of a new generation of students and faculty. The SPP intends to demonstrate the viability of the Scholars Portal vision with one vendor's products.

The libraries initially participating in the project are USC, University of California-San Diego, Dartmouth College, University of Arizona, Arizona State University, Iowa State University, and the University of Utah. Plans call for adding more ARL member libraries over the course of the three-year project. In that time participants will be able to determine how the FD-contributed development resources will be applied. At the end of the project participants will assess whether working with a single vendor resulted in the realization of the vision of the Scholars Portal Working Group. The libraries participating in the project sought and received ARL's ongoing involvement because they believe that this will spur all vendors--including but not limited to Fretwell-Downing--to work even harder to create or enhance products that serve the needs of research library communities.

Selected Vendors with Portal Products Mary E. Jackson The Scholars Portal Project is not the only portal activity in ARL member institutions. Other ARL members are working on similar projects with other vendors such as Cornell University's partnership with Endeavor and Boston College's use of Ex

Libris (USA)'s MetaLib. In February 2002, ARL surveyed its member libraries to identify the state of current or planned research library applications of portals. ARL recognized that there are many definitions and views of portals. The survey sought portals that include: 1) tools that enable the user to search across multiple sources and integrate the results of those searches, and 2) at least one kind of supporting service for the user (such as requesting retrieval or delivery of nondigital material, online reference help, etc.).

Seventy-seven ARL member libraries responded to the survey, with 16 responses fitting the definition of a portal. E-books and the local online catalog were the most frequently presented targets. Submitting online reference questions and interlibrary loan requests were the most common services offered. A complete analysis of the survey results may be found online (www.arl.org/access/scholars-portal/index.html).

The ARL survey confirmed that the definition and function of portals are still in flux. The following vendors offer portal products.

Auto-Graphics : AGent www.auto-graphics.com/ls_agent.html

Endeavor : ENCompass http://encompass.endinfosys.com/whatis/whatisENC2.htm

Epixtech: DigitaLink www.epixtech.com/products/digitalink.asp

ExLibris (USA) :MetaLib www.exlibris-usa.com/MetaLib/index.html

Fretwell-Downing :ZPORTAL & CPORTAL www.fdusa.com/products/zportal.html

Gaylord :Polaris PowerPAC www.gis.gaylord.com/Polaris/PAC/PowerPAC.asp

Innovative Interfaces, Inc. :Millennium Access Plus www.iii.com/html/products/p_map.shtml

The Library Corporation:YouSeeMore www.tlcdelivers.com/tlccarl/products/pacs/youseemore.asp

MuseGlobal :MuseSearch & Information Connection Engine (ICE)
www.museglobal.com/Products/index.html

Open Knowledge Initiative : http://web.mit.edu/oki/index.html

SIRSI :iBistro & iLink www.sirsi.com/Sirsiproducts/elibrary.html#what

VTLS: Chameleon iPortal www.vtls.com/Products/gateway

WebFeat : Knowledge Prism www.webfeat.org

Mary E. Jackson is Senior Program Officer for Access Services, Association of Research Libraries, Washington, DC Please visit http://libraryjournal.com for more information.

Glossary and Definitions

CAPs – Common Access Points

CT - CERL Thesaurus; the thesaurus of geographic, personal and imprint (printers' and publishers') names in the field of European printed books to 1830.

DFAS – Distributed Finding Aid Server

DTD – Document Type Description

DOI – Digital Object Identifier

Dublin Core metadata standards - interoperable online metadata standards that support a broad range of purposes and business models

EAD – Encoded Archival Description

EAMMS – Electronic Access to Medieval Manuscripts

HTML – Hypertext Mark-Up Language

HPB – Hand Press Book database, established by CERL, covering printed material of the period 1455 – c. 1830.

LCAP – LOCKSS Communication Protocol

LOCKSS – "Lots of Copies Keep Stuff Safe"

Metadata - is information about data or other information that provides a common set of terminology, definitions and information about values to be provided in information delivered across the Internet

MSP – Manuscript Studies Portal

MALVINE – Manuscripts And Letters Via Integrated Networks in Europe

MARC – Machine-Readable Cataloging

MASTER – Manuscript Access through Standards for Electronic Records

OAI – Open Archives Initiative

Open URL - A metadata transportation format; for linking OpenURL is the emerging standard.

RLG – Research Libraries Group

SGML – Standard Generalised Mark-up Language

SRW – Search/Retrieve for the Web

SRU - Search and Retrieve URL Service

TEI – Text Encoding Initiative

TEI/MASTER – A joint standard: TEI is global; MASTER is specifically for manuscripts

TEL –The European Library

XML – eXtensible Mark-up Language

Z39.50 – The International Standard, ISO 23950 Information Retrieval (Z39.50): Application Service Definition and Protocol Specification

ZING – Z39.50 International: Next Generation