

# Scaling and productising MOSAIC search and recommendation services

**Mark van Harmelen**

[mark@hedtek.com](mailto:mark@hedtek.com)

26 September 2010

**Hedtek Ltd**

[hedtek.com](http://hedtek.com)

# Table of Contents

Table of Contents .....	2
1 Introduction .....	3
2 MOSAIC search and recommendation .....	6
2.1 Use data levels and kinds of search and recommendation .....	6
2.2 A top level strategy for MOSAIC search and recommendation .....	7
3 Work so far: The prototype MOSAIC search engine .....	8
3.1 The search engine and its user interface .....	8
3.2 Social search and faceting.....	11
3.3 Broadening search: Reading lists and ‘more like this’ .....	11
3.4 The prototype’s architecture and implementation technology .....	12
4 Production deployment via incremental development phases.....	15
5 The next phase: recommendations .....	17
5.1 Architectural approach .....	17
5.2 Recommendations from the MOSAIC project .....	18
5.3 Requirements for the next phase .....	18
5.3.1 Implementation activities .....	18
5.3.2 Preparation activities .....	20
6 Architectural considerations and requirements .....	22
6.1 Use data repositories .....	22
6.2 Indexing.....	22
6.3 Index sizes .....	23
6.4 Building indexes and updates to indexes.....	25
6.5 Scaling and a reference architecture .....	25
6.5 User interfaces and usability.....	26

# 1 Introduction

The MOSAIC search and recommendation approach was a successful outcome of the MOSAIC project.<sup>1</sup> The MOSAIC search engine demonstrated personalised search based on user context and activity data. However, we are in an early stage of development of MOSAIC search and recommendation: Search has been prototyped, and a successful user evaluation has been performed. Recommendation has not been prototyped, though the University of Huddersfield has developed a local recommender service.

Through the MOSAIC project, we extended and refined our understanding of the approach's functionality and now consider that MOSAIC provides the following approaches

- MOSAIC search implements **social search**
- Individual resources are **prioritised** in the search results according to activity data, specifically use data derived from library circulation records. At a later stage this approach can be generalised to include data derived from access to electronic resources, from such as e-journals, VLEs and repositories.
- A user can **personalise** the search according to their own identity and results are then prioritised according to the user's social group, their class cohort.
- Users can **refine** search results using facets to discover who borrowed or accessed what in another year, on another course, at another institution, and/or at another progression level.

*See figure 2 below to see an example of faceting being used to yield a result illustrating as might be obtained for the two bullets above.*

- MOSAIC **social recommendation** is of various forms that include "Borrowers like you borrowed ..." and "Borrowers of this item next borrowed ..."

---

<sup>1</sup> See, inter alia:

van Harmelen, M. *TILE project: architectural proposals for creating context and enabling contribution*. 2008. <http://ie-repository.jisc.ac.uk/295/>

Kay, D, Chad, K, van Harmelen, M, Pattern, D, Miller, P and Harrop, H. *Making Our Scholarly Activity Information Count: The JISC MOSAIC Project*. 2010. <http://ie-repository.jisc.ac.uk/446/>

To expand more formally on the bullets above, there is a whole hierarchy of increasingly complex MOSAIC levels <sup>2</sup> where any successor level incorporates the abilities of its simpler levels

*simple*

-----

- |           |         |   |
|-----------|---------|---|
| <b>0</b>  | Search: | Simple counts of use affect search ranking<br>eg use promotes items   |
| <b>0+</b> | Search: | With limited faceting on borrower type in the LMS,<br>eg on progression levels which might be UG, PG, Staff |

*LMS only data sourced above, LMS and registry data required below*

- |          |                  |  |
|----------|------------------|--|
| <b>1</b> | Search:          | With faceting on institution, course, progression, academic year                 |
|          | Recommendations: | Borrowers like you borrowed<br>Borrowers like this borrowed                      |
| <b>2</b> | Recommendation:  | Borrowers of this item also borrowed   |
| <b>3</b> | Recommendations: | Borrowers of this item went on to borrow<br>Borrowers of this item next borrowed |

-----

*complex*

The most compelling benefits for the MOSAIC approach come with level 1 and 2 facilities. It was level 1 search that was prototyped in the MOSAIC project.

There is a boundary between the zero levels and the higher numbered levels. The former needs only information from the Library Management System (LMS), the latter from the student and staff registry. In the MOSAIC project institutional participants gathering trial data found it hard to do unite data from these sources<sup>3</sup>, so as an interim step in incremental development of the MOSAIC activity data approach, the next stage as recommended here includes the simplest possible implementation for real production use: To implement 0 or 0+ search in single institutional libraries. For information at this stage, the next stage recommendations also have some preparatory work for MOSAIC level 1 implementations.

---

<sup>2</sup> Only levels 0,1,2 were defined in the MOSAIC project, 0+ has been identified after the end of that project, and 3 was omitted from definition as unlikely to be attained in the MOSAIC project.

<sup>3</sup> This report recommends work in this area as a next step.

Looking at the general picture, the possible kinds of services not only encompass levels as above, but also the 'constituency' of the search engine, from providing a search/recommendation service in a single institution, or for best effect, across a cluster of institutions, or nationally.

Thus with five levels and three constituencies we have the following.

level	single institutions	clusters	nationally
0 item promotion in results based on use	A		
0+ one or possibly two-faceted search	A		
1 multi-faceted search borrowers like you borrowed  borrowers like this borrowed	B misses one facet	B	B
2 borrowers of this item also borrowed	C	C	C
3 borrowers of this item went on to borrow  borrowers of this item next borrowed			

Possible phases of activity are:

NEXT: Implement A and prepare for some/all of B

THEN: Implement some/all B and prepare for some of C

THEN: Implement some of C and prepare for more of C

This is an incremental approach that can be explored as need be, without commitment to large upfront design and investment in an area which still contains engineering and service unknowns. However the Huddersfield implementation of recommendations provides a practical production proof of the feasibility of using gathering and using level 2 data within a single institution, given that access is available to both LMS and student registry data.

**In what follows, MOSAIC level 1 search is introduced as an easy way to visualise the first sweet point in this incremental development.**

## 2 MOSAIC search and recommendation

### 2.1 Use data levels and kinds of search and recommendation

MOSAIC search and recommendation is based on activity data, specifically use data in the form of circulation records, possibly augmented with anonymised borrower information.

This use data is associated with loans of physical resources. There is no reason not to, in the longer term, also leverage attention data, that is comprised of information about access to electronic library resources. In the longer term this is highly desirable. However, in what follows we will only talk about loan data.

We refer to library circulation (loan & renewal) information as use data. Use data contains one use record per item loan. A renewal can be counted as a loan. Sets of use records may have different amounts of information in each record, according to the data level that applies to all the records in the set. The possible levels and their use appear below, and are discussed as if there is a cluster of participating libraries or a national service that pools their use data. Of course, a single institution approach is possible. The latter can be used as part of a growth path to larger clusters of participating institutions.

Level 0	Level 0 use records contain (where and) when the loan was made and the item borrowed.	Level 0 use data can be used to indicate popular loan items in a participating library or across all participating libraries. The effect is to promote popular items in search results or less usefully to make recommendations on gross usage data alone
Level 0+	As level 0 use records but with the type of borrower derived from LMS-held data.	Allows search with some faceting, eg on progression level, if progression level is held in the LMS in order to assign borrowing quotas. If the LMS holds progression level and course data this is a high-win situation that will be  very economic to exploit
Level 1	As level 1 use records with more borrower context information, indicating institution, course, progression level, and academic year of loan. This requires merging circulation data with student registry data.	Level 1 use data can be used to see, via facets, for a given search, what was borrowed in one or more of: a particular institution, a particular course, a particular progression level (or by staff), and in a particular academic year.  The MOSAIC demonstrator provides these search facilities.  The data allows recommendations of the form <i>borrowers like you borrowed ...</i> and <i>borrowers like this borrowed ...</i>

Level 2	As level 1 use records with an anonymised user ID	Level 2 use data enables recommendations of the forms <i>borrowers of this item also borrowed ...</i>
Level 3	As level 2 use records but with a sequence number for the loan item within the borrower's loan history.	Level 3 use data enables recommendations <i>borrowers of this item went on to borrow ...</i>

#### Notes

- This scheme expands a little on the earlier MOSAIC scheme, adding level 0+ and refining earlier level 2 descriptions to break out level 3 from level 2.
- Each level can also be used for the purposes of the lower levels
- All data is anonymised
- A definition of use data formats appears in Appendix 2 of the MOSAIC Final Report.
- Strategies to address data protection and anonymisation requirements are covered in section 7.1 of that document.

## 2.2 A top level strategy for MOSAIC search and recommendation

The MOSAIC Project Final Report provides some sage advice as to the development trajectory for MOSAIC search and recommendation:

“As the options have crystallized, it has become clear that the exploitation of use data should not be conflated with the potential for web-scale aggregation, with the derivation of deeper context or with the "open data" agenda. The feedback from HE libraries indicates that

- Activity data can be exploited at a local level as well as at the network level. Institutions can build on local systems (e.g. to provide management data) without making data public or aggregating beyond the institution.
- Linking library activity data with course data (thus providing rich context) is not without hurdles in terms of data sources (e.g. VLE, Registration systems) and data levels (Course or Module).

In order to build critical mass of adoption, based on interest, business case and confidence, it is therefore important not to undervalue the local use of library activity data in its simplest form (as per the MOSAIC Level 0 specification)....

In parallel, work can be done nationally to open up the benefits of network effect (via data aggregation) and/or an open data approach.”

### 3 Work so far: The prototype MOSAIC search engine

It is worth understanding the prototype search engine implemented during the MOSAIC project in order to gain a feel for MOSAIC search. Readers with a serious interest are also encouraged to read the CERLIM user survey and report to, eg, more deeply understand the need for reading lists.

The demonstrator can be accessed at <http://mosaic.hedtek.com/>

#### 3.1 The search engine and its user interface

The search engine implements facets that allow the user to narrow search results, refining those activity (borrowing) results by any combination of a particular Institution, Course, Progression level, and/or Academic year.

**This is called the *IPCA combination*.**

Faceting is illustrated by the screenshots in Figs 1 and 2 below.

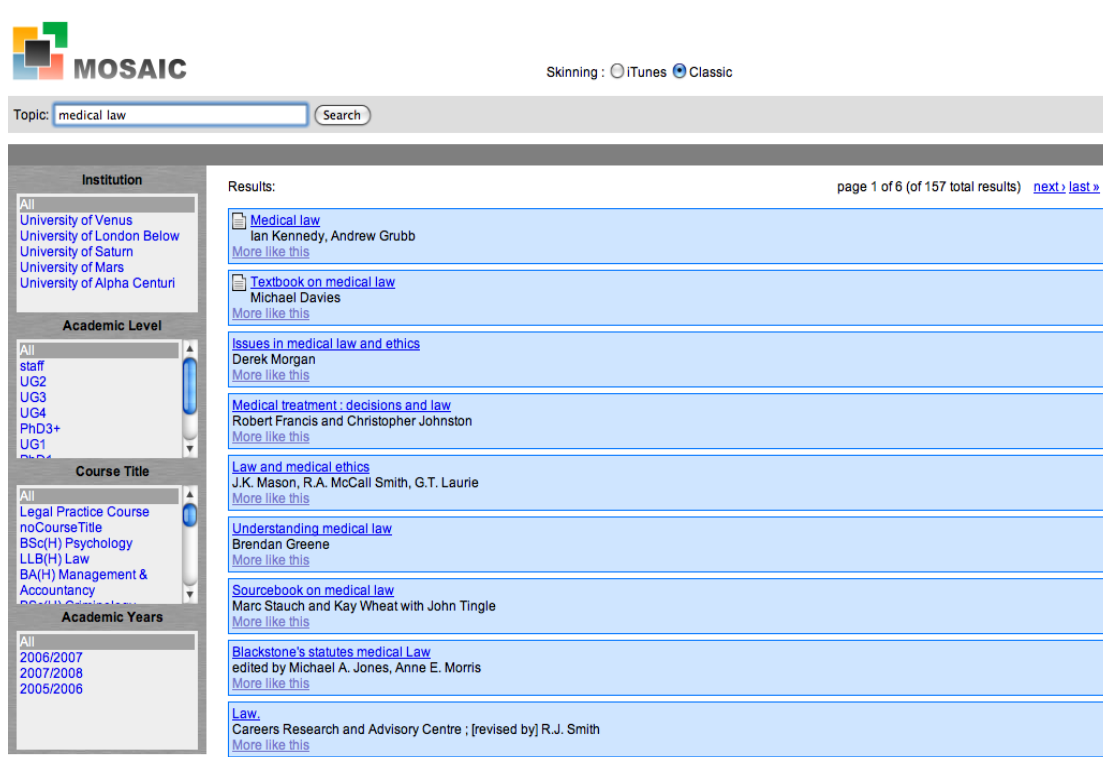


Figure 1: Basic (ie not-yet-faceted) search results for medical law

Topic:

UG2 x DipHE (BSc(Hons) option) Nursing Studies - Child x

**Institution**

University of Venus

---

**Academic Level**

All

UG2

---

**Course Title**

All

DipHE (BSc(Hons) option)

Nursing Studies - Child

---

**Academic Years**

2006/2007

Relevant results: page 1 of 6 (of 157 total results) [next](#) [last](#)

[Law and ethics](#)  
Judith Hendrick  
[More like this](#)

Other results:

[Medical law](#)  
Ian Kennedy, Andrew Grubb  
[More like this](#)

[Textbook on medical law](#)  
Michael Davies  
[More like this](#)

[Issues in medical law and ethics](#)  
Derek Morgan  
[More like this](#)

[Medical treatment: decisions and law](#)  
Robert Francis and Christopher Johnston  
[More like this](#)

[Law and medical ethics](#)  
J.K. Mason, R.A. McCall Smith, G.T. Laurie  
[More like this](#)

[Understanding medical law](#)  
Brendan Greene  
[More like this](#)

[Sourcebook on medical law](#)  
Marc Stauch and Kay Wheat with John Tingle  
[More like this](#)

[Blackstone's statutes medical Law](#)

**Figure 2:** Using faceting: the top darker result is for a particular combination of university, course, progression and academic year


The above screen shots illustrate the 'classic' shopping experience offered by many online retail sites, with facet / filter choices vertically arranged. The prototype also implements an alternate switchable 'iTunes' like interface, with facets / filter choices displayed horizontally, as shown below in Fig.3:

Topic:

Institution	Academic Level	Course Title	Academic Years
University of Venus	All UGz	All DIPHE (BSc(Hons) option) Nursing Studies - Child	2006/2007


Relevant results:

 page 1 of 6 (of 157 total results) [next](#) [last](#)

 [Law and ethics](#)  
Judith Hendrick  
[More like this](#)

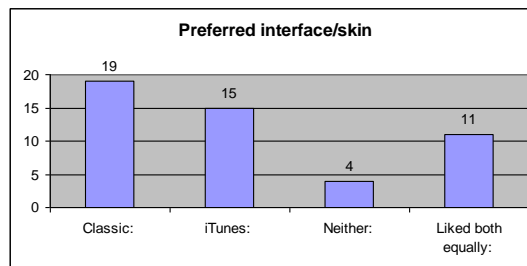
Other results:

 [Medical law](#)  
Ian Kennedy, Andrew Grubb  
[More like this](#)

 [Textbook on medical law](#)  
Michael Davies  
[More like this](#)

**Figure 3:** The iTunes like interface

In user testing with students CERLIM [2009] found a marginal student preference for the 'classic' interface. In contrast, those two members of the MOSAIC project team who are frequent iTunes users found the iTunes-like interface to be preferable.



**Figure 4:** User preferences as to interface style [CERLIM, 2009]

It is suggested that future interfaces concentrate on the classic interface if a choice has to be made; this echoes the standard faceted shopping experience. However, a switchable interface is almost no effort to implement, merely involving styling via CSS.

## 3.2 Social search and faceting

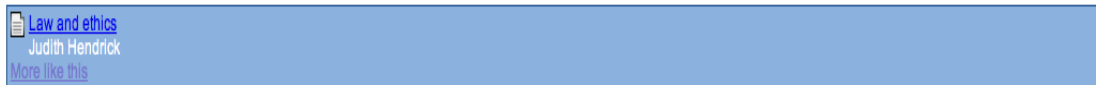
The MOSAIC prototype provides MOSAIC level 1 **social search as follows**:

1. Users can **personalise** the search according to their own ICPA identity; in a production system this could be automated for registered user.
2. Users can **refine** search results with any ICPA facets relevant at the time, e.g. to discover who borrowed what in another year, on another course, at another institution

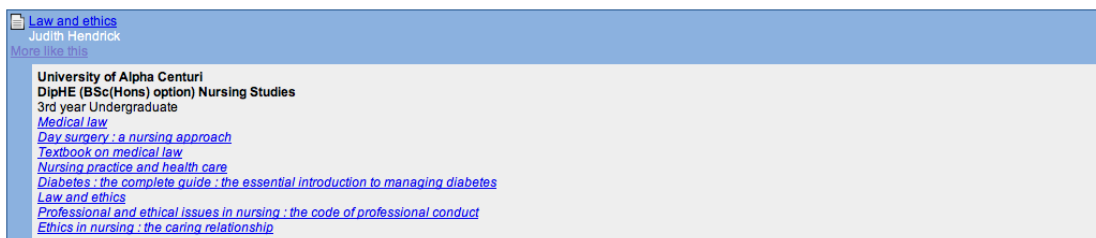
## 3.3 Broadening search: Reading lists and ‘more like this’

Because of some initial concern that the effect of ICPA-based recommendation would narrow choices, two extra mechanisms were introduced to help broaden choice:

- Reading lists become available for any book being recorded as part of a reading list to enable the learner to investigate related titles supplied by subject experts (the teachers/lecturers supplying the reading list). Icons were attached to items that are in the reading lists, clicking on one of them reveals the corresponding list – see Figs. 5 and 6. In some future implementation co-operative and ‘own’ list construction could be made available. In this, the work of the Open University’s TELSTAR project is interesting.
- A ‘More like this’ link was added to each returned search result. The link merely uses the item’s title as a search term and returned the matching results – see Fig 7. The result of this simple tactic is surprisingly effective.

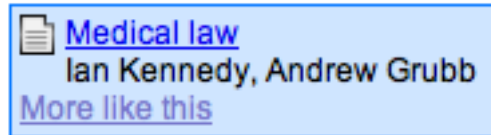


**Figure 5:** A reading list icon in the results



**Figure 6:** Clickable items in a reading list

Reading lists were highlighted in reviews as a powerful browsing and serendipitous discovery mechanism; therefore reading lists need to be made more accessible in the interface (without necessarily accessing lists via search results). This is an easily implemented improvement.



**Figure 7:** A 'More like this' link in an individual search results in the result set

### 3.4 The prototype's architecture and implementation technology

The basic architecture for the search facilities prototyped in the MOSAIC project was proposed in the TILE project<sup>4</sup> and comprised

- A mechanism to gather use data and construct a search engine index that incorporates the use data
- A search engine with customized matching and ranking algorithms
- A browser-based user interface providing faceted search

Two standard open source solutions are available to build custom search:

- Lucene is a search engine that may be extended to build custom search engines
- Solr builds on Lucene facilities to provide, amongst other things, a machine-to-machine web interface to its Lucene core.

The project chose to use and extend Solr to implement the search engine, and provided a front-end user interface using HTML, CSS, JavaScript and AJAX. For ease of implementation with immediately available data, we limited the search engine to index only ISBN-identified books.

As proposed in the TILE project, a key requirement of the search engine was to facet on any combination of Institution, Course, Progression level and/or Academic year (ICPA), so the search index records records were augmented as in this example:

```
Gore Vidal, The Selected Essays of Gore Vidal, ...  
<unique reference to catalogue entry>,  
  
<I, C, P, A, count> <I, C, P, A, count>, ...
```

For the MOSAIC prototype, we exploited the catalogue system from the Extensions for the Information Environment (EIE) project<sup>5</sup>, using the TILE architecture, thus enabling the records to

---

<sup>4</sup> <http://www.jisc.ac.uk/whatwedo/programmes/resourcediscovery/tile.aspx>  
<http://www.sero.co.uk/jisc-tile.html>

For the TILE roots of the architectural approach in this document see van Harmelen, M. *Creating Context & Enabling Contribution – Architectural Proposals: A guide to the TILE architecture and e-Framework SUMs*, 22 December 2008 <http://ie-repository.jisc.ac.uk/295/>

<sup>5</sup> <http://www.jisc.ac.uk/whatwedo/programmes/resourcediscovery/eie.aspx>

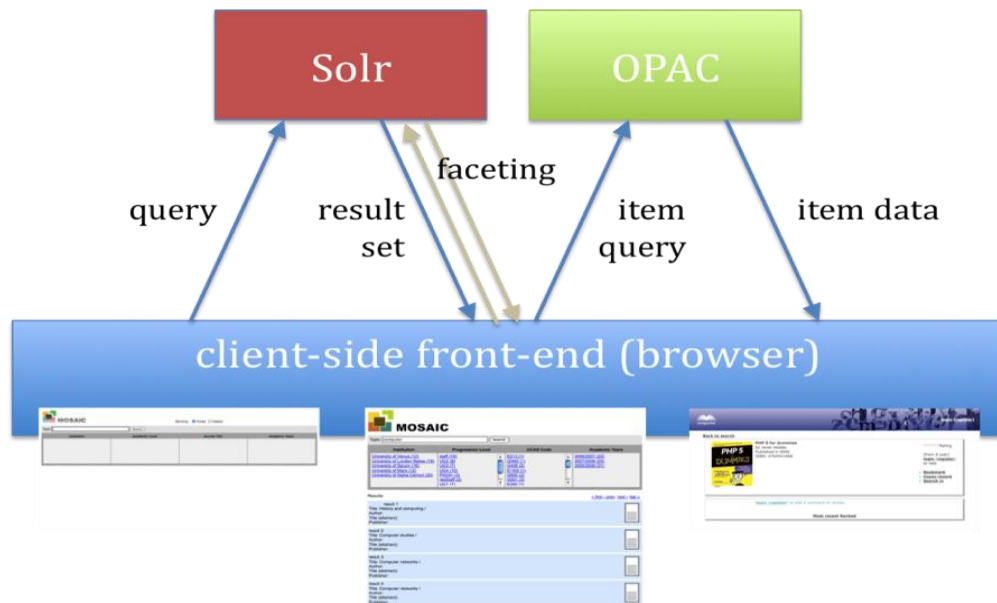
be enriched with catalogue metadata and social recommendation extensions. This allowed us to easily retrieve book records on the basis of ISBNs, and ISBNs were therefore used as the unique reference or identifier for the work. Progress does need to be made with identifiers both for description of a range of scholarly resources, and for use across institutions (options for identifiers are broader in single institutions, for example a catalogue holding number can be used).

Solr's Lucene core was modified to enable faceted search over result sets on the basis of <I, C, P, A, count> tuples in the result set.

The count in the <I,C,P,A,count> tuples was used to indicate how many times a book was borrowed by a given ICPA cohort, and could be used either to provide specialist ranking within search results (we used Lucene's default), and/or a facet count (which we dropped from the final prototype to simplify the user interface).

Pre-processing to build the search engine index was done by Python-based data manipulation to obtain Copac-derived book information for those books appearing in the use data. This derived information was then merged with the use records, and the EIE catalogue was re-built with those books appearing in the use data. With all of this in place the search engine operates as in Fig. 8 below:

---



**Figure 8:** Events in the search architecture when a search and facet-based refinement is performed to return a catalogue page as an end-result

## 4 Production deployment via incremental development phases

We recommend productising and deploying MOSAIC search in an incremental fashion. This starts with a first target of rollout of simple activity-data based search in single institutions. In later stages we move on to targets with greater functionality.

The pattern of development advocated here is an agile one, where each stage's activities are chosen before the stage commences. In general a stage should result in production, ie rolled out systems, and in preparation work that enables the next stage's rollout(s). As in any agile approach, knowledge gained, for example about requirements and ways of satisfying them or from the use of previously delivered systems, will affect what activities are undertaken in future stages.

Both production and preparation can be seen in the oval showing recommended activities for the next phase of MOSAIC development. See the *focus of next activities* in figure 9.

In Figure 9

- Preparatory activities are shown on the left of the figure.

As shown by arrows, preparatory activities feed into

- product development and rollout activities on the right and bottom of the figure.

To recap

- In the agile development we propose, we don't commit to any or all product development and rollout targets in figure 9 up front.
- Instead, informed by preparation, development experience and production system use, we choose what to perform at the beginning of each incremental phase of activity

Additionally, institutions who engage in simpler production activities in figure 9 will be able to move to later more complex implementations by selective upgrade of their production components.

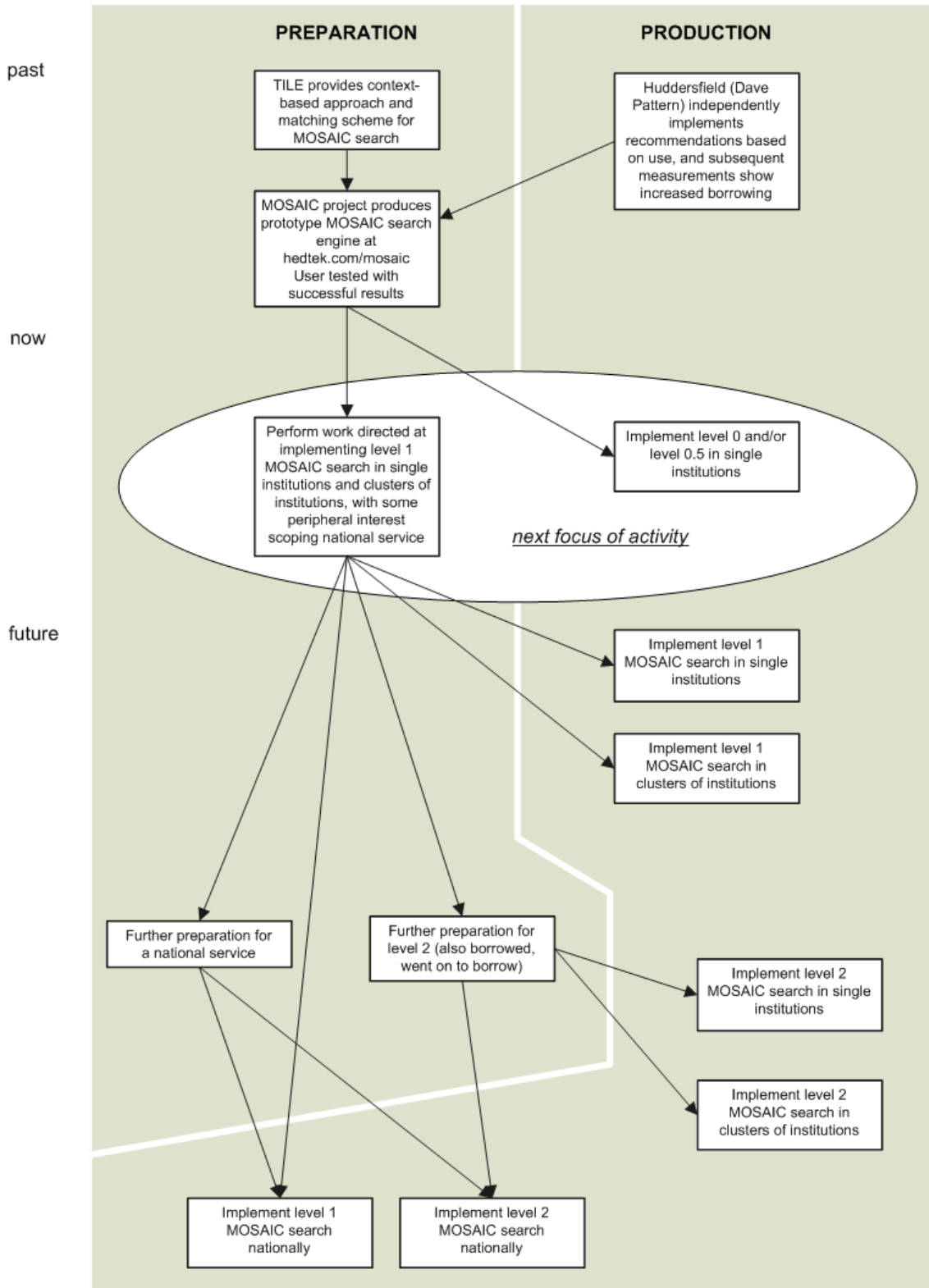


Figure 9: Stages in MOSAIC development and implementation

## 5 The next phase: recommendations

There is a spectrum of possibilities that form constituents of the next phase.

- The MOSAIC Final Report recommends some immediate actions; the most compelling of these in respect of technical development are listed in section 5.2.

These could usefully be part of next-phase activities (see Figure 9).

- The 'incremental phase recommendations' in this document are more architecturally and practice concerned, to head directly for a rollout of simple level 0 to level 0+ attention data gathering, construction of search facilities in some individual institutional libraries, and preparation for the next stage implementation and rollout.

These recommendations appear in section 5.3.

### 5.1 Architectural approach

We approach the architecture with the following three foundational principles:

- Search should be as fast as possible, within reasonable economic constraints.
- User interface and usability concerns are paramount.
- There should be a single scalable architecture for varying size installations, from single institutions to nationally.

Adopting the third principle leverages economies of scale; the consequence is that it should be an architecture that scales out, since it would be infeasible to scale this up for a national level system.

The most viable choice is a Lucene or Solr architecture that is replicated using Hadoop.

In fact, the index sizes are sufficiently small that for single institutions and small clusters a multi-core Hadoop installation can be used. As an indication for today's prices, a suitable multi-core machine would cost in the region of £3,000 to £4,000 + V.A.T.

Architectural factors and constraints are discussed in section 6. At this stage it is worth mentioning that the amount of time that use records are held for depends on the application for those records:

- Maximal returns are in providing facilities for undergraduates, for this purpose we recommend holding for five years of use records
- For collection curation purposes we recommend up to 15 years of use records

## 5.2 Recommendations from the MOSAIC project

The enabling activities (ie Recommended Immediate actions) from the Mosaic final report that resonate here are as follows, with their original numbering from that report. The normal font comment under each action is specific to this document.

- 1      **Publish case studies evidencing the range of opportunity and benefits**  
  
To arrive at this we need production installations, the first stage in incrementally more complex production installations is recommended below in section 5.3.
- 2      **Commission data extraction cookbooks and tools for popular systems**  
  
Requires data extraction systems to be built for popular library systems, initially for levels 0, 0+ and level 1 data.
- 4      **Establish a national data and services platform for use data**  
  
We can begin work on requirements for this, or implement it in an incremental fashion, since it is conceptually a very simple system for data storage and access. The eventual scale of the platform is a consideration, and this needs to be attended to, possibly in a phased manner.
- 6      **Commission an advisory report on corporate data use issues**  
  
As recommended here, this should simply concentrate on fair use notices as allowed for under the Data Protection Act, but others may want broader advice.
- 7      **Instigate development of aggregated service foundations**  
  
This is a recommendation is for a national service, which is not directly considered in this document but instead is mentioned as a potential target.  
  
However, the MOSAIC Final Report does have an incremental approach to this were it to be chosen as a target.

## 5.3 Requirements for the next phase

Please refer back to figure 9 to see the activities recommended here situated in the broader incremental approach.

### 5.3.1 Implementation activities

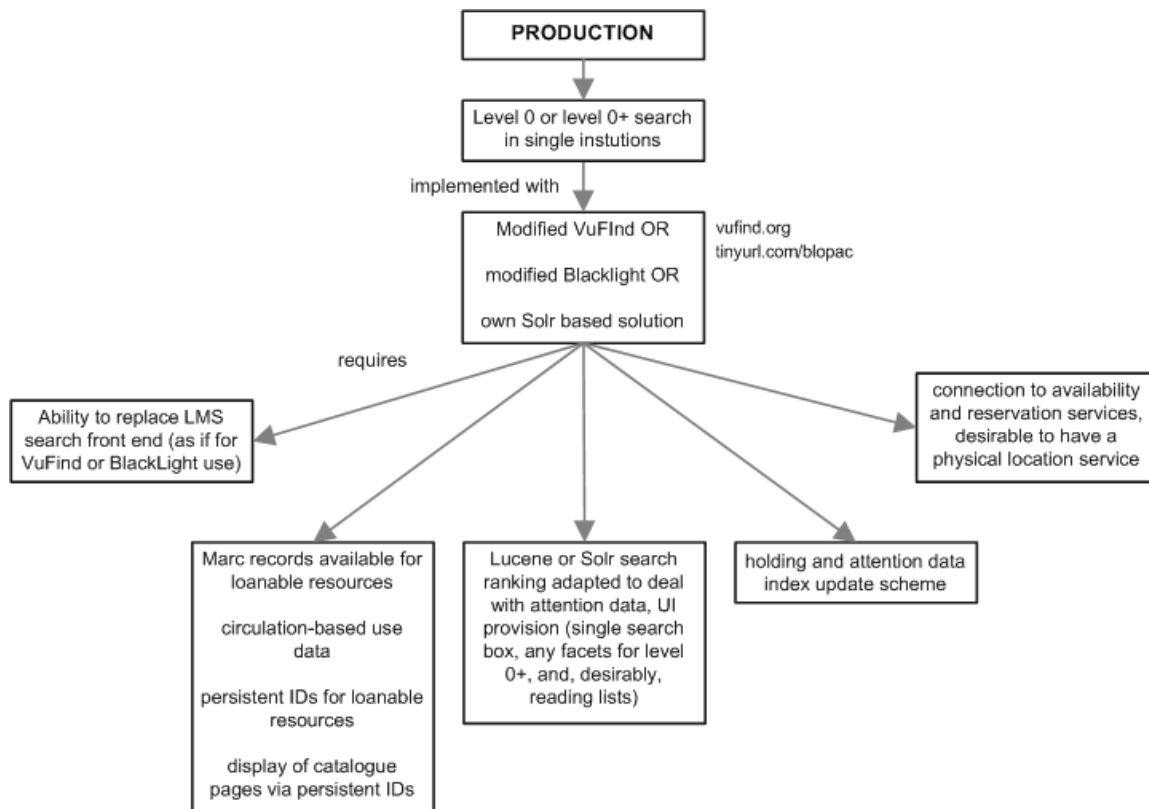
The obvious next phase is to avoid difficulties in merging data from library systems and registry systems, while preparing the community for MOSAIC and gaining experience with library systems

that allow for the supply of a custom search front end. The jiscLMS projects <sup>6</sup> are already experimenting with custom front ends based on VuFind <sup>7</sup> or BlackLight <sup>8</sup>, both of which use Solr.

Level 0 search is recommended for a few trial institutions. The framework they institute will allow search front-ends to be swapped easily to allow for level 1 search in a later phase.

Incorporation of reading lists as in the MOSIAC demonstrator (<http://mosaic.hedtek.com>) is highly desirable – CERLIM’s user survey and MOSIAC demonstrator usability study found users used and rated reading lists highly. Prior work with reading lists could be incorporated into MOSAIC search.

Elementary level 0 search could relatively easily become more MOSAC-level-1-like if the attention data collection leveraged information known about borrowers in the LMS. For example, the LMS might well hold UG, PG, staff status/progression level to determine borrowers’ loan quotas, and if easily extractable this could provide a first level of faceted search. See earlier sections of this document for more detail on this level, called level 0+.



**Figure 10:** Production activities recommended in the next phase

<sup>6</sup> <http://code.google.com/p/jiscLMS/>

<sup>7</sup> <http://vufind.org/>

<sup>8</sup> <http://www.lib.virginia.edu/digital/resndev/blacklight.html>

### 5.3.2 Preparation activities

These prepare for the subsequent phase of MOSAIC activities, needed in order to move on to provide level 1 services implementing one of MOSAIC's 'sweet spots', personalised social search.

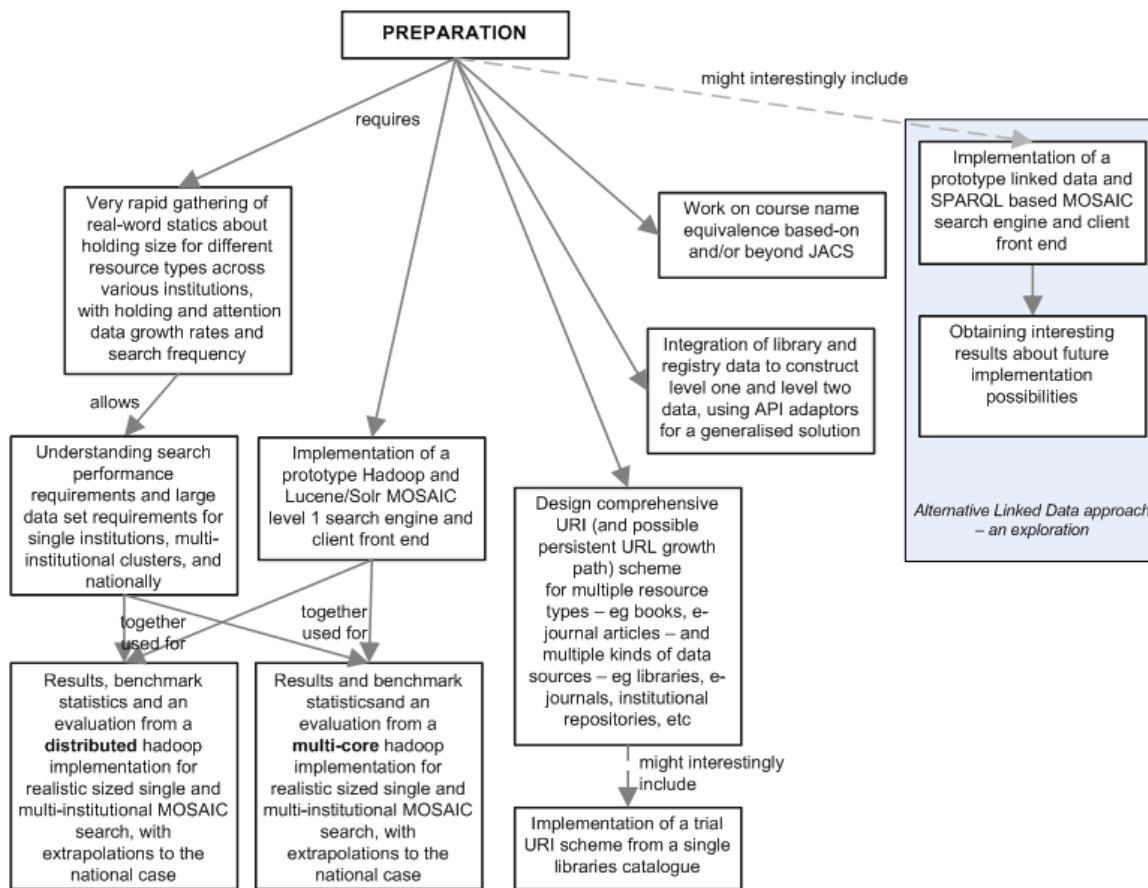


Figure 11: Preparation activities recommended in the next phase

One prioritisation of activities in figure 11 is

1. Statistics gathering
2. URIs/persistent URLs
3. Benchmarking
4. URI schemes
5. Generation of level 1 data
6. JACS refinement
7. Linked data and SPARQL experiments

For Activity 5, generation of level 1 data, we believe that most registry vendors are unlikely to supply suitable APIs, and some reverse engineering may be required.

## 6 Architectural considerations and requirements

### 6.1 Use data repositories

In an architecture as proposed here attention data (ie use data) is a valuable asset. Besides MOSAIC search and recommendation, we anticipate that use data can be used in other search and recommendation approaches, in collection management, spatial planning, and no doubt other attention-aware applications.

In the MOSAIC architecture a repository is used to store use data. For a single institution system the repository is part of the institution's intranet. Clusters have a repository that is created from use data from the co-operating cluster institutions, which would, in turn, be wise to maintain their own intra-institutional repository.

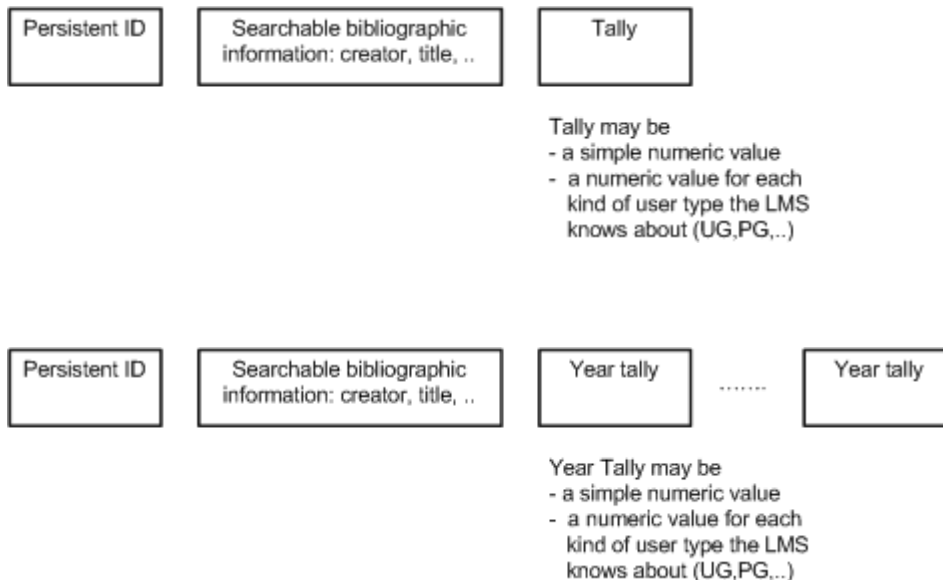
There is value in single institutions establishing institutional repositories for level 0 and 0+ data, they provide a growth path to collection of higher levels of use data.

With care, the same repository implementation could support one or a cluster of co-operating institutions, and level 0 though level 3 data.

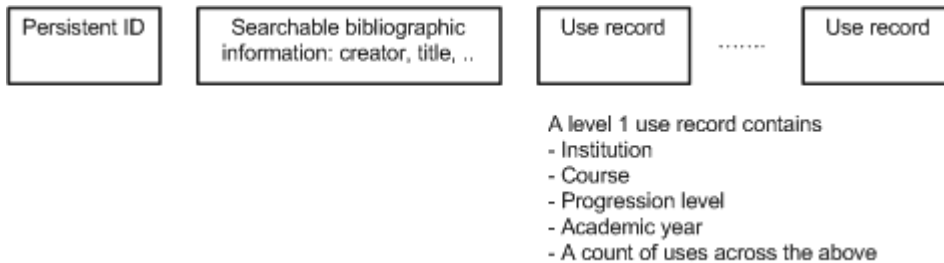
All data stored in the repositories is anonymised, this has been discussed at length elsewhere , eg the Mosaic Fila Report.

### 6.2 Indexing

There are different ways of organising indexes, with different consequences for the kinds of MOSAIC search facilities that may be supplied.



**Figure 12: Some level 0 indexing schemes (not exhaustive)**



**Figure 13: A level 0+ and above indexing scheme showing one of several different combinations of data in the use record**

### 6.3 Index sizes

At this stage one major step forward is a preliminary quantification of the size of a Lucene index under estimates of typical circumstances. Here we consider

- MOSAIC level 1 use data
- A small-to-medium sized institutional library with 0.7M loanable items
- 12 items borrowed by each of 20,000 students p.a. (12 items/student/year is a figure for Huddersfield’s current library use)
- Use data could be held for either five years, mainly directed at the purpose of assisting our best-returns target, undergraduates, or for fifteen years for collection curation purposes (eg to support “move all books that have not been borrowed for ten years from the stacks to long-term archival storage”).

The figure used below is 15 years. Divide the figures by 3 for the UG assistance case

Index sizes are presented for both a single institution and for a cluster of five institutions.

In the figure 12 the middle column is based on miserly sizes of catalogue and use records in the Lucene index, and right-hand side column is based on totally over-the-top sizes.

For a single institution index (in bytes)

<b>index size needed</b>	<b>1,352,000,000</b>	<b>21,200,000,000</b>
--------------------------	----------------------	-----------------------

For a five institution cluster index

<b>index size for cluster</b>	<b>6,760,000,000</b>	<b>142,000,000,000</b>
-------------------------------	----------------------	------------------------

Only keeping five years of use records (to hit the maximal use target of supporting UG search), and using a plausible holding item and use record size might yield an index size of 6GB–10GB. In

modern server implementations with memory sizes of up to 192GB, this leaves plenty of memory space to hold and process in-memory search results.

For reference the figures used were

loanable items	700,000	700,000
size in bytes	400	10,000
index for loanable items in bytes	280,000,000	7,000,000,000
java data structure overhead	280,000,000	7,000,000,000
<b>searchable catalogue space in bytes</b>	<b>560,000,000</b>	<b>14,000,000,000</b>
% of total index required	41	49
students	20,000	20,000
loans per student per year	12	12
loans per year	240,000	480,000
number of loanable items replicated	10,000	10,000
use records arising	10,000	10,000
use records epr annum	240,000	480,000
use record size	110	1,000
java data structure overhead	110	1,000
Index required	220	2,000
<b>use record space pa</b>	<b>52,800,000</b>	<b>960,000,000</b>
Years back data, N	15	15
<b>use record space N years</b>	<b>792,000,000</b>	<b>14,400,000,000</b>
% of total index required	59	51
<b>index size needed</b>	<b>1,352,000,000</b>	<b>28,400,000,000</b>
cluster size	5	5
<b>index size for cluster</b>	<b>6,760,000,000</b>	<b>142,000,000,000</b>

**Figure 14:** Some preliminary index size calculations

## 6.4 Building indexes and updates to indexes

Consider the scenario where a lecturer announces the importance of a some borrowable library resources close to an exam. Almost immediately students will start borrowing the resources. A student goes to the library without a list of these resources, and probably can determine what the works were based on her ICPA combo. If authenticated, she might do this by pressing a button called 'what are people like me borrowing?' Similarly a student might use the same mechanism to look at the most popular items in the list announced by the lecturer.

Both of these are examples of level 1 search with (close to) real time update of use records in the search index. We view frequent update of the search indexes as highly desirable for customer buy-in given that customers already have experience of an increasingly real time Web.

To achieve this, the service must allow incremental use data and holding updates. For use data, this will involve update to the use data repository and then to the index. Either a push or a pull architecture could be used to gather updates, though a real time update<sup>9</sup> requires push from the library system. Similarly, for real time update in cluster architectures there needs to be institutional push to a central use data repository.

Close to real time updates are probably sufficient, but it may be as easy or easier to offer a real-time update API. Solr offers the basis for such a facility.<sup>10</sup>

## 6.5 Scaling and a reference architecture

There has already been some consideration of clusters of institutional libraries forming a common level 1 search facility. This might become common, in say a metropolitan area, such as Greater Manchester, or across larger areas.

Three existing partnerships in the library domain are

- South West Wales Higher Education Partnership<sup>11</sup> – 3 institutions
- Scottish Confederation of University and Research Libraries<sup>12</sup>
  - some 23 institutions / libraries
- The M25 Consortium of Academic Libraries<sup>13</sup> – 59 institutions

As cluster size increases, so will the size of cluster's common search index, to the point where a single computer can no longer deal effectively with searches. Note, parenthetically, that there is a similar situation, of a single computer becoming unable to process searches in a reasonable time with increasing search frequency.

As cluster sizes increase (or as numbers of searches increase) scaling out – partitioning the search across more computers – becomes the only economic response.<sup>14</sup>

---

<sup>9</sup> Push is not needed for close to real time update

<sup>10</sup> <http://lucene.apache.org/solr/tutorial.html#Updating+Data>

<sup>11</sup> <http://www.swwhep.ac.uk/>

<sup>12</sup> <http://scurl.ac.uk/>

<sup>13</sup> <http://www.m25lib.ac.uk/>

The infrastructure for scaling out is provided by Hadoop, which is open source. Hadoop is conventionally used with Lucene for large scale search. Hadoop enables parallel Lucene processes to independently perform parts of a search before combining the resultant multiple search results into a single list of search results for user consumption.

Hadoop and Lucene traditionally run on a number of loosely (Ethernet) coupled computers, but recent advances have seen Hadoop run on multi-core architectures within a single computer. Without the multi-core advances, a likely scenario might be the use of Solr or Lucene for single institution search, and Hadoop and Lucene for clusters. However, the multi-core advances are interesting, because they raise the possibility of single reference architecture for both single institutions and clusters. For single institutions the Hadoop-based architecture runs on a multi-core machine. Then, as necessitated by scale-out demands the physical infrastructure can be implemented using the same Hadoop-based architecture running on multiple computers.

Very high performance Hadoop implementations (and Google's map-reduce implementation) are known to use other techniques, eg distributing part of the reduction of multiple search results, having a particular master architecture, and caching common search results. It is anticipated that the reference architecture hypothesized above might not scale to a national architecture, which might need to implement further performance enhancements.

## 6.6 User interfaces and usability

A trend in library search user interfaces is moving to a simple search box, with subsequent results refinement via facets. Which in turn fits nicely with the ICPA approach outlined above.

Surfacing information and services in the search results is highly desirable in order to enhance usability. The following might be easily obtainable directly in the search result, or by interacting with the result eg hovering over it, clicking an icon or selecting from a drop down menu: More information about the item, copies available as reference or for loan, the ability to reserve a copy, shelf location and its position in the library, the ability to SMS details to the borrower's phone, and so on.

Incorporation of reading lists as in the MOSIAC demonstrator (see section 3) is highly desirable – CERLIM's user survey and MOSIAC demonstrator usability study found users used and rated reading lists highly.

---

<sup>14</sup> The alternative of scaling up – buying a single more powerful computer – rapidly becomes economically infeasible with increasing scale and/or performance demands.