

# Mosaic data collection – A guide

Mark van Harmelen for the Mosaic Team, 14 May 2009

[markvanharmelen@gmail.com](mailto:markvanharmelen@gmail.com)

1	Introduction.....	2
2	Terminology: Courses and modules .....	2
3	Desired data .....	3
4	Reading lists.....	4
5	Use records.....	5
5.1	Use records that represent library loans.....	5
5.2	Level 1 library loan use records.....	5
5.2.1	Basic data: Institution, year and dates.....	5
5.2.2	Resource data .....	6
5.2.3	User context data.....	6
5.3	Level 2 use data .....	7
5.4	DPA requirements.....	7
5.5	Examples of use records .....	8
6	Use record format .....	9
6.1	Use record name value format .....	9
6.1.1	Basic data: Institution, year and dates.....	9
6.1.2	Resource data .....	9
6.1.3	User context data.....	10
6.2	Use record XML format.....	12
6.2.1	Start .....	12
6.2.2	Basic data: Institution, year and dates.....	12
6.2.3	Resource data .....	12
6.2.4	User context data.....	14
6.2.5	End record, more records .....	15
6.2.6	End all records .....	16
7	Reading List formats.....	17
7.1	Reading list name=value format .....	18
7.1.1	Institution and year .....	18
7.1.2	Module .....	18
7.1.3	Resources .....	18
7.2	Reading list XML format.....	19

# 1 Introduction

*Please read this document in conjunction with the letter inviting participation in the MOSAIC project experiment with user activity data generated by library and learning systems in UK HE.*

We are seeking as rich a data set as possible for use in the Mosaic project. We expect that different institutions will be able to supply different amounts of this data, and during your data generation activities we would very much like to engage in a dialogue with you as you go through the process of finding out just what data you can supply.

The dataset has various kinds of components:

- Reading lists
- Use records derived from library circulation/loan systems (LMSs)
- Use records derived from virtual learning environments (VLEs)
- OpenURL resolver / ERM derived information

In this document we concentrate on the first two of these components.

We will provide a follow up document with specific data formats for exported data.

We only seek data that is sufficiently anonymised so as not to require registration or other actions under the Data Protection Act (DPA) - more about this below.

## 2 Terminology: Courses and modules

In what follows we use **course** to refer to a course as an entity for which an institution grants a degrees or certificates. Two examples might be *BSc Computer Science at the University of Manchester*, or a *PhD in Geology at the University of Southampton*.

A **module** is an entity contributing to a course, so we might have *CS2034 Database Technology*, a second year Computer Science Module at the University of Manchester.

### 3 Desired data

We use symbols to indicate the importance of information to us in our current work.

- ▶ Fields that are absolutely essential
- ▷ Fields that we would really like to have
- Fields that are desirable if easily available

*If you can only supply some data please discuss it with us, there is considerable flexibility, for example, we will be very happy with use records where the resource is only identified by an ISBN, ISSN, DOI etc. and lacks other resource information.*

## 4 Reading lists

You may hold reading lists in Word or HTML formats, which are fine for us, but so are more structured forms - such as LMS, VLE or other structured records.

The data is

- ▶ The institution – e.g. *University of Manchester*
- ▶ The reading list's module ID – *CS3024*
- ▶ The reading list's module name – *Database Technology*
- ▶ The academic year the module is being offered in – *2008/2009*
- ▷ The progression level most students are at when they take the module – *UG2*
- ▷ The course(s) the module is offered for, as UCAS code <sup>1</sup> and course name pair(s) – *FIG1, BSc Chemistry with Mathematics*
- ▶ The reading list – a list of the recommended items

Some notes:

- If you are supplying a text format reading list that is published on the web we need the text as a file please, rather than the URL at which it is stored.
- If you are supplying catalogue entries, we'd prefer catalogue entries in MODS rather than MARC format.
- Exports from Talis List, Talis Aspire or Ex Libris Reading Room are very welcome.

---

<sup>1</sup> UCAS code background information  
[http://www.ucas.ac.uk/he\\_staff/datamanagement/jacs/coursecodingprinciples](http://www.ucas.ac.uk/he_staff/datamanagement/jacs/coursecodingprinciples)

## 5 Use records

An example use record might contain circulation information that *a first year Chemistry undergraduate at the University of Hull borrowed a book “Organic Chemistry: A Primer” in academic year 07/08*. This is just an example, and there could be other data in the record.

Use records could come from more sources than library circulation systems.

- Please talk to us about VLE sourced data that you may be able to supply.

### 5.1 Use records that represent library loans

We are expecting that library loans will be the most common kinds of data, so we illustrate their use records here. It is a small step to imagine use data for other sources, e.g. VLEs.

For library use records that represent loans, there are two kinds of data that we are interested in. We would very much like to obtain both kinds of data.

- Level 1 Use Data may be used to personalize search for resources like library items, promoting results for popular items according to various criteria.

This data is to be released under the Open Data Commons License.

- Level 2 Use Data may be used to provide more complex information about resource use, that is not derived from a single record; for example
  - *‘borrowers who borrowed this also borrowed ...’*
  - *‘users of this went on to ...’*.

We don’t intend to publish this more esoteric data.

We describe below the fields we are interested in for use records that represent library loans.

### 5.2 Level 1 library loan use records

#### 5.2.1 Basic data: Institution, year and dates

- ▶ Institution identifier – the full name of your institution - without any preceding ‘the’ – *Oxford Brookes University, University of Oxford*.
- ▶ Academic year – formatted as yyyy/yyyy - *2007/2008*
- ▶ Record date - the date the record was extracted – formatted as yyyy.mm.dd – *2009.05.10*
- ▶ Item source – *LMS*

Note on Item sources: These are identified by a string, e.g., LMS for library management system, OUR for OpenURL Resolver, VLE for virtual learning environment, etc. We only consider LMS here, so please contact us for information about use records from other sources. It is our intention that the same or very similar format should work for all item sources.

### 5.2.2 Resource data

- Media type – e.g. *Book, e-Book, Journal, Article, Presentation*
- ▶ Author or editor – *A.N. Author*
- ▶ Title – *Global Pandemics in the Twentieth Century*
- ▶ Global ID type and its value, if the item has an ISBN, ISSN, SICI, or DOI please supply it, it is more important than the author/editor and title – *ISBN 12345678*
- ▷ Reliable local ID – unique ID from your source system, such as internal catalogue reference which must be persistent for updating purposes - *0123456789*
- Persistent URL - for the resource or its catalogue entry – *http://www.lib.inst.ac.uk/cat/012345t789.html*
- Publisher – *John Wiley and Sons*
- Publication Year – *1978*
- Journal Issue - for a Journal related record volume, issue number and month – *volume 12, issue 3, July*
- acquisition date
- creator

### 5.2.3 User context data

For students:

- ▷ Course UCAS code, please supply where available – *FIG1*
- ▷ Course JACS codes in letter three digit form where available – *F101, G123*
- ▶ Course – the user's course name as a string – *BSc Chemistry with Mathematics*
- ▶ Progression level – *FD, UG1, UG2, UG3, UG4, MD, PhD1, PhD2, PhD3+*

For staff:

- ▶ Staff type, if this is available – *acStaff, resStaff, adminStaff* (or if not available, simply use the designator *staff*)

## 5.3 Level 2 use data

Level 2 use data contains a GUID (a large random number) to represent the user making the withdrawal. This should replace any traceable identifier such as the Institutional Student ID or Library User ID.

There will be many records with a single GUID to represent an anonymised user's use of many resources over time. For DPA purposes, the dates in all records for a particular GUID are replaced by a sequence number, and any 'singleton' records are removed.

If anyone simply wants to supply only level 2 use data, then we can create the level 1 data that we distribute from their level 2 data.

## 5.4 DPA requirements

We will supply code to help you to meet the anonymisation requirements of the DPA.

For level 1 data this will strip out 'singleton' records.

For level 2 data this will strip out 'singleton' records, replace names with GUIDs, and replace dates of loans with sequence numbers (sequenced 1, 2, 3, ... for each GUID).

Thus

- We are interested in the platforms you might have to run this code (Windows or Linux, with version information).
- Alternately, we could supply access to a web service to do the anonymisation. Please let us know if this is of interest.

## 5.5 Examples of use records

Three examples of the kinds of data that are interesting to us are tabulated below.

Institution Name	University of Huddersfield	University of Huddersfield	University of Huddersfield
Academic year	2008/2009	2008/2009	2008/2009
Record date	2009.04.06	2009.03.31	2009.02.01
Item Source	LMS	VLE	OUL
Title	The Waves	Woolf & the cult of modernism – lecture	Woolf & the cult of modernism
Media Type	Book	Presentation	e-Journal
Global ID	ISBN 98765		ISSN 98765
Reliable local ID	LMS123	VLE999	ERM987
Resource URL		<a href="http://www.hud.ac.uk/123.html">www.hud.ac.uk/123.html</a>	<a href="http://www.openurlhere">www.openurlhere</a>
Author	Virginia Woolf	Prof. A. Smith	A.B.Smith
Publisher	Bloomsbury		Wolf Society Journal
Publication Year	1927		2005
Journal Issue			15.04
Course	English	English	English
UCAS code	1234	1234	1234
Progression level	UG1	UG1	UG3
Modules list		Female Authors	

## 6 Use record format

The following shows the data we need for a use record, e.g. a loan of a book, the use of a VLE record, a reference to an item via an OpenURL resolver.

We can parse two formats, name value pairs, or XML.

### 6.1 Use record name value format

Basic format details:

- One name=value pair per line, starting column 1, no blanks on either side of the = sign, the value terminated at the end of the line (i.e. no multi-line values, spaces allowed within the value, spaces after the last non-space character on the line ignored)
- The name value pairs that constitute a use record appear on subsequent lines, and are terminated by an empty line, that has no content characters, merely a new line character

Illustrative examples of name value pairs appear below

#### 6.1.1 Basic data: Institution, year and dates

- ▶ Institution identifier – the full name of your institution - without any preceding ‘the’ – *Oxford Brookes University, University of Oxford.*  
`institution=Oxford Brookes University`
- ▶ Academic year – formatted as yyyy/yyyy  
`academicYear=2007/2008`
- ▶ Record date - the date the record was extracted – formatted as yyyy.mm.dd  
`extractedOn=2009.06.20`
- ▶ Item source  
`source=LMS`

Note on Item sources: These are identified by a string, e.g., LMS for library management system, OUR for OpenURL Resolver, VLE for virtual learning environment, etc.

#### 6.1.2 Resource data

- ▶ Resource global ID type and its value – if the item has an ISBN, ISSN, SICI, or DOI please supply it,; this is more important than the author/editor and title. Please use the capitalised global ID type ISBN, ISSN, SICI or DOI. Depending on how use records are constructed we can see one case where multiple ISBNs might be associated with a record, so if this is the case please supply them all; otherwise we only expect one global ID per use record.

Some examples are

ISBN=0123456789

ISSN=0378-5955

SICI=0095-4403(199502/03)21:3<12:WATIIB>2.0.TX;2-J

DOI=10.1000/j.1365-313X.2008.03660.x

- ▷ Media type – choose from book, ebook, journal, article, ppt, doc, jpg, wav, mpg etc.  
media=book
- ▶ Author or editor  
author=A.N. Author  
editor=A.N. Editor
- ▶ Title  
title=Global Pandemics in the Twentieth Century
- ▷ Reliable local ID – unique ID from your source system, must be persistent for the item being used  
localID=0123456789
- Catalogue URL  
catalogueURL=http://www.lib.inst.ac.uk/cat/012345.html
- Resource URL - for an electronic resource; this might be a non-redirected URL, or a redirectable URL such as an OpenURL or a DOI redirection  
resourceURL=http://www.lib.inst.ac.uk/res/something.html
- Publisher  
publisher=John Wiley and Sons
- Publication Year  
published=1978
- Journal Issue - for a Journal related record volume, issue number and month –  
*volume 12, issue 3, July*  
volume=12  
issue=3  
month=July

### 6.1.3 User context data

For students and staff, level 2 data only, this will be anonymised using software we supply to you

- ▶ User identity – we would prefer if the user name could be translated into a GUID by the time it reaches the software we supply, but we may, if requested, write the software to do the initial anonymisation of names with GUIDs  
user=38a52be4-9352-453e-af97-5c3b448652f0
- ▶ Date of use – this will be transformed into a per-user sequence number by our software  
useDate=2008.04.27

For students, in both level 1 and level 2 data:

- ▷ Course code – Where available please supply either or both of the UCAS code (followed by) comma-separated JACS codes in the ‘letter three digits’ format  
UCAS=F1G1  
JACS=F102,G144,Z123
- ▶ Course name – the user’s course name as a string  
courseName=BSc Chemistry with Mathematics
- ▶ Progression level – please choose from the following  
F (for foundation year), UG1, UG2, UG3, UG4, M, PhD1, PhD2, PhD3+  
progression=UG2

For staff, in both level 1 and level 2 data:

- ▶ Staff type – if this is available, please choose from  
acStaff, resStaff, adminStaff (or if not available, simply use staff)  
progression=acStaff

## 6.2 Use record XML format

We use a simple XML format: Names used in the name-value format are used as element tags in the XML, and values are the same as defined in the name-value format section above.

### 6.2.1 Start

```
<useRecordCollection>
  <useRecord>
```

### 6.2.2 Basic data: Institution, year and dates

```
<from>
```

- ▶ Institution identifier – the full name of your institution - without any preceding ‘the’ – *Oxford Brookes University, University of Oxford.*

```
<institution>
  Oxford Brookes University
</institution>
```

- ▶ Academic year – formatted as yyyy/yyyy

```
<academicYear>
  2007/2008
</academicYear>
```

- ▶ Record date - the date the record was extracted

```
<extractedOn>
  <year>2009</year>
  <month>6</month>
  <day>20</day>
</extractedOn>
```

- ▶ Item source – LMS, VLE, etc

```
<source>
  LMS
</source>
```

```
</from>
```

### 6.2.3 Resource data

```
<resource>
```

- ▷ Media type – choose from book, ebook, journal, article, ppt, doc, jpg, wav, mpg etc.

```
<media>
  book
</media>
```

- ▶ Global ID type and its value – if the item has an ISBN, ISSN, SICI, or DOI please supply it; this is more important than the author/editor and title. Please use the capitalised global ID type ISBN, ISSN, SICI or DOI as type attributes. Depending on how use records are constructed we can see one case where multiple ISBNs might be associated with a record, so if this is the case please supply them all; otherwise we only expect one global ID per use record.

```
<globalID type="ISBN">  
    0123456789  
</globalID>
```

or

```
<globalIdCollection>  
    <globalID type="ISBN">  
        0123456789  
    </globalID>  
</globalIdCollection>
```

Any angle brackets in a global ID, such as may be found in a SICI, must be translated into &lt; and &gt;

```
<globalID type="SICI">  
    0095-4403(199502/03)21:3&lt;12:WATIIB&gt;2.0.TX;2-J  
</globalID>
```

- ▶ **Author**  
<author>  
 A.N. Author  
</author>  
or editor  
<editor>  
 A.N. Editor  
</editor>

- ▶ **Title**  
<title>  
 Global Pandemics in the Twentieth Century  
</title>

- ▷ **Reliable local ID** – unique persistent ID from your source system, such as an internal catalogue number  
<localID>  
 0123456789  
</localID>

- **Catalogue URL** - for the resource's catalogue entry  
<catalogueURL>  
 http://www.lib.inst.ac.uk/cat/012345.html  
</catalogueURL>

- **Resource URL** - for an electronic resource; this might be a non-redirectioned URL, and OpenURL, or a DOI redirection, i.e. any URL that locates or

redirects to the resource

```
<resourceURL>  
    http://www.lib.inst.ac.uk/res/something.html  
</resourceURL>
```

– **Publisher**

```
<publisher>  
    John Wiley and Sons  
</publisher>
```

– **Publication Year**

```
<published>  
    <year>1978</year>  
</published>
```

– **Journal Issue - for a Journal related record volume, issue number and month –  
*volume 12, number 3, July***

```
<journalData>  
    <volume>12</volume>  
    <number>3</number>  
    <month>6</month>  
</journalData>
```

```
</resource>
```

## 6.2.4 User context data

For students and staff, level 2 data only, this will be anonymised using software we supply to you

```
<context>
```

- ▶ User identity, we would prefer if the user name could be translated into a GUID by the time it reaches the software we supply, but we may, if requested, write the software to do the initial anonymisation of names with GUIDs

```
<user>  
    38a52be4-9352-453e-af97-5c3b448652f0  
</user>
```

- ▶ Date of use e.g. that a library loan was made.  
This will be transformed into a per user sequence number by our software

```
<useDate>  
    <year>2009</year>  
    <month>6</month>  
    <day>20</day>  
</useDate>
```

The following for students, in both level 1 and level 2 data:

- ▷ Course codes – Most HE courses have UCAS codes, and they are very important to the work we propose. We would appreciate as many use records with UCAS codes as possible. If available, we are also interested in a breakdown into letter and three digit format JACS codes, supplied individually, not as a comma separated list (as was specified for the

```

name=value format).
<courseCodeCollection>
  <courseCode type="ucas">F1G1</courseCode>
  <courseCode type="jacs">F102</courseCode>
  <courseCode type="jacs">G144</courseCode>
  <courseCode type="jacs">Z123</courseCode>
</courseCodeCollection>

```

If there is only one courseCode then simply use

```

<courseCode type="ucas">
  F1G1
</courseCode>

```

- ▶ **Course name**– the user’s course name as a string

```

<courseName>
  BSc Chemistry with Mathematics
</courseName>

```
- ▶ **Progression level**, please choose from the following
 F, UG1, UG2, UG3, UG4, M, PhD1, PhD2, PhD3+ (F is for foundation year)

```

<progression>
  UG2
</progression>

```

For staff, in both level 1 and level 2 data:

- ▶ **Staff type**, if this is available, please choose from
 acStaff, resStaff, adminStaff (or if not available, simply use staff)

```

<progression>
  acStaff
</progression>

</context>

```

### 6.2.5 End record, more records

```

</useRecord>
<!-- more useRecords here if need be -->

```

## 6.2.6 End all records

```
</useRecordCollection>
```

## 7 Reading List formats

The data is divided in three parts

- The institution and year the reading list is from
- The module that the reading list is from, together with any course and progression information that can be supplied
- The items that compose the reading list, these are called resources below

We would like to discuss formats with you, please, if you can only provide reading lists in formats that are not described below.

## 7.1 Reading list name=value format

### 7.1.1 Institution and year

As 6.1.1 p9.

### 7.1.2 Module

- ▶ The reading list's module ID  
`moduleId=CS3024`
- ▶ The module name  
`moduleName=Database Technology`

We do not know the extent to which institutions will be able to supply the course(s) for which the module is offered. If this information is available it is valuable. In the name=value format, a reading list may only be associated with one module and course combination, and the reading list should be repeated for any other courses where the module is offered. If supplied this information is as in 6.1.3 p10, reproduced below.

- ▷ Course code – Where available please supply either or both of the UCAS code (followed by) comma-separated JACS codes in the 'letter three digits' format  
`UCAS=F1G1`  
`JACS=F102, G144, Z123`
- ▶ Course name – the user's course name as a string  
`courseName=BSc Chemistry with Mathematics`
- ▶ Progression level – please choose from the following  
F (for foundation year), UG1, UG2, UG3, UG4, M, PhD1, PhD2, PhD3+  
`progression=UG2`

### 7.1.3 Resources

- ▶ The reading list – a list of the recommended items

Ideally the format is

Starting with an empty line,  
A list of resources in the format of 6.1.2, p9, separated by empty lines

However, other formats are possible, such as an associated word file.

## 7.2 Reading list XML format

```
<readingListCollection>

  <readingList>

    <from>
      <!-- as 6.2.2, p12 -->
    </from>

    <module>
      <moduleId>
        CS3024
      </moduleId>
      <moduleName>
        Database Technology
      </moduleName>
      <courseAndProgressionCollection>
        <courseAndProgression>

          <!-- as courseCodes, courseName, and
          progression under 6.2.4, p14 -->

          <courseCodeCollection>
            <!-- ... -->
          </courseCodeCollection>
          <courseName>
            BSc Chemistry with Mathematics
          </courseName>
          <progression>
            UG3
          </progression>
        </courseAndProgression>
        <!-- ... repeat courseProgression etc -->
      </courseAndProgressionCollection>
    </module>

    <resourceCollection>
      <resource>
        <!-- as 6.2.3, p12 -->
      </resource>
      <!-- ... repeat resource as needed -->
    </resourceCollection>

  </readingList>

  <!-- ... repeat readingList as needed -->
</readingListCollection>
```

- If there is only one course and progression then substitute element `courseAndProgression` for `courseAndProgressionCollection` above
- If there is only one course code then substitute the element `courseCode` for `courseCodeCollection` above