Exploiting activity data in the academic environment

Content of the Activity Data website at

http://www.activitydata.org/

Compiled by:

Tom Franklin (Franklin Consulting)
Helen Harrop (Sero)
David Kay (Sero)
Mark van Harmlen (HedTek)

On behalf of the University of Manchester

Funded by JISC

Note

This report is a copy of the website. As such, it was written to be read as a web site, and there is some duplication in the text and the flow may not always be appropriate for a report.

Despite this we hope you will find it useful.

Contents

Home page

Successful organisations are increasingly collecting and making use of their activity data. Exploiting your institutions **activity data** allows you to understand and support your users more effectively and manage your resources more efficiently. Three examples illustrate how you could benefit from exploiting the data

- Identifying and supporting at risk students earlier to reduce the number of students leaving courses or failing can be achieved through understanding patterns of behaviour from the students (eg through the use of the VLE, accessing library resources, attendance data).
- Providing recommendations for resources that support learning and research by using the results of other people's searches that relate to the users' current search.
- Understanding how resources are actually being used so that it becomes possible to plan more effectively.

This web site synthesises the work of the JISC funded activity data projects in order to help you to understand how you might benefit from exploiting activity data. Beyond the benefits it discusses the legal considerations that you need to be aware of (primarily data protection and data licensing) before looking in some detail at how to actually exploit your activity data.

The site also contains a set of guides that provide an overview of particular topics and a set of detailed "recipes" that explain how to undertake some of the detailed technical tasks for particular systems that were used in the projects.

To get an overview you will probably find the following sections most useful.

- Introduction
- Benefits of using activity data
- Data protection
- Guides

For a more detailed understanding you will find it useful to additionally look at the following sections:

- Collecting, processing and presenting activity data
- Lessons learnt by the projects
- Licensing and sharing activity data
- Recipes

This work was undertaken by Tom Franklin, Helen Harrop, David Kay and Mark van Harmlen on behalf of the University of Manchester.

This online resource was produced by the JISC-funded Activity Data Synthesis Project at the School of Computer Science, University of Manchester. The project team consisted on Tom Franklin, Helen Harrop, David Kay and Mark van Harmelen.

Introduction

The programme reported here sought to develop robust evidence of practice and of the business case as well as technical foundations to establish positively that the HE-specific exploitation of activity data has merits for both the institution and the user. The role of the Synthesis Project has been to support, cohere and leverage this work in the form of documentation and dissemination activities to fulfil the objectives of the call, notably:

- Identifying approaches or technical **solutions** that can be rolled out more widely.
- Supporting the development of **knowledge and skills**.
- Consolidating evidence of valid **business cases** for institutions to engage.
- Taking account of other UK projects (notably in Business Intelligence) and elsewhere.

The following sections establish the definition and landscape for the use of activity data in higher education:

Background

The recommendations of the 2010 MOSAIC project report are strongly embedded in the Activity Data programme: “The project encountered genuine interest in the value of activity data. Students demonstrated intuitive recognition and expressed few reservations … University librarians and lecturers typically balanced recognition of value with the necessary challenges regarding data availability and service business case … 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 … In parallel work can be done nationally to demonstrate the benefits of the network effect (aggregation) and open data.”

These recommendations suggested that the development of user and management services based on activity data is highly desirable to the UK HE sector and to individual institutions in terms of service economy, effectiveness and user satisfaction. The July 2010 JISC event confirmed the view that this should not be left to ‘web scale’ operations, notably content services (such as Amazon), discovery services (such as Google) or global domain equivalents. The UK HE sector has the ability to gather ‘context’ data (e.g. ‘Other students on this course / similar courses also read), which could uniquely add value for the user, the course designer, the lecturer and the library, learning resource and repository manager. The event further emphasised the value in identifying what data exists or could exist in other systems (eg VLE, Repository, publisher services), thus establishing the institutional potential of this type of business intelligence.

What is Activity Data?

Put simply, activity data is the record of any user action (online or in the physical world) that can be logged on a computer. We can usefully think of it falling in to three categories:

- **Access** - logs of user access to systems indicating where users have travelled (eg log in / log out, passing through routers and other network devices, premises access turnstiles).
- **Attention** - navigation of applications indicating where users have been are paying attention (eg page impressions, menu choices, searches).
- **Activity** - ‘real activity’, records of transactions which indicate strong interest and intent (eg purchases, event bookings, lecture attendance, book loans, downloads, ratings).

Attribution (knowing who the user is) is invaluable in analysing and building services around activity data. This allows us to tie activity together (i.e. by the same person or, more uncertain but still of interest, from the same IP address). Whilst knowing who people are is a hazardous proposition in the online world, the veracity and utility of activity analysis is greatly enhanced by:

- **Scale** (network effect) - which highlights patterns of activity in spite of exceptions (such as a shared family login to Amazon).
- **Context** - which adds detail to identity, such as an area of interest (eg enrolled course, current module, occupation, family size)

Analytics might be best defined as 'what we do with activity data' - analysing patterns of known importance (eg people who do this do that) or more broadly looking for clues and exploring data to identify the stories it may have to tell. Analysis may involve combining multiple data sources, some of which may not be activity data (eg Exam results). On account of the scale of data involved (large numbers of records) such analysis and subsequent presentation can be assisted by a range of visualisation tools.

Why should you care?

The rationale is straightforward. Activity data is one of the tools that may assist in business improvement, both generally and specifically (being particularly useful in addressing key issues such as retention, progression and the quality of the student experience).

That business improvement may, for example, relate to:

- **Efficiency** - the data tells suggests that a process is broken
- **Effectiveness** - the data provides indications of customer behaviour / progression that might be used in near- real time or as Performance indicators.

See Benefits of using activity data for a more detailed discussion.

What are the particular benefits of Activity Data?

Activity data is special on account of the way it is collected and the nature of the evidence

- It is derived in real time, automatically and effortlessly (once the systems are set up - such as tills in a supermarket or analytics for a website)
- It represents a particular type of evidence - relatively undeniable (it really happened) and authentic (compare a user survey about activity)
- It can be collected at scale - unlike traditional mechanisms such as surveys and interviews
- It can be combined - potentially opening up complex narratives could not otherwise be readily captured or analysed

However we should be aware that activity data tells its own particular lies - we can never be certain who a person is or why they browsed a series of pages or whether they even intended to read the book they borrowed. Activity data is just part, but a special part, of the range of evidence available to businesses (and potentially to the users themselves) about the user experience, alongside such a surveys, interviews and counselling.

What is Activity Data useful for?

We have considered the nature of activity data and the types of analysis for which it may be particularly useful. These apply to a range of domains, processes, key performance indicators and problem spaces in the work of an educational institution, just as they do in a supermarket or an insurance company. Historically, it is interesting to note the range of business domains identified in a 2005 North American survey in which 213 users of 'educational analytics' reported their data sources (http://www.educause.edu/ers0508). There is strong emphasis on administrative systems rather than academic systems, with a VLE- proxy making only 6th place and no mention of Library systems.

System	% of respondents using the system
Student information system	93.0%
Financial system	84.5%
Admissions	77.5%
HR system	73.7%
Advancement	36.2%
Course management system	29.5%

The JISC Activity Data programme is based on the premise that new opportunities for capturing and exploiting activity data have opened up in the Web 2.0 world, not least involving student and researcher facing systems. Here are 10 example uses, many of which are exemplified in the projects and themes introduced in this synthesis website with relevant projects listed:

Student Facing

- Student recruitment
- Student retention
    - STAR-Trak: NG
- Student choice and progression

Academic Performance

- Teaching & Learning quality
- Research impact
    - AEIOU
- Research collaboration
- Resource recommendation
    - AEIOU
    - RISE
    - SALT
    - Using OpenURL Activity Data

Process Related

- Business Process improvement
    - AGtivity
- IT Systems optimisation
- Scholarly Resource management

What are the dangers?

The thematic explorations and project uses that are the focus of this synthesis website indicate both the significant opportunity and the range of issues to be addressed by institutions, domain practitioners and IT professionals dealing with activity data in any of its guises in the educational setting.

Firstly, this is 'big data'. It can involve very large numbers of records, sometimes diverse in format and potentially from many systems - even within a single domain such as a library. Use of activity data therefore raises challenges of:

- Storage
- Processing and Analysis
- Presentation

Secondly, this is potentially dangerous data. Implementers, analysts and users must remain vigilant regarding the following issues that are explored in detail in this synthesis:

- Quality of data - are there issues of veracity and completeness?
- Critical mass of data - is it at least statistically significant?
- Applications involved - do they act as a distraction therefore not telling the real story of service use?
- Aggregation - is the story best told locally or through aggregation (eg at faculty, institution, UK sector levels)
- Not least, legal compliance - are obligations regarding Data Protection and privacy being met?

Benefits of using activity data

The use of activity data has the potential to bring significant benefits to organisations that make effective use of it to enhance core business processes. In the commercial world businesses have been exploiting their activity data for over 20 years. For some of these businesses it is critical to their business model. Before looking at the benefits in higher education It may be worth listing a few of the benefits that have been achieved in the commercial world to indicate the range of mature uses of activity data that already exist.

One of the earliest to exploit activity data was Barclaycard who used it to identify unusual patterns of behaviour by card users which could indicate that the card had been stolen and was being used fraudulently. This lead to significant savings through the early identification and stopping of stolen cards.

Amazon use your searches and purchases to recommend other items that you might be interested in by showing people who bought what you are looking at also bought these items and through their people bought these items together. Clearly, they are interested in increasing sales and have been exploiting the data to give a more personalised service to their customers.

Google uses information from your searches to determine what adverts to display to you. for instance, if you search for a camera you are likely to see adverts for cameras for the next few days.

In higher education there are several ways in which activity data can be used to provide real benefits to institutions. These include:

- Supporting student success
- Improving services and systems

Supporting student success

Enhancing the student experience and supporting student success are increasingly important for all colleges and universities. There are many ways in which this can be done, and here we will look at some of the ways that activity data can be used to support students. There are two key ways in which activity data can be used to support students

- Identifying good and poor patterns of online behaviour and supporting students to improve their performance. This is often called academic analytics.
- Personalising the ways in which they access and use online resources based on their activity.

Supporting study skills

Students vary hugely in their study skills, with some arriving at university with excellent independent learning skills and others needing much support to achieve appropriate study skills. Whether in the virtual world or the real world there are patterns of behaviour which are more effective than others. One of the purposes of higher education is to support students to achieve effective study skills. If we can identify effective patterns and ineffective patterns by the traces that they leave in the log files and matching these to outcomes then we can help students to enhance their learning. There are a number of advantages to using activity data to support improving the experience. These include:

- The work can be undertaken fully or partly automatically. The application can look at all the information automatically and flag up students with patterns of behaviour that give cause for concern to either the tutor or the student, possibly with suggestions of things that they can do to help themselves.
- The system can look across all the courses or modules that the student is taking. Activity that might not cause concern if limited to one course might be of concern if it is occurring in several of them.
- Because the system is looking at very fine grained data from a number of sources it may be able to spot issues or problems earlier than a tutor would and so through earlier remediation prevent problems from becoming greater.

The STAR-Trak: NG project at Leeds Metropolitan University has developed a system that allows students and staff to view their attendance and their use of the VLE in comparison to measures of performance so that they can see if they are high / medium or low users.

On a slightly different note the LIDP has been demonstrating the relationship between the use of library resources and degree outcome.

For project details see:

- Benefits of LIDP
- Benefits of STAR-Trak: NG

Personalising interactions

The other key way in which activity data can be used to support student success and enhance the learning experience is to personalise the information that students are presented with. In the introduction to this section we saw that Amazon uses the results of searches to offer additional material that their customers may be interested in buying. In a similar fashion the Surfacing the Academic Long Tail and projects are using searches undertaken by other people to provide users with resources that may be of particular interest to them that they might not otherwise find. This can help to increase the number and range of resources that students. However, care is needed when undertaking personalisation as students wanted to understand how the recommendations are derived. For instance the students were interested in the grades of the students making the recommendations, as discussed by focus groups undertaken by Activity data to Enhance and Increase Open- access Usage.

For project details see:

- Benefits AEIOU
- Benefits of LIDP
- Benefits of RISE
- Benefits of SALT

Benefits of Activity data to Enhance and Increase Open-access Usage (AEIOU)

In order to identify the benefits of the work that AEIOU had undertaken some focus groups which established the value of recommendation systems to users. Though, it may be important to note that the undergraduate and postgraduate students had different priorities and different views about who they wanted recommendations from.

Undergraduates

The focus group comprised six undergraduate students, three studying level 1 courses, 3 at level 2. Two had previously studied several modules up to level 3. The students were studying a range of subjects. The group were asked if they would make use of recommendations.

There was a general consensus that ratings and reviews from other students would be beneficial (because 'other people's experiences are valuable') especially if it was known which module the student leaving the rating had done, and how high a mark they had got for their module.

Postgraduates

This focus group was made up of five postgraduate students (one of whom was also a member of staff) studying a range of different subjects through arts, science, social sciences and educational technology. The main feedback was that:

- Students use citation information as a form of recommendation
- Students are wary of recommendations when they do not know the recommender eg tutor recommendations are valued
- It was felt that recommendations specific to a module should be fed through to that module's website eg for good databases

- Students would appreciate recommendations of synonyms when searching our collections e.g. stress/anxiety
- Resources from the institutional repository are trusted as authors can be contacted (this comment from a student who is also a member of staff)

Reflections on the comments in the focus groups

Knowing the provenance of a recommendation is clearly important and that seems to be a clear difference between academic recommendations and an 'amazon-type' purchasing recommendation. There is a critical element of trust that is needed. You could characterise it as 'I don't know whether to trust this information until I know more about who or where it comes from' That implies a good level of academic caution about the quality of resource recommendations. So that is possibly a qualification to our hypothesis

> ***"That recommender systems can enhance the student experience in new generation e-resource discovery services"***

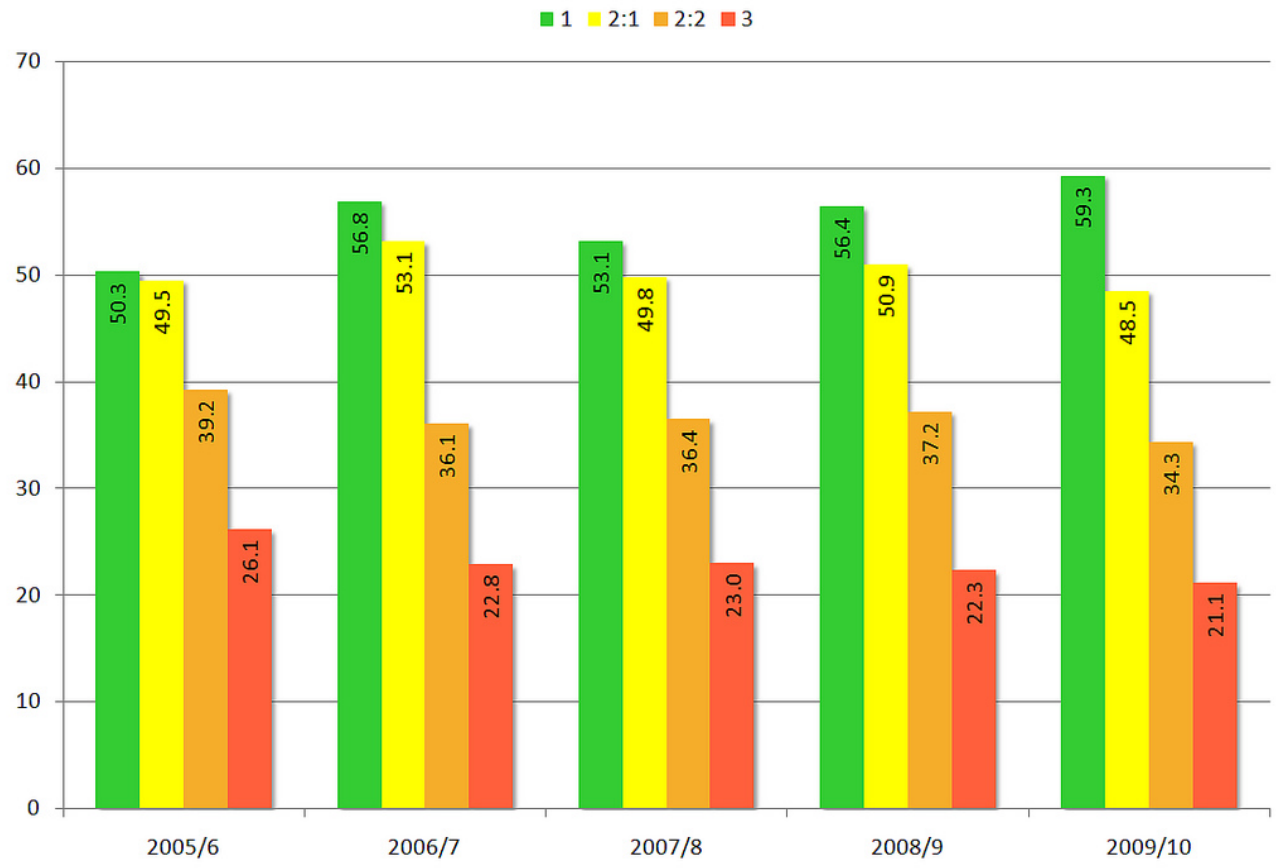
> ***'Qualification 1 … as long as it is clear where the recommendations come from and users have trust in their quality'***

While both undergraduates and postgraduates demonstrated the importance of trusting the source of the recommendation there were differences in how that trust might be established. For undergraduates trust came from the success of their peers making the recommendations, whilst postgraduates wanted recommendations from people that they trust. This may be to do with undergraduates having a proxy for trust that they are willing to use.

Benefits of Library Impact Data Project (LIDP)

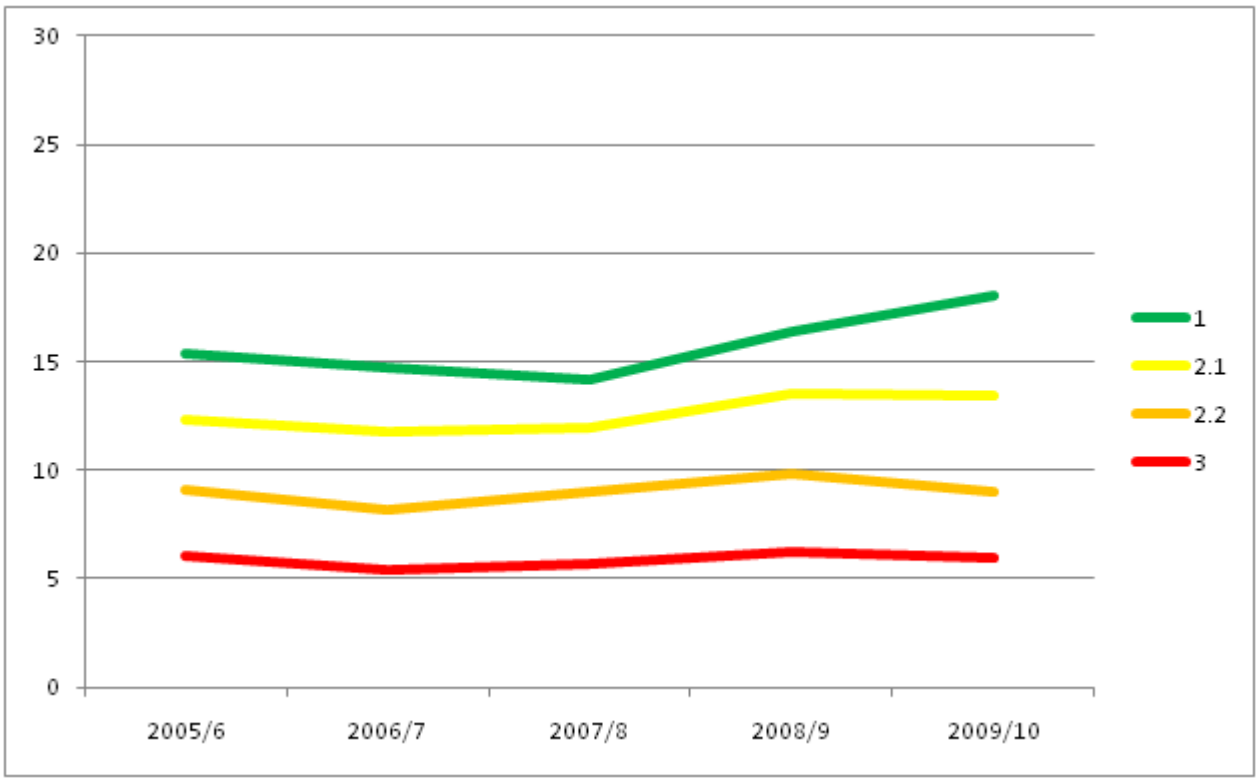
Among other things the LIDP project looked at student borrowing of books over the life of their degree. Because the use of e-resources has only been logged since 2005 the following discussion only looks at book loans.

The following graph shows the average number of books borrowed by undergraduate students who graduated with an specific honour (1, 2:1, 2:2 or 3) in that particular academic year. The data is based on approximately 3,000 students graduating each year.



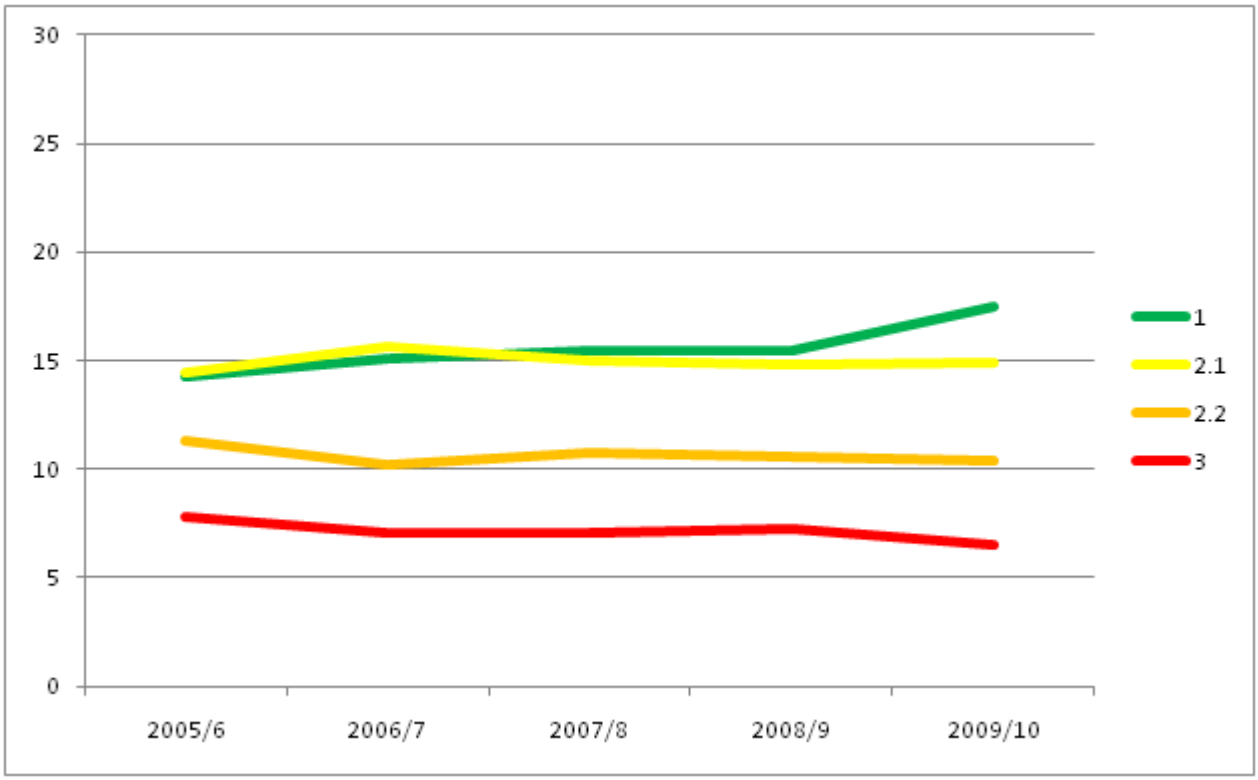
Number of books borrowed by students graduating in each year broken down by degree result

The correlation between eventual outcome and the number of books borrowed is established quite early, even in the first year there is already a threefold difference in book borrowing between students who will get a first and a third!

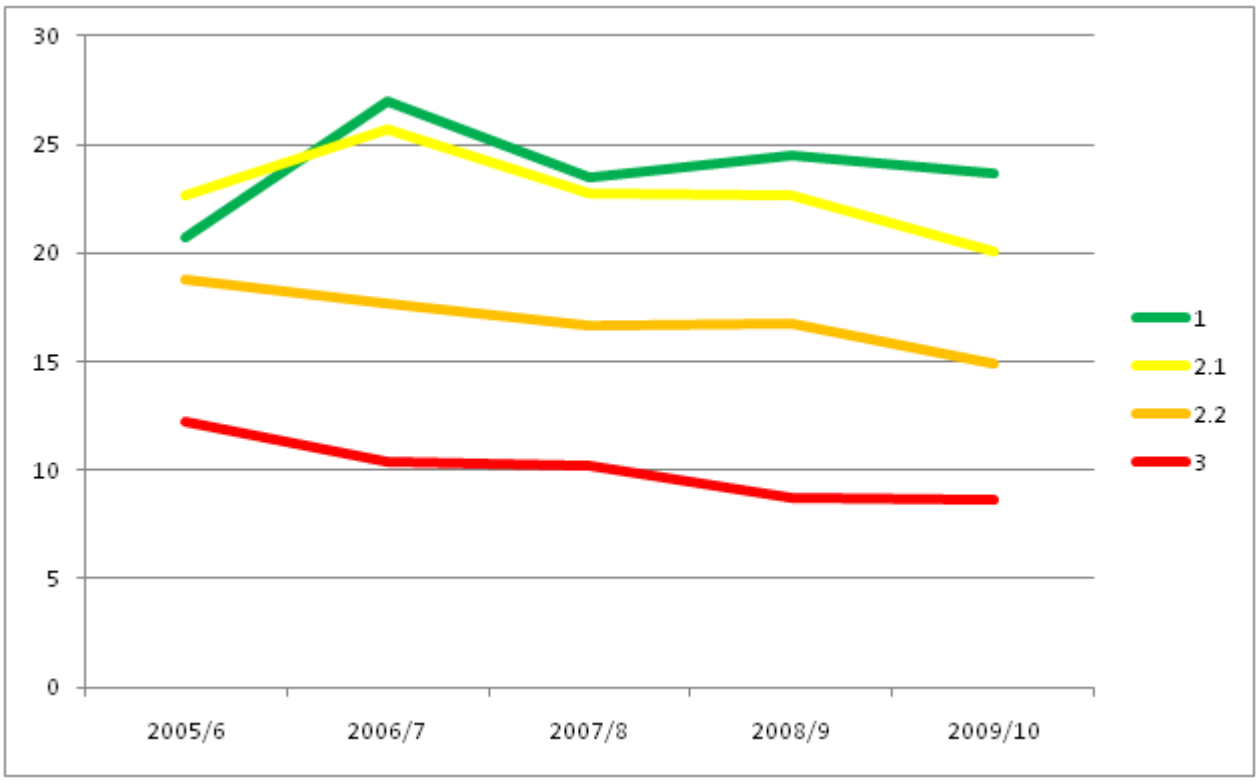


Number of books borrowed in their first year by students graduating in each year broken down by degree result

As can be seen this is repeated in each of the subsequent years though note that the borrowing levels of those students getting a first and two one are similar:



Number of books borrowed in their second year by students graduating in each year broken down by degree result



Number of books borrowed in their third year by students graduating in each year broken down by degree result

Benefits of Recommendations Improve the Search Experience (RISE)

The benefits from the RISE project are that it is easier to locate resources and that resources that users might not otherwise have found become visible. Evaluation showed that overall, user reaction was positive towards recommender systems. All of the students interviewed during the 1:1 sessions said that they liked the idea of getting recommendations on library e- resources. 100% said they preferred course related recommendations mainly because they would be seen to be the most relevant and lead to a quick find.

> *"I take a really operational view you know, I'm on here, I want to get the references for this particular piece of work, and those are the people that are most likely to be doing a similar thing that I can use.*" - postgraduate education student

There is also an appreciation that recommendations may give a more diverse picture of the literature on a topic:

> *"I think it is useful because you run out of things to actually search for….You don't always think to look in some of the other journals… there's so many that you can't know them all. So I think that is a good idea. You might just think "oh I'll try that". It might just bring something up that you 'd not thought of.*" - postgraduate psychology student

People using similar search terms often viewed was seen in a good light by some interviewees who lacked confidence in using library databases:

> *"Yes, I would definitely use that because my limited knowledge of the library might mean that other people were using slightly different ways of searching and getting different results." -* undergraduate English literature student

Their users also suggested the benefits could be enhanced through the following improvements:

- Make the recommendations more obvious, using either a right hand column or a link from the top of the search results.
- Indicate the popularity of the recommendations in some way, such as X percentage of people on your module recommended A. OR 10 People viewed this and 5 found it useful.
- Indicate the currency of the recommendations. If something was recommended a year ago it may be of less interest than something recommended this week.
- In order for students to trust the recommendations, it would be helpful to be able to select the module they are currently doing searching for. This would greatly reduce cross-recommendations which are generated from people studying more than one module.
- Integrate the recommendations into the module areas on the VLE.
- When asking students to rate a recommendation, make it explicit that this rating will affect the ranking of the recommendations, eg "help us keep the recommendations accurate. Was this useful to you?"

Benefits of STAR- Trak: NG

Full implementation of STAR-Trak will lead to the following benefits:

- Reduction in non-completion rates and increase in student learning performance
- Reduction in student administration time spent by teaching staff
- The ability to model and undertake scenario analysis using Business Intelligence (BI) applications and the data warehouse cubes (a type of database structure for BI) containing the activity data
- The creation of a longitudinal repository of student activity data that over time might otherwise be lost

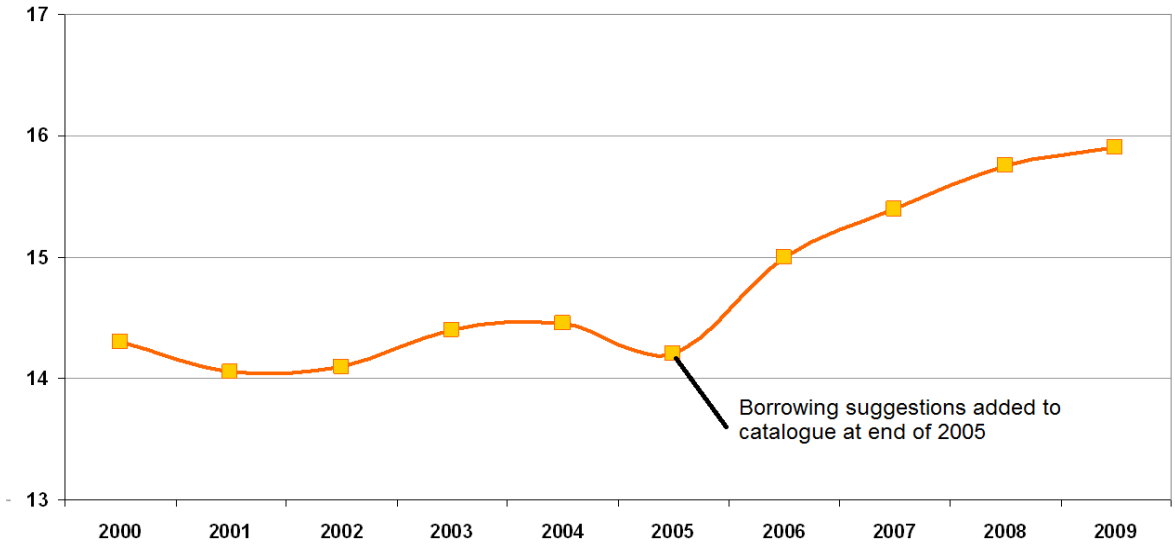
- A platform to support harvesting & analysis of data from resource discovery and management tools

Service and system improvement

Understanding how services are being used is one of the first steps in improving the service that can be offered to users. There are, of course, many ways which can be used to understand how services are used and how users would like to see them improved. The most obvious are to ask the users and service providers, and both of these are commonly done. However the use is then being remembered and interpreted, so this can be supplemented by using activity data to understand how users have actually behaved. The EVAD project was set up to increase understanding of how the VLE is being used at Cambridge University with the aims of providing senior stakeholders with better information on virtual learning environment usage and identifying signatures that relate to skilled usage (both by staff and students) in order to help to spread good practice both amongst staff in how to use it in their teaching and to students on how to use it in their learning. The approach here is very different to that discussed in supporting student success where the intention is to provide personal advice. Here the aim is to look at a higher level, looking at patterns rather than individuals. The results of this can lead to improved promotion and management of the VLE.

Similarly, the AGtivity project looked at using the activity data produced by the Access Grid Advanced Video Conferencing nodes to help improve the both the use of nodes in order to make more effective use of resources and to diagnose possible problems that may exist at any node. For instance, a pattern of very short conferences may indicate that there is a problem connecting to other centres. By improving the service as a result of using the activity data they also hope to reduce the amount of travel that participants undertake thereby saving both costs and the amount of CO2 produced.

Another way in which activity data can be used to directly support service improvement is through the use of recommendation services to help users to discover resources that will be of most value to them. This has already been discussed in terms of supporting student success. The SALT project set about offering users of the John Ryland's University Library at Manchester University recommendations based on books that other people had used; but rather than simply offer the most popular books that others had borrowed they set about offering relevant titles that had been more rarely borrowed. This has the dual benefit of offering users resources that they would be unlikely to find otherwise, and it ensures that the resources which might otherwise very rarely be used are used more often. Work at the University of Huddersfield (pre-dating the LIDP) has shown the impact the recommender services can have on the use of library resources.



average number of books borrowed per active borrower per calendar year (2009 predicted)

In Wales there was concern that materials in the Welsh Repository Network (WRN) were not being as widely used as they might be as the logs showed that most people were coming in from a Google search, locating a paper and moving on elsewhere. They wished to increase the time people spent in the repository once they had arrived and the number of papers that they looked at with the intention of

increasing the impact of research undertaken within the member universities. Comments from the evaluation report included:

> *"It could encourage wider use than materials in our own repository - it has real value."*
>
> *"It is good for a small University with only a small number of items in its repository as it can send users to other repositories to relevant items."*
>
> *"[It] helps to expose things in smaller less well known collections."*
>
> *"Allows you to discover items which may not be picked up at the top of the search results for the terms you have used."*

Effective use of activity data can encourage use of services to support the core functions of the institution by increasing the productivity of learners and researchers and empowering them to make more effective use of a wide variety of services.

Referring to a related JISC funded project Martyn Harrow, director of information services at Cardiff University, said: "The strength of the RAPTOR tool at a time when education budgets are being squeezed is in providing the evidence needed for academic schools to assess the e- resources subscriptions that are in place. Universities using this system will be able to prove the impact of the e- resources they provide, and ensure that they continue to deliver the best possible value for money for individual academic schools and entire institutions.

For further details see:

- Benefits of AEIOU
- Benefits of SALT
- Benefits of AGtivity

Benefits of Activity data to Enhance and Increase Open-access Usage (AEIOU)

The AEIOU project identified a number of benefits that could arise from the project including:

- **Increasing repository content usage** - The project anticipates that the addition of the recommendation service will lead to increased visits to the repositories and, hence, a greater awareness of ongoing Welsh research interests and improved collaboration between Welsh Institutions.
- **Raising the profile of Welsh research and enhancing the reputation of the individual institutions** - Through the increased use of the repository awareness of the research being undertaken by the institutions will increase resulting in a higher reputation for the research that is conducted in the member institutions. This may lead to increased collaboration opportunities.
- **Greater number of deposits** - Increasing the number of visits to a repository may bring with it the added advantage of encouraging academics to deposit more of their research outputs, in the knowledge that by so doing they are increasing the visibility of their work.
- **Compliance with funding body mandates** - If a researcher's work has been externally funded, depositing would be complying with the mandate of many sponsors, including the Research Councils, who, in return for providing research grants, require academics to deposit their research in a publicly accessible repository.

Benefits of Exploiting Access Grid Activity Data (AGtivity)

The AGtivity project identified the following benefits:

- **Planning utilisation for specific Physical Rooms.** Rooms are built according to space audits and proposed usage plans, and although planned diary schedules are used to give a good estimate of usage, recording the actual usage should be more useful. It is hypothesised that extra activities can easily out number sessions which were not cancelled.

- **Improving quality of service.** Indirect evidence is being looked for in the statistics before and after QA (quality assurance) tests thus providing potential evidence to the usefulness of this activity.

They have produced four case studies to illustrate the benefits:

- A tale of two rooms
- Testing, testing, testing
- To book or not to book?
- CO2 - loads of it

Case Study 1: A Tale of Two Rooms

Presented here is story of two different groups of researchers each providing postgraduate training courses; one for Computer Science and the other for Mathematics. The computer science graphics group AIG (Advanced Interface Group) in the University of Manchester ran a monthly local group under the banner of the ACM SIGGRAPH Manchester Professional Chapter http://manchester.siggraph.org/. Using the Access Grid Support Centre (http://www.agsc.ja.net) they provided a recording and annotation system that was sponsored by the JISC funded scientific visualization network vizNET (http://www.viznet.ac.uk). The Mathematicians run a consortium of 19 universities across the UK that share postgraduate lectures throughout the first two semesters of each year funded by the EPSRC project MAGIC (http://maths.dept.shef.ac.uk/magic/index.php).

Room 1: School of Computer Science: Room1.10, Kilburn Building, e-Science NorthWest which has 82 active users booking the site, and has seating for up to 30 (20 comfortably). It was QA tested last on 12/04/2011

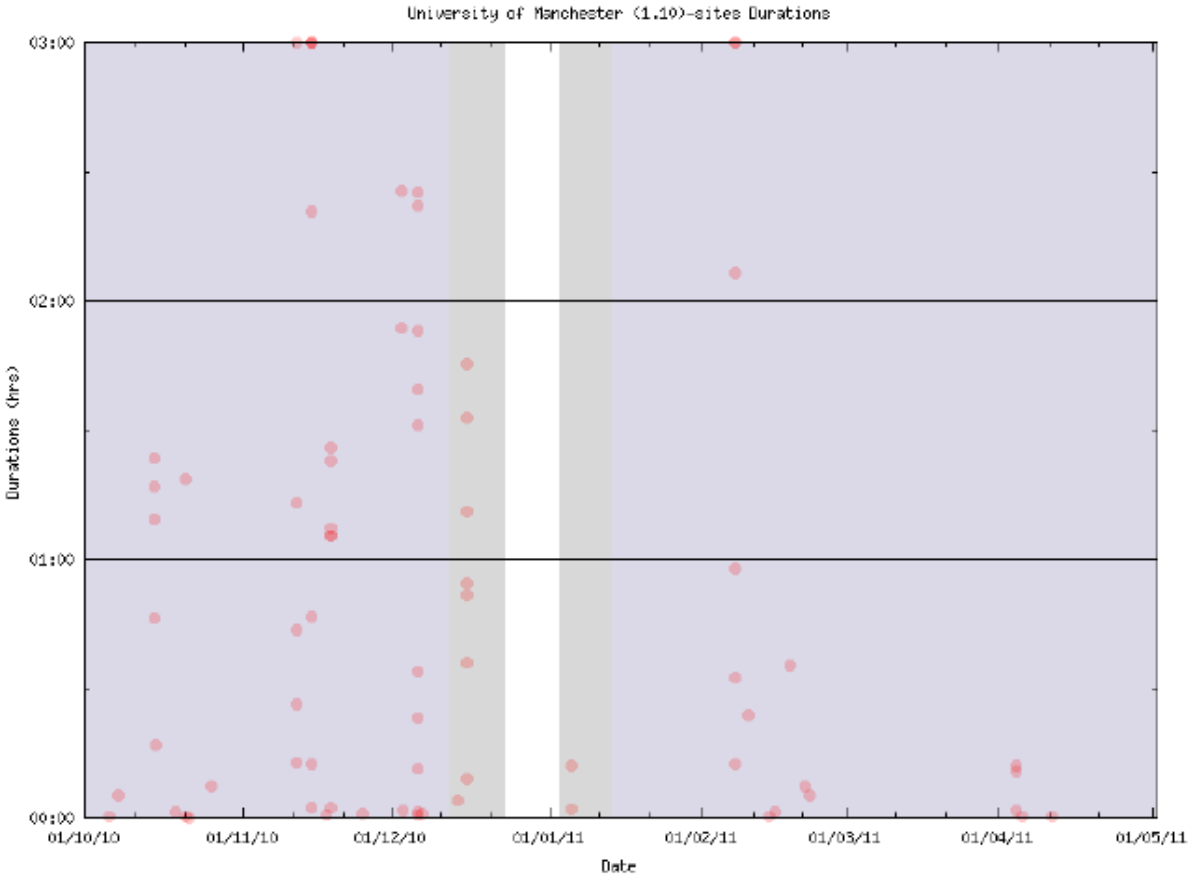
Room 2: School of Mathematics: Room 1.213, Alan Turing Building, Mathematics. It is only bookable by prior arrangement and has three active booking users for this reason. It can seat 20 comfortably, but is a part of 18 other similar rooms. It was QA tested last on 17/01/2011

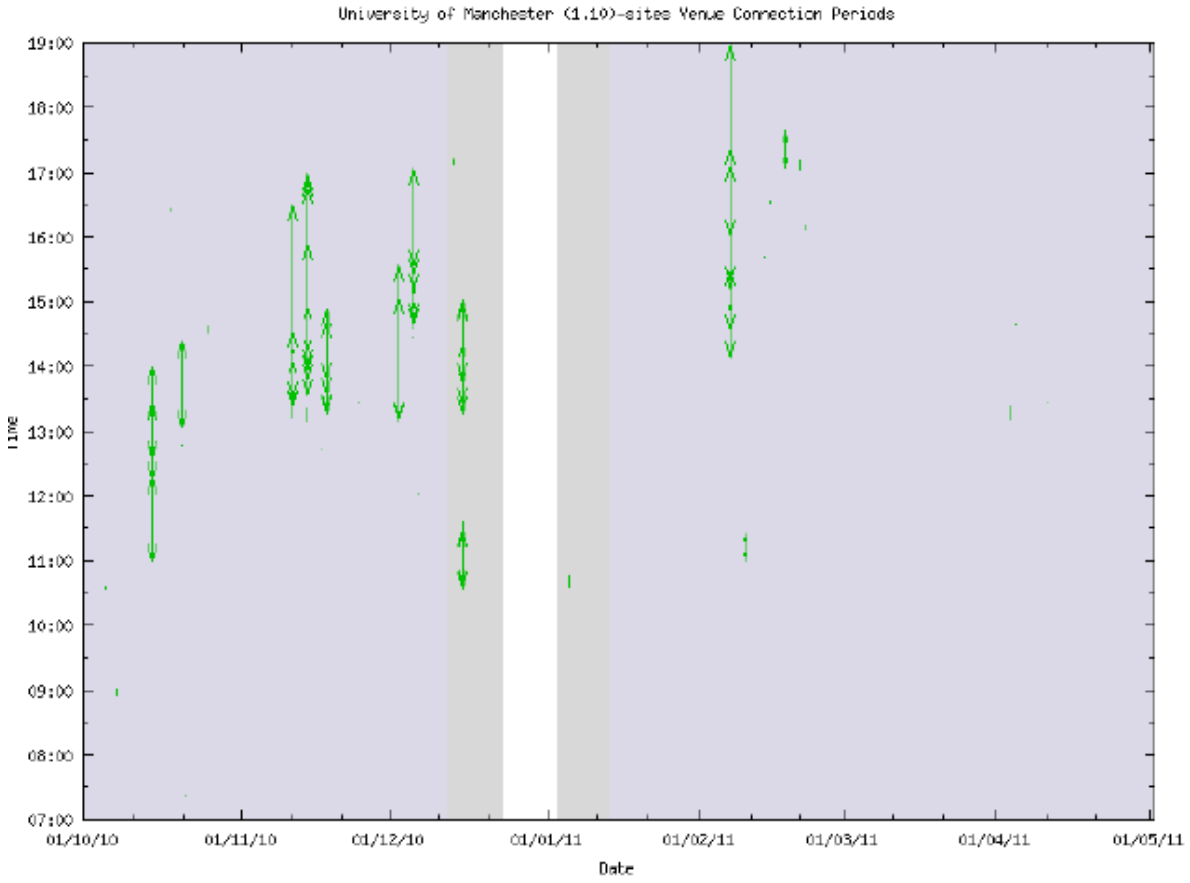
School of Computer Science AGtivity:

A summary list of events for the period of 2005-2011: Over the last six years they have hosted 72 different events with over 981 local attendees, and via the Access Grid video conferencing links there have been 61 unique external sites join. As a research experiment 15 of the lectures have been recorded and some annotated with slide changes and text comments.

Although this is a very strong group supporting numerous PhD students in terms of statistics this does not show up on the analysis for two reasons; one it is too fine grain being monthly activity and secondly occurred over seven years pre-dating the high resolution recording period when data could be captured.

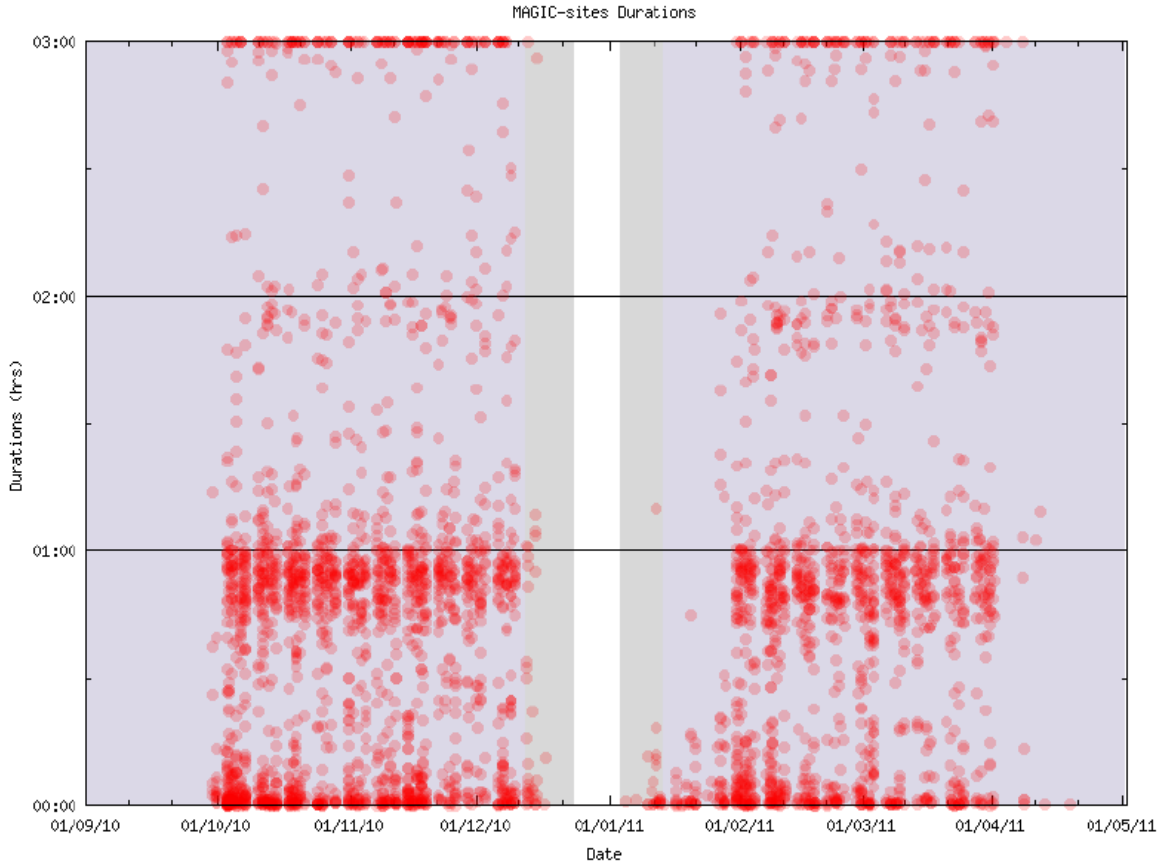
The following shows the statistics for Rm1:10 and indicates regular afternoon use.

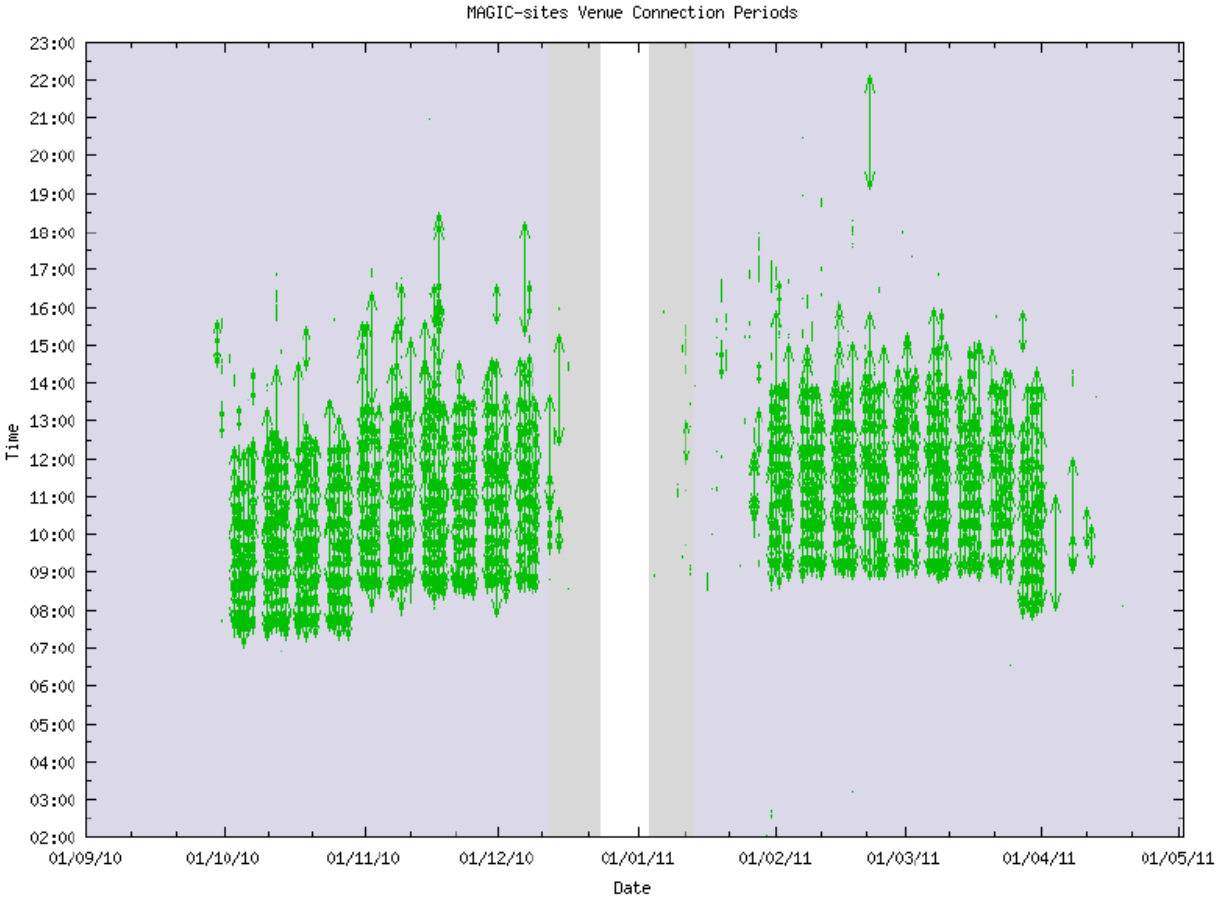

University of Manchester (1.10)-sites Durations
Durations (hrs)
03:00
02:00
01:00
00:00
01/10/10
01/11/10
01/12/10
01/01/11
01/02/11
01/03/11
01/04/11
01/05/11
Date


University of Manchester (1.10)-sites Venue Connection Periods
Time
19:00
18:00
17:00
16:00
15:00
14:00
13:00
12:00
11:00
10:00
09:00
08:00
07:00
01/10/10
01/11/10
01/12/10
01/01/11
01/02/11
01/03/11
01/04/11
01/05/11
Date

School of Mathematics AGtivity:

A summary list of events for the last 12 months: Over the last academic year MAGIC nodes were responsible for 4255 sessions, which in reality involved about 450 hour lectures. The room is occasionally outside of the lecture timetable which can be seen from the statistics. Also there is a high level of testing between the 19 universities, shown with lots of short events, as well as other sites which keep their video conferencing nodes on throughout the day.





Three observations that can be made to help room planning:

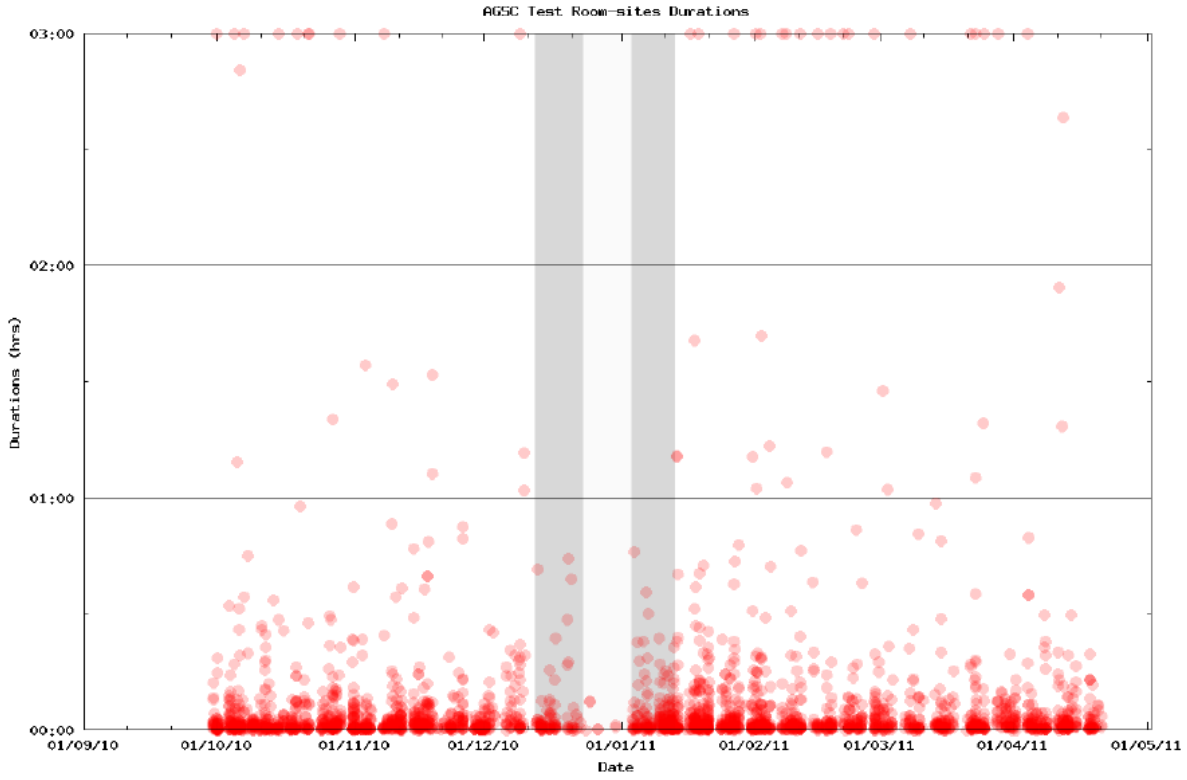
- It is clear to see how physical rooms and their virtual venues are used outside of known core time tabled events.
- It is also clear to see gaps when core timetable events do not occur.
- Use of this room is extremely efficient with consecutive back-to-back events.

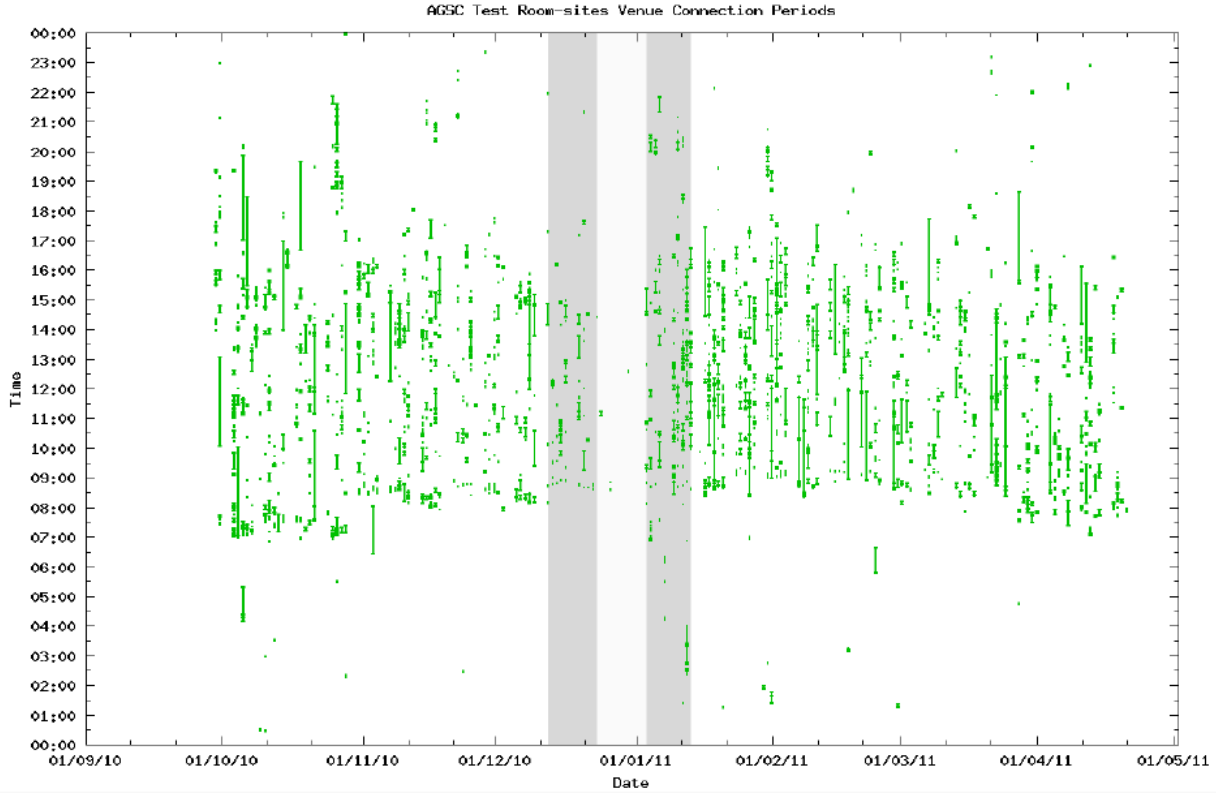
Both groups are active and approach postgraduate activities in different ways. One anecdotal statement to make is Computer Science mainly have meetings in the afternoon, whereas Mathematics lectures are in the morning - so we could merge the activities within one physical room. Both forms of use are important and could co-exist - which is not a usual planned idea in room allocation.

Case Study 2: Testing, testing, testing

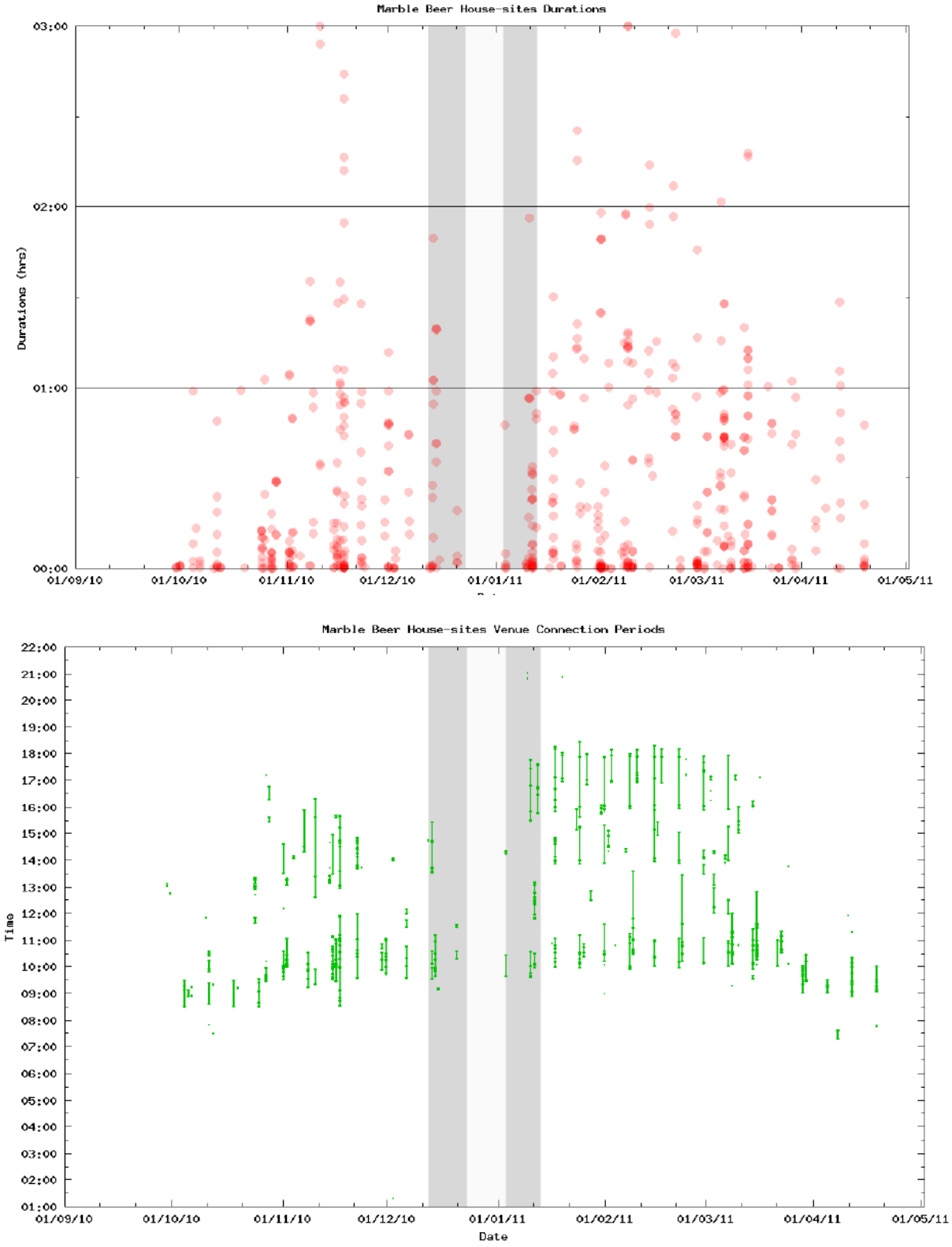
One of the most important roles of a support centre is testing.

The official AGSC Test Room is used almost solely for very short mostly self- test sessions, which is what it is designed for. The reason for this is that it broadcasts continual music and has a fixed delay causing any sound you send to be played back a few seconds later. This is a 'popular' service and as expected over a recent seven month period there have been 2,776 visits, averaging only 1 minute and 4 seconds and these meetings are scattered throughout the day fairly evenly.





There is also a weekly open session which anyone can attend that occurs at 10am-11am every Tuesday, in the Marble Beer House Virtual Venue. We can also explore this virtual venue visually and for the same seven month period there have been 597 visits, averaging 28 minute 55 seconds. The 10am slot is clearly visible in the graphs, but it is also used regularly throughout other parts of the day, mainly during the working hours of 9-5.



A key component was to analyse the data when cross correlation with QA data. **Quality Assurance** process is a step-by-step activity any video conferencing node can undertake to confirm all components of a room and the installed software/hardware are working well.

We have been looking in the activity data for evidence that the QA has a noticeable effect. Directly this has been difficult to establish, but there are stats for comments and complaints on one of the lecture series. Over the 2011/2011 lecture series there were 32 modules consisting of 433 lectures delivered. Half way through a full QA test was carried out on all the participating nodes. Prior to this there were 48 instances of poor audio, that were then resolved. In the future it could be possible to monitor audio quality levels within the statistics.

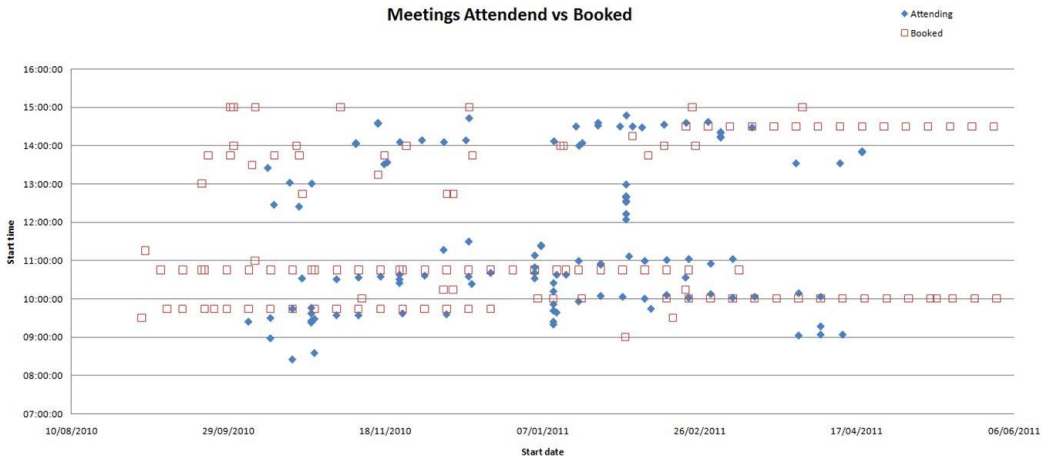
A longer term aim would be to automatically detect when QA tests should occur. This could be due to;

- Noticeable intermittent connections within the activity data - this has not been not seen in sample available, or
- Extended period of lack of use; as opposed to a regular check.

Case Study 3: To Book or Not To Book

For activity data mining to be useful it should involve cross-analysis with other databases. As well as the Access Grid activity data showing actual connections from video conferencing meetings, there is also an official booking system. Looking at the global statistics it can be see that only about 20% of meetings are booked through the official system when compared to the total number of actual meetings that have occurred. This value could be quite a lot less as it assumes all officially booked meeting actually have occurred.

The question is how many of these booked meetings are really used within groups. We chose a couple of research groups that have indicated to us that they actively use both the Access Grid sessions in conjunction with the official booking service. This means the statistics should correlate. The following plot shows the start times from the official booking times with the start times of real meetings that have occurred at a single specified physical room node. This illuminates a set of interesting observations and highlights the trust people put into booking systems equating to actuality.



This leads to some illuminating observations:

- In this case most booked meetings occurred as expected with a few not occurring. Often the real meetings started a few minutes before the booking ones.
- Extra meetings often occurred outside of the booking system as expected from the global statistics.
- Follow on meetings were often not officially booked and (informally) by interview were often initiated by voice during a previous meetings.
- Extensions of meetings were also common - ad hoc meeting sessions were often organised during the meetings.
- The graph shows a test session and workshop when there were a series of short meetings on 9th January 2011.

This was an initial test data mining of the statistics and further analysis could easily be carried out, for example analysing the change in the length of meetings as they progress across a series could be revealing.

Case Study 4: CO2 - loads of it

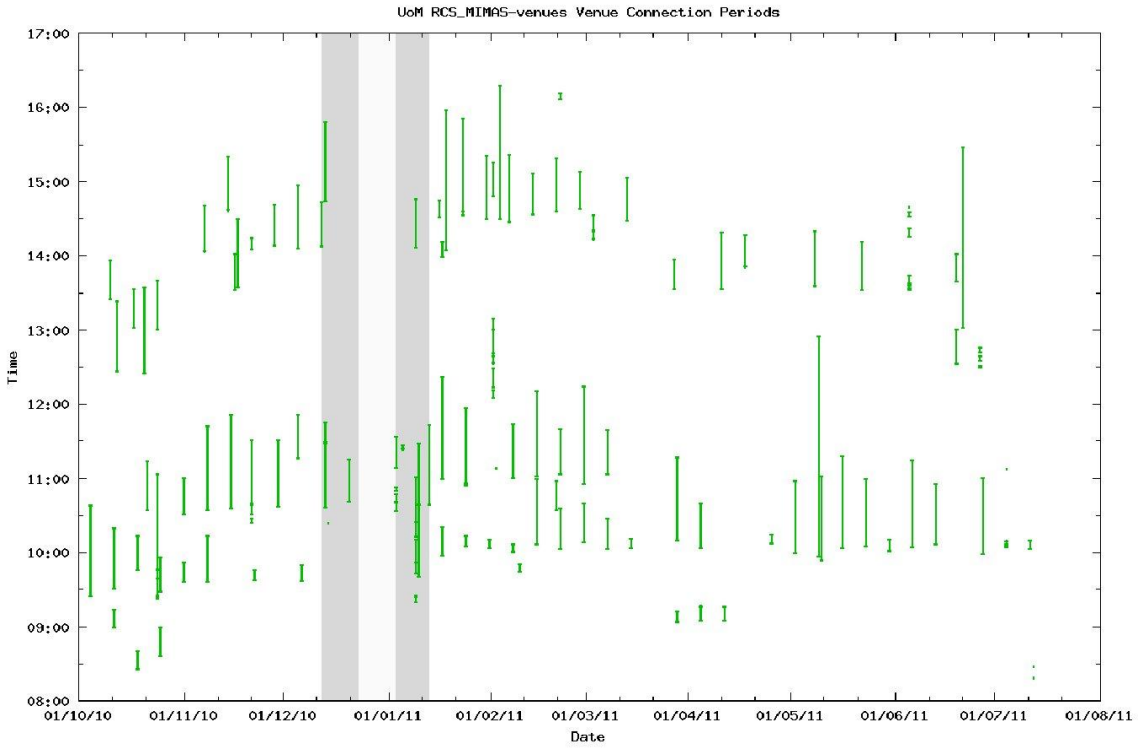
One of the continual requests we have when considering the use and analysis of the activity of video conferencing sessions is to clarify and quantify the CO2 savings. This is a lot harder than initially

considered as often meetings may be planned to be regular, but some may not actually exist; and some may not exist and actual attendance may vary from plan.

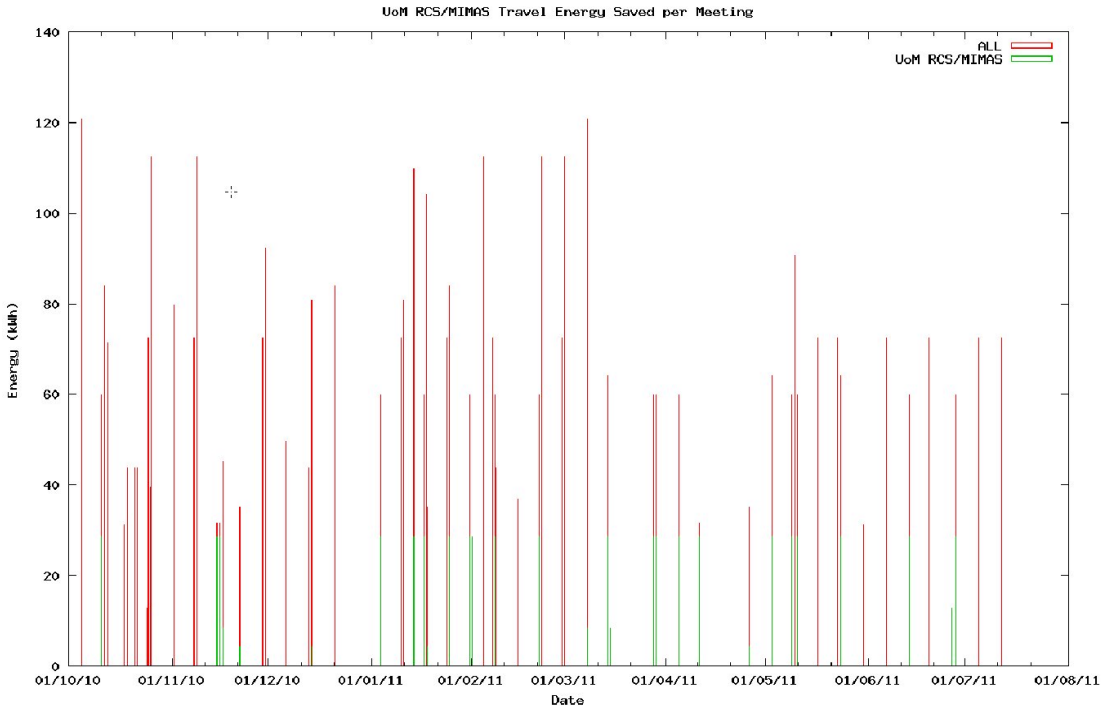
We are now able to mash-up the actual meetings occurred, cross correlated this with virtual venue meeting spaces to define the other participating physical nodes; and then use open access distance calculations to specify equivalent travel distances. All the major physical nodes in the UK needed to be geo-tagged, which was done for visualization purposes over the last couple of years and this data can be reused.

An Example Test-case: The CO2 savings reports have been calculated in KWh as this is now considered more universal. It is also simple on the web for a reader to find a calculator, depending on fuel capture type, which can then inversely calculate the CO2 values saved, if needed. A sample graph, for the RCS/MIMAS physical AG room node, is shown below that illustrates the process. The activity data for meetings within this room is about 65% work related to the NGS (National Grid Service) meetings. The list of all the meetings from the physical room node and then correlates these with the virtual venues can be given. An estimate of the number of people meeting in the same virtual venue, is used to assume the cost of a face- to- face meeting at the cheapest physical location, when travelling by car (1KWh is about 10 miles driving).

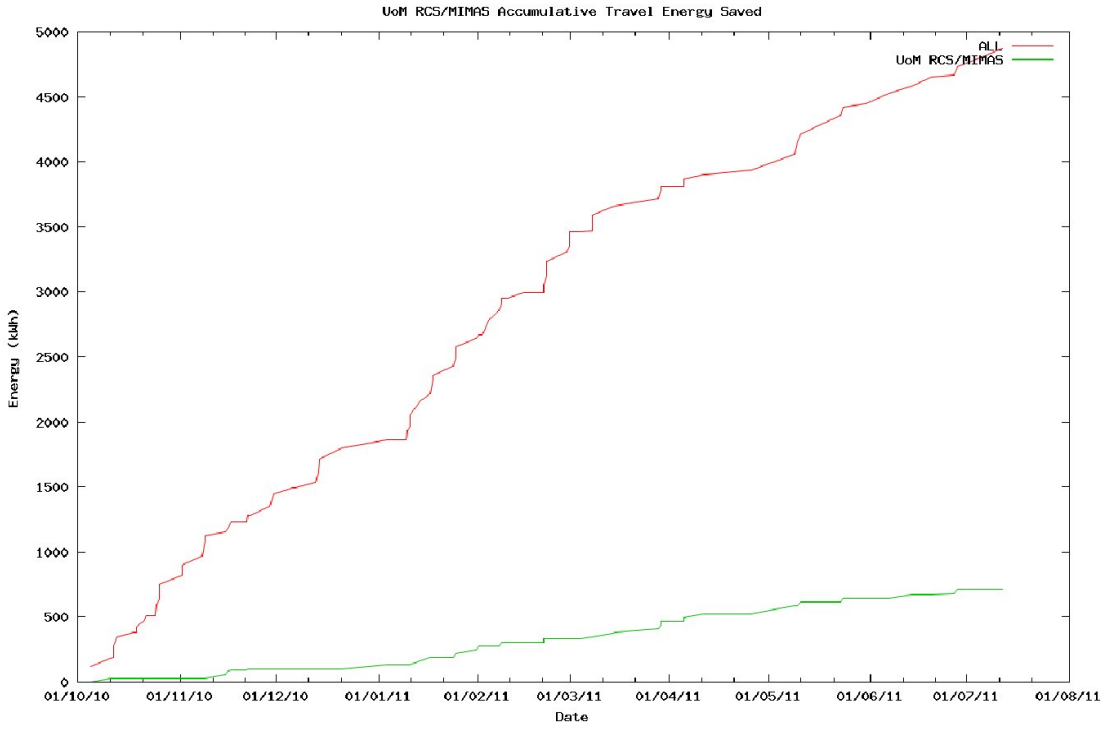
The first graph shows all the meetings in a period just under ten months.



Over this time there were 147 meetings within 18 virtual venues, with the average length of these meetings being 29 minutes 12 seconds. For each meeting the energy saving can be calculated:



and finally a cumulative energy saving can be displayed.



We could say the AG video conferencing sessions over the last ten months, within that room have saved approaching 50,000 miles of car travel! (Although this is not really an accurate statement).

There are many reasons why this is not a true saving.

- Meetings may not have occurred, if the equipment was not there. A budget is required to meet face- to-face
- There may have been fewer possibly longer face-to-face meetings reducing the KWh
- Costs of the physical room nodes and equipment have a kWh footprint
- Other even cheaper methods may have been used for example email or phone

Overall a multi-user and mixed purpose physical room node even with minimal Access Grid usage is likely to have an impact on the environment. Further work will need to be done in order to calculate if and how significant this is.

Benefits of Surfacing of the Long Academic Tale (SALT)

The SALT project offers users the potential to find interesting and useful sources that they would be unlikely to come across otherwise through the use of recommendations which focus on rarely borrowed items. This can help users to broaden the scope of the items that they discover and also means that resources may be better used than they otherwise would.

The SALT project also ran some focus groups that looked at the potential benefits that users thought that could accrue if data from additional libraries was also aggregated. These include:

- Recommendations could **surface, and hopefully increase the usage of, hidden collections.** Obviously circulation data is only going to offer a partial solution to this problem of discoverability (i.e. many 'hidden gems' are of course non-circulating) but nonetheless, we believe that the long tail argument borne out by Chris Anderson can also hold true for libraries - that the collective 'share' or recommendation of items can turn the Pareto Principle on its head. For libraries this means being able **to demonstrate the value of the collections** by pointing to increased usage. It might also give libraries a **better sense of what *is* of value to users, and what perhaps is not.**

- For users, particularly those in the humanities, a recommender function can help providing **new routes to discovery based on *use* and *disciplinary contexts*** (not traditional classification). In other words, what you are viewing through 'recommenders' are patterns of real usage, how other users with similar academic interests are aggregating texts. This is particularly useful for **finding conceptually related groupings of texts that cut across different disciplines**, and which will not ordinarily sit together in a standard results set.

- It also means we can **support humanities users in their preferred mode of discovery, powering 'centrifugal searching' and discovery through serendipity**. The downstream benefits of this concern the emergence of **new, original research**, new knowledge and ideas.

A further workshop with staff from the libraries at Manchester and Leeds Universities identified the following potential benefits of using activity data aggregated across a number of different libraries. These include:

- Aggregated activity data could support activities such as stock weeding by revealing collection strengths and allowing librarians to cross check against other collections.

- By combining aggregated collection data and aggregated activity data, librarians will see a fuller picture. This means they can identify collection strengths and recognise items that should be retained because of association with valued collections. We thought about this as a form of "stock management by association." Librarians might treat some long- tail items (eg items with limited borrowing) with caution if they were aware of links/associations to other collections (although there is also the caveat that this would not be possible with local activity data reports in isolation)

- Aggregated activity data could have benefits for collection development. Seeing the national picture might allow librarians to identify related items - "if your collection has this, it should also have…"

- This could also inform the number of copies a library should buy, and which books from reading lists are required in multiple copies.

- Thinking more outside the box, we thought it might also inform digitisation decision- making - i.e. if you digitised this, you might also consider digitising…

- Aggregated activity data could inform stock purchase - allow librarians to see what's new, what's being used elsewhere and therefore what's worth buying.

- This could also have benefits when discussing reading lists and stock purchases with academic staff, and thus enhance academic engagement.

Lessons learnt

One of the key purposes of the activity data programme was that other universities and colleges should be able to learn lessons from the work of the projects in order to make it easier for others to build on the work of the projects. Here we draw together some of the key lessons learnt by the projects and either reported by themselves or observed by us as part of our synthesis work.

The lessons learnt can be divided into the following areas:

Legal issues (data protection, privacy, anonymisation)

All the projects addressed issues around data protection and how they could ensure the privacy of the users of the systems. While there is no simple solution to this, one of the key points was would the data be traceable back to an individual? Clearly, some information would make this easy. For instance names and user IDs are easy to trace to an individual. There was however considerable discussion about whether or not IP addresses are personal information, and can be used to identify people. While in many cases it will not be possible (for instance because they are using a shared computer or the request is coming via a proxy and a single IP address may appear for a whole institution) in other cases it may be easy as users are using a fixed IP address.

- AEIOU wrote "In order to comply with recent changes to ICO code of practice we have been advised that as a minimum requirement we should include text in the header or footer of repository web pages and a link to a Data Privacy Policy that clearly informs users about how their data is being used and whether it is passed to third parties". For further details see here.
- LIDP suggest that "you should ensure you discuss privacy issues with your institute's legal advisor" For further details see here.
- RISE has developed a separate Privacy policy to cover use of activity data as it was felt that the standard OU Privacy policy was not sufficiently explicit regarding the use of data for this purpose. The newly developed privacy policy is available at http://library.open.ac.uk/rise/?page=privacy
- SALT, on the other hand, found that concerns over anonymisation and data privacy are not remotely shared by the users we spoke to. While we might question this response as potentially naive, this does indicate that users trust libraries to handle their data in a way that protects them and also benefits them.
- STAR-Trak's basic position is that they are simply consuming data that already exists in their IT ecosystem – nothing new is created and suggest that there is nothing innovative in this data that requires us to treat it differently from similar data elsewhere. For further details see here.

Recommended action

- Ensure that you have requested permission from users for whatever you wish to do with the data. See Informing users about data collection guide.
- Ensure that data is held securely.
- Ensure that any data that is publishes is suitably anonymised. See Anonymising data guide.

Data (scale, quality)

Several of the projects commented on the importance of data quality. There are several aspects to this including the completeness of the data, how up to date the data is, the signal noise ratio and the actual information itself.

- AEIOU commented that the data can contain a lot of noise which can impact on performance, and that data will grow rapidly, so you need to think big from the start. For further details see here.
- AGtivity suggest that you should log as much data and for as long as possible in order to capture long- term trends, and note that scripting and regular expressions are key to extracting useful

information from the large quantity of unrelated data. Future work could and should consider further recipes and advice.

- LIDP warn of the importance of retaining data for later analysis and that this needs to be thought about early. They also warn that you need to consider what the data means particularly around the use of e- resources. For further details see here.
- The OpenURL project also commented that release of data facilitates unanticipated use, which meant that the effort expended in sharing OpenURL knowledge was worthwhile as it expanded the data fields that have been made available and enabled an analysis of the quality of the data set by others that was not originally anticipated. They also found it necessary to have large volumes of data in order to identify sessions. For further details see here.
- RISE were also concerned with retention, but they also had to consider the licensing of some of the other (cataloguing data) that they needed to link to and expose in order to create the recommendations.
- Whilst SALT had 10 years of data available to them they found that a good recommendation service could be based on as little as five weeks worth of data, so it may not be necessary to have huge volumes of data. For further details see here.
- STAR-Trak which was rather different from other projects in that it brought together data from multiple sources and not just log files noted the importance of retaining data for use and of having a suitable data model. For further details see here.

Recommended action

- Ensure that the data you need is actually being collected. See Which events are significant guide
- Ensure that the data you need is retained for as long as you need it.
- Ensure that the data is of suitable quality and sufficiently complete.
- Consider the amount of data that will be generated and how you will process it.
- Consider filtering out data that you do not need in order to improve performance.

Understanding and presenting the data

Several of the projects discussed the importance of presenting the data in ways that are meaningful to users. AGtivity commented that one should consider visualisations and statistics results for users at an early stage and then be adaptable and agile to change. User requirements change and one should with them.

Recommended action

- Consider using visualisations to present the data. See Bringing activity data to life guide.

Recommender systems

Recommendations based on user activity can be very powerful, as demonstrated by Amazon amongst others. The number of recommendations offered can have a powerful bearing on the usefulness of the recommendations. Too few and nothing useful may turn up; too many and it can be overwhelming. AEIOU suggest that 5 or 6 is about right.

- AEIOU besides looking at the number of items to present also discuss what a session might be and where to but recommendations. For further details see here.
- RISE found that there are several types of recommendation that can be used based on data in proxy log files, but that you need access to bibliographic data as well. For further details see here.
- SALT noted that if you allow very rare events to provide recommendations then they may have little to do with the current item, though they might be of interest. For further details see here.

Recommended action

- Ensure that you can relate log file data to bibliographic data.
- Consider using personal data to determine what course users are on to focus recommendations.
- Consider the most appropriate number of items to recommend.
- See What's behind a recommendation? guide.

Legal: Activity data to Enhance and Increase Open-access Usage (AEIOU)

The IP Address identifies the computer from which the request originated and is used to provide the notion of a user session. Although this may not directly identify a user (e.g. the computer maybe shared publicly), in terms of Data Protection Act (DPA), IP addresses may constitute personal data if an individual user can be identified by using a combination of that IP address and other information. This applies even when personal data are anonymised after collection.

New European legislation came into force from May 26th 2011 and The Information Commissioner's Office (ICO) Code of Practice has been revised. The Code now clearly states that in many cases IP addresses will be personal data, and that the DPA will therefore apply. These changes also apply to the use of cookies and methods for collecting and processing information about how a user might access and use a website. An exception exists for the use of cookies that are deemed "strictly necessary" for a service "explicitly" requested by a user. In general, the regulations advise that an assessment should be made on impact to privacy, whether this is strictly necessary and that the need to obtain meaningful consent should reflect this.

We also need to consider that the AEIOU project is aggregating and processing data (that includes IP Addresses) originating from other institutional Repositories with no direct end-user relationship. The Using OpenURL Activity Data project has addressed this by notifying institutions that sign up for their OpenURL resolver service. We have no explicit agreement with the partners involved in the current project but aim to review their existing privacy policies should the service be continued. For example, do policies for storing and processing user data include repository reporting software and Google analytics and should users be made aware of this through the repository website?

The current cookie policy for Aberystwyth University can be found here

In order to comply with recent changes to ICO code of practice we have been advised that as a minimum requirement we should include text in the header or footer of repository web pages and a link to a Data Privacy Policy that clearly informs users about how their data is being used and whether it is passed to third parties (e.g. Google). Where possible, they should also be given the option to opt out of supplying personal information (IP address) to the Recommendation service. This would not affect them receiving recommendations but their information would not be stored or processed as part of the service.

Legal: Library Impact Data Project (LIDP)

One of the big issues for the project was to ensure we were abiding to legal regulations and restrictions, and continue to do so. You should ensure you discuss privacy issues with your institute's legal advisor, records manager and/or ethics committee. As detailed earlier we made efforts to ensure there is:

- Full anonymisation of both students and universities so that neither can be identified via the data. We contacted JISC Legal prior to data collection to confirm our procedures were appropriate, and additionally liaised with our Records Manager and the University's legal advisor.
- We have excluded any small courses in public reports or open access release to prevent identification of individuals i.e. where a course has less than 35 students and/or fewer than 5 of a specific degree level.

- To notify library and resource users of our data collection. We referred to another data project, EDINA, which provides the following statement for collaborators to use on their web pages:

Legal: Student Tracking And Retention (Next Generation): STAR-Trak: NG

Our basic position is that we are simply consuming data that already exists in our IT ecosystem – nothing new is created. That is not quite true as we have added functionality to capture interactions between student and tutor, extra- curricular activities and use of the system creates new data. However there is nothing innovative in this data that requires us to treat it differently from similar data elsewhere.

Of course what has changed is that data that was previously hidden or only available with difficulty and/or to certain people is now available to a wider set of people, and we have had to consider the privacy implications of that. To recap on the basic STAR-Trak functionality, it allows the student to see their own information from the student record system and activity data from the systems that they use to support their learning activities. It also allows tutors, module and course leaders to view this information on a course, module or student basis. We made the decision that trying to map relationships between staff and students as a basis for deciding who can see what, while great in theory, was not going to be sustainable in practice. These relationships may or may not be defined in external systems and they can be quite fluid. The combination of these two factors makes maintaining the relationships potentially a heavy administrative burden that would not be sustained, resulting in the system becoming unusable over time.

As an alternative we have implemented what we call “social controls”. In simple terms this has two elements to its operation:

- it allows a student to block all or any member of staff from seeing anything bar their very basic data (their name and what course they are on)
- any member of staff [authorised to use the system] can explicitly define in the application a relationship with a student (personal tutor, module tutor etc) and by doing so they can then view the student’s details (subject to 1 above). However a record is created in the application noting the relationship and this can be audited.

This is further strengthened by participation in STAR-Trak being fully voluntary on the students part.

We view this control system as an innovative experiment to be validated under strict trial conditions. We have already identified several enhancements that may improve the level of confidence in such a system. As per previous post we are still working hard to get formal commitment to running a year-long pilot of STAR-Trak and this continues to move in the right direction, albeit slowly (from the project’s perspective!). It is only in practice that we will see how successful we have been in developing a usable and useful application that meets data compliance requirements. As Chuck Reid famously said “In theory, there is no difference between theory and practice. In practice, there is”.

Data: Activity data to Enhance and Increase Open-access Usage (AEIOU)

Clean up!

Your data may contain a lot of noise which makes processing less efficient and results less relevant e.g. filter robots and web crawlers used by search engines for indexing web sites and exclude double clicks. Try using queries to identify unusually high frequencies of events generated by servers and flag these.

Think Big!

Your activity data will grow very quickly and things (e.g. SQL Queries) that took a few milliseconds will take tens of seconds. Use open source technologies that are tuned for Big Data (e.g. Apache Solr and Mahout) or process data offline and optimise your database and code - see Deploying a massively scalable recommender system with Apache Mahout.

Initially we used SQL queries to identify items that had been viewed or downloaded by users within specific 'windows' or session times (10, 15 and 30 minutes). Refinements were made to rank the list of recommended items by ascending time and number of views and downloads. We have a requirement that the service should respond within 750 milliseconds, if not the client connection (from the repository)

will timeout and no recommended items are displayed. The connection timeout is configured at the repository and is intended to avoid delays when viewing items.

Unsurprisingly, queries took longer to run as the data set grew (over 150,000 events) and query time was noticeably influenced by the number of events per user (IP address). Filtering out IP addresses from robots, optimising the database and increasing the timeout to 2 seconds temporarily overcame this problem.

However, it was clear that this would not be scalable and that other algorithms for generating recommended items maybe required. A little research suggested that Apache Mahout Recommender / Collaborative filtering techniques were worth exploring. We are currently testing Recommenders based on item preferences determined by whether or not an item has been viewed (boolean preference) or the total number of views per item. Item recommenders use similarities which require pre-processing using a variety of algorithms (including correlations). An API also exists for testing the relevance of the recommended items and we will be using this over the next few weeks to assess and refine the service.

Data: Library Impact Data Project (LIDP)

Forward planning for the retention of data. Make sure all your internal systems and people are communicating with each other. Do not delete data without first checking that other parts of the University require the data. Often this appears to be based on arbitrary decisions and not on institutional policy. You can only work with what you are able to get!

Beware e-resources data. We always made it clear that the data we were collecting for e-resource use was questionable, during the project we have found that much of this data is not collected in the same way across an institution, let alone 8! Athens, Shibboleth and EZProxy data may all be handled differently – some may not be collected at all. If others find that there is no significance between e- resources data and attainment, they should dig deeper into their data before accepting the outcome.

Using OpenURL Activity Data

analysis of a large body of data is essential in distinguishing what constitutes a real user session from noise.

it is possible to use the OpenURL Router activity data to identify sessions and to make recommendation links between articles. However, data for requests via proxies had to be excluded from the data set for recommendations, leaving out a large amount of potentially valuable data.

Data: Recommendations Improve the Search Experience (RISE)

If you want to make use of activity data then you need to make sure that you retain it for an appropriate period of time. Our EZProxy log files were routinely destroyed after a few months because they were not being used. Using the data to provide recommendations provides justification for keeping the data (but you still need to ensure you think about when you delete that data from the recommendations system)

You need some bibliographic data and it isn't always easy to get from the log files or from the systems you use. And when you get it you can't always store it locally due to licensing restrictions. But you need article titles, journal titles and dates for example so you can show users a sensible recommendation. Users need enough information to be able to judge the potential value of the recommendation.

Data: Surfacing the Academic Long Tail (SALT)

You don't necessarily need a significant backlog of data to make this work locally. Yes, we had ten years worth from JRUL, which turned out to be a vast amount of data to crunch. But interestingly in our testing phases when we worked with only 5 weeks of data, the recommendations were remarkably good. Of course, whether this is true elsewhere, depends on the nature and size of the institution. But it's certainly worth investigating.

If the API is to work on the shared service level, then we need more (but potentially not many more) representative libraries to aggregate data from in order to ensure that recommendations aren't skewed to represent one institution's holdings, course listings or niche research interests, and can support different use cases (i.e. learning and teaching).

Data: Student Tracking And Retention (Next Generation): STAR-Trak: NG

Review own retention strategy (whether explicit or implicit) and create a sense of urgency around the need to provide solutions that will assist in its success

Develop a canonical data model for the domains you are interested in. It surprised us that even within the same department colleagues had different interpretations of data ! We have not formally developed such a model, but have laid the foundations for its development (which is the subject of a different programme) through workshops.

Understand what data is critical to understanding retention. For us it is a subset of the student record data (such as demographics, entry qualifications, whether they came through clearing) and attendance data. We suspect that this data will give us around 90% of the information that we need. The other activity data is almost the icing on the cake – but clearly we need to evaluate this over time.

- our domain knowledge was insufficient,
- the metaphor we were using to identify students at risk was not best- suited to the task,
- the focus of the application was wrong
- we needed to put the student in control of the viewing of their data to maintain appropriate data privacy,
- we need to work on the ease of use and intuitiveness of the application

Internal domain knowledge: we assumed that the business (including IT) understood its data. However, while all parties did understand their data, they all had a more or less different understanding! As we are trying to develop an application that will be useful across the sector, we also had to make intelligent guesses about how other HEIs might in practice construct their ontologies. It was outside of the project scope to investigate this formally, however informal contacts were useful in this area.

Recommender systems: Activity data to Enhance and Increase Open-access Usage (AEIOU)

- When working with multiple repositories identify the repository each recommended item is from - NB this should be included in the results.
- Consider appropriate weightings for views and downloads. You may wish to consider downloads as having a higher weighting.
- Think about the number of recommendations. There are differences of opinion on this, but 5 or 6 seem about right.
- Think about the location of recommendations on item page - above or below? Maybe useful to have a link at the top of the record to take you to the recommendations at the bottom of the record.
- What makes a session where it is likely that the user is looking for related items? Consensus is that 30 minutes is about right.

Recommender systems: Recommendations Improve the Search Experience (RISE)

Here are a few things that we have learnt about EZProxy log files and how you can use them to make recommendations.

1. What is in the log file determines what you can do

You are restricted by what is stored within the log file as you need certain elements of data to be able to make recommendations or even to get other data from elsewhere to improve the data you collect. So you need user IDs to be able to get course information and you need bibliographic data of some form (or at

least a mechanism to get bibliographic data related to the data you have in the log file). Essentially you need hooks that you can use to get other data.

2. How you use EZProxy determines what you see in the log file

At the OU we link as many systems through EZProxy as we can. This includes our Discovery Service from Ebsco. The big implication is that Discovery Services aggregate content so the EZProxy log files show the Discovery Service as the provider. Our log files are full of Ebsco URLs and have far fewer resource URLs from other providers.

3. You can make basic recommendations from the proxy log files

You can make a very simple recommendation from a log file. There is a high chance that if a user looks at two resources, one after the other, then there is a relationship between the resources. If you store that connection as a relationship then that can form the basis of a recommendation 'These resources may be related to resources you have viewed'. The more people that look at those two resources one after another the more that reinforces that relationship and recommendation.

4. You need some bibliographic data for your recommendations

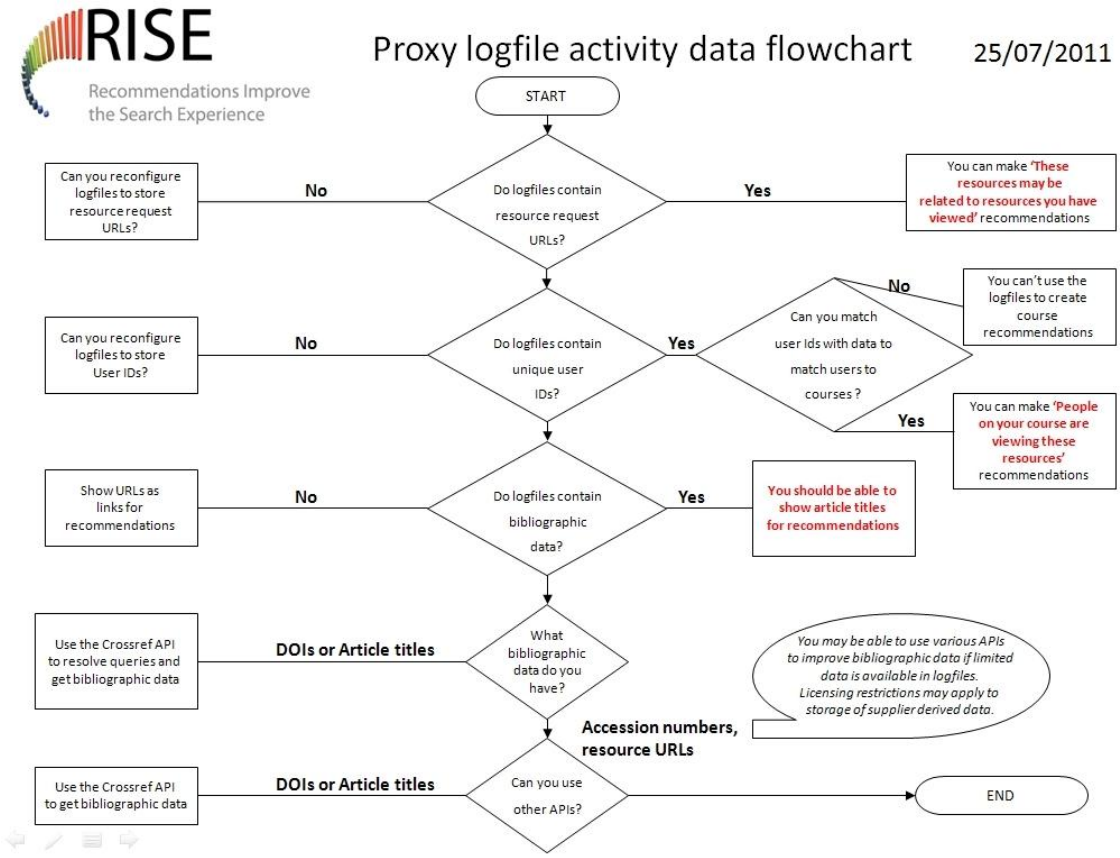
To show a recommendation to a user you really need to have something like an article title to display. Otherwise the user can't easily judge what the recommendation is about. For RISE we've used the Ebsco Discovery API to retrieve some suitable metadata and then passed that to the Crossref API to get bibliographic data that we can store in the system. The approach is to a great extent determined by what you have in your log files and what systems you can access.

5. You can get other data to make other types of recommendations

You can enhance your log file data as long as you have key bits of data you can use. So if you have a user ID or logon that matches up with your student information system then you can relate activity to other things such as the course being studied.

Proxy log file flowchart

To summarise the things that we have found with our EZProxy log files RISE has put together the following flowchart.



A lower threshold may throw up ‘long tail’ items, but they are likely to not be deemed relevant or useful by users (although they might be seen as ‘interesting’ and something they might look into further). Set a threshold of ten or so, as the University of Huddersfield has, and the quality of recommendations is relatively sound.

High level guidance

Guides provide short focused descriptions of topic areas that are essential to know about when thinking about or undertaking an activity data project.

While there is some variation in the structure of guides, they are uniformly prefaced by a short topic description, a description of the general problem area that the guide addresses, and options for solutions in the area.

The guides that appear here are ordered in the order that topics might have to be considered while formulating and starting up an activity data/recommender project.

The guides are:

Building a business case

The problem:

Getting senior management buy in for projects which make use of activity data to enhance the user experience or management of facilities is key if projects are to get the go ahead in the first place and become a sustainable service in the long term. There is a lack of persuasive business cases to refer to in the public realm. This guide gives some high level advice for the effective development of a solid business case.

In the current programme, activity data is being used to enhance the learner experience through recommending additional material, effectively manage resources and increase student success by helping them improve their online practices. Each of these is a powerful strategic benefit.

The solution:

The most important thing to remember when developing a business case is that its purpose is to persuade someone to release resources (primarily money or staff time) for the proposed activity. The person who will have to make the decision has a wide variety of competing requests and demands on the available resources, so that what they need to know is how the proposed project will benefit them.

The answer to this question should be that it helps them move towards their strategic goals. So the first thing that you need to find out is what their strategic goals are. Typically these are likely to include delivering cost savings, improving the student experience or making finite resources go further. You should then select one (or at most two) of these goals and explain how the project will help to meet this goal (or goals). Aligning the project to many goals has the danger of diluting each of them and having less impact than a strong case for a single goal.

Structure of a business case:

- Title

- Intended audience
- Brief description
- Alternative options
- Return on investment
- Costs
- Project plan
- Risks
- Recommendation

Do not 'over egg the pudding' in terms of understating the costs and risks or overstating the benefits. If the costs or benefits are not credible then the business case may be rejected as it appears to be not offering realistic alternatives.

The benefits should be realistic and quantifiable and, wherever possible, the benefits should be quantified in monetary terms. This allows the decision maker to compare the benefits and costs (which can usually be expressed in monetary terms), and so clearly see the return on investment, and compare this business case with other calls on their funding and staff.

Taking it further:

If the sector is to build a higher level picture of the business cases for exploiting activity data and also for pursuing the path towards open data then it is important to share knowledge of what works in terms of convincing key decision makers to give sustained support to using activity data.

The programme has produced some example business cases which can be used to understand the type of information that it is sensible to include, and which may form the basis for your business case. However, the business case must relate to the local circumstances in which you are writing it, and the audience for which you are writing it.

Additional resources:

Guidance and templates

- http://www.omafra.gov.on.ca/english/busdev/facts/02-023.htm
- http://www.impactonthenet.com/bizcase.html
- http://www.ogc.gov.uk/documentation_and_templates_business_case.asp
- http://www.klariti.com/Business-Case-Template/Sample-Business-Case- Template.shtml

Examples and further reading

- Full guide to writing a business case
- Open data in the library context: http://helibtech.com/Open+Data
- Exploring the business case for open metadata: http://discovery.ac.uk/businesscase/
- Business Case for New Office-Automation Equipment: http://www.impactonthenet.com/bc-oae.html
- Business case recommending no change: http://bit.ly/hMu95R [pdf]

Guidance on creating a business case

Purpose of Business case

The most important thing to remember when developing a business case is that its purpose is to persuade someone to release resources for the proposed activity. This will primarily be money or staff

time in most cases. It makes little difference whether this is re-allocating staff to a new project, employing new staff or contracting for the service.

The person who will have to make the decision has a wide variety of competing requests and demands on the available resources, so that what they need to know is how the proposed project will benefit **them**. Note that while this should be about benefits to them in their role they are people with particular interests and desires (including keeping their job and their next promotion). The question that they need the answer to is why should I use the resources on this project rather than some of the others?

The answer to this question should be that it helps them move towards their strategic goals. So the first thing that you need to find out is what their strategic goals are. These could be around cost savings (very likely in the current economic client), improving the student experience, increasing the use of resources. You should then select one (or at most two) of these goals and explain how the project will help to meet this goal (or goals).

You may feel that the project can address many goals, and it is possible that this is true, but you need to choose the ones that are central to the person you are writing the case for. Aligning the project to many goals has the danger of diluting each of them and having less impact than a strong case for a single goal.

The business case is intended to enable the decision maker to make an informed decision based on the (potential) benefits, the costs and the risks and how this particular project would further their strategic goals. It doesn't require academic rigour, but it does require evidence

Structure

A business case, like a bid for funding or a student essay, must answer the questions that the person reading it has. A typical business case of the type we need here will have the following structure:

- **Title**. Should be short and descriptive (eg Proposal to use data from the VLE and SRS to identify students at risk of dropping out).
- **Intended audience.** While this would not normally form part of a business case this would, it will be helpful for other people reading this business case and considering how they might use it for themselves. Essentially, the intended audience should be the budget holder who will benefit from the project. Why would the librarian want to pay for a project that only benefits student services when there are plenty of projects that they would like to fund of benefit to the library? Examples might include head of student service, librarian, PVC for teaching and learning.
- **Brief description** of what is being proposed. This should focus on the effect that will be achieved when the result is delivered. It is best to avoid technical descriptions, and to give the reader some concrete benefits from the project. (See panel for a brief example). The level of detail should be appropriate to the size of the project and the audience that it is intended for. The description will be fleshed out further in the main part of the business case, the idea is that the reader can quickly understand what the proposal covers and provides some context for all that follows which will include the details needed to support an informed decision.

  Brief description

  For a VLE activity data project

  The University of Wigan has a significant problem with non-completion by students. Currently 21% of all undergraduate students fail to complete, with half of those dropping out in the first year. Evidence demonstrates that early identification of students at risk of dropping out followed by active intervention could reduce this to 18% gaining the University £3million per year in additional HEFCE and student income. This project will use data that is collected by the attendance system and VLE to automatically identify students displaying at risk patterns of behaviour, thereby enabling student services and personal tutors to focus their efforts where they will have the greatest impact.

- **Alternative options**. There are always alternative options that could be implemented to achieve the same business goal, and it is important to show that you have considered them and explain

why the proposed option is superior. Note that these are not technical alternatives (or at least not only technical alternatives), but alternative approaches to achieving the same benefit that this project is seeking to achieve. For instance, an alternative to using activity data for the early identification of students at risk might be reports from tutors. The important thing in these business cases is to demonstrate that you have considered alternatives and have valid reasons for the choice that you made. This may be because the cost is lower, the benefits are higher or the risks are lower.

For each alternative you should

- Provide a brief description of the approach, highlighting the key differences from the approach proposed
- Describe the benefits of the alternative approach (lower cost, less staff development, fewer risks, no legal implications…..)
- Describe the costs and risks associated with the alternative
- Summarise the reasons for rejecting this approach in favour of the selected approach.

Do not "over egg the pudding" in terms of overstating the costs and risks or understating the benefits. If the alternatives are not credible then the business case may be rejected as it appears to be not offering realistic alternatives.

- **Benefits.** This is where you outline the benefits of the project. Remember that the business case is aimed at a particular person (role), who will be funding the project either with cash or by allocating staff time. They are primarily interested in benefits that address **their** strategic goals. Or, to put it another way, why would they pay for a project if the benefits fall elsewhere, they would expect them to pay for the project. It therefore may be

  The benefits should be realistic and quantifiable. If your project will, if successful, reduce student drop-out from 28% to 27% then that is what you should claim. If you over claim you might be asked to deliver that, and then when you deliver what is actually very worthwhile the project may be seen as failure as it did not deliver what was promised. It may even be worth under claiming the benefits (so long as they still exceed the costs), as delivering more than promised is usually seen as a good thing. Wherever possible the benefits should be quantified in monetary terms, this allows the decision maker to compare the benefits and costs (which usually can be expressed in monetary terms), and so see the return on investment.

Return on investment
Formally this can be calculated as
ROI = (benefit - cost) / cost
But more usefully it can be considered to be the amount of time taken for the investment (cost) to be covered by the benefit.

- **Costs.** This can be expresses in financial terms or in terms of staff effort (which can easily be turned into financial costs. A breakdown of the main headings is useful. If you want to relate the costs directly to the project plan then this may be more appropriately put after the project plan.
- **Project plan.** You have all produced these in the past, so I don't think there is any need to go into any great details. I would expect a fuller description of the project, the main tasks involved and how much effort or cost each will take.
- **Risks.** This is similar to the type of risk register that you might include in a JISC bid, though it is useful to give an indication of the cost that risk will incur if it occurs. This could be expressed financially (pounds) or as effort (person days).

Risk	Owner	Probability	Cost	Amelioration
Data formats incompatible	IT manager	Low	5 days	Map between formats
Sued for breach of privacy	Data protection officer	Very low	>£10,000	Ensure agreements are in place and signed by students
....	....	....	....	....

- **Recommendation.** This is likely either to be to take action (ROI is positive) or not (ROI is negative).

Additional information

Guidance and templates

- http://www.omafra.gov.on.ca/english/busdev/facts/02-023.htm
- http://www.impactonthenet.com/bizcase.html
- http://www.ogc.gov.uk/documentation_and_templates_business_case.asp
- http://www.klariti.com/Business-Case-Template/Sample-Business- Case- Template.shtml

Examples

- Business Case for New Office-Automation Equipment - http://www.impactonthenet.com/bc-oae.html
- Business case recommending no change - http://bit.ly/hMu95R

Sharing activity data: Legal issues

Topic

This guide discusses concerns which arise when activity data is shared outside the institution where the data was gathered.

The problem

If you want to share activity data with others then you have to make sure that two aspects are addressed, that you have the right to do so, and that you then distribute the data in an appropriate fashion.

In order to share data you need to have to have the right to do so, Typically this means that you own the intellectual property, and if so, in order for others to legally use the data they need to do so under an suitable license to use that intellectual property.

- If the data subjects might be able to be identified (i.e. you are realising full data rather than anonymised data or statistical summaries) then the data subjects need to have been informed that sharing can happen when they agreed to the data being collected (and they had a real ability to opt out of this).

Intellectual property rights (IPR)

It is likely that you will own the data produced by any systems that you are running, though it may be necessary to check the licence conditions in case the supplier of the systems is laying any claim to the data. However, if the system is externally hosted then it is also possible that the host may lay some claim to the log-file data, and again you may need to check with them.

- JISC Legal has a section addressing copyright and intellectual property right law

Licensing activity data

Any data automatically comes with copyright. Copyright is a form of IPR. You need to licence the data in order for other people to legitimately use your IPR, in this case, the data. There are a wide variety of

types of licence that you can use, though the most common is likely to be some form of creative commons licence.

Guidance is available from a wide variety of places including:

- JISC OSS Watch has a section on IPR and licensing.
- The Licensing Open Data: A Practical Guide by Naomi Korn and Charles Oppenheim
- The JISC sponsored IPR and licensing module that can be found at http://www.web2rights.com/SCAIPRModule/ . Within that you might be particularly interested in:
- Introduction to licensing and IPR http://xerte.plymouth.ac.uk/play.php?template_id=352
- Creative Commons license: http://xerte.plymouth.ac.uk/play.php?template_id=344

Data protection

Data protection, which addresses what one may do with personal data, is covered by the Data Protection Act (1998) . Some material which discusses data protection is:

- JISC Legal has a section on data protection
- Edinburgh University has produced a useful set of a definitions

An alternative approach to addressing the needs of data protection is to anonymise the data (see also the guide on anonymising activity data).

Informing users about data collection

The problem:

In planning the OpenURL Router activity data project EDINA became aware that by processing activity data generated by the Router service being, which is used by around 100 HE institutions, it effectively acts as a 'data processor'. Even the act of deletion of data constitutes processing so it is difficult to avoid the status of data processor if activity is logged. In the project, EDINA is collecting, anonymising and aggregating activity data from the Router service but has no direct contact with end users. Thus, it can discharge its data protection duties only through individual institutions that are registered with the Router.

The solution:

After taking legal advice, EDINA drafted a paragraph to supply to institutions that use the OpenURL

Router service for them to add into their institutional privacy policies.

"When you search for and/or access bibliographic resources such as journal articles, your request may be routed through the UK OpenURL Router Service (openurl.ac.uk), which is administered by EDINA at the University of Edinburgh. The Router service captures and anonymises activity data which are then included in an aggregation of data about use of bibliographic resources throughout UK Higher Education (UK HE). The aggregation is used as the basis of services for users in UK HE and is made available so that others may use it as the basis of services. The aggregation contains no information that could identify you as an individual."

EDINA wrote to the institutional contacts for the OpenURL Router service giving them the opportunity to 'opt out' of this initiative, i.e. to have data related to their institutional OpenURL resolver service excluded from the aggregation. Institutions opting out had no need to revise their privacy policies. Fewer than 10% of institutions that are registered with the OpenURL Router opted out and several of those did so temporarily, pending revision of their privacy policies.

Taking it further:

If you plan to process and release anonymised activity data, you may use the EDINA example as the basis of a paragraph in your own privacy policy - in consultation with your institution's legal team. If your institution has already incorporated the paragraph because you are registered with the OpenURL Router, you may simply amend it to reflect the further activities that you undertake.

Additional resources:

- The research undertaken by EDINA and the advice received prior to adopting this approach: http://edina.ac.uk/projects/Using_OpenURL_Activity_Data_Initial_Investigation_2011.pdf
- The University of Edinburgh's Data Protection policies and definitions: http://www.recordsmanagement.ed.ac.uk/InfoStaff/DPstaff/DataProtection.htm http://www.recordsmanagement.ed.ac.uk/InfoStaff/DPstaff/DPDefinitions.htm
- The University of Edinburgh's Website Privacy policy: http://www.ed.ac.uk/about/website/privacy-policy
- JISC Legal's 'Data Protection Code of Practice for FE & HE' [2008]: http://www.jisclegal.ac.uk/Portals/12/Documents/PDFs/DPACodeofpractice.pdf
- Information Commissioner's Office's 'Privacy by design' resources: http://www.ico.gov.uk/for_organisations/data_protection/topic_guides/privacy_by_design.aspx

Which events are significant?

Which events are significant?

Topic

This guide is about compiling (in tabular form) a compendium of activity data to assist in planning what recommenders may be possible in a given institution or group of institutions that share activity data.

The problem

E H Carr famously explored the difference between 'facts of the past' and significant 'historical facts' (What is History? 1961), comparing cake burnings in general with royalty-related cake burnings. We are faced with a comparable challenge of judgement in determining what to select from the mass of data arising from the use of computer systems. Given an area of interest (such as student success or utilisation of subscribed resources), the problem of selecting significant facts or 'events' can be split in to two along the lines suggested by Carr:

- **What exists?** What event data are we collecting or what could we collect?
- **What is significant?** What event data is worth preserving?

Before we can address those questions, it may help to have a sense of the possibilities. Put simply, event data records **any** user action (online or in the physical world) that can be logged on a computer, ideally containing a reliable form of user identification. We can usefully think of it in three categories:

- **Access** – logs of user access to systems indicating where users have travelled (e.g. log in / log out, passing through routers and other network devices, premises access turnstiles)
- **Attention** – navigation of applications indicating where users have been are paying attention (e.g. page impressions, menu choices, searches)
- **Activity** – 'real activity', records of transactions which indicate strong interest and intent (e.g. purchases, event bookings, lecture attendance, book loans, downloads, ratings)

The solution

Given this definition, you may already be concluding that you do in fact have a lot of data in your systems, not all of which is likely to tell useful stories. The first step is therefore to determine the types of event data that may be of relevance. Perhaps start with a checklist along these lines:

Problem space – Analysis of learning resource usage by undergraduate students

Event Category	Event	Logged by	Logged now?	User ID	Annual Volume	Value for my purposes
Access	Student logs on to VLE	VLE	Yes	Standard University ID	2.2m	Possibly just noise
Access	Student logs on to cash	Cash	Yes	Standard	410k	Out of scope

	card top up	Card		University ID		
Attention	Student loads module resource page	VLE	Yes	Standard University ID	1.4m	YES
Activity	Student downloads local learning resource	VLE	Not sure	Standard University ID	360k	YES
Activity	Someone downloads a JORUM resource	Jorum	Yes	University IP address	5k	YES
Activity	Student borrows a book	LMS	Yes	LMS Borrower ID	310k	YES but ID needs mapping

Taking it further

It may be useful to undertake a more general exercise with your 'team'. Use a similar table (e.g. the first six columns). Don't start with the problem to be addressed but rather work together to compile the list of every example of event data that exists, could or should exist. One you have the list, you can add extra 'value' columns for different analytical purposes – for example a VLE team might add columns for such as student success, local content utilisation, usage patterns by faculty, VLE performance optimisation. You can then do a second pass assessing the significance of each event for each purpose – perhaps initially in a scale of yes / possibly / no.

Additional resources

- The LIDP project considered a range of library 'events' - http://library.hud.ac.uk/blogs/projects/lidp/about/ and http://www.slideshare.net/gregynog/bw-dave- pattern-lidp
- The EVAD project investigated to range of event logging in the Sakai VLE - http://vledata.blogspot.com/
- The JISC Business Intelligence projects have taken a wider sweep of what events might be useful to derive business intelligence - http://www.jisc.ac.uk/whatwedo/programmes/businessintelligence/
- Educause reports a survey of the data being used for 'academic analytics' - http://www.educause.edu/ers0508

Identifying library data sources

Topic

This guide lists some of the sources of attention data that are available in libraries.

The problem

Libraries wishing to build a picture of user attention may face the challenge of identifying the appropriate data. This, in turn, depends on the purpose of use. The latter may range from collection management (clearing redundant material, building 'short loan' capacity), through developing recommender services (students who used this also used that, searched for this retrieved that, etc), to providing student success performance indicators.

Libraries use a range of software systems through which users interact with premises, services and resources. The library management system (LMS) system is far from the only source of data, and the OPAC and the LMS circulation module represent increasingly partial views of user attention, activity and usage.

Breaking the problem down. In this guide considers some of the variety of sources available within library services. You may have additional sources to consider.

Some typical data sources

Libraries already working with activity data have identified a range of sources for the purposes of Collection Management, Service Improvement, Recommender Services and Student Success. Potential uses of data will be limited where the user is not identified in the activity (flagged as 'no attribution' below).

Some key examples are:

Data Source	What can be counted	Value of the intelligence
Turnstile	Visits to library	Service improvement, Student success
Website	Virtual visits to library (no attribution)	Service improvement
OPAC	Searches made, search terms used, full records retrieved (no attribution)	Recommender system, Student success
Circulation	Books borrowed, renewed	Collection management, Recommender system, Student success
URL Resolver	Accesses to e-journal articles	Recommender system, Collection management
Counter Stats	Downloads of e-journal articles	Collection management
Reading Lists	Occurrence of books and articles - a proxy for recommendation	Recommender system
Help Desk	Queries received	Service improvement

Further consideration

Here are some important questions to ask before you start to work with user activity data:

- Can our systems generate that data?
- Are we collecting it? Sometimes these facilities exist but are switched off
- Is there enough of it to make any sense? How long have we been collecting data and how much data is collected per year?
- Will it serve the analytical purpose we have in mind? Or could it trigger new analyses?
- Should we combine a number of these sources to paint a fuller picture? If so, are there reliable codes held in common across the relevant systems – such as User ID?

Additional resources

Consider also the Guides on Student Success and Data Strategies

Library Impact Data Project (LIDP) led by the University of Huddersfield.

What's behind a recommendation?

Topic

This guide is about types of recommendations that might be desirable to use in different situations, together with notes about the kinds of activity data needed to make the recommendations

The problem

Recommendations, such as those embedded in services like Amazon and iTunes, are typically based on attention (navigation, search) and activity (transaction) data, though they may also take account of user generated ratings. Different types of recommendation need different levels of data.

For example,

'People who bought that also bought this'

requires less data to recommend than either

'People who bought that bought this next'

or

'People like you bought this'.

The initial problem is to identify, for a given service what kind of recommendations might be offered.

These recommendations will have particular activity data demands as to the activity data to be collected, preserved, collated and processed. You should check availability of the data sources, possibly using a pre-prepared list of activity data sources (as recommended elsewhere in these guides). You should also check if you can use the data within the bounds of your corporate Privacy Policy (again please see elsewhere in these guides).

The solution

Services need to make a list of the recommendations that they would like to offer, identifying the data required for each.

To save you doing that from scratch, here is a list of broad yet distinct recommendation types and the generic data from which they might be derived.

Generic recommendation types

[A] People who did that also did this – a list of all transactions (e.g. book circulation records) with anonymised user IDs and no further user details

[B] People who looked at that also looked at this – a variation on [A] based on 'attention', perhaps signified by retrievals of full details catalogue pages

[C] People who did that did this next / first – as per [A] but this will need a time stamp

[D] - People like you also did this - 'Like you' can be in a different league of difficulty, depending on how you define 'like you'; you can cheat and treat it as a variation on [A] where likeness is based on activity. However, in education we might expect 'like you' to be based on a scholarly context – such as a Course or even a Module or perhaps a discipline and a level (e.g. Undergraduate Physicists, First Year Historians) – all of which implies personal data, with each transaction linked to a user and their context.

[E] - People like you rated this – A variation on [D] that is possible if your system collects ratings such as stars or a 'liked' / 'useful' indicator.

[F] - New items that may interest you – This 'awareness' service requires no information about other users, simply using a 'precise-enough' catalogue code (e.g. Subject heading or Dewey class in libraries) to link the user's previous activity to new items.

The following are particularly interesting in libraries, though they could be supported elsewhere

[G] People who searched using this term ended up viewing / borrowing in these subject areas – building a mapping from search terms to the eventual classification of choice may help with accessing classifications that differ from the key words used in teaching; this does not need data about individual users

[H] The following alternative titles are available – at its simplest, this can use such as Dewey codes or reading lists to recommend alternatives; whilst this does not have to involve individual user data, the recommendation could be enhanced based on 'people like you'.

Taking it further

The first steps are to determine

- What sort of recommendations you could offer in order to assess the user experience
- Which events you would need to capture in order to generate those recommendations

- What the recommendation algorithm is, or how feasible it is to design an algorithm, possibly experimentally
- What supporting personal context data would be desirable
- Whether capturing additional data would enable to support a wider range of recommendation types at a later stage

Additional resources

As illustrated in Dave Pattern's presentation, the University of Huddersfield has identified a range of recommendation types that can be derived from library activity data – http://www.slideshare.net/gregynog/bw-dave-pattern-lidp

The RISE (http://www.open.ac.uk/blogs/RISE/) SALT (http://salt11.wordpress.com/) projects also focused on recommender services based for library resources. The OpenURL project developed a recommender as a by-product of other work (http://edina.ac.uk/cgi- bin/news.cgi?filename=2011- 08-09-openurldata.txt).

Websites such as Amazon illustrate a range of activity-based recommendations; you might consider what data lies behind each type. For your own account, consider what they need to know about you and about other users – www.amazon.co.uk

Strategies for collecting and storing activity data

Topic

This guide discusses factors which impinge on what activity data to gather and if it should be aggregated before storage.

The problem

Activity data sets are typically large and require processing to be useful. Decisions that need to be made when processing large data sets include selecting what data to retain (e.g. to analyse student transactions and not staff) and what data to aggregate (e.g. for the purposes in hand do we require records of all books or simply a count of books borrowed by each student).

If we are being driven by information requests or existing performance indicators, we will typically manipulate (select, aggregate) the raw data early. Alternatively, if we are searching for whatever the data might tell us then maintaining granularity is essential (e.g. if you aggregate by time period, by event or by cohort, you may be burying vital clues). However, there is the added dimension of data protection – raw activity datasets probably contain links to individuals and therefore aggregation may be a good safeguard (though you may still need to throw away low incidence groupings that could betray individual identity).

The options

It is therefore important to consider the differences between two approaches to avoid, on one hand, loosing useful data by aggregation, or, on the other hand, unnecessarily using terabytes of storage.

Directed approach

Start with a pre-determined performance indicator or other statistical requirement and therefore selectively extract, aggregate and analyse a subset of the data accordingly; for example

- Analyse library circulation trends by time period or by faculty or …
- Analyse VLE logs to identify users according to their access patterns (time of day, length of session).

Exploratory approach

Analyse the full set (or sets) of available data in search of patterns using data mining and statistical techniques. This is likely to be an iterative process involving established statistical techniques (and tools), leading to cross- tabulation of discovered patterns, for example

- Discovery 1 – A very low proportion of lecturers never post content in the VLE
- Discovery 2 – A very low proportion of students never download content

- Discovery 3 – These groups are both growing year on year
- Pattern – The vast majority of both groups are not based in the UK (and the surprise is very low subject area or course correlation between the lecturers and the students)

Additional resources

Directed approach – Library Impact Data Project (LIDP) had a hypothesis and went about collecting data to test it

Exploratory approach - Exposing VLE data was faced with the availability of around 40 million VLE event records covering 5 years and decided to investigate the patterns.

Recommender systems (a particular form of data mining used by such as supermarkets and online stores) typically adopt Approach 2, looking for patterns using established statistical techniques - http://en.wikipedia.org/wiki/Recommender_system and http://en.wikipedia.org/wiki/Data_Mining

Anonymising data

Topic

This guide discusses anonymisation of activity data. This is required before activity data can be distributed to others..

The problem

The Data Protection Act legislates that one cannot release personal data to other people without the data subjects' permission. Much of the activity data that is collected and used contains information which can identify the person responsible for its creation. Examples of personal information are username, library number, IP address of the computer that a user is using, or other information including patterns of behaviour that can identify individual users.

Therefore where information is to be released as open data, consideration needs to be given to anonymising the data. This may also be required for sharing data with partners in a closed manner depending on the reasons for sharing and the nature of the data, together with any consent provided by the user.

The options

Two main options exist if you want to share data.

The first is to **only share statistical data** . As the Information commissioner recently wrote "Some data sharing doesn't involve personal data, for example where only statistics that cannot identify anyone are being shared. Neither the Data Protection Act (DPA), nor this code of practice, apply to that type of sharing."

The second is to **anonymise the personal data** so that it cannot be traced back to an individual. This can take a number of forms. For instance, some log files store user names while other log files may store IP addresses, where a user uses a fixed IP address these could be traced back to them. anonymising the user name or IP address through a purpose-specific algorithm would prevent this. A further problem may arise where rare occurrences might be able to be used to identify an individual. For instance a pattern of accessing some rare books could be linked to someone with a particular research interest. Small counts are often omitted, e.g. loans made by a people attending a course with a low class size.

Taking it further

If you want to take it further then you will need to consider the following as a starting point:

- Does the data you are considering releasing contain any personal information?
- Are the people that you are sharing the data with already covered by the purpose the data was collected for (eg a student's tutor)?
- Is the personal information directly held in the data (user name, IP address)?
- Does the data enable one to deduce a user identity (only x could have borrowed those two rare books – so what else have they borrowed)?

Additional resources

- See also Sharing activity data: Legal issues.
- Data sharing code of practice, the Information Commissioner's office, 2011 http://www.ico.gov.uk/~/media/documents/library/Data_Protection/Detailed_specialist_guides/data_sharing_code_of_practice.pdf
- Three of the projects have also been addressing this issue:
    - Library Impact Data Project (LIDP)
    - Surfacing the Academic Long Tail (SALT)
    - Using OpenURL Activity Data

Bringing activity data to life

Topic

This guide discusses the role of visualisation tools in the exploration of activity data.

The problem

Activity and attention data is typically large scale and may combine data from a variety of sources (e.g. learning, library, access management) and events (turnstile entry, system login, search, refine, download, borrow, return, review, rate, etc). It needs methods to make it amenable to analysis.

It is easy to think of visualisation simply as a tool to help our audiences (e.g. management) 'see' the messages (trends, correlations, etc) that we wish to highlight from our datasets. However experience with 'big' data indicates that visualisation and simulation tools are equally important for the expert, assisting in the formative steps of identifying patterns and trends to inform further investigation, analysis and, ultimately, identify (combinations of data) that enable the recommendation function that you need.

The options

Statisticians and scientists have a long history of using computer tools, which can be complex to use. At the other extreme, spreadsheets such as Excel have popularised basic graphical display for relatively small data sets. However, a number of drivers (ranging from cloud processing capability to software version control) have led to a recent explosion of high quality visualization tools capable of working with a wide variety of data formats and therefore accessible to all skill levels (including the humble spreadsheet user).

Taking it further

YouTube is a source of introductory videos for tools in this space, ranging from Microsoft Excel features to the cloud based processing from Google and IBM to tools such as Gephi, which originated in the world of version control. Here are some tools recommended by people like us:

- **Excel Animated Chart** - http://www.youtube.com/watch?v=KWxemQq10AM&NR=1
- **Excel Bubble Chart** - http://www.youtube.com/watch?v=fFOgLe8z5LY
- **Google Motion Chart** - http://code.google.com/apis/chart/interactive/docs/gallery/motionchart.html
- **IBM Many Eyes** - http://www.youtube.com/watch?v=aAYDBZt7Xk0 Use Many Eyes at http://www-958.ibm.com/software/data/cognos/manyeyes/
- **Gapminder Desktop** - http://www.youtube.com/watch?v=duGLdEzlIrs&feature=related See also http://www.gapminder.org/
- **Gephi** - http://www.youtube.com/watch?v=bXCBh6QH5W0&feature=related
- **Gourse** - http://www.youtube.com/watch?v=E5xPMW5fg48&feature=related

Additional resources

To grasp the potential, watch Hans Rosling famously using Gapminder in his TED talk on third world myths - http://www.youtube.com/watch?v=RUwS1uAdUcI&NR=1

UK-based Tony Hirst (@pyschemedia) has posted examples of such tools in action – see his Youtube channel - http://www.youtube.com/profile?user=psychemedia . Posts include **Google Motion Chart** using Formula 1 data, **Gourse** using EDINA OpenURL data and a demo of **IBM Many Eyes** .

A wide ranging introduction to hundreds of visualisation tools and methods is provided at http://www.visualcomplexity.com/vc/

Enabling student success

Topic

This guide discusses the role of activity data in supporting student success by identifying students at risk who may benefit from positive interventions to assist in their success.

The problem

Universities and colleges are focused on supporting students both generally and individually to ensure retention and to assure success. The associated challenges are exacerbated by two factors: Large student numbers and as teaching and learning becomes more 'virtualised' and the effects of the current economic climate on funding.

Institutions are therefore looking for indicators that will assist in timely identification of such as 'at risk' learners so they can be proactively engaged with the appropriate academic and personal support services.

The options

Evidence is accumulating that activity data can be used to identify patterns of patterns of activity that indicate 'danger signs' and sub-optimal practice. Many students at risk can be automatically identified by matching them on indicators of poor performance. Typically such systems provide alerts or some kind of indicator (eg a traffic light using a green-amber-red metaphor). This approach forms part of the field of 'learning analytics', which is becoming increasingly popular in North America.

Well-chosen indicators do not imply a cause and effect relationship, rather they provide a means to single out individuals using automatically collected activity data, and perhaps based on co- occurring indicators (e.g. Students who do not visit the library in Term 1 and who also do not download content from the VLE are highly likely to be at risk).

Taking it further

Institutions wishing to develop these capabilities may be assisted by this checklist:

- Consider how institutions have developed thinking and methods. Examples from the JISC Activity Data programme appear in the resources below.
- Identify where log information about learning –related systems 'events' are already collected (e.g. Learning, library, turnstile and logon / authentication systems);
- Understand the standard guidance on privacy and data protection relating to the processing and storage of such data
- Engage the right team, likely to include key academic and support managers as well as IT services; a statistician versed in analytics may also be of assistance as this is relatively large scale data
- Decide whether to collect data relating to a known or suspected indicator (like the example above) or to analyse the data more broadly to identify whatever patterns exist
- Run an bounded experiment to test a specific hypothesis

Additional resources

- Three projects in the JISC Activity Data programme investigated these opportunities at

    - EVAD
    - LIDP
    - STAR-Trak
- More about Learning Analytics appears in the 2011 Educause Horizon Report - http://www.educause.edu/node/645/tid/39193?time=1307689897
- Academic Analytics: The Uses of Management Information and Technology in Higher Education, Goldstein P and Katz R, ECAR, 2005 - http://www.educause.edu/ers0508

Data protection issues

It is essential that you consider data protection as this is a legal issue governed by the Data Protection Act 1998 and various codes produced by the information commissioner. There are two key areas that you will have to consider when using activity data:

- Data can only be used for purposes which it had been collected (as notified to the user).
- Where information is being made public (released as open data) the information needs to be suitably anonymised.

Data protection

Whenever data is collected it is a requirement of the data protection act that the user gives their permission. Typically students and staff do this when they join the university through the various policies that they sign up to, however you need to ensure that the permission that they are giving is appropriate to whatever you are trying to do.

The Surfacing the Long Academic Tail (SALT) project found that the users that they consulted expressed no concern about the collection of activity data and its use in the recommender. In fact, they found, most assumed this data was collected anyway and encouraged its use in this way, as ultimately it is being used to develop a tool which helps users to research more effectively and efficiently.

The following advice is available:

- Activity Data Synthesis Guide to Informing users about data collection
- JISC Legal has produced a guide to Data Protection Code of Practice for FE & HE [2008]
- The Office of the Information Commissioner has produced a paper Privacy by design resources http://www.ico.gov.uk/for_organisations/data_protection/topic_guides/privacy_by_design.aspx

Consent management

The Data Protection Act 1988 requires organisations collecting data to gain permission from people prior to collecting that information. It is therefore important to ensure that any consent that your users (staff, students, prospective students and any other visitors) give appropriate consent for the information that is being collected.

The OpenURL Activity Data project sought legal advice on consent management which led them to asking universities using the OpenURL router to seek consent from their users, while the Student Tracking And Retention: STAR-Trak: NG project enables users to take explicit control of what their users can see.

The following further information and advice is available:

- Guide to Informing users about data collection
- JISC legal has produced a briefing Consent Management: Handling Personalisation Data Lawfully from which we have abstracted some of the key details.

Anonymisation

Personal data cannot be shared except for the purposes that the user originally agreed that it could be collected for. However, if the data is anonymised such that it cannot be traced back to an individual then it is no longer personal data and can be published. Therefore, if data is going to be opened up, or shared, perhaps to create a better recommender system, then one approach is to anonymise the data. This can be done in a number of different ways depending on the nature of the data and the purpose it will be used for.

It should be noted that a recent appeal court judgement suggested that the critical issues is "does the 'other information' (if provided to the hypothetical member of the public) add anything to the statistics which would enable them to identify the underlying individuals? If the answer is no, the statistics are not personal data. The underlined words are important: if identification can be achieved from the 'other information' in isolation (rather than when added to the statistics) then the statistics themselves are truly anonymous, and are not personal data."

Within this program, there have been a number of different approaches to anonymisation in different projects. For instance LIDP wrote:

> "Our data involves tying up student degree results with their borrowing history (i.e. the number of books borrowed), the number of times they entered the library building, and the number of times they logged into electronic resources. In retrieving data we have ensured that any identifying information is excluded before it is handled for analysis. We have also excluded any small courses to prevent identification of individuals eg where a course has less than 35 students and/or fewer than 5 of a specific degree level."

Note that they have both anonymised any identifiers that might be used to identify students, but have also excluded small courses, setting the lower limit on course size that they use at 35.

The following advice is available:

- Guide to Anonymisation
- Further information on publishing anonymised data
- Data sharing code of practice, the Information Commissioner's office, 2011 http://www.ico.gov.uk/~/media/documents/library/Data_Protection/Detailed_specialist_guides/data_sharing_code_of_practice.pdf
- Projects have also taken their own advice as summarised below:
- AEIOU were concerned with the ability to identify users from IP addresses.
- AGtivity - At present only sites that agree to receiving their data have been analysed. For more public documents specific names have been removed within for example the case studies.
- EVAD took a cautious approach to data protection, with the ability to release more data to partners at a later date if this seemed appropriate.
- LIDP have produced summary statistical information from which individuals cannot be identified.
- OpenURL
- RISE - have altered the Open University's privacy policy and anonymise the data
- STAR-Trak- NG

You should also look at the section on licensing and sharing of activity data.

Consent management

JISC Legal have a produced a report Consent Management: Handling Personalisation Data Lawfully which addresses the following questions:

- Must an institution always get the consent of learners if it wants to process their information in a new and innovative way?
- What is the best means of administering the individual consent of learners to various processing activities that will occur with their data?
- What laws apply to the vast quantities of user activity data that are generated as learners participate in and engage with online resources?

Their recommendations are:

For Institutions Generally

- Institutions should raise awareness of the risks to those engaging with and providing personal data to external third party organisations.
- Where there is no reasonable likelihood of an individual being identified personalisation data can be considered anonymous data and not subject to the DPA 1998. Consent to its use therefore is not required.

- Where consent is obtained, the data subject should be left in no doubt that they are giving their consent - consent should be specific and informed.
- Ensure that details of online services that are required as part of a student's learning are described up- front - probably at the time registration.
- Where learning providers process users' information that is indicative of their online activity and interests this should be considered personal data.
- Any profiling of natural persons or other value-added services should only be carried out after expressed consent has been obtained from the user.
- When processing is **not necessary** then personally identifiable data must be processed only with the consent of the individual involved.
- Where processing is **necessary** in order for an agreed relationship to take place and this processing is clearly described in advance to the learner then further consent to the particular processing is not necessary.
- Where learner's personal data is provided to external online services as part of their learning there is an onus on the institution - as data controller - to ensure that the learner understands that the transfer is necessary for the purposes of their learning.
- Where institutions are using social network services as data processors then the institution is obliged to ensure that this processing is data protection compliant.
- Data protection compliance should be designed into systems that are processing personal information from the start. In many cases conducting a Privacy Impact Assessment can be a useful method of gauging the privacy risks to individuals.
- If storing or processing of personal information including personalisation information cannot be justified as necessary, then it should not take place at all.
- In the Information Commissioner's view, provided there is no likelihood of a significant adverse effect on the individual as a result of processing their information, the specific consent of the data subject will not always be required.
- It is recommended that sensitive personal data of learners is not provided to external service providers without the explicit consent (in writing) of the individual learners involved.

For Information Technology Staff

- Attributes that do not reveal personally identifiable information should be used wherever possible.
- Data should be anonymised by removing all personal identifiers wherever possible.
- It is necessary to have in place the technical and organisational measures necessary to ensure that anonymous data cannot be reconstituted to become personal data.
- A collection notice describing who is in control of the processing and what use will be made of the personal data is required to satisfy the fair processing information element of consent.
- Attributes should be designed not to collect or categorise information according to racial or ethnic origin or gender, for example, unless there are clearly justified and lawful reasons for such collection or categorisation.
- The **eduPerson** specification is recommended as the basis for the metadata standard for the attribute set that will be stored for end users since this is well developed and implemented in some institutions' directory servers already.

For External Service Providers

- Users should be fully informed by means of a collection notice of all profiling and personalising processing that involves their usage activities.
- Clear informed consent obtained from the individuals concerned can satisfy the legal obligations contained in the DPA 1998.
- The use of systems that anonymise personalisation data fits well with the approach that information that identifies individuals should be minimised or avoided.

Activity data to Enhance and Increase Open-access Usage (AEIOU)

The IP Address identifies the computer from which the request originated and is used to provide the notion of a user session. Although this may not directly identify a user (e.g. the computer maybe shared publicly), in terms of Data Protection Act (DPA), IP addresses may constitute personal data if an individual user can be identified by using a combination of that IP address and other information. This applies even when personal data are anonymised after collection.

New European legislation came into force from May 26th 2011 and The Information Commissioner's Office (ICO) Code of Practice has been revised. The Code now clearly states that in many cases IP addresses will be personal data, and that the DPA will therefore apply. These changes also apply to the use of cookies and methods for collecting and processing information about how a user might access and use a website. An exception exists for the use of cookies that are deemed "strictly necessary" for a service "explicitly" requested by a user. In general, the regulations advise that an assessment should be made on impact to privacy, whether this is strictly necessary and that the need to obtain meaningful consent should reflect this.

We also need to consider that the AEIOU project is aggregating and processing data (that includes IP Addresses) originating from other institutional Repositories with no direct end-user relationship. The Using OpenURL Activity Data project has addressed this by notifying institutions that sign up for their OpenURL resolver service. We have no explicit agreement with the partners involved in the current project but aim to review their existing privacy policies should the service be continued. For example, do policies for storing and processing user data include repository reporting software and Google analytics and should users be made aware of this through the repository website?

The current cookie policy for Aberystwyth University can be found here

In order to comply with recent changes to ICO code of practice we have been advised that as a minimum requirement we should include text in the header or footer of repository web pages and a link to a Data Privacy Policy that clearly informs users about how their data is being used and whether it is passed to third parties (e.g. Google). Where possible, they should also be given the option to opt out of supplying personal information (IP address) to the Recommendation service. This would not affect them receiving recommendations but their information would not be stored or processed as part of the service.

The AEIOU privacy policy:

Recommendation service

In order to provide a recommendation service (lists of items frequently viewed together) for this repository, usage data is collected and processed. Data is collected using software that tracks each web site visitor via their IP address and this is stored in an encrypted form to protect the users' privacy. The usage data may also be used for reporting purposes, i.e. most highly viewed article, most downloaded pdf file, etc. We do not gather any other personal information.

[Optional - if repository uses Google Analytics]

Tracking Cookies

This repository website also collects data using tracking Cookies*. The use of cookies to track web site usage is a widely used and well established technique. Each tracking cookie provides information to show how a visitor moves from page to page, but no personal data is stored by the tracking system. The tracking system used on this repository web site is called Google Analytics.

As the service provider, Google also has access to the data (for more information visit the Google Privacy Policy web page).

Rejecting Cookies

You can set up your web browser to reject all cookies and if you decide to do this you will still be able to use the repository web site. The documentation for your web browser will provide you with instructions on how to disable cookies.

* A cookie is a small piece of information in the form of a text file, sent by a web server and stored on the computer of a visitor to a web site. It can then be read back later by the web server when required.

Exposing VLE data (EVAD)

This cautious approach meant hiding any data that would identify an individual user, site names and anything that might link a session back to a particular user. We settled on a hashing algorithm to use to obscure any such items of data yielding a string that can be determined uniquely from the value; we also used a salt to prevent inversion of the hash through exhaustive search of short inputs.

At this stage, we also looked at some sample data to reinforce our decisions.

The decision on what to hash was straightforward in many cases such as concealing any field with Site, User Name, URL or Content in it. Some things were less clear cut. For instance, the skin around a site could be used to identify a Department. The Session Ids that we looked at appeared to be numeric and we decided there was little risk in leaving this in its raw state. However, later testing revealed that, in some instances and points of time, this has included a user identifier so we agreed to hash this. It is worth remembering that the hashing algorithm is consistent so even though the value of the Session Id has been changed, it can still be used to link the Event and Session tables.

Using OpenURL Activity Data

The Using OpenURL Activity Data project sought legal advice which resulted in the following:

> "The Router currently generates activity data, i.e. when a user clicks on a link to view a paper and their request is routed through the Router, the Router logs capture the IP address of the computer making the request along with the information in the request eg the OpenURL of the paper that has been requested. These are not currently collected and processed but the service envisaged here would involve anonymising IP addresses and creating an aggregation both of which constitute processing. We have been advised to err on the side of caution by seeking to ensure that collection and processing of activity data is disclosed to users. As the OpenURL Router service is middleware with no direct relationship with end users, we cannot expect end users to be aware that they are using an EDINA service and that EDINA is potentially processing their personal data (subject to the EDINA privacy policy). Instead, disclosure of this activity would have to be made by the institution whose end-users are routed through the Router, eg in the institution or library's privacy policy. There are currently 90 institutions registered with the Router.
>
> "Our advisor suggested that EDINA post notification on the page where institutions register to use the Router which tells the institution that EDINA collects activity data and intends to include them in an anonymised form in such an aggregation and tells those institutions that they should advise
>
> "their users of this activity (eg through their privacy policy). It would be important to offer the option for institutions to opt out of the aggregation. The 90 institutions that have already registered should be advised of the aggregation by email and also given the option to opt out. Data from those registered institutions which opt out will be excluded from the aggregation. (Identifying data related to a specific institution and deleting it constitutes processing of data but, again, this would be done to avoid inclusion in an aggregation and to protect the wishes of the institution with regard to data that may be construed as personal data. Thus, we seek to minimise the risk that privacy is breached). Clearly, data generated before this process of disclosure is undertaken cannot be included in the aggregation. "

Full details can be found in their report: Using OpenURL Activity Data: Legal and policy issues - initial investigation.

Recommendations Improve the Search Experience (RISE)

Within the RISE database personal data is stored and processed in the form of the Open University Computer User account name (OUCU). The OUCU is generally a 5 or 6 character alphanumeric construction (e.g. ab1234) that is used as the login for access to OU systems. This OUCU is stored within

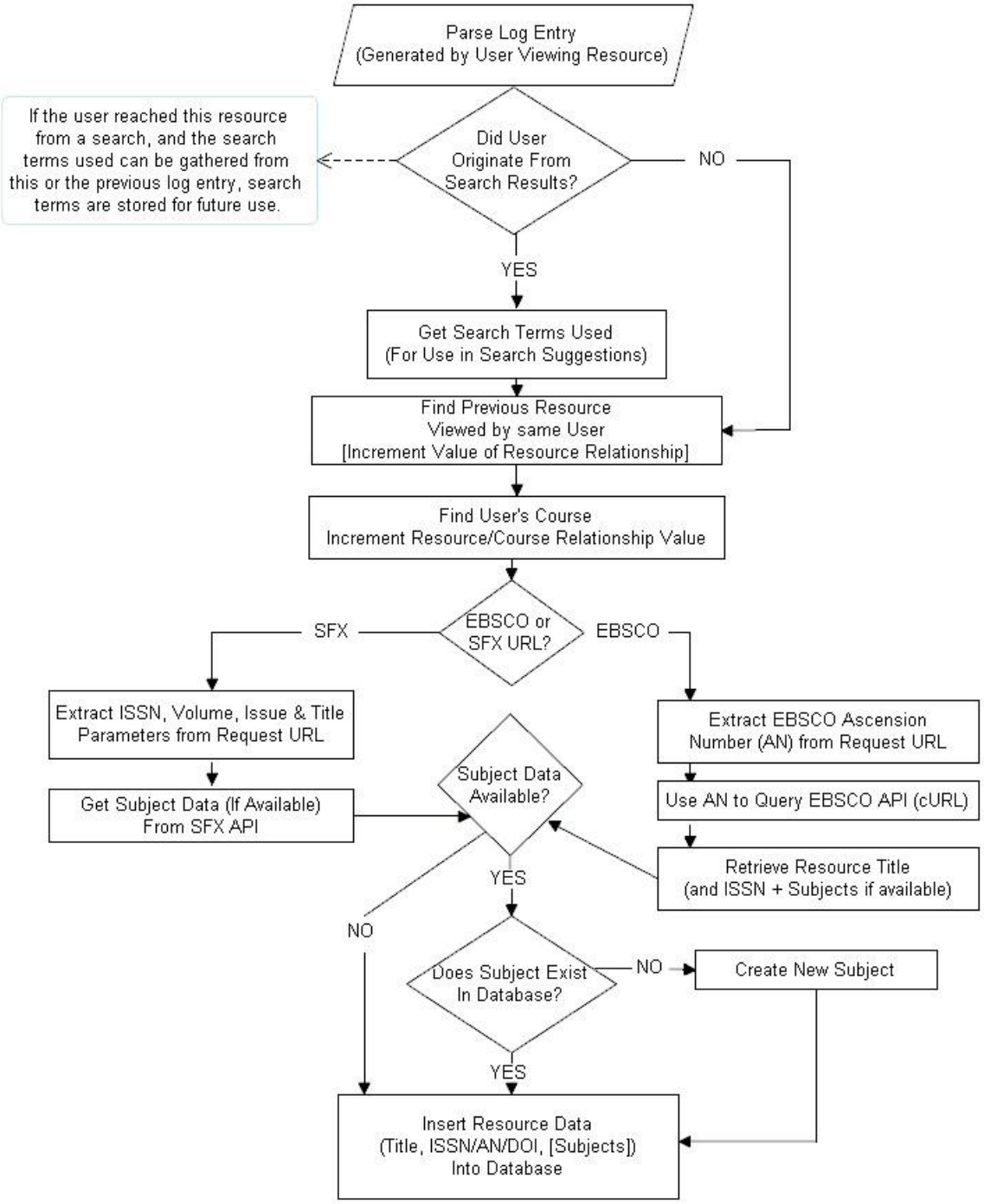
the EZProxy log files that are ingested into the RISE database and is also tracked by the RISE interface to allow searches to be related to users.

This OUCU is used within the RISE system for two purposes:

- To be able to make a connection between a search and a module of study associated with the searcher, to allow recommendations based on module; and,
- To be able to remove all searches for a particular user from the recommendations database at their request.

Processing takes place using a file of data from internal systems to add the module(s) being studied by matching the OUCU in the RISE database with the OUCU stored by internal systems. The data on which module is being studied is added into the RISE database. As each new OUCU is added to the database a numerical userID is assigned. This is a simple incremental integer.

The RISE database stores details of which electronic resources are accessed by the user and the search terms used to retrieve that resource (for searches carried out through the RISE interfaces)



Privacy approach

The RISE project has developed a separate Privacy policy to cover use of activity data as it was felt that the standard OU Privacy policy was not sufficiently explicit regarding the use of data for this purpose. The newly developed privacy policy is available at http://library.open.ac.uk/rise/?page=privacy

One of the challenges with using EZProxy data is that the EZProxy log files contain records from links in several different systems as we link as many systems as possible through EZProxy. So this privacy policy has also been linked from the Library SFX and Ebsco Discovery Search interfaces.

As well as explaining how their data will be used the policy provides a mechanism for users to ask for their data to be removed from the system and for their data not to be recorded by the system. This opt-out approach has been cleared by the Open University Data Protection team.

The EZProxy log files that are used within the system provide a particular challenge to an opt-in approach. Access to this system is simply through expressing a URL with libezproxy.open.ac.uk within the URL string e.g. http://portal.acm.org.libezproxy.open.ac.uk/dl.cfm This URL then redirects the user through the EZProxy system. These links can exist in many different systems.

Data on accesses to electronic resources is still required to be kept within log files to allow the library to comply with licensing restrictions for the electronic resources to allow the library to track any abuse of license conditions. An opt- out could only be applied to the usage data element of the personal data.

Users do not login to the EZProxy system directly but are faced with a standard Open University login screen to authenticate if they are not already recorded as being logged in.

Future privacy changes

An opt-in approach may be required to comply with the new EU directive on 'cookies'. Conceivably this may be achievable by redirecting all EZProxy links through an additional authentication process and asking users to agree to storing their usage data. This acceptance could be stored at the server- side although this introduces a further single-point of failure that could block access to electronic resources. Alternatively a cookie approach could be taken along with asking the user to accept the cookie.

Student Tracking And Retention: STAR-Trak: NG

The STAR-Trak project wrote:

> **Data**: In an earlier incarnation of our proposal to JISC we postulated creation of an activity data set that might form the basis for a future national schema and dataset. The ultimately very good advice (thank you!) we received was that was too ambitious a goal for this project, and so our data is for internal use only, removing the need to consider anonymisation. We have not managed to run a "live" pilot during the term of the project, and so there is no live data in the application. Moving forwards our basic position is that we are simply consuming data that already exists in our IT ecosystem - nothing new is created. That is not quite true as we have added functionality to capture interactions between student and tutor, extra-curricular activities and use of the system creates new data. However there is nothing innovative in this data that requires us to treat it differently from similar data elsewhere.
>
> Of course what has changed is that data that was previously hidden or only available with difficulty and/or to certain people is now available to a wider set of people, and we have had to consider the privacy implications of that. To recap on the basic STAR-Trak functionality, it allows the student to see their own information from the student record system and activity data from the systems that they use to support their learning activities. It also allows tutors, module and course leaders to view this information on a course, module or student basis. We made the decision that trying to map relationships between staff and students as a basis for deciding who can see what, while great in theory, was not going to be sustainable in practice. These relationships may or may not be defined in external systems and they can be quite fluid. The combination of these two factors makes maintaining the relationships potentially a heavy administrative burden that would not be sustained, resulting in the system becoming unusable over time.
>
> As an alternative we have implemented what we call "social controls". In simple terms this has two elements to its operation:
>
> 1) it allows a student to block all or any member of staff from seeing anything bar their very basic data (their name and what course they are on)
>
> 2) any member of staff [authorised to use the system] can explicitly define in the application a relationship with a student (personal tutor, module tutor etc) and by doing so they can then view the student's details (subject to 1 above). However a record is created in the application noting the relationship and this can be audited.
>
> This is further strengthened by participation in STAR-Trak being fully voluntary on the students part.
>
> We view this control system as an innovative experiment to be validated under strict trial conditions. We have already identified several enhancements that may improve the level of confidence in such a system. As per previous post we are still working hard to get formal commitment to running a year-long pilot of STAR-Trak and this continues to move in the right

direction, albeit slowly (from the project's perspective!). It is only in practice that we will see how successful we have been in developing a usable and useful application that meets data compliance requirements. As Chuck Reid famously said "In theory, there is no difference between theory and practice. In practice, there is".

Licensing and sharing activity data

There is much that can be gained from sharing activity data in order to gain a wider picture. For instance, the SALT project believes that information from 10 University libraries would probably be sufficient to provide a useful recommender system for all universities based on their methodology. In order to share data you will need to ensure that you have either anonymised the data or requested suitable permissions from the data subjects. You will also have to consider the license that you will offer the data under and how you will publish the data.

In this section we cover:

- Licensing of data
- Publishing

Licensing of data

Licensing considerations regarding activity data need to take account of the motivations of the institution generating the data as well as general legal and sector principles

Motivations for generating activity data are discussed in the introduction. Significantly, when considering licensing for wider use, some institutions will be focused on benefits to individual students, including retention, success and progression; such data may therefore be traceable to a recognisable individual and not suitable for wider use and re-use unless anonymised.

Much activity data will however be anonymised in a manner that opens the options of wider use through licensed release within and beyond the institution. The following considerations apply to such data

- **Are there any benefits from releasing activity data?** - The work of this programme suggests there may be as the data may tell more significant stories when aggregated within and / or beyond the institution
- **Are there legal data protection / privacy issues?** - Not if it is anonymised (see Anonymisation guide)
- **Will the institution be undermining its competitive advantage or its reputation?** - That's a business case decision
- **Will members of staff feel or actually be compromised?** - Even if the data is anonymised in terms of staff identification, that may still be the case as such as course codes point to faculties and individuals; albeit sensitive, this is a business case issue relating to corporate mission and performance.
- **What will students say?** - They probably expect the institution to be leveraging this data to their advantage; they should require the institution to adhere to the law and its own policies in terms of privacy; they might request access to their own data for all sorts of reasons, such as using it in other services.

For further information see:

- Licensing considerations
- Licenses used

Licensing Considerations

So you have decided that there is a case for making the data available for use and re-use

- **Is licensing necessary when activity data is released as open data?** - Yes, otherwise potential users will be in doubt what they may or may not do with it. This is at the heart of the Discovery principles set out by JISC in 2011 (http://discovery.ac.uk/profiles/principlesprofile/), which already have traction amongst the UK HE library, archives and museums community

- **Are the Discovery principles wholly applicable to activity data?** - The SALT project observed that 'whilst there are common principles between the sharing of activity data and bibliographic data there are also some differences. In particular, activity data is unique to that particular institution and is generated from the behaviour of individuals within the institution.' This might influence the specific choice of open data license.
- **Is licensing still necessary when the data is released for internal use only?** - Yes, as the same applies - that potential users will be in doubt what they may or may not do with it.
- **Does licensing involve a lot of effort and legal advice?** - It should not, as there is a range of appropriate and recommended open data licenses available from Creative Commons and Open Data Commons (see Guide) legal advice can be focused on the choice of one of these licenses in the context of the Discovery principles.
- **Which license should we be using?** Creative Commons and Open Data Commons open data licenses provide opportunity to apply the conditions you believe necessary, notably attribution obligations and use restrictions (such as no commercial use). However, the Discovery principles make it clear that such restrictions can be counter productive.
- **Where can we get advice?** The institutions that licensed activity data in this project are an important source of advice (see licenses used). You should also study the practical guide to licensing open data prepared for the Discovery initiative by Naomi Korn and Charles Oppenheim (2011) - see http://discovery.ac.uk/files/pdf/Licensing_Open_Data_A_Practical_Guide.pdf. JISC Legal can also provide further advice.

Licenses used

The table shows the licenses used by each of the projects.

Project	License Adopted	Commentary
AEIOU	n/a	This project principally focused on using activity data in a consortium recommender service rather than opening it up for wider use. However, this still required the project to address issues regarding the reuse of data and statistics linked to IP addresses (as per OpenURL and RISE). See here for further details. AEIOU has also developed a Privacy Policy to cover the use of activity data in a recommender service - see http://cadair.aber.ac.uk/dspace/handle/2160/7202
AGtivity	Decision pending	The project originally aimed for all the data to be licensed for free use by the community. Grouped statistics are being published openly and now available in reports from Janet UK (not as open data). Individual statistics although being captured with permission, are only being sent to those with node ownership. The aim is to convince people that the gains from the hypothesis questions, are worth a small risk of lack of privacy and security. See here for further details.
EVAD	Licensing planned - choice of licence pending	EVAD has developed the Privacy Policy for the data and consulted with the university Legal Services team on releasing anonymous logging data under an Open Licence. The data has been released but the license choice is pending. However they also considered the Privacy Policy for the site and consulted with our Legal Services team on releasing anonymous logging data under an Open Licence.
OpenURL	Released under ODC-	In May 2011, as agreed with the service partners, EDINA made the first release of the OpenURL Router Data under the Open Data Commons - Public Domain Dedication Licence (ODC-PDDL)

	PDDL	with Attribution Sharealike Community Norms. Details of the data set, sample files and the data itself is available at OpenURL Router Data. Further data will be made available monthly. EDINA also developed a privacy policy for use by service partners.
LIDP	Plans to use ODbL	In the near future, the project (which involves data from several institutions) hopes to start releasing data using the Open Data Commons Attribution License covering attribution and share- alike rights for data.
RISE	Licensing intended - perhaps ODC-PDDL	There is a commitment to investigate the potential of making the RISE activity data available under an open licence. The project notes that similar data has already been released by the OpenURL project at EDINA (above). The project has also developed a privacy policy. See here for further details.
SALT	To be released under CC-BY	In joining the SALT project, the John Rylands University Library agreed to make its loan data available for use by the SALT recommender service and to make the data available to others under an appropriate licence. The University Librarian has therefore agreed that JRUL anonymised loan data will be made available under a Creative Commons attribution only licence (CC BY). See here for further details.
STAR-Trak-NG	n/a	Open licensing is not appropriate for this project, which is focused on individual learner support. The project has worked extensively on the privacy implications of its use cases and has developed a 'social controls' model whereby a student can determine which staff can access their activity data. See here for further details.
UCIAD	n/a but licensing relating to the individual raised	Like STAR-Trak-NG, UCIAD is focused on providing benefits to the individual user and therefore open release by the institution is not a priority and furthermore may require the data to be watered down. Therefore licensing is not planned. However the project recommends consideration from the user 'my data' perspective, See here for further details.

Activity data to Enhance and Increase Open-access Usage (AEIOU)

The AEIOU project is aggregating activity data generated by users (both registered and anonymous) who download or view an item held in an institutional repository. The data used to describe this activity is represented by an OpenURL Context Object which is stored and processed to provide the shared Recommendation Service and includes the Request IP address.

Data Protection & Privacy Issues

The IP Address identifies the computer from which the request originated and is used to provide the notion of a user session. Although this may not directly identify a user (eg the computer maybe shared publicly), in terms of Data Protection Act (DPA), IP addresses may constitute personal data if an individual user can be identified by using a combination of that IP address and other information. This applies even when personal data are anonymised after collection.

New European legislation came into force from May 26th 2011 and The Information Commissioner's Office (ICO) Code of Practice has been revised. The Code now clearly states that in many cases IP addresses will be personal data, and that the DPA will therefore apply. These changes also apply to the use of cookies and methods for collecting and processing information about how a user might access and use a website. An exception exists for the use of cookies that are deemed "strictly necessary" for a

service "explicitly" requested by a user. In general, the regulations advise that an assessment should be made on impact to privacy, whether this is strictly necessary and that the need to obtain meaningful consent should reflect this.

We also need to consider that the AEIOU project is aggregating and processing data (that includes IP Addresses) originating from other institutional Repositories with no direct end-user relationship. The Using OpenURL Activity Data project has addressed this by notifying institutions that sign up for their OpenURL resolver service. We have no explicit agreement with the partners involved in the current project but aim to review their existing privacy policies should the service be continued. For example, do policies for storing and processing user data include repository reporting software and Google analytics and should users be made aware of this through the repository website?

The current cookie policy for Aberystwyth University can be found here

In order to comply with recent changes to ICO code of practice we have been advised that as a minimum requirement we should include text in the header or footer of repository web pages and a link to a Data Privacy Policy that clearly informs users about how their data is being used and whether it is passed to third parties (eg Google). Where possible, they should also be given the option to opt out of supplying personal information (IP address) to the Recommendation service. This would not affect them receiving recommendations but their information would not be stored or processed as part of the service.

Anonymisation & Re-use of data

We will make data available to individual partners and hope to provide a reporting service (based on the activity data) so that institutions can view usage statistics in a National context. We also hope to publicly release the data with regard to personal data encryption and licensing outlined below. Ideally, we would like to release OpenURL Context Object data as XML but in the short term this will be made available in CSV format.

The JISC Usage Statistics Review looked at European legal constraints for recording and aggregating log files and noted that the processing of IP- addresses is strongly regulated in certain countries (e.g. Germany) and that current interpretation maybe ambiguous. In such cases, they advise that "To avoid legal problems it would be best to pseudonymize IP- addresses shortly after the usage event or not to use IP-addresses at all but to promote the implementation of some sort of session identification, which does not record IP- addresses"

Currently, we are encrypting the IP addresses using an MD5 hash algorithm recommended in Knowledge Exchange Usage Statistics Guidelines so that personal data is anonymised. Although MD5 is a relatively fast and efficient algorithm it has been shown to have security vulnerabilities and other more secure methods for encryption (eg SHA-1 & SHA-2) are recommended. If this becomes an issue we could release data with stronger encryption or replace the IP address with a system identifier as suggested above. Removing the IP address would, however, compromise the ability to aggregate data.

The Knowledge Exchange Usage Statistics Guidelines also point out that when the IP address is obfuscated, information about the geographic location is lost. They therefore recommend using the C-Class subnet part of the IP address which will give a regional (network) location but can not identify a personal computer. This would be appropriate where activity data is used for reporting statistics.

Licensing

Document outputs, software and any data that is released will be licensed according to the IPR section in the project plan which said:

> "All document outputs will be released under a Creative Commons license (attribution CC BY) which is recommended for maximum dissemination and re-use of materials. Data will be released under ODC Public Domain Dedication and Licence (PDDL) and software is released under the Apache foundation license (version 2.0) which allows for modification and distribution of source code. The aim is to promote re-use of all outputs and encourage collaborative development across both nonprofit and commercial organisations"

Exploiting Access Grid Activity Data (AGtivity)

The AGtivity project wrote:

> **"Open source**: We described originally that all scripts developed in this project will be available on an open source basis, licensed for free

non-commercial use and development and will be available to the UK HE and FE community in perpetuity. This has been formulated a bit more with the description of 'recipes' (see previous blogs) describing the automated process to data mine and produce pdf files with in- built visualisations.

There is an issue with privacy of the data sets.

- We originally aimed for all the data to be licensed for free use by the community.
- And the grouped statistics are being published openly that sum access times for example, and now available in reports from Janet

Individual statistics although being captured with permission, are only being sent to those with node ownership. It has been found that with data mining evaluations can be carried out that could judge people

The aim is to convince people that the gains from the hypothesis questions, are worth a small risk of lack of privacy and security.

The collection of global stats reporting required by Janet are now under review and it is hoped that these reports form use-cases will inform the process and insist on a finer degree of reporting. The new SLAs for a combined video conferencing support service to be held at the University of Edinburgh is under negotiations to be confirmed in August 2011.

Privacy and anonymisation: At present only sites that agree to receiving their data have been analysed. For more public documents specific names have been removed within for example the case" studies.

Exposing VLE activity data (EVAD)

We are now in a position to make an initial release of our VLE logging data. This blog details the process we went through to get to this stage and the decisions that we made along the way.

We were helped in this process by David Evans from Cambridge Computer Lab (http://www.cl.cam.ac.uk/~de239/) who is an expert in data privacy issues. We also considered the Privacy Policy for the site and consulted with our Legal Services team on releasing anonymous logging data under an Open Licence.

There are 3 files that we will be releasing:

- Sakai Event File - Individual events such as a site visit which occur within a Session (held as a separate file for each Academic Year)
- Sakai Session File - Session details which include browser information and session times
- Sakai Site File - Details of the VLE Site being accessed

A tarball file can be obtained from here - however this is a 4GB file that will expand to over 12GB when uncompressed.

Our first step was to provide David Evans with the database schema for these files so he could consider, based on discussions with us, which fields might contain sensitive data. We then discussed the implications of making these fields public and what information they might expose. This was of course a balancing act between releasing data from which others could draw value and not giving away details of what a known individual might have used the VLE system for.

We decided on a first data release which would use a cautious approach to data privacy. Should other institutions find this data of interest, we can work with them to reveal more information in the area of interest to them in a manner that does not compromise individuals.

This cautious approach meant hiding any data that would identify an individual user, site names and anything that might link a session back to a particular user. We settled on a hashing algorithm to use to obscure any such items of data yielding a string that can be determined uniquely from the value; we also used a salt to prevent inversion of the hash through exhaustive search of short inputs.

At this stage, we also looked at some sample data to reinforce our decisions.

The decision on what to hash was straightforward in many cases such as concealing any field with Site, User Name, URL or Content in it. Some things were less clear cut. For instance, the skin around a site could be used to identify a Department. The Session Ids that we looked at appeared to be numeric and we decided there was little risk in leaving this in its raw state. However, later testing revealed that, in some instances and points of time, this has included a user identifier so we agreed to hash this. It is worth remembering that the hashing algorithm is consistent so even though the value of the Session Id has been changed, it can still be used to link the Event and Session tables.

The Session Server holds both the Host Name and software module Id. We decided to hash the Host Name, in case this might reveal something about the network's internal structure, but leave the numeric part (software module id) alone as it reveals nothing other than which particular instance of software processed that session. We discovered that the format of this field had changed over time so we needed a mildly complex function to detect and extract from both formats of the Session Server.

The Session User Agent may provide the Browser Name and Version and the Operating System and Version of the computer used for the session. However this is a free-form field which may be changed by the user. There was a danger that this could identify a given user. A visual inspection showed at least one College Name, some company names and some school names within this field which could present a risk. Ideally we would extract and expose data such as the Browser Name but as this is a free-form field this in non-trivial. We therefore took the decision to hash this field.

As a final sanity check, we revisited some sample data from each of the tables once they had been hashed to satisfy ourselves that there was no raw data left that might possibly contain a user identifier.

Recommendations Improve the Search Experience (RISE)

Project code licensing

By the end of the project the RISE code, covering the data ingestion processes and recommendation code will be made available via Google Code at http://code.google.com/p/rise- project/. After consideration of suitable open source licenses it has been decided to use the standard license for Google Code GNU GPL v3 http://www.gnu.org/licenses/gpl.html. This has previously successfully been used to release previous project code created by the OU for JISC projects.

Open data licensing

Discussions with the Open University Rights team have identified that we are able to release data from EZProxy, from search terms used within RISE, and covering the general subjects covered by OU courses. An appropriate license for this content would be CCZero. This owes much to the previous work of the Lucero project in paving the way for the open release of data.

What data could be included?

What became apparent during the project was that most of the EZProxy request URLs linked through to EBSCO (the reason being that we link our EBSCO Discovery Solution through EZProxy) and that there was very little bibliographic data within the log files. We discovered that we could use the EBSCO accession number to retrieve bibliographic data but that we weren't licensed to store that data in the RISE database yet alone release it openly. We found an alternative source of article level metadata (from Crossref) that we could store locally, but again licensing terms prohibit its inclusion within an open data set.

A conversation was had with JISC Legal, who advised that if restrictions are placed on database vendors, these are usually passed on to subsequent users. Restrictions may not necessarily be just in relation to copyright. If the database vendor is using third party material (i.e. obtained from elsewhere) there will very likely be a purchasing agreement/contract/ as well as a licensing agreement (where the copyright position is stated) between the parties stating what the vendor may do with the data. The vendor would then need to impose the same conditions on the customer, so as not to breach their agreements with the party from where they obtained the material. So it could be breach of contract terms as well as breach of copyright depending on the agreements.

There is some difference of opinion between Rights experts about the position with article level metadata about whether it could be used and released. Commercial providers assert in their terms and conditions that you cannot reuse it or share it and libraries are in a position where they have signed license agreements that contain those clauses. This is an area where agreement about the exact legal position with regards to article level metadata should be established. Not having openly available and reusable

article level metadata would be a distinct barrier to establishing useful and usable datasets of article level activity data.

Advice from JISC Legal on the copyright issues around metadata, directed us to a quote from their paper Licensing Open Data:

> " Where there has been substantial investment in the selection and or presentation of the content of datasets they may attract copyright as well as database right if it was created after 27 March 1996 and if there has been evidence of creative effort in selecting or arranging the data. A database might have copyright protection in its structure if, by reason of the selection or arrangement of its contents, the database constitutes the author's own intellectual creation. Copyright protection of individual data, including records and metadata that have been "expressively" written or enriched may also subsist in the structure of the database if that structure has been the subject of creativity."

So in terms of what we could release openly we are left with a dataset that contains URLs that link to EBSCO, search terms entered through RISE and course subjects.

Type	**Data**
Basic data: Institution, year and dates	institution name academic Year extracted date source
Resource data	Resource URL
User context data	anonymised User ID timestamp Students subject Students level eg [FE, UG1, UG2, UG3, UG4, M, PhD1, PhD2, PhD3+ Staff
Retrieved from	SearchTerm

The dataset includes relationships between resource records in the dataset but there is no easy way of being able to relate that resource to a DOI or article title. And that leaves the dataset as being potentially of use to other EBSCO Discovery Solution customers but no one else. **So at this stage we have reluctantly decided that we won't be able to release the data before the RISE project ends.** Further work would be needed to review other data sources such as the Mendeley or OpenURL router data to see if they could provide some relevant article level metadata.

Surfacing the Academic Long Tail (SALT)

The SALT Project wrote the following regarding licensing:

> "In joining the SALT project, the JRUL agreed to make its loan data available for use by the SALT recommender service and to make the data available to others.
>
> "The second step was to agree the terms on which this data would be released. JISC and the Resource Discovery Taskforce argue, in the context of bibliographic data, that the most open licence possible should be used and that restrictions, such as restricting use to non- commercial activities, should only be applied if the impact of this is fully understood. They also strongly recommend that institutions use … standard licences now widely available rather than developing their own. [See here]
>
> "Whilst there are common principles between the sharing of activity data and bibliographic data there are also some differences. In particular, activity data is unique to that particular institution and is generated from the behaviour of individuals within the institution.
>
> "Rather than waiving all rights, therefore, a recommendation was made to the University Librarian that JRUL activity data be licensed for use outside of the University and that this be done using the most open licence available.

> "The University Librarian has now agreed that JRUL anonymised loan data will be made available under a Creative Commons attribution only licence (CC BY)."

Student Tracking And Retention: STAR-Trak: NG

The STAR-Trak project wrote:

> "**Software**: The STAR-Trak application has been developed in PHP (using the CodeIgniter framework) with an Oracle database at the back end. We took some advice early on from OSS Watch as we wanted to keep open the potential for releasing STAR- Trak code as open-source, and our understanding is that nothing we have done by way of development or contractually with our external developers stops us from achieving this goal. However we do carefully need to consider if that is the best way forward for the application the sector and our University - and we will be seeking further advice in due course.
>
> **Data**: In an earlier incarnation of our proposal to JISC we postulated creation of an activity data set that might form the basis for a future national schema and dataset. The ultimately very good advice (thank you!) we received was that that was too ambitious a goal for this project, and so our data is for internal use only, removing the need to consider anonymisation. We have not managed to run a "live" pilot during the term of the project, and so there is no live data in the application. Moving forwards our basic position is that we are simply consuming data that already exists in our IT ecosystem - nothing new is created. That is not quite true as we have added functionality to capture interactions between student and tutor, extra-curricular activities and use of the system creates new data. However there is nothing innovative in this data that requires us to treat it differently from similar data elsewhere.
>
> Of course what has changed is that data that was previously hidden or only available with difficulty and/or to certain people is now available to a wider set of people, and we have had to consider the privacy implications of that. To recap on the basic STAR-Trak functionality, it allows the student to see their own information from the student record system and activity data from the systems that they use to support their learning activities. It also allows tutors, module and course leaders to view this information on a course, module or student basis. We made the decision that trying to map relationships between staff and students as a basis for deciding who can see what, while great in theory, was not going to be sustainable in practice. These relationships may or may not be defined in external systems and they can be quite fluid. The combination of these two factors makes maintaining the relationships potentially a heavy administrative burden that would not be sustained, resulting in the system becoming unusable over time.
>
> As an alternative we have implemented what we call "social controls". In simple terms this has two elements to its operation:
>
> 1) it allows a student to block all or any member of staff from seeing anything bar their very basic data (their name and what course they are on)
>
> 2) any member of staff [authorised to use the system] can explicitly define in the application a relationship with a student (personal tutor, module tutor etc) and by doing so they can then view the student's details (subject to 1 above). However a record is created in the application noting the relationship and this can be audited.
>
> This is further strengthened by participation in STAR-Trak being fully voluntary on the students part.
>
> We view this control system as an innovative experiment to be validated under strict trial conditions. We have already identified several enhancements that may improve the level of confidence in such a system. As per previous post we are still working hard to get formal commitment to running a year-long pilot of STAR-Trak and this continues to move in the right direction, albeit slowly (from the project's perspective!). It is only in practice that we will see how successful we have been in developing a usable and useful application that meets data compliance requirements. As Chuck Reid famously said "In theory, there is no difference between theory and practice. In practice, there is".

User-Centric Integration of Activity Data (UCIAD)

UCIAD wrote:

"Deciding on licensing and data distribution is always challenges where talking about data which are intrinsically personal: activity data. Privacy issues are of course relevant here. We cannot distribute openly, or even on proprietary basis, data that relate to users' actions and personal data on our systems. Anonymisation approaches exist that are supposed to make users un- identifiable in the data. Such approaches however cannot be applied in UCIAD for two main reason:

Such anonymisation mechanisms are only guarantied in very closed, controlled environment. In particular, they assume that it is possible to completely characterise the dataset, and that integration with other datasets will not happen. These are two assumption that we can't apply on our data as it is always evolving (in ways that might make established parameters for anonymisation suddenly invalid) and they are meant to be integrated with other data.

The whole principle of the project is to distribute the data to the user it concerns, which means that the user is at the centre of the data. Anonymising data related to one user, while giving it back to this user makes of course not sense. More generally, anonymisation mechanisms are based on aggregating data into abstracted or averaged values so that individual users disappear. This is obviously in contradiction with the approach taken in UCIAD.

The issue with licensing data in UCIAD is actually even more complicated: what licence to apply to data exported for a particular user? The ownership of the data is not even clear in this case. It is data collected and delivered by our systems, but that are produced out of the activities of the user. We believe that in this case, a particular type of license, that give control to the user on the distribution of their own data, but without opening it completely, is needed. This is an area that we will need to put additional work on, with possibly useful results coming out of the mydata project.

Of course, despite this very complicated issue, more generic components of UCIAD can be openly distributed. These include the UCIAD ontologies, as well as the source code of the UCIAD platform, manipulating data according to these ontologies."

Publishing (anonymisation and sharing)

Anonymisation

Several projects have published data. Due to constraints imposed by the obligation of Universities to keep much student data confidential, and the legal requirements of the Data Protection Act (1998) data needs to be anonymised before publication, eg by encrypting or removing User Identifiers and/or IP addresses that could be used to identify an individual or by publishing summary or statistical information from which it is not possible to derive individual information. See the guide on anonymisation

Outside of the UK, there have been some concerns over the re- identification of users from anonymised data. The Information and Privacy Commissioner of Ontario, has produced a paper entitled *Dispelling the Myths Surrounding De- identification: Anonymization Remains a Strong Tool for Protecting Privacy*. Its introduction states:

"The goal of this paper is to dispel this myth - the fear of re- identification is greatly overblown. As long as proper de-identification techniques, combined with re-identification risk measurement procedures, are used, de-identification remains a crucial tool in the protection of privacy. De-identification of personal data may be employed in a manner that simultaneously minimizes the risk of re- identification, while maintaining a high level of data quality. De- identification continues to be a valuable and effective mechanism for protecting personal information, and we urge its ongoing use."

Sharing

There are many reasons for sharing data, but the primary one at the moment seems to be that it enables other people to explore the data and come up with other ways in which it can be used to support institutional work.

What projects have done

All the projects have looked at sharing their data, and some have decided that is possible to do so in some form.

- AEIOU are currently only sharing data with partners, but have looked at anonymisation and methods of sharing.

- AGtivity - At present only sites that agree to receiving their data have been analysed. For more public documents specific names have been removed within for example the case studies.
- EVAD - have published the data in anonymised form.
- LIDP - are intending to publish statistical data only.
- OpenURL Activity data - have anonymised and are sharing the data collected during the project.
- RISE - anonymise the data and publish it using a schema based on that produced by Mosaic.
- SALT - have not yet decided how to publish their data.
- STAR-Trak data is for internal use only, removing the need to consider anonymisation.

Data Sharing Advice from the Information Commissioner's Office

The Information Commissioner's Office publishes a UK code of practice on data sharing, *Data sharing code of practice*.

A few quotes provide a flavour of the document's contents:

> "As I said in launching the public consultation on the draft of this code, under the right circumstances and for the right reasons, data sharing across and between organisations can play a crucial role in providing a better, more efficient service to customers in a range of sectors - both public and private. But citizens' and consumers' rights under the Data Protection Act must be respected."
>
> "Organisations that don't understand what can and cannot be done legally are as likely to disadvantage their clients through excessive caution as they are by carelessness."
>
> "the code isn't really about 'sharing' in the plain English sense. It's more about different types of disclosure, often involving many organisations and very complex information chains; chains that grow ever longer, crossing organisational and even national boundaries."

The code covers activities such as a school providing information about pupils to a research organisation.

> "By 'data sharing' we mean the disclosure of data from one or more organisations to a third party organisation or organisations, or the sharing of data between different parts of an organisation. Data sharing can take the form of:
>
> a reciprocal exchange of data;
>
> - one or more organisations providing data to a third party or parties;
> - several organisations pooling information and making it available to each other;
> - several organisations pooling information and making it available to a third party or parties;
> - different parts of the same organisation making data available to each other."
>
> "When we talk about 'data sharing' most people will understand this as sharing data between organisations. However, the data protection principles also apply to the sharing of information within an organisation - for example between the different departments of a local authority or financial services company."
>
> "When deciding whether to enter into an arrangement to share personal data (either as a provider, a recipient or both) you need to identify the objective that it is meant to achieve. You should consider the potential benefits and risks, either to individuals or society, of sharing the data. You should also assess the likely results of not sharing the data. You should ask yourself:
>
> - What is the sharing meant to achieve? ...
> - What information needs to be shared? ...
> - Who requires access to the shared personal data? ...

- When should it be shared? ...
- How should it be shared? ...
- How can we check the sharing is achieving its objectives? ...
- What risk does the data sharing pose? ...
- Could the objective be achieved without sharing the data or by anonymising it ? …
- Do I need to update my notification? …
- Will any of the data be transferred outside of the European Economic Area (EEA)? …"

> "Whilst consent will provide a basis on which organisations can share personal data, the ICO recognises that it is not always achievable or even desirable. If you are going to rely on consent as your condition you must be sure that individuals know precisely what data sharing they are consenting to and understand its implications for them. They must also have genuine control over whether or not the data sharing takes place."

The report goes on to discuss where consent is most appropriate and what other conditions allow sharing (p14- 15), with some examples of what is permissible.

> "The general rule in the DPA is that individuals should, at least, be aware that personal data about them has been, or is going to be, shared - even if their consent for the sharing is not needed. However, in certain limited circumstances the DPA provides for personal data, even sensitive data, to be shared without the individual even knowing about it."
>
> "The Data Protection Act (DPA) requires organisations to have appropriate technical and organisational measures in place when sharing personal data."
>
> "It is good practice to have a data sharing agreement in place, and to review it regularly, particularly where information is to be shared on a large scale, or on a regular basis." What is to be covered by such an agreement is outlined on p25.

Things to avoid doing appear on p34. Paraphrased, these are:

- Misleading individuals about whether you intend to share their information.
- Sharing excessive or irrelevant information about people.
- Sharing personal data when there is no need to do so.
- Not taking reasonable steps to ensure that information is accurate and up to date before you share it.
- Using incompatible information systems to share personal data, resulting in the loss, corruption or degradation of the data.
- Having inappropriate security measures in place.

Other useful sections are:

- Section 14 on data sharing agreements p41-43.
- Section 15 provides a data sharing checklist p46.
- A case study on p55 covers the use of research using data from other organisations.

Activity data to Enhance and Increase Open-access Usage (AEIOU)

AEIOU will make data available to individual partners and hope to provide a reporting service (based on the activity data) so that institutions can view usage statistics in a National context. We also hope to publicly release the data with regard to personal data encryption and licensing outlined below. Ideally, we would like to release OpenURL Context Object data as XML but in the short term this will be made available in CSV format.

The JISC Usage Statistics Review looked at European legal constraints for recording and aggregating log files and noted that the processing of IP- addresses is strongly regulated in certain countries (e.g. Germany) and that current interpretation maybe ambiguous. In such cases, they advise that "To avoid legal problems it would be best to pseudonymize IP-addresses shortly after the usage event or not to use IP-addresses at all but to promote the implementation of some sort of session identification, which does not record IP-addresses"

Currently, we are encrypting the IP addresses using an MD5 hash algorithm recommended in Knowledge Exchange Usage Statistics Guidelines so that personal data is anonymised. Although MD5 is a relatively fast and efficient algorithm it has been shown to have security vulnerabilities and other more secure methods for encryption (e.g. SHA-1 & SHA-2) are recommended. If this becomes an issue we could release data with stronger encryption or replace the IP address with a system identifier as suggested above. Removing the IP address would, however, compromise the ability to aggregate data.

The Knowledge Exchange Usage Statistics Guidelines also point out that when the IP address is obfuscated, information about the geographic location is lost. They therefore recommend using the C-Class subnet part of the IP address which will give a regional (network) location but can not identify a personal computer. This would be appropriate where activity data is used for reporting statistics.

Exploiting Access Grid Activity Data (AGtivity)

There is an issue with privacy of the data sets.

AGtivity originally aimed for all the data to be licensed for free use by the community.

- And the grouped statistics are being published openly that sum access times for example, and now available in reports from Janet
- Individual statistics although being captured with permission, are only being sent to those with node ownership. It has been found that with data mining evaluations can be carried out that could judge people

The aim is to convince people that the gains from the hypothesis questions, are worth a small risk of lack of privacy and security.

The collection of global stats reporting required by Janet are now under review and it is hoped that these reports form use-cases will inform the process and insist on a finer degree of reporting. The new SLAs for a combined video conferencing support service to be held at the University of Edinburgh is under negotiations to be confirmed in August 2011.

Exposing VLE activity data (EVAD)

The EVAD project were helped in this process by David Evans from Cambridge Computer Lab (http://www.cl.cam.ac.uk/~de239/) who is an expert in data privacy issues. They also considered the Privacy Policy for the site and consulted with our Legal Services team on releasing anonymous logging data under an Open Licence.

There are 3 files that they will be releasing:

Sakai Event File:

- Individual events such as a site visit which occur within a Session (held as a separate file for each Academic Year)
- Sakai Session File - Session details which include browser information and session times
- Sakai Site File - Details of the VLE Site being accessed

A tarball file can be obtained from here - however this is a 4GB file that will expand to over 12GB when uncompressed.

Our first step was to provide David Evans with the database schema for these files so he could consider, based on discussions with us, which fields might contain sensitive data. We then discussed the implications of making these fields public and what information they might expose. This was of course a

balancing act between releasing data from which others could draw value and not giving away details of what a known individual might have used the VLE system for.

We decided on a first data release which would use a cautious approach to data privacy. Should other institutions find this data of interest, we can work with them to reveal more information in the area of interest to them in a manner that does not compromise individuals.

This cautious approach meant hiding any data that would identify an individual user, site names and anything that might link a session back to a particular user. We settled on a hashing algorithm to use to obscure any such items of data yielding a string that can be determined uniquely from the value; we also used a salt to prevent inversion of the hash through exhaustive search of short inputs.

At this stage, we also looked at some sample data to reinforce our decisions.

The decision on what to hash was straightforward in many cases such as concealing any field with Site, User Name, URL or Content in it. Some things were less clear cut. For instance, the skin around a site could be used to identify a Department. The Session Ids that we looked at appeared to be numeric and we decided there was little risk in leaving this in its raw state. However, later testing revealed that, in some instances and points of time, this has included a user identifier so we agreed to hash this. It is worth remembering that the hashing algorithm is consistent so even though the value of the Session Id has been changed, it can still be used to link the Event and Session tables.

The Session Server holds both the Host Name and software module Id. We decided to hash the Host Name, in case this might reveal something about the network's internal structure, but leave the numeric part (software module id) alone as it reveals nothing other than which particular instance of software processed that session. We discovered that the format of this field had changed over time so we needed a mildly complex function to detect and extract from both formats of the Session Server.

The Session User Agent may provide the Browser Name and Version and the Operating System and Version of the computer used for the session. However this is a free-form field which may be changed by the user. There was a danger that this could identify a given user. A visual inspection showed at least one College Name, some company names and some school names within this field which could present a risk. Ideally we would extract and expose data such as the Browser Name but as this is a free-form field this in non-trivial. We therefore took the decision to hash this field.

As a final sanity check, we revisited some sample data from each of the tables once they had been hashed to satisfy ourselves that there was no raw data left that might possibly contain a user identifier.

Library Impact Data Project (LIDP)

LIDP describes its anonymisation process as:

> "One of the big issues for the project so far has been to ensure we are abiding to legal regulations and restrictions. The data we intend to utilise for our hypothesis is sensitive on a number of levels, and we have made efforts to ensure there is full anonymisation of both students and universities (should our collaborators choose to remain so). We contacted JISC Legal prior to data collection to confirm our procedures are appropriate, and additionally liaised with our Records Manager and the University's legal advisor.
>
> "Our data involves tying up student degree results with their borrowing history (i.e. the number of books borrowed), the number of times they entered the library building, and the number of times they logged into electronic resources. In retrieving data we have ensured that any identifying information is excluded before it is handled for analysis. We have also excluded any small courses to prevent identification of individuals eg where a course has less than 35 students and/or fewer than 5 of a specific degree level.
>
> "To notify library and resource users of our data collection, we referred to another data project, Using OpenURL Activity Data project"

The data has been published and is available under the Open Data Commons Attribution License from http://eprints.hud.ac.uk/11543/ which has been completely anonymised as follows:

The data contains final grade and library usage figures for 33,074 students studying undergraduate degrees at UK universities.

Background

Each of the 8 project partners provided a set of data, based on the initial data requirements document. Not all partners were able to provide data for e- resource logins and library visits, but all were able to provide library loans data.

In order to ensure anonymity:

- the 8 partners are not named in the data release, instead they have been allocated a randomly selected name (from LIB1 to LIB8)
- the names of schools and/or departments at each institution have been replaced with a randomly generated ID
- the year of graduation has been removed from the data
- where a course had less than 30 students, the course name has been replaced with a randomly generated ID
- some course names have been "generalised" in order to remove elements that may identify the institution

Grades

The awarded degree has been mapped to the following code:

- **A** = first (1)
- **B** = upper second (2:1)
- **C** = lower second (2:2)
- **D** = third (3)
- **E** = pass without honours

Library Usage

Where supplied by the project partner, the following library usage data measures are included:

- **ISSUES** = total number of items borrowed from the library by that student (nb this may include renewals)
- **ERES** = a measure of e-resource/database usage, e.g. total number of logins to MetaLib or Athens by that student
- **VISITS** = total number of times that student visited the library

Other Notes

- each graduate has been allocated an randomly generated unique ID
- where the course/school/department name was not supplied, it has been replaced with N/A
- where the measure of library usage was not supplied by the partner, the value is blank/empty

Using OpenURL Activity Data

The Using OpenURL Activity Data Project wrote of anonymisation:

> "The main issue to address relates to data protection; if the activity data collected include anything that may be construed as personal information, processing of those data must comply with Data Protection legislation. Although OpenURL Router and Resolver activity data tend not to include the name, ID or email address of an individual, they usually include IP addresses. In fact those IP addresses are important to this project as data in an aggregation are often useful only if they give some indication of the activity of an individual user within a discrete session (i.e. if such activity may be inferred from the data). The IP address links the user to the activity.
>
> "The Information Commissioner advises that IP addresses may constitute personal data in terms of the Data Protection legislation if an individual user can be identified by using a combination of that

IP address and other information which is either (i) held by the organisation collecting the data; or (ii) publicly accessible. This applies even when personal data are anonymised after collection as they are, nevertheless, personal data when collected. Clearly, IP addresses do not always constitute personal information as they identify a computer not an individual and the IP address may be dynamic rather than static; the static IP address of a computer used by a single individual, when combined with other information about that person, would be personal data while the IP address of a computer in an Internet café would not. However, the Information Commissioner advises in *Personal Information Online: Code of Practice*, 'When Using OpenURL Activity Data: Legal and Policy Issues - Initial Investigation you cannot tell whether you are collecting information about a particular person, it is good practice to treat all the information collected as though it were personal data …'

"We have explored the steps that must be taken in order to collect the data required for the aggregation without exposing the collecting organisation to legal risk through breach of data protection legislation. We sought to review the status of IP addresses, whether their collection and processing must be disclosed and how such disclosure may be made.

"We sought to further explore the legal and policy issues through dialogue with the university's legal advisor and staff within the University's Records Management Department who have expertise in data protection. We also included questions in our questionnaire (see below) to elicit the respondents' understanding of the privacy issues related to the activity data collected in their institutions.

"Our advisors confirmed that IP addresses may constitute personal data if the organisation holds or can easily obtain information that links these addresses to an individual. Online information service providers are likely to have information about individuals that may be linked with IP addresses. An email, for example, indicates the email address of the sender, and often includes the IP address in the Internet header information. So, if a service provider receives an email from an individual (in whatever context) and that individual also uses a service being provided by that same service provider, which is collecting IP addresses, the latter may be deemed personal data, the processing of which must be consented to by the data subject. The implications of this are far reaching. It suggests that online information service providers can hardly avoid collecting personal data as IP addresses and other personal information are routinely communicated through use of the web and email. Clearly, in most instances, there is no intent to 'process' data (in the ordinary sense of the word). However, as the definition of 'processing' in the Data Protection legislation is wide enough to include deletion, this is difficult to avoid. The law has been overtaken by technology. Our aim is to stay within the spirit of the law by protecting the privacy of individuals whilst operating an online service and thus minimising the risk to the service provider."

And the project decided that before being made available in any form, the data are anonymised to remove data that may identify an individual institution or individual person. Then the data file is made available 'as is'. We identify this as Level 1. It includes resolver redirect requests and those "lookup" requests where no institution is identified. It excludes the button requests as these identify an institution. It may be used by anybody for any purpose that they believe will be useful, such as for analysis or to create services for UK Higher and Further Education. The Level 2 data file contains data that have been further processed, i.e. all extraneous data are removed leaving only redirection data. EDINA uses these data as the basis of a prototype recommender service for UK HE and makes them available for others to use.

We distinguish three levels of data as follows:

Level	What's this?	What has been processed?	Is it available?
0	Original log file Data	No processing (contains identifiable IP addresses and institutions)	No
1	Anonymised Data	IP addresses are encrypted using an algorithm and institutional identifiers are anonymised	Coming soon
2	Anonymised Redirect Data	A subset of the Level 1 data, containing only entries that redirect to a resolver	Yes

The data are made available under the Open Data Commons (ODC) Public Domain Dedication and Licence and the ODC Attribution Sharealike Community Norms. Please read the terms of the PDDL and the Attribution Sharealike Community Norms before using the data.

The sample files are made available in two forms for initial analysis. The first is tab-delimited csv format, the format in which the data files are made available. The second is formatted to be displayed in spreadsheets, such as MS Excel, so that an understanding can be gained of the type of data available, and the full data (in csv format only).

- Sample data
- Full data

Recommendations Improve the Search Experience (RISE)

One of the aspirations of the RISE project is to be able to release the data in our recommendations database openly. So we've been thinking recently about how we might go about that. A critical step will be for us to anonymise the data robustly before we make any data openly available and we will post about those steps at a later date.

RISE took the LIDP-like approach to anonymisation:

- Remove user names.
- Remove all records for courses with less than x students.
- Replace the course code with a generic subject

Once we have a suitably anonymised dataset our current thinking is to make it available in two ways:

- as an XML file; and,
- as a prepopulated MySQL database.

The idea is that for people who are already working with activity data then an XML file is most likely to be of use to them. For people who haven't been using activity data and want to start using the code that we are going to be releasing for RISE then providing a base level of data may be a useful starting point for them. We 'd be interested in thoughts from people working with this type of data about what formats and structures would be most useful.

XML format
For the XML format we've taken as a starting point the work done by Mark van Harmelen for the MOSAIC project and were fortunately able to talk to him about the format when he visited to do the Synthesis project 'Recipes' work. We've kept as close to that original format as possible but there are some totally new elements that we are dealing with such as search terms that we need to include. The output in this format makes the assumption that re-users of this data will be able to make their own subject, relationship and search recommendations by using the user/resource/search term relationship within the XML data.

Proposed RISE record XML format

Start

<useRecordCollection>

<useRecord>

Basic data: Institution, year and dates

<from>

<institution>Open University

</institution>

<academicYear>2010/2011

</academicYear>

<extractedOn>

<year>2011</year>

<month>4</month>

<day>19</day>

</extractedOn>

<source>OURISE

</source>

</from>

Resource data

<resource>

<media>Article

</media>

<globalID type="DOI">10.1007/s00521-009-0254-2

</globaLID>

or

<globalID type="ISSN">09410643

</globaLID>

or

<globalID type="EDSN">12345678 [Ebsco Accession number]</globaLID>

<author>Cyr, Andre

</author>

<title>AI-SIMCOG: a simulator for spiking neurons and multiple animats' behaviours

</title>

<resourceURL>http://www.???.??/etc

</resourceURL>

<journalTitle>Nature

</journalTitle>

<published>

<year>2009</year>

</published>

<journalist>

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

</journalData>

</resource>

User context data

<context>

<user> anonymised UserID

</user>

<sequenceNumber>1 [Note: sequence number already stored within database]

</sequenceNumber>

</useDate>

For students: [propose to map to a subject]

<courseCode type="subject">Engineering

</courseCode>

<progression>UG2 [F, UG1, UG2, UG3, UG4, M, PhD1, PhD2, PhD3+ (F is for foundation year)]

</progression>

For staff

<progression>Staff

</progression>

</context>

Retrieved from

<retrievedFrom>

<searchTerm>artificial intelligence

</searchTerm>

</retrievedFrom>

End record, more records

</useRecord>

<!– more useRecords here if need be –>

Surfacing the Academic Long Tail (SALT)

SALT wrote the following about anonymisation:

> "The first step was to anonymise the data. This was partly done by removing all details about an individual for each loan transaction apart from a single user ID field which provides a system generated ID unique to that system. Following discussion with colleagues on the project it was then agreed that student course details would also be removed to eliminate the small risk that an individual could be identified in this way."

An open licensing statement from JRUL which means the data can be made available for reuse (They have yet to determine how to make this happen, given the size of the dataset).

MOSAIC

The MOSAIC Project, the precursor to the Activity Data Programme, took a simple approach to activity data anonymisation in circulation data.

For any circulation data, resources which were loaned only once or twice had their records immediately removed from the circulation data. The project referred to this as removing singletons and doubletons.

Beyond this the project posited several levels of detail in its specification of activity data formats. Different levels could be used for different purposes, and were numbered numerically, 0 to 2. The 'lower' levels contained less detail than the upper levels, and could only be used for simple recommendations. The

highest level, level 2, could be used for complex recommendations like "users who borrowed this resource went on to borrow ...".

User identity was simply removed from the lower levels. User identity was removed, but replaced with an identifying but anonymous GUID for the highest level data.

See Appendix 3 in the MOSAIC Final Report Appendices.

Sharing data

Several of the projects have made data available. The table below shows what they have done.

XML has been used as a data interchange technology in different projects.

The MOSAIC Project published a comprehensive guide to sharing resource circulation data for various purposes. See Appendix 3 in the MOSAIC Final Report Appendices. Three levels of activity data description are specified there, characterised by the amount of detail in each level.

- The MOSAIC approach was reused in the RISE Project with some project- specific modifications as discussed.

For more general guidance see a summary of guidance from the Information Commissioner's Office.

The following projects also shared their data some or all of their data

- SALT
- OpenURL Activity Data

Collecting, processing and presenting activity data

In order to work with activity data it is necessary to collect the data, it then needs to be processed and this will depend on the types of data being collected and the purposes for which it is being collected. Finally the data is likely to have to be visualised in some way in order to help users to understand the large volumes of data. We look at each of these in turn.

- Collecting activity data Here there are issues to do with the data needed for analysis tasks, its quality and completeness, its formats, and any specialist tools are required to collect it.
- Processing activity data to obtain useful results Two simple topics under this heading are aggregation and filtering of data. A more complex areas is the construction of recommendation algorithms, both to recommend resources and to identify students at risk. There is some other 'catch all' analysis that might be undertaken.
- Presenting the results of activity data use Primarily questions of visualisation of activity data for diverse purposes, including early exploration. There are also questions of optimal user interfaces to recommender systems.

Collecting

When collecting data there are several issues that you will have to consider in order to ensure that the information that you are collecting is appropriate for the purposes that you are collecting it for. These include what data you should be collecting and its quality. For instance, the data may be coming from a variety of different sources in which case you may need to match users up between systems. This may be easy as there is a common user id in the systems, or require some mapping exercise if not. You will also need to check that the systems are actually logging the types of event that you are interested in. When the logs were turned on they may not have been set to collect data that you need. You will also need to consider the format that you collect the data in, which will depend on how you want to process the data. Each of these is discussed:

- The data to collect.
- Data quality and completeness.
- Formats for collection.
- Data collections tools that can sometimes be useful.

We have also produced a guide and a number of recipes on collecting and extracting data.

Guides

- Strategies for collecting and storing activity data

Recipes

- Extraction of tracking data from DSpace
- Extract authentication counts from the EZproxy Authentication and Access Software
- Extract circulation data from the Horizon Library Management System
- Extract entry statistics from the Sentry library gate entry system
- Extract trace data from log files
- Extract user ID and course code from CIRCE MI
- Extract anonymised loan data from Talis Alto

The data to collect

The type of activity data that needs to be collected depends on what the data is to be used for. Often however, there are limitations imposed by proprietary systems that do not have APIs or published database formats to enable data collection.

In the Activity Data Programme, the major uses for activity data fell into:

- recommendations for library resources and other resources
- learning analytics to provide recommendations as to students at risk, so as to improve their results at university

The following table summarises the data collection strategies employed:

Project	purpose	source	data description
AEIOU	Recommendation system for research repositories	DSpace repositories	Viewing and download information
AGtivity	Improve the videoconferencing service	Log data from videoconferencing bridges	Using AG Toolkit and IACOM bridges
EVAD	Management information and to identify "at risk" students	Sakai	All the events from the Sakai event table
LIDP	Demonstrate correlation between library usage and student success	Library turnstile system Library management system EZProxy service Student record system	Library and student records data See collecting the data for a detailed description of the data and how it was collected at De Montfort University and the University of Lincoln.
RISE	Recommender systems to enhance the student experience	EZProxy service student record system Ebsco Discovery Solution SFX	Library and student records data
STAR-Trak	Manage interventions with students at risk	CMIS - timetable; attendance Banner - personal; course; grades Honeywell - security gates User - notes Help Desk - Search; incidents Xstream Portal/VLE -	

		usage; assignment status /mark Symphony - search; loans; fines Resource Manager - AV loans; fines Gmail - email, apps Agresso - fees payment	
SALT	Support humanities research by surfacing underused "long tail" library materials through recommendations	Talis Alto	Circulation data
UCIAD	Integration of activity data spread across various systems in an organization, and exploring how this integration can both benefit users and improve transparency in an organization	Diverse web servers	Server logs
OpenURL	Publication of anonymised OpenURL data for third party us	OpenURL Router	OpenURL requests

General limitations on collected data:

- Data quality: Some of the data may be corrupted.
- Data completeness: Some of the data may be missing.
- In some cases data volume may be large (eg EVAD had very large log files generated by Sakai)

LIDP

We use LIDP as a case study in data collection. Library use data was collected from seven partners.

Specific requirements for the data collection follow.

For a specific academic year (eg 2009/10), extract details for each graduating student, e.g.

Item	**example data**
academic year of graduation	2009/10
course title	Software Development
length of course in years	3
type of course	post grad
grade achieved	2:1
school/academic department	School of Computing & Maths
number of items borrowed from library	
either the total number borrowed by that student -	50 items during the 3 years of the course
or separate values for each academic year -	11 items in 2007/8, 16 in 2008/9 and 23 in

number of visits to the library

either the total number of visits by that student

or separate values for each academic year

number of logins to e-resources (or some other measure of e-resource usage)

either the total number of logins made by that student

or separate values for each academic year

Notes

- As per the UK Data Protection Act, the extract shouldn't include information that identifies, or could be used to identify, a named individual (eg their name or a campus network ID).
- For privacy reasons, you may also wish to exclude data from courses that had small numbers of students (eg less than 5 students).
- Ideally, the data extract should be in a machine processable format such as Excel, XML or CSV.

University of Lincoln's collection procedures

Lincoln's data consists of fully anonymised entries for 4,268 students who graduated from the University of Lincoln with a named award at the end of the academic year 2009/10 - along with a selection of their library activity over three years (2007/08, 2008/09, 2009/10).

The library activity data represents:

- The number of library items (book loans etc.) issued to each student in each of the three years; taken from the *circ_tran* table within our SirsiDynix Horizon Library Management System (LMS). We also needed a copy of Horizon's borrower table to associate each transaction with an identifiable student.
- The number of times each student visited our main GCW University Library, using their student ID card to pass through the Library's access control gates in each of the three years; taken directly from the Sentry access control/turnstile system. This data applies *only* to the main GCW University Library: there is no access control at the University of Lincoln's other four campus libraries, so many students have '0′ for this data.
- The number of times each student was authenticated against an electronic resource via AthensDA; taken from Portal server access logs. Although by no means all e-resources are accessed or retrieved via an Athens authentication, it serves as a proxy for e-resource usage more generally.

Data on PC and network logins was also desirable, but was generally not available.

Student data, including the 'primary key' of the student account ID, was supplied by the Registry department from the University's QLS student records management system.

Once the data above was collected it was merged into a large CSV file by database manipulation. Producing CSV can also be done using Excel (see here), but there are limitations to data set sizes that can be processed using Excel.

De Montfort University collection procedures

What was included in DMU's data?

We have sent in some data for our part of the LIDP. The data was collated from different sources: along with the basic student performance was network logins; Athens usage; loans from our Talis Library Management System and visits to the Kimberlin library from the Access Control System.

Network logins

Students can login to the library network from each of the libraries (Kimberlin, Charles Frears, Law and the Eric Wood Learning Zone, and also from Student Services in the Gateway Building). However, someone could use their own laptop and wifi in the library without appearing in these figures.

Athens usage

There are over 100 resources to which students have access via their Athens account. The figures would underestimate how much use is made of electronic resources as some require other usernames/passwords and others, available through the UK Access Management Federation, ask for Athens details, but these transactions are not included in these statistics. Use of some major resources for Law, notably Westlaw, might be significantly under counted here.

Loans count

The loans count covers borrowing from all of the libraries. Items borrowed include books as well as videos, slide sets and other loanable material. Although we do have a postal loans service for distance learners, it is a pretty good bet that most borrowing events took place during a visit to a library.

Access Control data

Visitors to the Kimberlin library have to present their student card or other library ID to gain access. This is also true for the Learning Zone in the Eric Wood building, but not for the Law and Charles Frears libraries. Library use counts for law and health students would be significantly under counted in this area.

There are no other services based in the library building. Well, there is a cafe and the Learning Zone acts as a social space for students wanting to meet each other, but everyone coming through the door can be seen as wanting to visit the library and not some other agency.

What else could we have included?

DMU has a Citrix service called 'the Library Applications Portal' which students need to login to. The Portal makes tools like Microsoft Word and SPSS available, so may be a more reliable indicator of academic activity than just logging onto the library network would be.

JACS codes.

JACS codes in used in Higher Education to standardise analysis of the subjects being studied. They enable a comparison of course related to subjects like 'Forensic and Archaeological Science' to take place, using the subject term, or just a code number like '027'. The National Student Survey (NSS) results include a breakdown by JACS code and would be a useful way of linking subject to degree result, via the Programme Code that appears in both spreadsheets.

How comprehensive is the data?

It is possible to imagine a law student regularly visiting the library in the Hugh Aston building, using the book stock (and perhaps more the printed journals) without borrowing any, relying on research available through HeinOnline and Westlaw and writing up on their own wifi enabled laptop. Such a person could easily gain a first class degree without much activity showing up in the data we are able to compile.

Data quality and completeness

The collection of activity data is not at this time a normal procedure, and thus institutions have, naturally enough, not concentrated on maintaining and assuring the quality of applicable data sources. While this currently presents some problems, we view this as a transient situation that will improve over time.

A general picture is portrayed by the EVAD Project at Cambridge:

> "First, we've had to get our data (first, catch your rabbit..), from when we started using CamTools (our current institutional VLE) to December 31 2010. This involved retrieving archived data, which didn't go as smoothly as we 'd hoped. We had to do some restoration of corrupted data, and we 're missing about two weeks of data as well. This just illustrates the problems of dealing with data that's collected but not looked at very often."

Other projects also ran into problems as discussed here

- AGtivity
- EVAD
- LIDP

Exploiting Access Grid Activity Data (AGtivity)

AGtivity reported some problems

- Data corruption due to human data entry errors: The data sets we received had numerous human errors and inconsistency. People at times re- entered names but often slightly differently. Association tags had to be created in order to correlate data entries and create meaningful statistics.
- Ontologies for cross-linking were not always used consistently or at all: For example virtual venue names were used for the same purpose and thus should have an ontology to describe their use as well as just their unique name.
- Automated recording and log data at the correct level of detail: Data had been logged for the last seven years, but is not complete and only over the last 10 months had detailed data logs been recorded. This data is more finely-grained than currently needed, leading to some problems.
- Merging data from other sources is not trivial: A key useful component in processing activity data is to cross- merge data sets. AGtivity used multiple sets which mostly correlated for merging purposes, but again human mistyping errors were an issue.
- A minor issue was to have unified time stamps - AGtivity spent time reconciling BST and GMT time stamps.

Exposing VLE data (EVAD)

EVAD reported on the problems involved in large data sets, and the data massaging they had to perform in order to obtain a data set from which the project could work:

> "A lot of our time so far has gone into marshalling our data and getting it into a workable form, properly indexed and so forth. This means that it should now be comparatively quick to process the data into chart form - no setting it going and leaving it overnight now!
>
> "The Sakai event table, which is most of our data, has given us 240 million rows of data for the last 5 years (a little over, actually). Only 70m of those turned out to be useful, as the rest were mostly things like the 'presence' event. We turned that event off in 2007 because it was causing unacceptable server load. Basically, our VLE is made up of a series of sites, and the presence event was a row written to the database every 5 seconds for every logged in user, saying which site they were currently looking at, so that an up-to- date list of who was in the site was displayed. So you can guess that this presence event generated an awful lot of data until we turned it off.
>
> "We also have 500M rows of Apache web data logs, telling us who visited which pages using which web browser and operating system. This is currently not something we're looking at so much (beyond a quick analysis of which web browsers we need to do our most thorough testing against), but it will be most useful when we're looking at which of our help resources have been most visited.
>
> "For our sets of data, we've been breaking it up by academic year (well, academic financial year - the year end is 31st July), and by week, so that we can see the fluctuation in usage. (We're starting to break it up by day as well, but this takes a long time to index.)"

Library Impact Data Project (LIDP)

LIDP which used library data from seven universities ran into problems at some universities. Liverpool John Moore's University (LJMU) reported:

> "Energised by the initial project team meeting, the LIDP team at LJMU set about gathering the required data in order to make our contribution to the project. Having already had a few discussions we were fairly confident that we would be able to gather the required data. We had

access to student records and numbers, Athens data, library usage data from Aleph and we were aware that our security system (gate entry into the library) kept a historic record of each individual's entry and exit into the buildings which are serviced through the swipe card entry system. We just needed to pull all this together through the unique student number.

"Getting this particular bit of data from our Academic Planning department was straightforward. An anonymised list of all 2010 graduating students, along with their programme of study and degree classification, and most importantly their unique id number was duly provided.

"**Hurdle one** - Our security system does indeed log and track student movement through the use of the swipe card entry system, but we are unable to get the system to report on this. All data from the system is archived by the system supplier and is subsequently not readily available to us. This means that entry into the Learning Resource Centres is not going to be something we can report upon on this occasion.

"**Hurdle two** - Our Network team systematically delete student network accounts upon graduation, which means the record that links an individual's unique student ID number, Athens account, security number and Library barcode is not available for the students whose library usage we wished to analyse!

"There were about 4,500 students who graduated from LJMU in 2010 with undergraduate degrees, but unfortunately, by the time I got to speak to the network manager, 3,000 of these had been deleted, as is our institutional practice and policy.

"The upshot of all this is that we are only going to be able to provide data for a third of the potential students that we could have provided data for if we had thought to ask these questions earlier on. But at least we are still able to contribute."

The project then reflected on the data more widely

Following on from De Montfort's blog post about the nature of their data submission, we've been thinking a bit more about what we could have included (and indeed what we might look at when we finish this project).

We've already been thinking about how we could incorporate well established surveys into data consideration (both our own internal data collection, such as our library satisfaction survey, and external surveys). While our biggest concern is getting enough data to draw conclusions, qualitative data is naturally a problematic area: numerical data 'just' needs obtaining and clearing for use, but getting some information from students to find out why they do or don't use resources and the library can be quite complicated. Using other surveys outside of the project focus groups could be a way of gathering simple yet informative data to indicate trends and personal preferences. Additionally, if certain groups of students choose to use the library a little or a lot, existing surveys may give us feedback on why on a basic level.

We also may want to ask (and admittedly I'm biased here given my research background!) what makes students choose the library for studying and just how productive they are when they get here. Footfall has already clearly demonstrated in the original project that library entries do not necessarily equate to degree results. Our library spaces have been designed for a variety of uses, for social learning, group study, individual study, specialist subject areas. However, that doesn't mean they are used for those purposes. Footfall can mean checking email and logging on to Facebook (which of course then links back to computer log in data and how that doesn't necessarily reflect studying), but it can also mean intensive group preparation eg law students working on a moot (perhaps without using computers or resources other than hard copy reference editions of law reports).

If we want to take the data even further, we could take it deeper into borrowing in terms of specific collection usage too. Other research (De Jager, K (2002) has found significant correlations between specific hard copy collections (in De Jager's case, examples include reference materials and short loan items) and attainment, with similar varying relationships between resource use and academic achievement across different subjects. If we were to break down collection type in our borrowing analysis (particularly where there may be special collections of materials or large numbers of shorter loan periods), would we find anything that would link up to electronic resource use as a comparison? We could also consider incorporating reading lists into the data to check whether recommended texts are used heavily in high attainment groups…

De Jager, K. (2002), “Successful students: does the library make a difference?” Performance Measurement and Metrics 3 (3), p.140- 144

Recipes

The project also produced two recipes on extracting the data needed for the project.

- Extract circulation data from the Horizon Library Management System
- Extract entry statistics from the Sentry library gate entry system

Data formats

The advantages of standard and common data formats are self-evident.

When gathering and, particularly, sharing activity data with others (possibly as open data) it is useful to be able to make use of standard and/or common data formats. We note the existence of the OpenURL Context Object and the formats advanced by the MOSAIC project. The choice of a representation will be dependent on the precise content of the data being collected.

Further information is provided by discussions of the issues relating to appropriate formats:

- JISC Usage Statistics Review: The JISC Usage Statistics Review Project is aimed at formulating a fundamental scheme for repository log files and related information.
- KE Usage Statistics Guidelines: Guidelines for the aggregation and exchange of Usage Data. Includes guidelines for the exchange of usage statistics from a repository to a central server using OAI-PMH and OpenURL Context Objects.

Some useful formats:

- MOSAIC Usage Data Formats: These provide three levels of detail in the production of usage data for recommenders.
- OpenURL Context Object: A simple specification to embed OpenURL references in HTML.
- PIRUS2: A format for usage reports at the individual article level that can be implemented by any entity (publisher, aggregator, repository, etc.).
- OA-Statistics: Uniform standards for the gathering, exchange, and analysis of usage data from Open- access documents.

The AEIOU Project provides an example of use of data formats.

AEIOU

AEIOU decided to use the OpenURL Context Object format. The project describes this as follows:

> "How best to represent the activity data we're gathering and passing around? Several projects (PIRUS2, OA- Statistics, SURE, NEEO) have already considered this and based their exchange of data (as XML) on the OpenURL Context Object - the standard was recommended in the JISC Usage Statistics Final Report . Knowledge Exchange have produced international guidelines for the aggregation and exchange of usage statistics (from a repository to a central server using OAI-PMH) in an attempt to harmonise any subtle differences.
>
> "Obviously then, OpenURL Context Objects are the way to go but how far can I bend the standard without breaking it? Should I encrypt the Requester IP address and do I really need to provide the C- class Subnet address and country code? If we have the IP addresses we can determine subnet and country code. Fortunately the recommendations from Knowledge Exchange realised this and don't require it.
>
> "So for the needs of this project where we're concerned with a closed system within a National context, I think I can bend the standard a little and not lose any information. I can use an authenticated service. I also want to include some metadata - the resource title and author maybe.
>
> "So here's the activity data mapped to a Context Object:

- Timestamp (Request time) Mandatory
- Referent identifier (The URL of the object file or the metadata record that is requested) Mandatory
- Referent other identifier (The URI of the object file or the metadata record that is requested) Mandatory if applicable
- Referring Entity (Referrer URL) Mandatory if applicable
- Requester Identifier (Request IP address - encrypted possibly!) Mandatory
- Service type (objectFile or descriptiveMetadata) Mandatory
- Resolver identifier (The baseURL of the repository) Mandatory"

Data collection tools

Collecting data may range in ease between being simple, eg copying a log file, to being rather complex, eg adding modules to repositories as performed by AEIOU. The latter is discussed here.

The PIRUS2 Project has produced a patch for DSpace (and EPrints) for capturing activity data and either making it available via OAI-PMH or pushing it to a tracker service. AEIOU patched DSpace code to make usage data available for harvesting. However, the project wanted to avoid harvesting via OAI- PMH so looked closer at the tracker code. This code provides a solution that uses Spring's dependency injection framework to create a listener on the DSpace Event service to capture downloads, with some programming this was extended to capture item views. This is all packaged into a single JAR file to modify a DSpace system's behaviour.

Processing

Once collected, activity data is processed for one of several purposes, of which we cover two important reasons:

- Recommending resources.
- Identifying students at risk in a learning analytics approach.

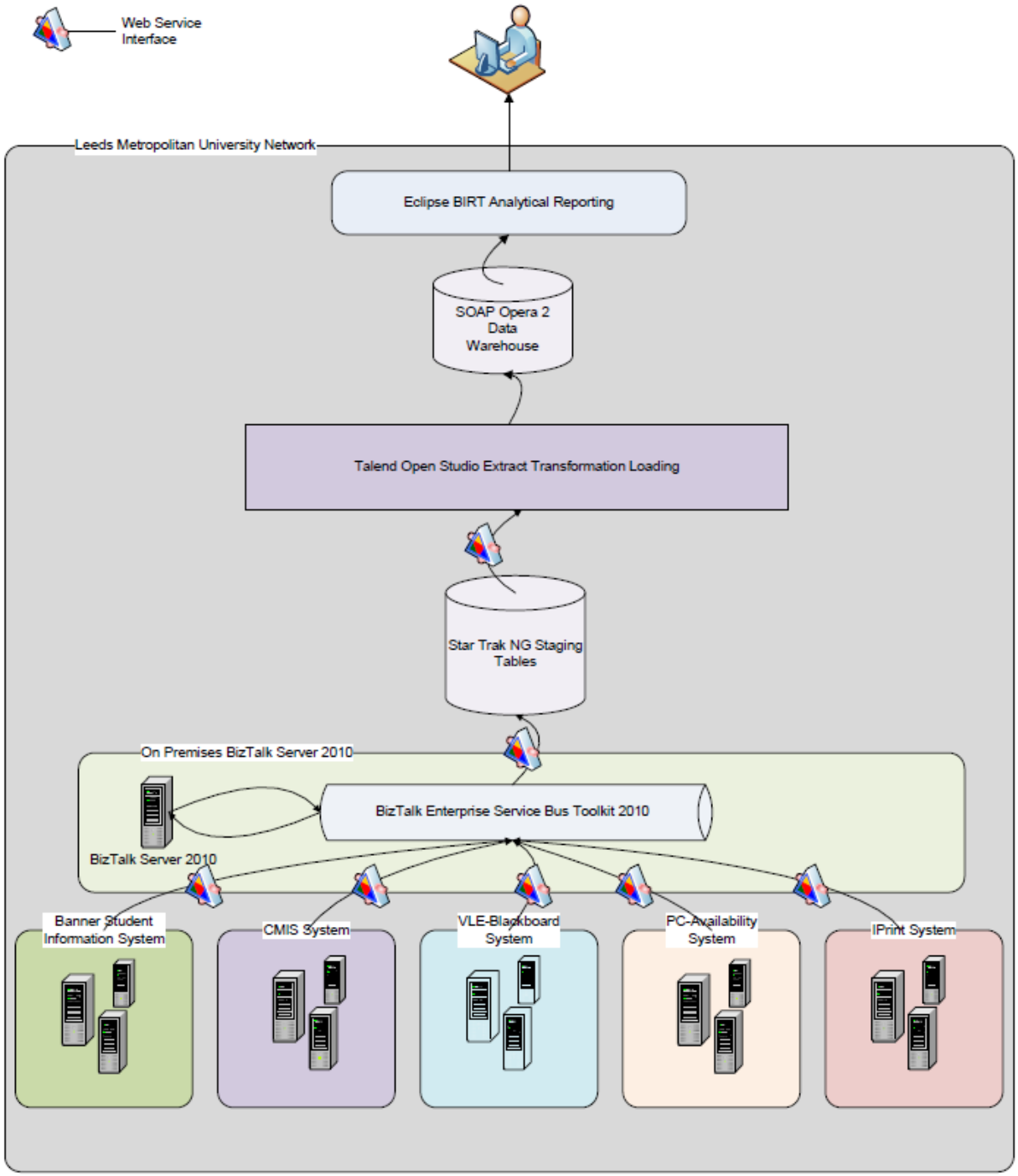
Processing is discussed under these headings:

Aggregation of multiple data sets

Aggregation of data from different sources to form useful records requires a common key between records in order to determine a relationship.

The STAR-Trak project looked in some detail at the data that it would extract from each source and produced the STAR-Trak Data Warehouse Schema which describes each source and the data that should be used. The diagram below shows the way in which the sources are aggregated using an enterprise service bus (and then presented to the user).

th



The RISE Project took the approach of using the log files to provide the base layer of data and then queried other systems to pull in further information to enhance this data. Through a combination of ISSNs, DOIs and other techniques they are able to add in data such as journal tiles and article titles.

The SALT Project aggregated their data as follows:

The process begins with data extracted from the Talis library management system (LMS) at JRUL in CSV format. This data is parsed by a PHP script, which separates the data into two tables in a MySQL database:

- face="Cambria"The bibliographic details describing an item go into a table called ‘items’.
- face="Cambria"The loan specific data, including borrower ID, goes into a table called ‘loans’.

A further PHP script then processes the data into two additional MySQL tables:

- face="Cambria"‘nloans’ contains the total number of times each item has been borrowed.
- face="Cambria"‘nborrowers’ contains, for each combination of two items, a count of the unique number of library users to have borrowed both items.

Additional processing is performed on demand by the web API:

- face="Cambria"When called for a given item, say item_1, the API returns a list of items for suggested reading, where this list is derived as follows.
- face="Cambria"From the ‘nborrowers’ table a list of items is compiled from all combinations featuring item_1.
- face="Cambria"For each item in this list the number of unique borrowers, from the ‘nborrowers’ table, is divided by the total number of loans for that item, from the ‘nloans’ table, following the logic used by Dave Pattern at the University of Huddersfield.
- face="Cambria"The resulting values are ranked in descending order and the details associated with each suggested item are returned by the API.

The LIDP project aggregated data from a wide variety of sources, varying with what each participating university was able to provide. This has already been discussed under collecting data for the LIDP project.

Recipes

- Stitching together library data with Excel

Filtering data sets

The process of creating recommendations is an example of selecting items from a set of items. As such it is an example of filtering; although often rather complex filtering, with behaviour like 'recommend the two items in geomorphology most borrowed by student X's class cohort.'

Typically recommenders apply some selection criteria to pick out which items to recommend, and then order the results, and display a certain number or all of the results. E.g.

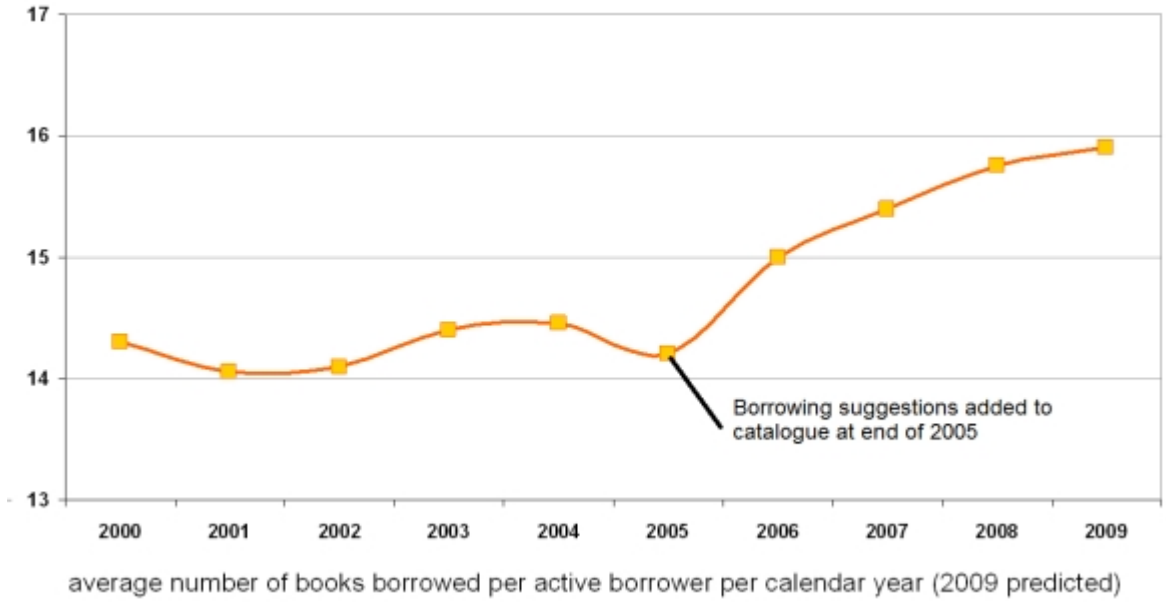
- As above, find all items on geomorphology borrowed by a student's class cohort, and order them according to the number of times borrowed. Then display just the top two.
- Or, in a system to discover students at risk, find all students whose activity matches the risk indicator, order the results in some way (alphabetically, or in order of perceived risk) and display all of them in some way. Raising an alert is much the same as displaying.

The algorithm to produce recommendations may be relatively easy to implement (for example the RISE Project produced three different recommenders during the time available for the Programme). Alternately, the algorithm may, as in the case of the Facebook Edge Rank Algorithm, be exceedingly complex.

Considerable variation in recommender algorithms is possible. For example the RISE Project used a third party system to perform some of it's processing by passing a parameter to the EBSCO API so that the results listed only those items where full text was available. The earlier MOSAIC Project performed its recommendations by a customisation of index table formats and the ranking algorithm in a Solr search engine.

Recommendation algorithms

Recommending resources is extremely powerful for users. For instance, it is essential to Amazon's business model where they use people who bought X also bought Y to encourage increased business. Similarly, Huddersfield University has seen an increase in library borrowing since they introduced a recommender system to the OPAC.



This clearly demonstrates the value that users gain from having recommendations.

The MOSAIC Project produced a recommender system before this Programme, details appear from page 49 onwards in the MOSAIC project's final report.

The LIDP project produced a recommender that also predated this programme, details appear here.

Different approaches to recommendation were used in three of the Programme's projects:

- AEIOU
- LIDP
- RISE

Activity data to Enhance and Increase Open-access Usage (AEIOU)

Initially The AEIOU project used SQL queries to identify items that had been viewed or downloaded by users within specific 'windows' or session times (10, 15 and 30 minutes). Refinements were made to rank the list of recommended items by ascending time and number of views and downloads. They have a requirement that the service should respond within 750 milliseconds, if not the client connection (from the repository) will timeout and no recommended items are displayed. The connection timeout is configured at the repository and is intended to avoid delays when viewing items.

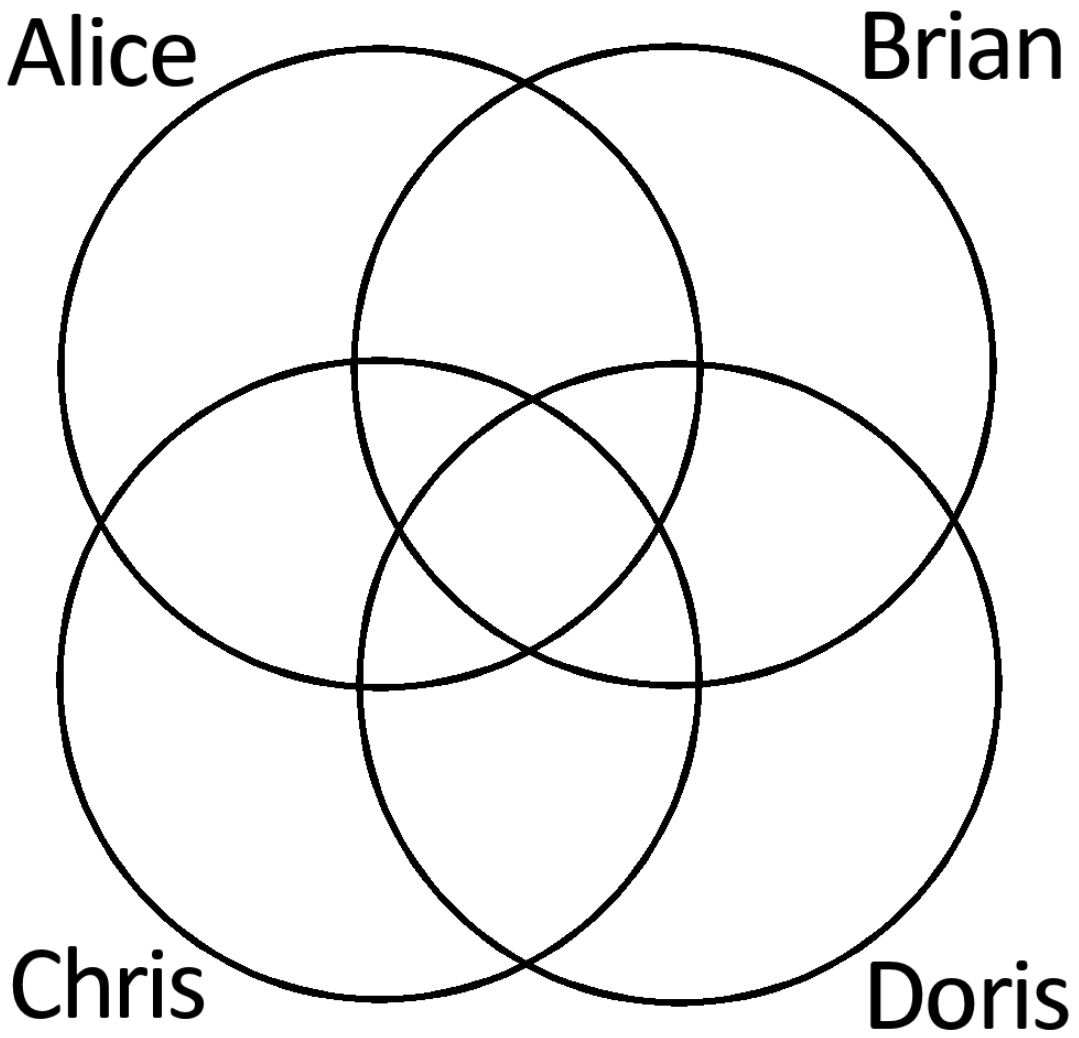
Unsurprisingly, queries took longer to run as the data set grew (over 150,000 events) and query time was noticeably influenced by the number of events per user (IP address). Filtering out IP addresses from robots, optimising the database and increasing the timeout to 2 seconds temporarily overcame this problem.

However, it was clear that this would not be scalable and that other algorithms for generating recommended items maybe required. A little research suggested that Apache Mahout Recommender / Collaborative filtering techniques were worth exploring. They are currently testing recommenders based on item preferences determined by whether or not an item has been viewed (boolean preference) or the total number of views per item. Item recommenders use similarities which require pre-processing using a variety of algorithms (including correlations). An API also exists for testing the relevance of the recommended items and we will be using this over the next few weeks to assess and refine the service.

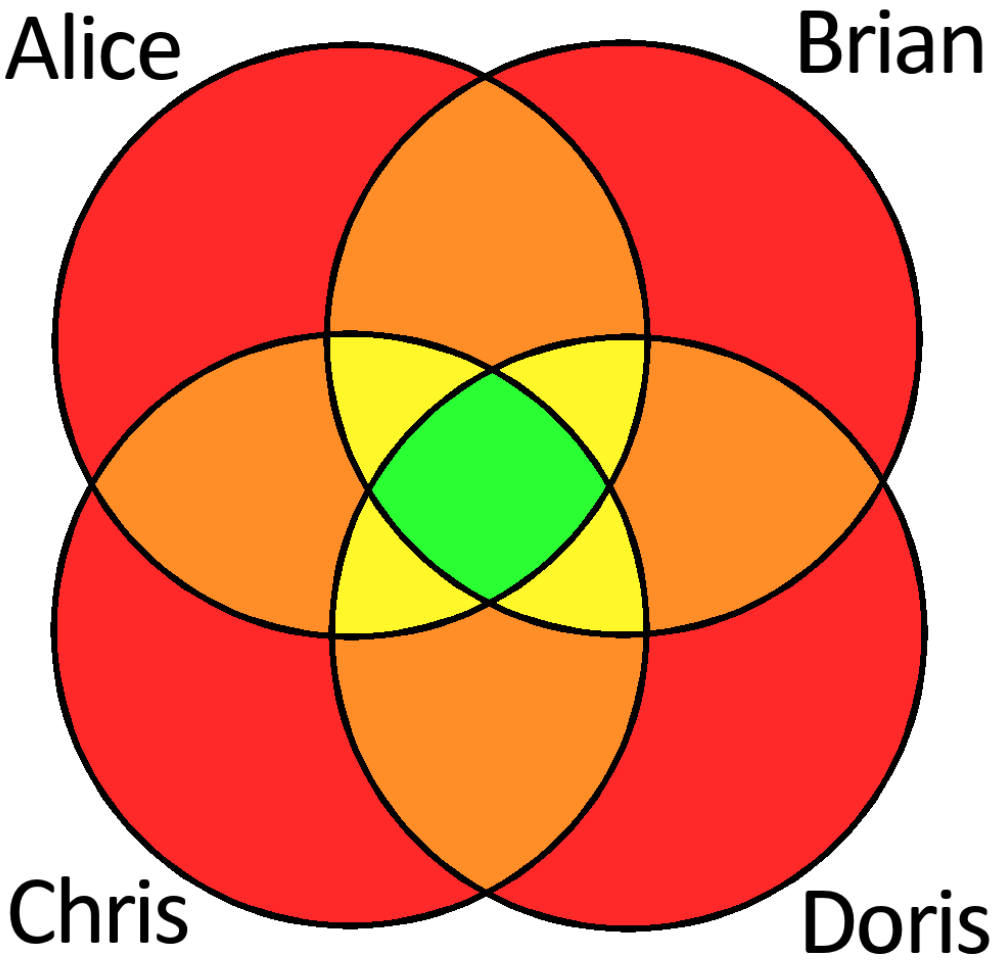
Library Impact Data Project (LIDP)

The Huddersfield team have also been developing recommender systems to generate the "people who borrowed this, also borrowed…" suggestions in our OPAC and whether or not there are privacy issues with generating them. As Dave Pattern wrote:

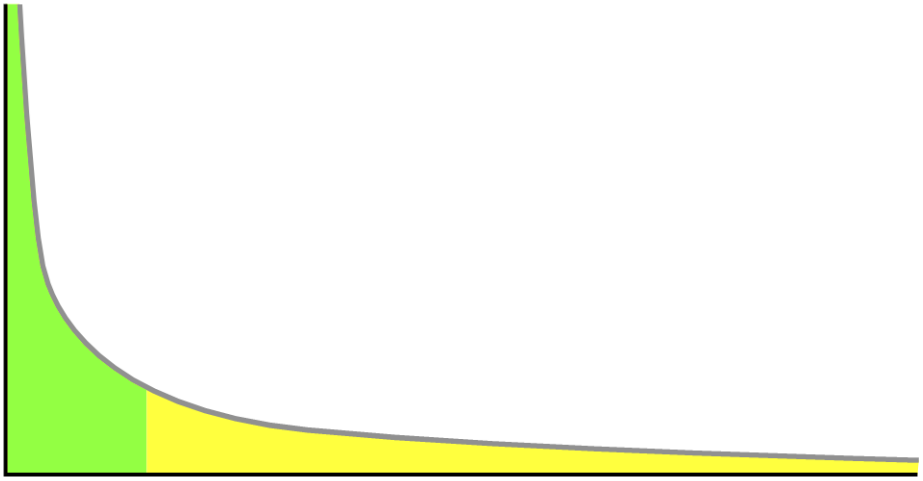
“To generate recommendations for book A, we find every person who's borrowed that book. Just to simply things, let's say only 4 people have borrowed that book. We then find every book that those 4 people have borrowed. As a Venn diagram, where each set represents the books borrowed by that person, it'd look like…



“To generate useful and relevant recommendations (and also to help protect privacy), we set a threshold and ignore anything below that. So, if we decide to set the threshold at 3 or more, we can ignore anything in the red and orange segments, and just concentrate on the yellow and green intersections…



“There'll always be at least one book in the green intersection — the book we're generating the recommendations for, so we can ignore that.



“If we sort the books that appear in those intersections by how many borrowers they have in common (in descending order), we should get a useful list of recommendations. For example, if we do this for "Social determinants of health (ISBN 9780198565895), we get the following titles (the figures in square brackets is the number of people who borrowed both books and the total number of loans for the suggested book)…

- Health promotion: foundations for practice [43 / 1312]
- The helping relationship: process and skills [41 / 248]
- Skilled interpersonal communication: research, theory and practice [31 / 438]
- Public health and health promotion: developing practice [29 / 317]
- sociology of health and illness [29 / 188]
- Promoting health: a practical guide [28 / 704]
- Sociology: themes and perspectives [28 / 612]
- Understanding social problems: issues in social policy [28 / 300]
- Psychology: the science of mind and behaviour [27 / 364]
- Health policy for health care professionals [25 / 375]

"When we trialled generating suggestions this way, we found a couple of issues:

- More often than not, the suggested books tend to be ones that are popular and circulate well already — is there a danger that this creates a closed loop, where more relevant but less popular don't get recommended?
- The suggested books are often more general — eg the suggestions for a book on MySQL might be ones that cover databases in general, rather than specifically just MySQL

"To try and address those concerns, we tweaked the sorting to take into account the total number of times the suggested book has been borrowed. So, if 10 people have borrowed book A and book B, and book B has only been borrowed by 12 people in total, we could imply that there's a strong link between both books.

"If we divide the number of common borrowers (10) with the total number of people who've borrowed the suggested book (12), we'll end up with a figure between 0 and 1 that we can use to sort the titles. Here's a list that uses 15 and above as the threshold…

- Status syndrome : how your social standing directly affects your health [15 / 33]
- Values for care practice [15 / 61]
- The study of social problems: seven perspectives [18 / 90]
- Essentials of human anatomy & physiology [15 / 81]
- Human health and disease [15 / 88]
- Social problems: an introduction to critical constructionism [15 / 90]
- Thinking about social problems: an introduction to constructionist perspectives [21 / 127]
- The helping relationship: process and skills [41 / 248]
- Health inequality: an introduction to theories, concepts and methods [20 / 122]
- The sociology of health and illness [29 / 188]

"…and if we used a lower threshold of 5, we'd get…

- Status syndrome : how your social standing directly affects your health [15 / 33]
- What is the real cost of more patient choice? [5 / 12]
- Interpersonal helping skills [5 / 12]
- Coaching and mentoring in higher education : a learning-centred approach [6 / 15]
- Understanding social policy [5 / 13]

- Managing and leading in inter-agency settings [11 / 29]
- Read, reflect, write : the elements of flexible reading, fluent writing, independent learning [5 / 14]
- Community psychology : in pursuit of liberation and well-being [6 / 20]
- Communication skills for health and social care [9 / 32]
- How effective have National Healthy School Standards and the National Healthy School programme been, in contributing to improvements in children's health? [5 / 18]

"If you think of the 3 sets of suggestions in terms of the Long Tail, the first set favours popular items that will mostly appear in the green ("head") section, the second will be further along the tail, and the third, even further along.

"As we move along the tail, we begin to favour books that haven't been borrowed as often and we also begin to see a few more eclectic suggestions appearing (eg the "How effective have National Healthy School Standards…" literature based study).

"One final factor that we include in our OPAC suggestions is whether or not the suggested book belongs to the same stock collection in the library — if it does, then the book gets a slight boost."

Recommendations Improve the Search Experience (RISE)

The RISE MyRecommendations search system is going to provide three types of recommendations:

1. Course-Based "People on your course(s) viewed"

This type of recommendation is designed to show the user what people studying their module have viewed recently. At the moment this only picks up the first module that a student is studying but we are planning a future enhancement that will include all the modules that are being studied with a feature to allow users to flag which module they are currently looking for resources for. The recommendations are generated by analysing the resources most viewed by people studying a particular module.

2. Relationship Based "These resources may be related to others you viewed recently"

These recommendations are generated for a resource-pair. For example, if users commonly visit resource B after viewing resource A, the system considers this to be a relationship, and will recommend resource B to people viewing resource A in the future. As the system doesn't host resources internally, it instead looks at a user's previously viewed resources (most recent), and then checks for the most often viewed resources by users who've also viewed the same (most recent) resources.

3. Search Based "People using similar search terms often viewed"

We have limited data on search terms used, from the EZProxy log files so we are using the searches carried out in MyRecommendations to build search recommendations. Using this we associate search terms used with the resources most often visited as a result of such a search. For example, if people searching for 'XYZ' most often visit the 50th result returned from Ebsco, this part of the recommendation algorithm will pick up on this. Hence in future when people search for 'XYZ', that particular result will appear top of the list of recommendations for users in a "People searching for similar search terms often viewed" section.

Providing Recommendations

The recommendation types outlined in previous posts are generated based on the logged-in user's credentials, and resources which may have relationships with said credentials.

Relationships are stored in tables within the database, an example of a relationship stored within one of these tables is as follows:

Field Name:	Value:
course_id	32
resource_id	54645
value	14

This example depicts a relationship between resource 54645 and course 32, having a value of 14. This means people on course 32 have given resource 54645 a value of 14. Values are assigned based on resource views and subsequent relevancy ratings (if available).

For example, a user following such a recommendation would increment the above relationship to 15 by simply following the link, indicating at least the resource title was somewhat interesting to the user.

If this user then chooses to rate, the following rating choices would result in the respective 'Value' after rating.

Rating Choice	Resulting Relationship Value	Change Logic, after +1 for resource visit
Very Useful	16	+1
Somewhat Useful	15	0
Not Useful	13	-2

From the table above, it is evident that the logic is weighted in favour of the resource being relevant (i.e. the value will always go UP by at least 1, unless 'not useful' is selected), this is based on the theory above that the resource link appearing useful enough to follow gives some indication that the resource is more likely to be relevant than irrelevant. Giving the 'Not Useful' rating a weight of - 2 means the least useful resources can still 'sink' and appear less frequently. This approach also ensures the actions of those users choosing not to provide feedback can still be utilized by the recommendations engine.

It is also important to note that the ratings are purely for the actual relevance to a particular user, and not on the quality of the resource itself. This ability to rate both relevance and quality may be implemented in the future.

There is a relationship table for each of the three recommendation types, all are controlled and manipulated by actions of users in the RISE system, and the access logs generated by other library systems.

Learning analytics

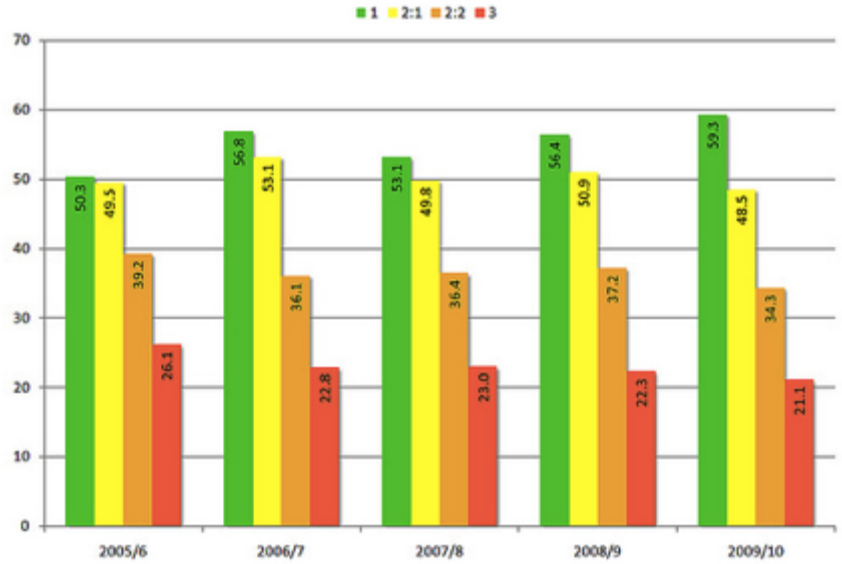
Learning analytics produce suggestions as to students who appear to be performing badly, and may benefit from a positive intervention designed to help them in their academic studies. Thus at first sight learning analytics may seem to involve the same kinds of recommendation engines as resource recommendation. However, the basic approach is somewhat divergent.

Basically the approach works by finding activity data that differentiates passing students from students at risk of failing or dropping out of university. This involves correlating historic activity data sets with the results achieved by past cohorts of students, looking for statistically significant measures. Once found these activity data sets are called differentiators.

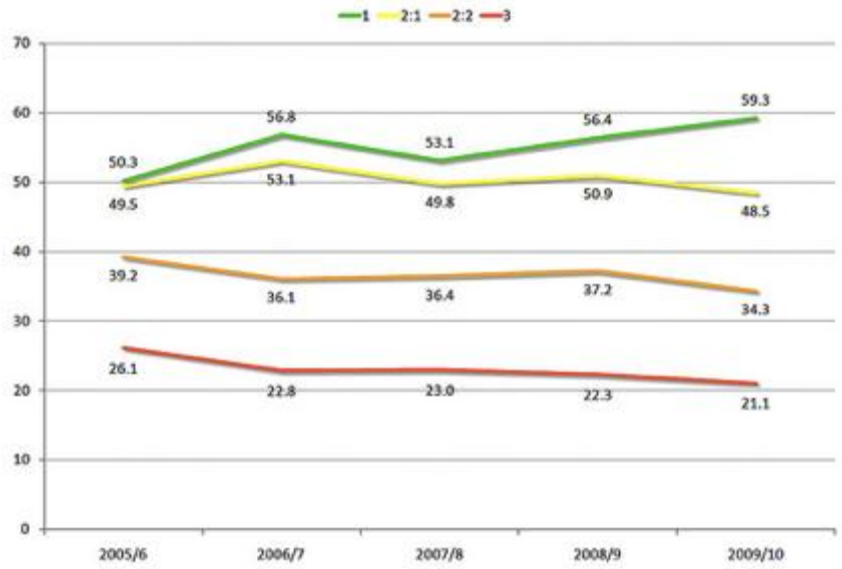
Once differentiators (or, using greater sophistication, patterns in differentiators) have found, the differentiators can be used to identify students ho are in current cohorts and at risk.

The two projects interested in learning analytics were LIDP and STAR Trak: NG.

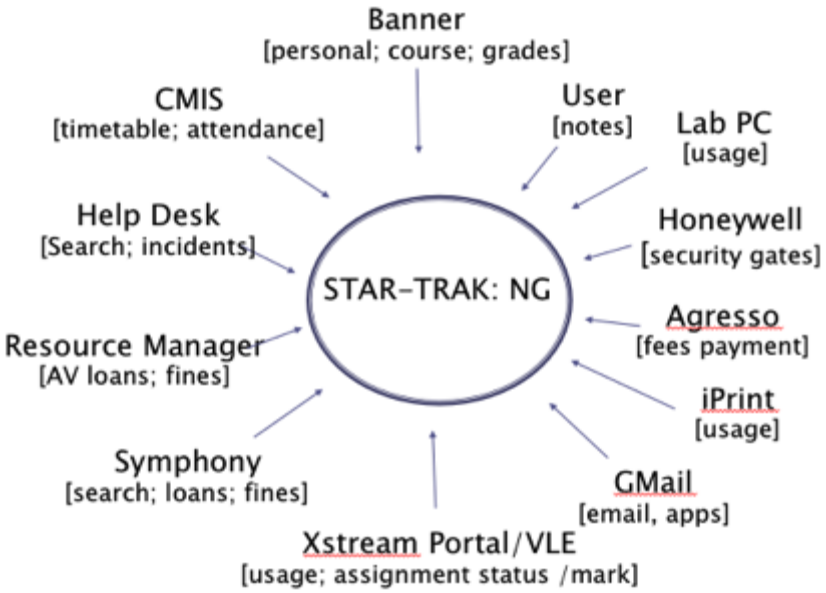
LIDP found a differentiator in the number of library books borrowed by students who achieved different degree classifications (1, 2.1, etc). The project takes care to point out that this is not a causal relationship, but rather is only based on statistically significant correlations. Two graphs expressing this are:



and



STAR Trak: NG used more sources of data in its learning analytics algorithms:



However, STAR Trak: NG have not published the algorithm they used to produce differentiators from these sources.

Further analysis

Various kinds of analysis can be performed on activity data, often in an exploratory manner, to get a feel for the data, before using it in other ways, eg for recommender systems.

Two blog posts point to this in connection with OpenURL data, Tony Hirst's original post on *NIX command line processing of the data, and Mark van Harmelen's Ruby riposte.

Analysis can also include manual processing, as has been done by the Exposing VLE data project who wrote:

> "We've observed sites being used for: teaching, research, administration, social activities and testing (using sites as a sandbox to try things out before updating a site that's already being used by students or researchers). More specifically, we've seen sites used for teaching lecture courses, whole degree programs, small-group teaching, language learning. We've seen sites used to organise research projects, from PhD theses up to large international collaborations. CamTools has been used to administer part of the college applications process, and for university and college societies, and to organise conferences. But unless a human looks at a site, we've got no way of deducing this from the data (we don't capture extensive metadata on site creation).
>
> "So, how do we categorise a site?
>
> "Currently, sites which are associated with a specific lecture course or a degree course are tagged with metadata on creation. This is a relatively new procedure, so only sites active from October 2010 are tagged. However, signalling that a site is no longer in active use for teaching (because a new site has been created for the new academic year, for example), is harder. The case I just mentioned can be done by editing the metadata, because we will have information that there is a new site; but if a lecture course has been discontinued, we can't currently update that.
>
> "For other sites, we have to rely on manual inspection. What is the site called? How many people are in it? What documents are stored there? Which tools does it use? From this information, we can usually deduce what the site is used for."

Presentation

Presentation of the results of recommendation processes is usefully split into two areas:

- Visualisation of activities derived from activity data sets: Here three projects contributed to the body of knowledge developed in the Programme.
- Consideration of user interfaces for recommender systems: This is mostly about displaying the outputs from recommender systems.

visualisation

Because activity data is relatively hard to interpret in its raw forms, it may be necessary to use visualization tools to make sense of activity data. This is particularly true when exploring activity data to see what it contains during the early stages of designing and building a recommender or learning analytics system.

Some of the necessity for visualisations becomes apparent when we think about the properties of activity data: It is potentially large scale, and may combine data from a variety of sources (eg library management systems, virtual learning environments, access management systems) and events (turnstile entries, system authentication actions, searches, downloads, etc.).

In the Programme, some projects used or considered using visualisation tools. Tools used or considered by the projects include:

Business intelligence tools

- Business Intelligence and Reporting Tools (BIRT) is an Eclipse-based open source reporting system for web applications, especially those based on Java and Java EE. BIRT has two main components: A report designer based on Eclipse, and a runtime component that you can add to your application server. BIRT also offers a charting engine that lets you add charts to your own application.
- Pentaho is an open source Business Intelligence suite. So far impressions are not good: EVAD say of Pentaho "[Have] been working on setting up a Pentaho instance for most of the last week, and hasn't yet [sic] managed to get a significant improvement on what Excel provides, though it's taken considerable effort to get this far. Pentaho requires various modules to be installed, but its documentation is rather incomplete, especially the documentation for creating aggregate tables. Aggregate tables are essential when dealing with large volumes of data - we have over 10M rows of Sakai event data, so without aggregate tables, every time we try to look at a large section of the dataset, we run out of resources. So thus far, our suggestion would be that if you want business information software, you may be better off paying for a commercial product."

Projects

Several of the projects made effective use of visualisation; details of what projects did can be found below:

- AGtivity
- LIDP
- EVADOpenURL Activity Data

General visualisation tools

- Gnuplot is a portable command-line driven graphing utility for linux, OS/2, MS Windows, OSX, VMS, and many other platforms.
- Gephi
- GraphViz
- Visual Complexity offers links to a comprehensive set of visualisation tools

General tools

- Excel pivot tables
- Open office pivot tables

- noSQL databases for custom programs include BigTable, Cassandra cluster, MongoDB

See also:

Guides

- Bringing activity data to life

Recipes

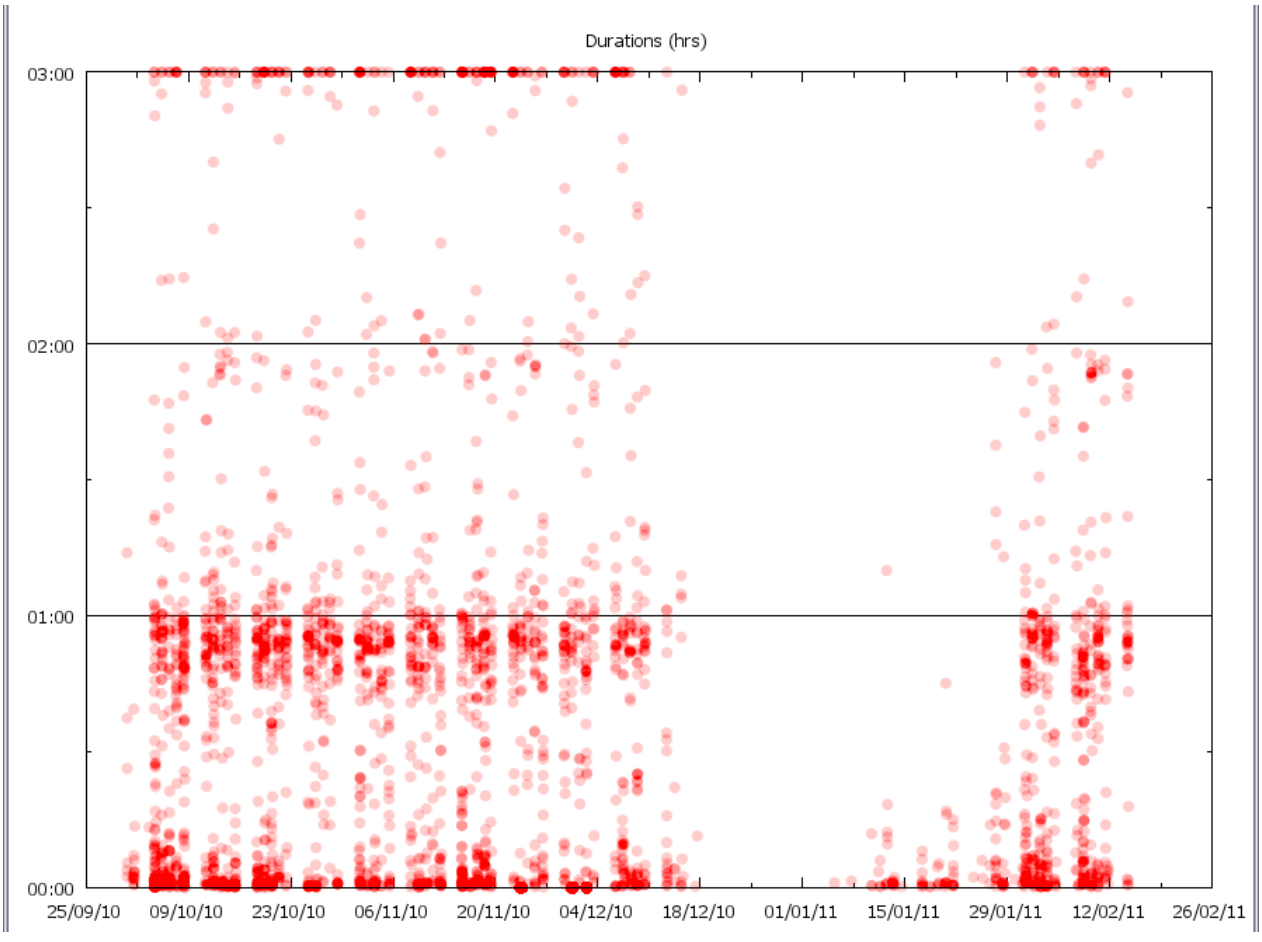
- How to Pivot data in Open Office Spreadsheet
- Visualising OpenURL Referrals Using Gource
- Producing PDF Reports Programmatically
- Plotting Calendar Data with GNUplot
- OpenURL Router Data: Total Requests by Date

Exploiting Access Grid Activity Data (AGtivity)

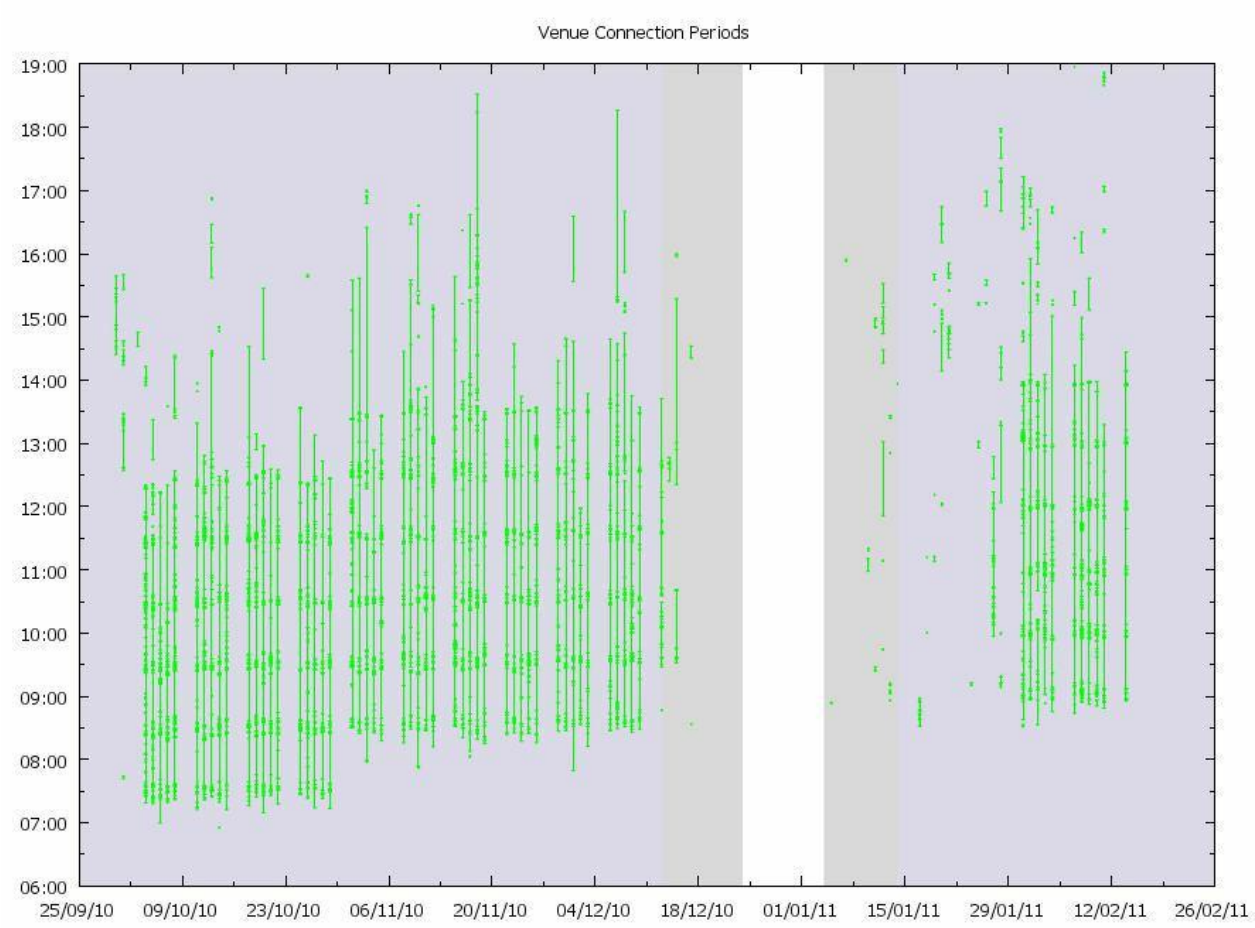
From AGtivity we have a range of new results for the user communities. This is the start for visual themes to create an 'autobiography' for video conferencing room nodes: Initial linking to attendance rates through remote lectures.

A virtual venue called 'MAGIC' is used for postgraduate lectures. Two data mining views are shown below that start to highlight individual course components.

The first shows duration values over time, allowing general usage to be considered (also incorporating test periods and extra- curricular items that are seen in order to aid quality control):



The second graph shows connection periods.

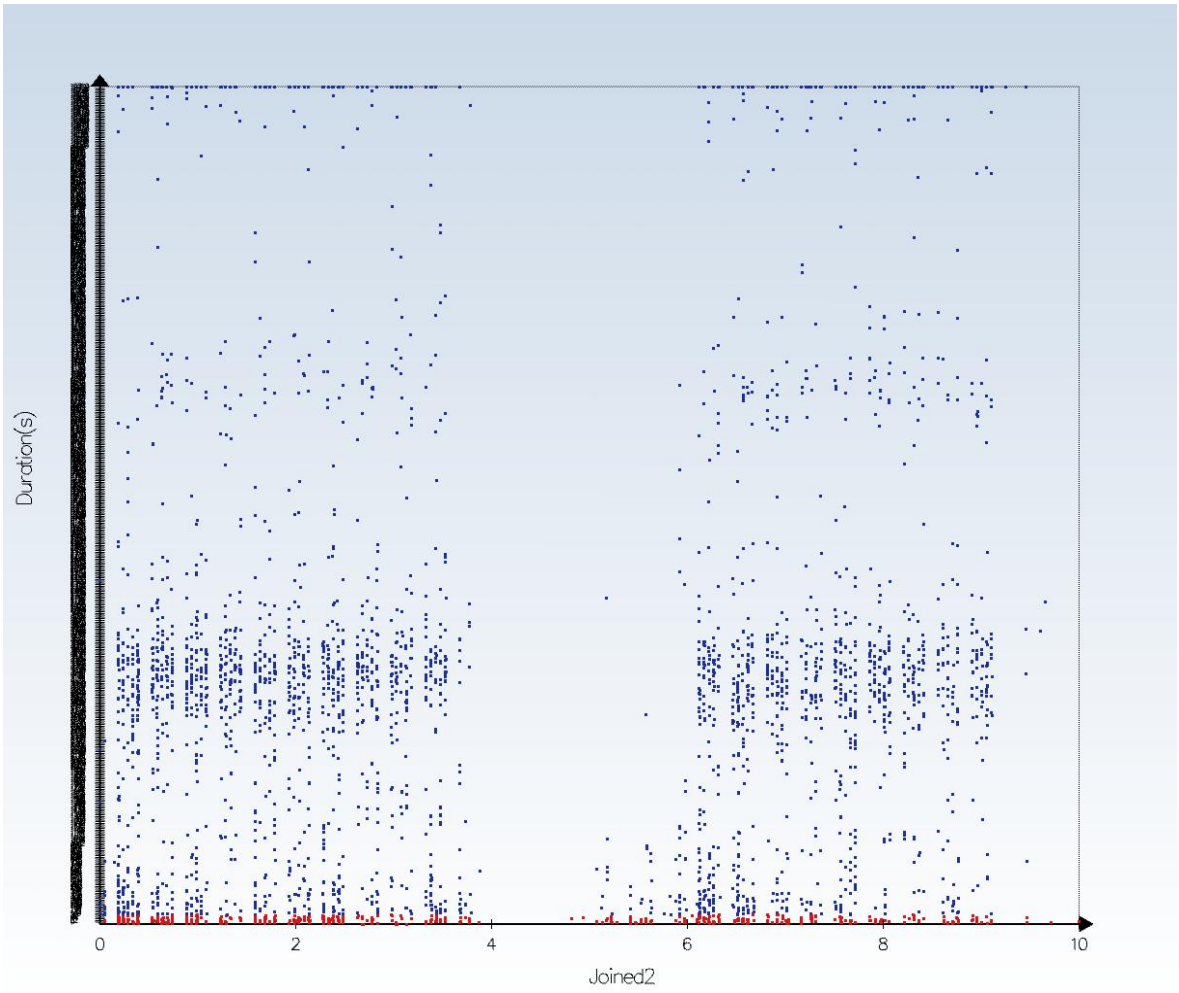


Cancelled lectures, rapidly increasing and/or poor attendance, and extra lectures are visible.

Examples of visualisation choices

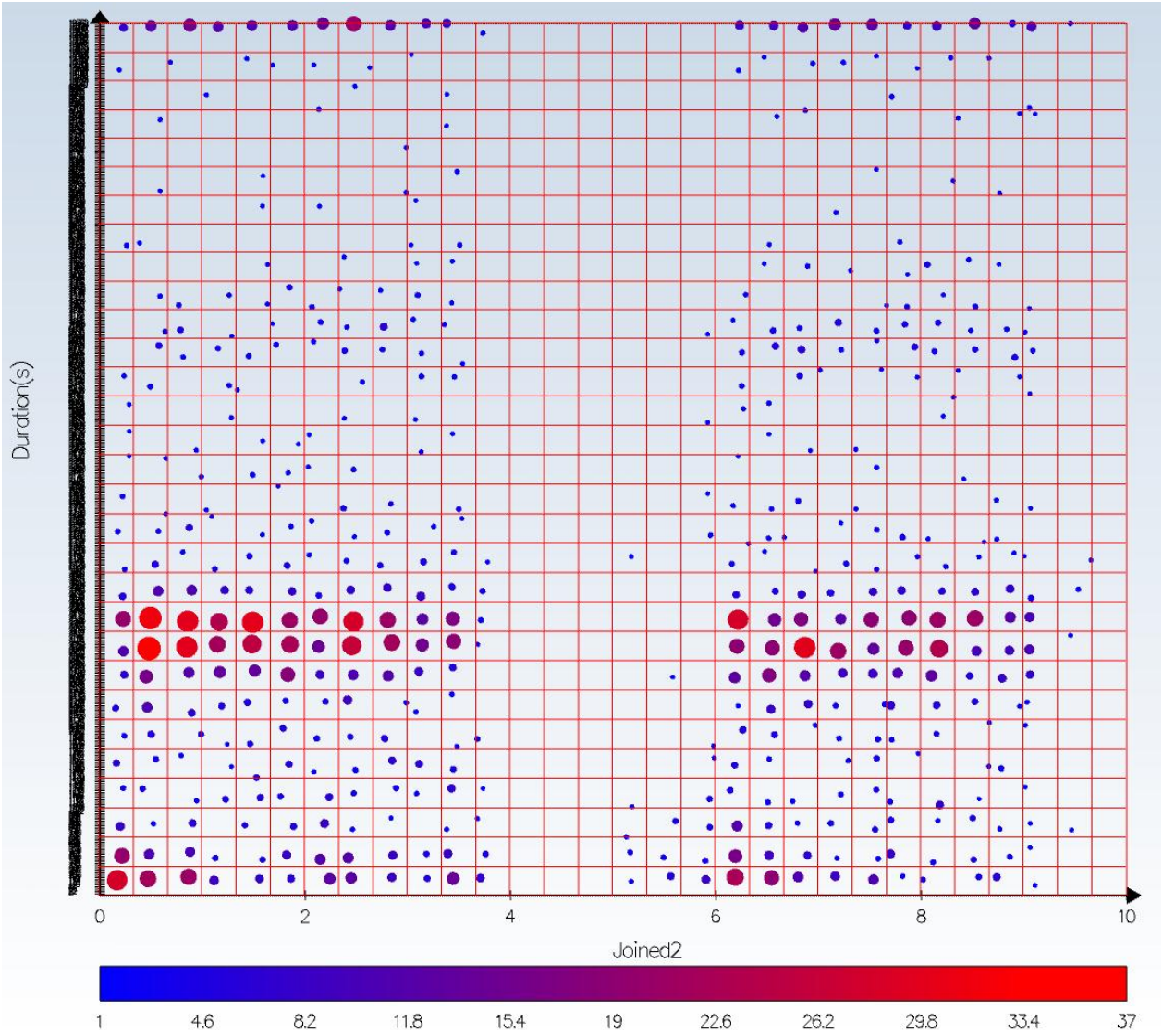
Various visualisation styles are shown below for AGtivity data, produced using code written by Stephen Longshaw.

The first graph shows duration mapped against date and time for a series of events. A single dot represents a single event; either a lecture, or an associated course meeting. As shown semester 1 and 2 are clearly visible, separated by a gap.



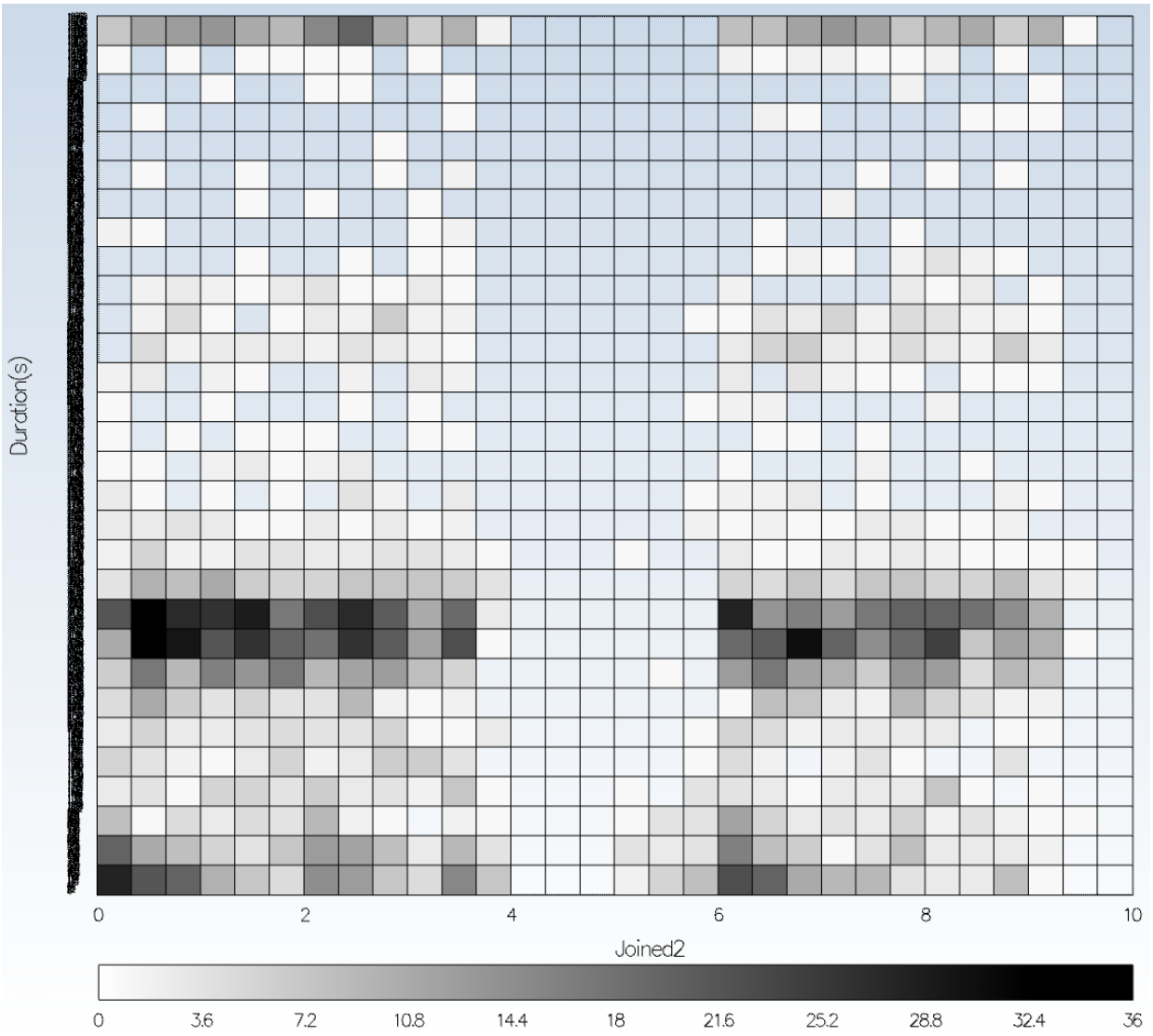
The elements in red indicate very short test sessions and were omitted from later analysis.

There are too many points to really make informed decisions; some could appear on top of others; to help resolve this AGtivity used a technique called binning (i.e. clustering) the points shown by the grid in the following graph. The dots are replaced by a single circle within each grid square with a size that is proportional to the number of data points in that grid square.



The colour also represents the number of points and from this some quantifiable statistics can be estimated. This shows the majority of sites attended for about one hour, but a significant proportion attended for multiple concurrent hour-long lectures. Note the graph stops after 3 hours.

An alternative technique to using coloured circles is to colour the boxes themselves according to the occurrence values.



There are many different options that could be considered. It is an open debate as to which style of visualisation a particular user will prefer, but a key process is always to consider choice in presentation and visualisation mode; and to see which method of presentation may provide the best insight.

Exposing VLE activity data (EVAD)

The EVAD project has demonstrated a number of visualisations including the use of Gephi to show connections between sites within the VLE (See How to work with Gephi to visualise a network of Sites connected by Users for details on how to do this)

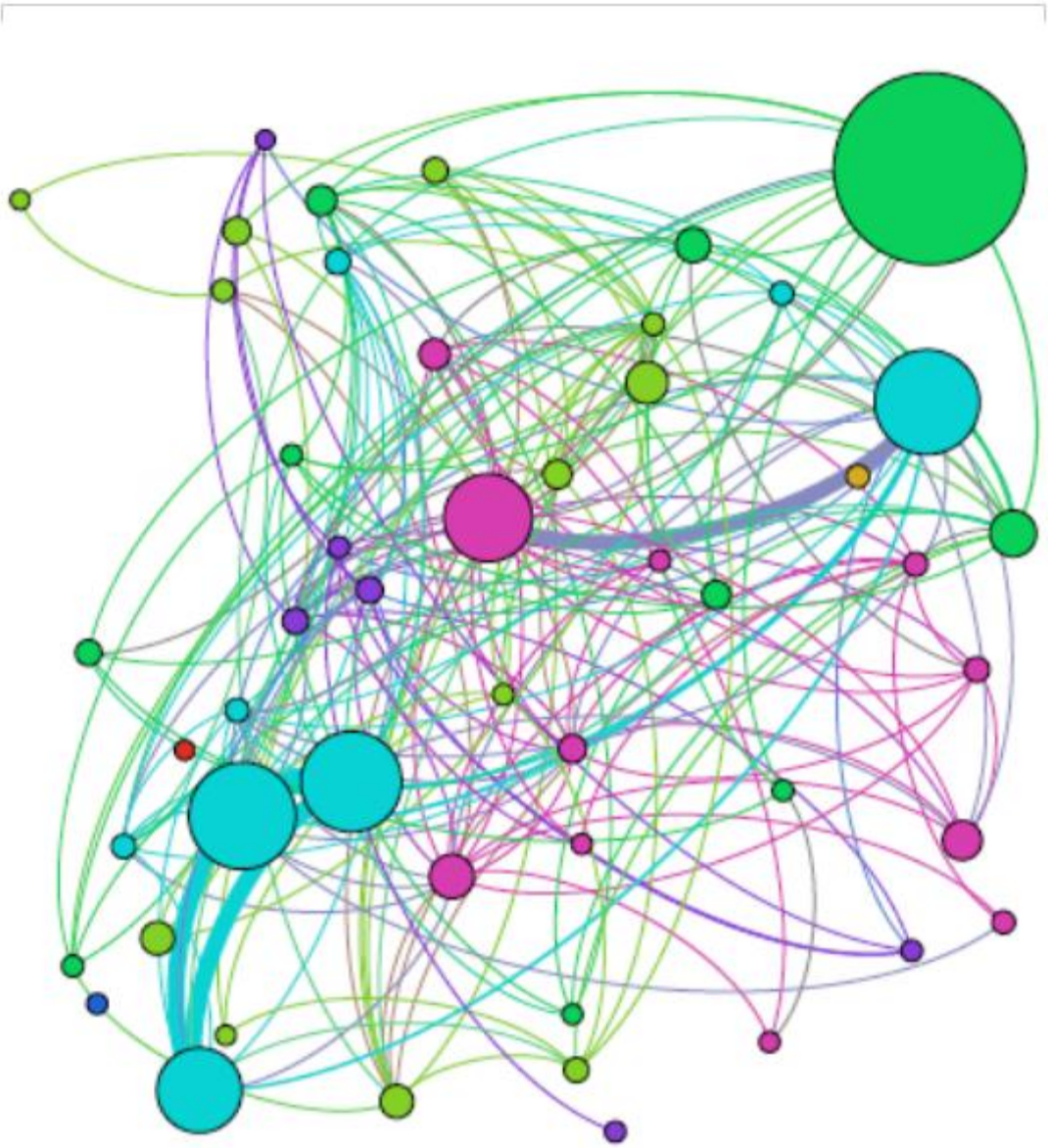


Illustration 1: Gephi network connecting the Top50 Sites from one of our Departments

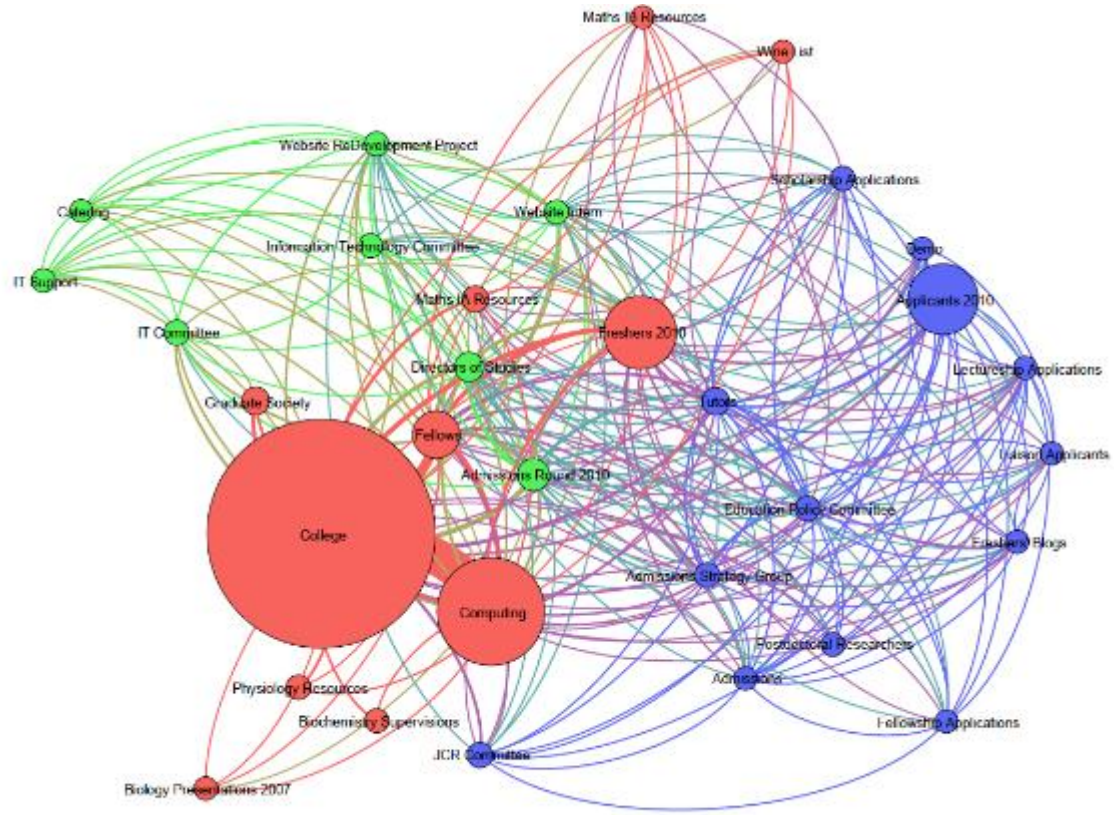
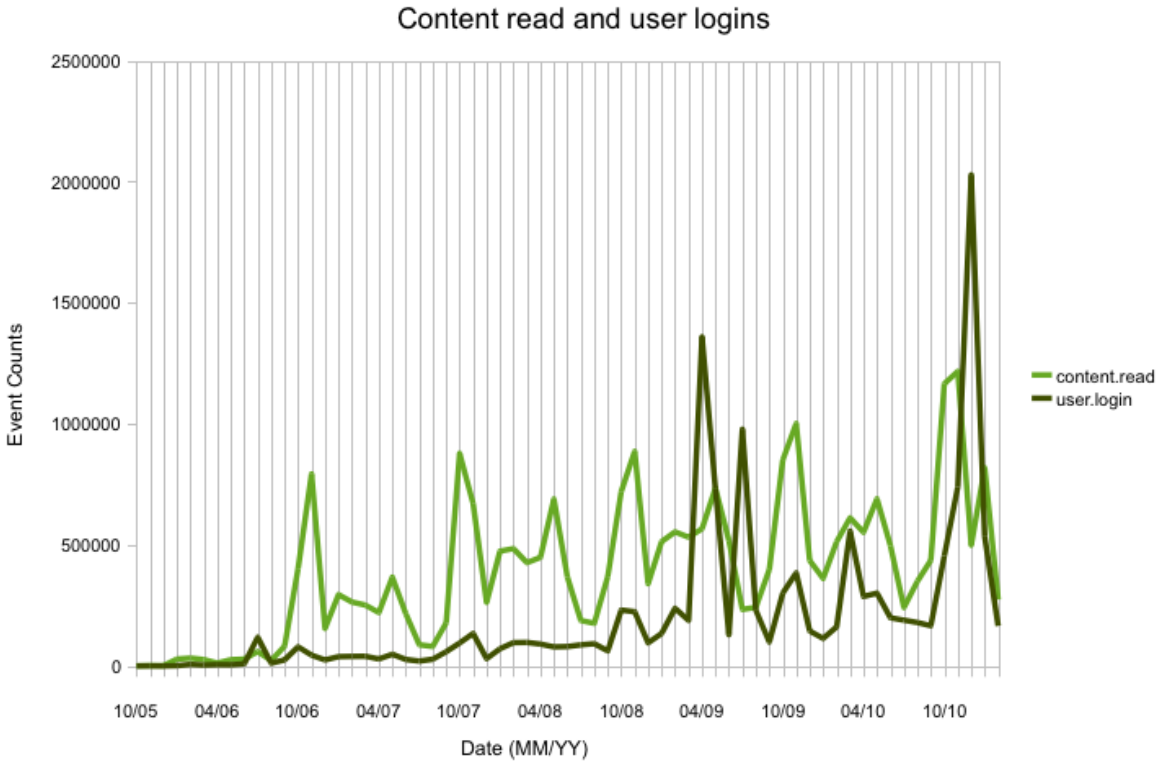


Illustration 2: Gephi visualisation of Sites used by one of our Colleges connected by shared users

They also used visualisation to demonstrate patterns of VLE usage:

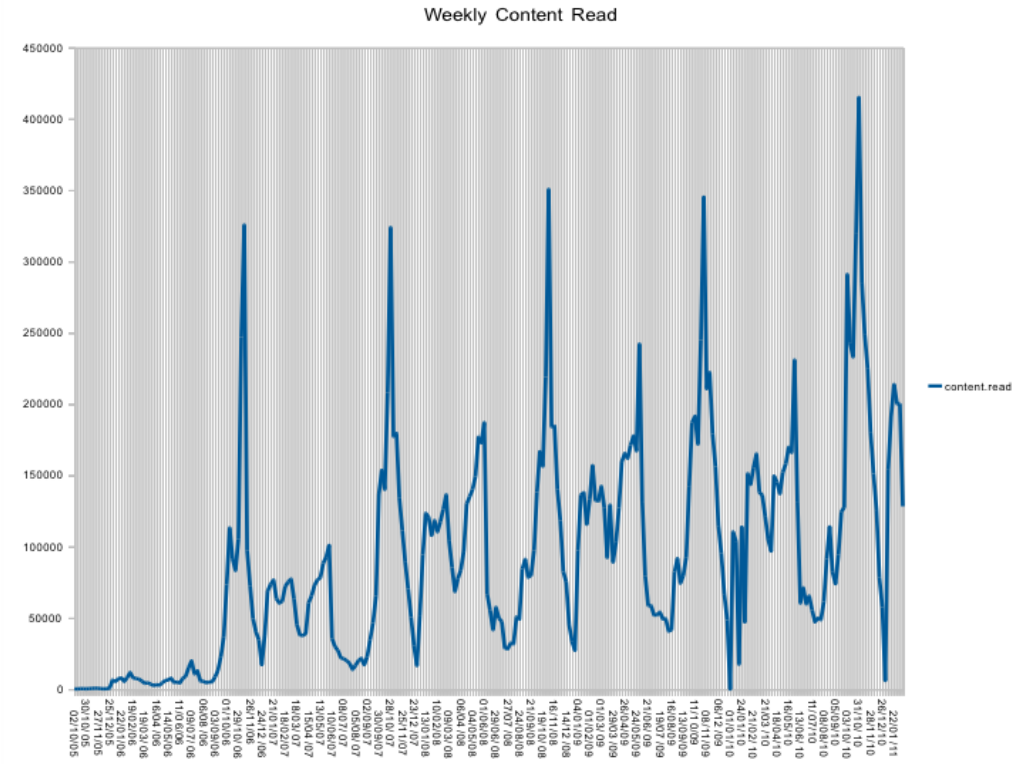
The top 2 events were found to be reading content (green - 44% of events logged) and user logins (dark green - 22% of events logged). The chart below shows monthly counts for these events.



Whilst the content read shows a regular peak in Oct/Nov with a secondary peak around May, the logins

do not show such a clear pattern. Both show a clear gradual underlying increase over the years. More investigation would be needed to find out exactly what is causing the peaks in logins. This is especially the case since it bears little relation to the number of user logouts recorded (see later graph).

The following shows a weekly count for content read giving more granularity of when exactly the peaks are. The main peaks are repeated in late October with secondary peaks in Feb and mid- June.



This considers weekly counts over the last year. (Feb 2010-2011).

Term dates were:

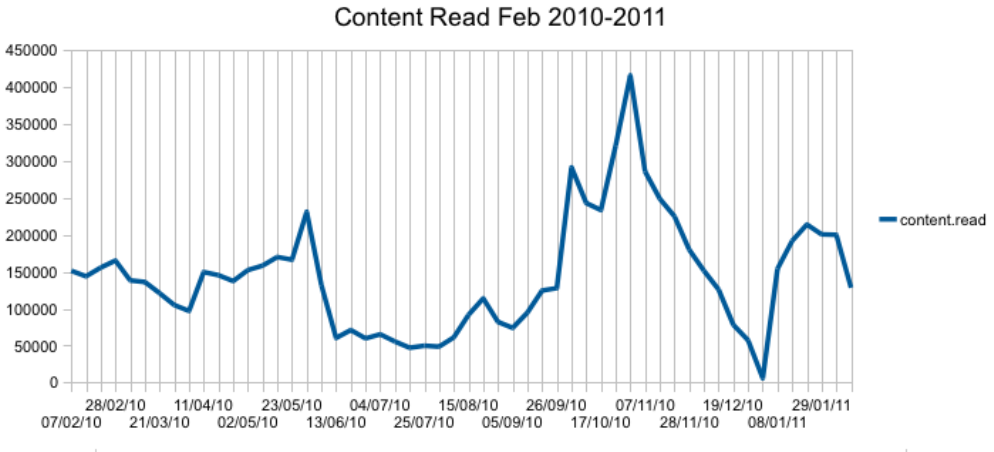
12th January - 12th March 2010. Easter (5th April 2010) shows a slight dip.

20th April - 11th June 2010. A gradual increase with a peak at the end of May.

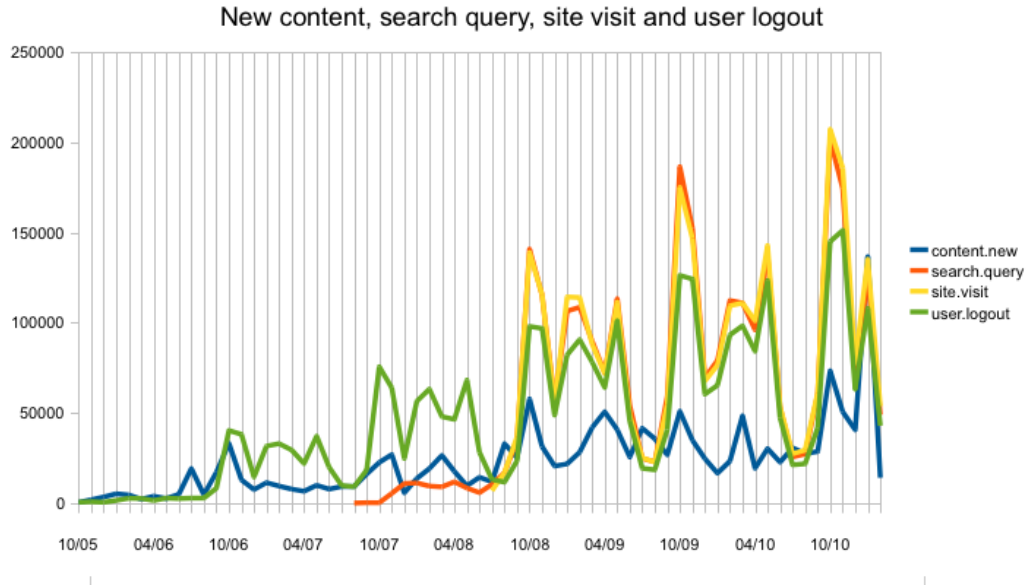
This coincides with the examination period. There is a long dip over the summer vacation period.

5th October-3rd December 2010 First peak is early October with the highest peak at the end of October. This peak at the end of October might not be expected and it would be interesting to find out more about what is going on here. There is a large dip over the Christmas period

18th January - 18th March 2011. There is a wide peak during January and early February.



Although content read does have a peak at the start of the Academic Year in October, there is a higher peak in November. It would be interesting to investigate if there is an explanation for this peak half way through the term – such as a large number of students starting assignments which rely on Camtools.

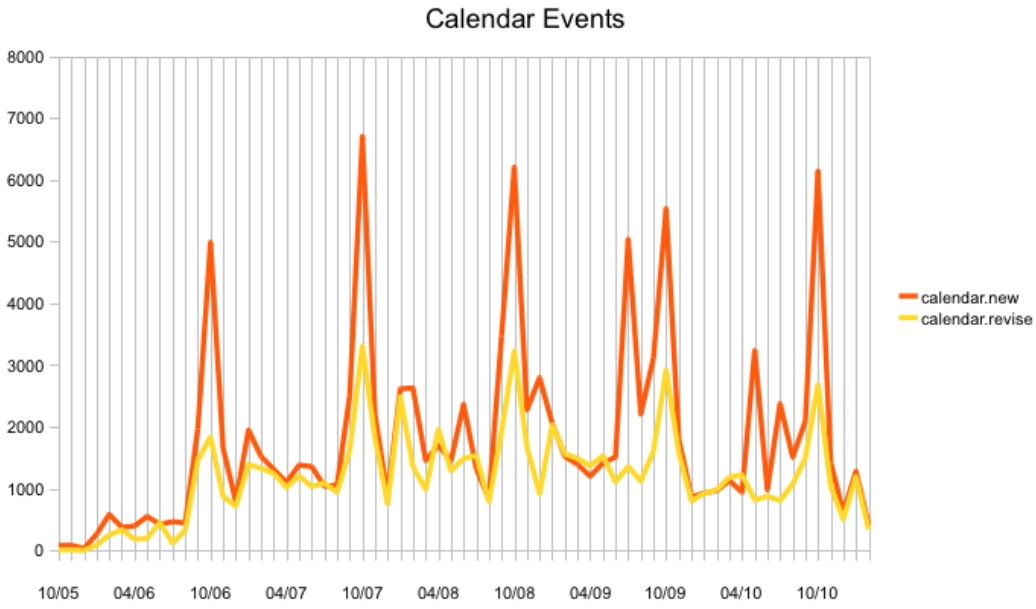


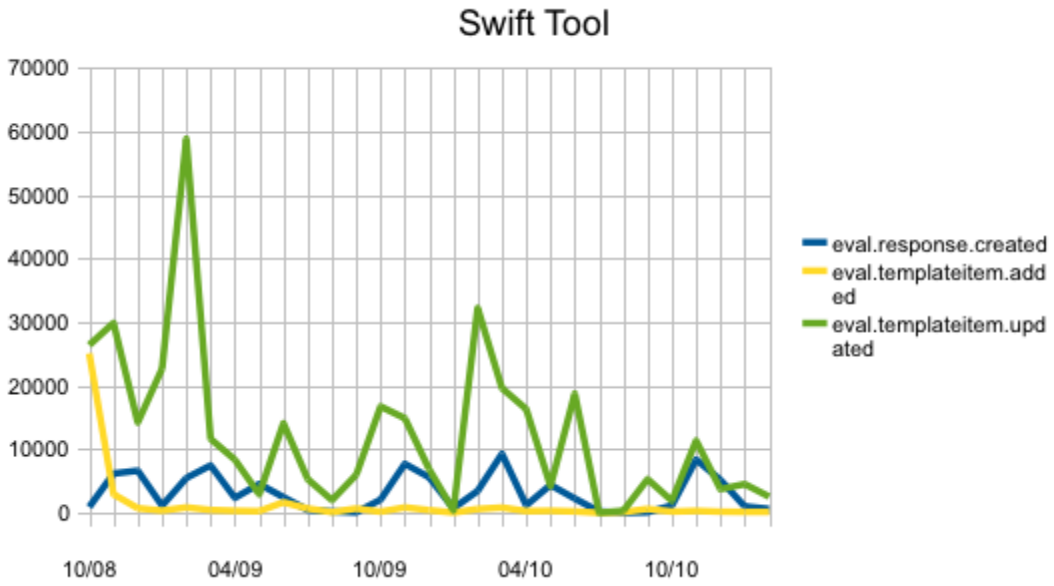
This shows the next active events in terms of counts. Note the scale is roughly 10% of that for viewing content and logging in. However the overall peaks for site visits, search and user logout very much mirror those for viewing connect and logins showing consistent periods of peak activity corresponding to term dates.

As one might expect, lots of new content (blue line) has been added each October, at the start of the academic year, but there is a more recent and much higher peak in January of this year.

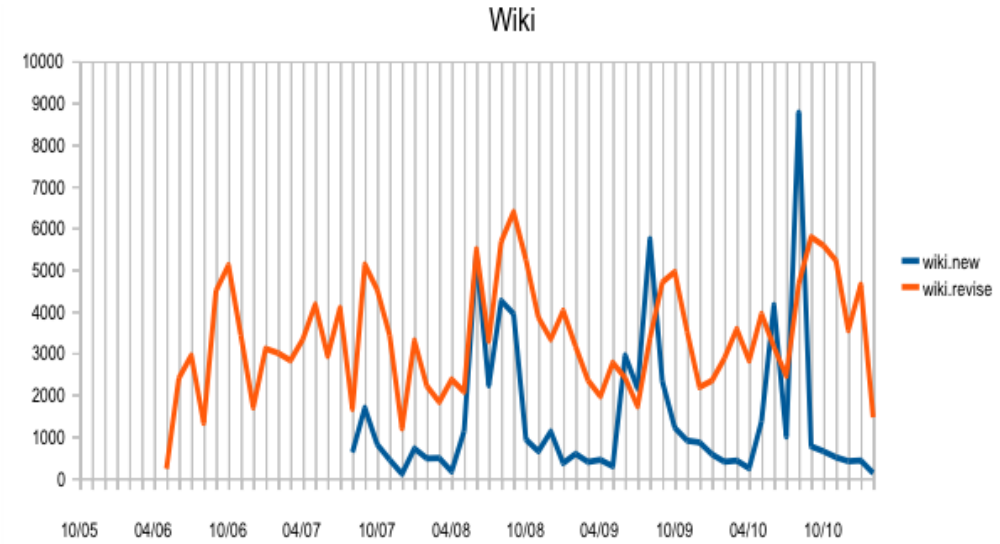
The following series of charts show analysis of the next highest peaks broken down by event type. Note the scale is several times smaller that that for reading content.

The calendar event is logged when users create (red line) or revise (yellow line) site calendars which may be used to display lecture times, assignment submission dates and so on. There is a clear October peak for both creating and revising calendars.

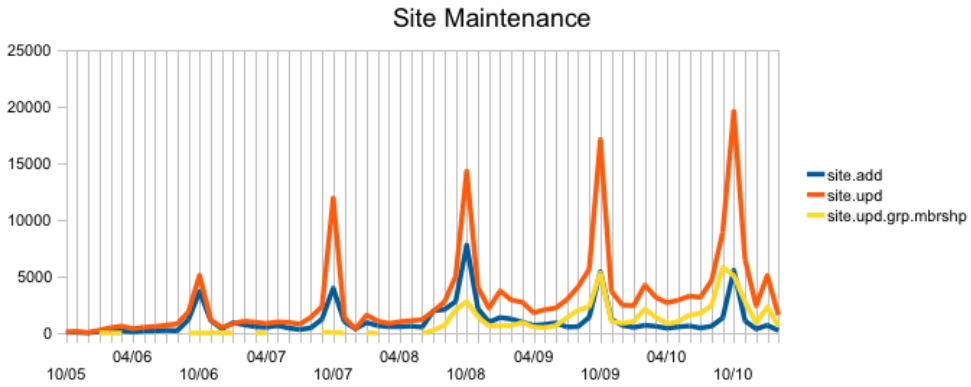




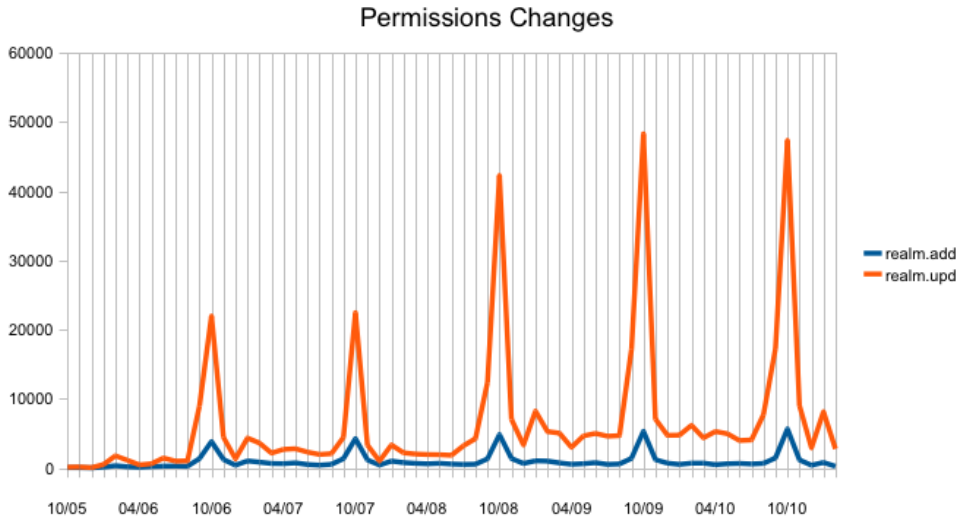
This tool has been added in the last few years and is used to for student surveys. The green update peaks in February would correspond to lecturers preparing surveys at this point which are returned by students roughly a month later (blue line).



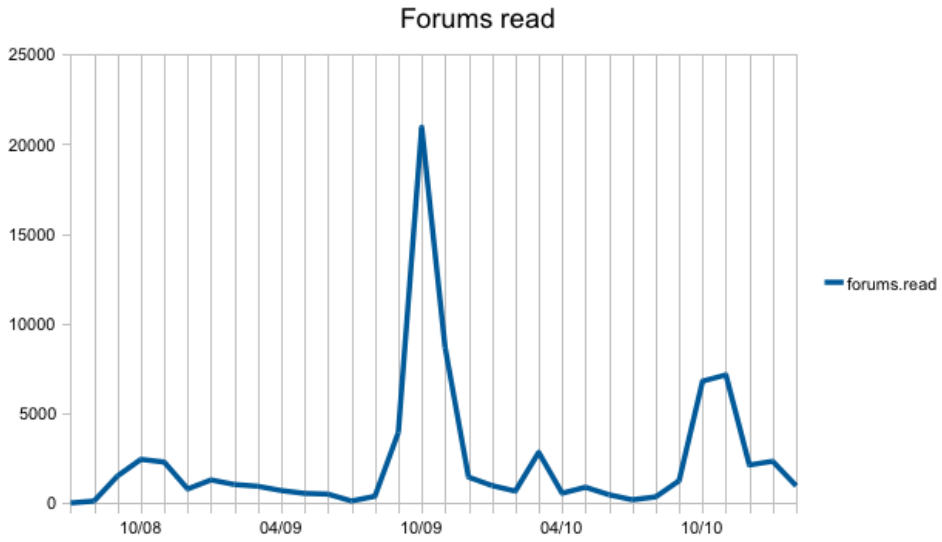
The counts for Wiki activity show a peak in new material just prior to the October start of year with high revision activity at the start of October.



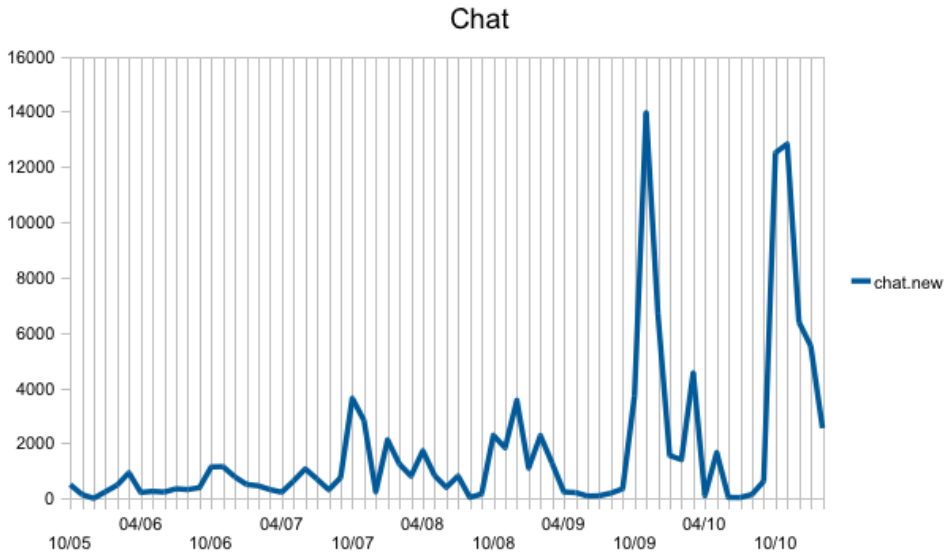
Site Maintenance demonstrates a clear October peak with a secondary peak in Jan/Feb. There is clearly a gradual increase in the number of sites being updated (red line) whereas the number of new sites (blue line) has levelled off.



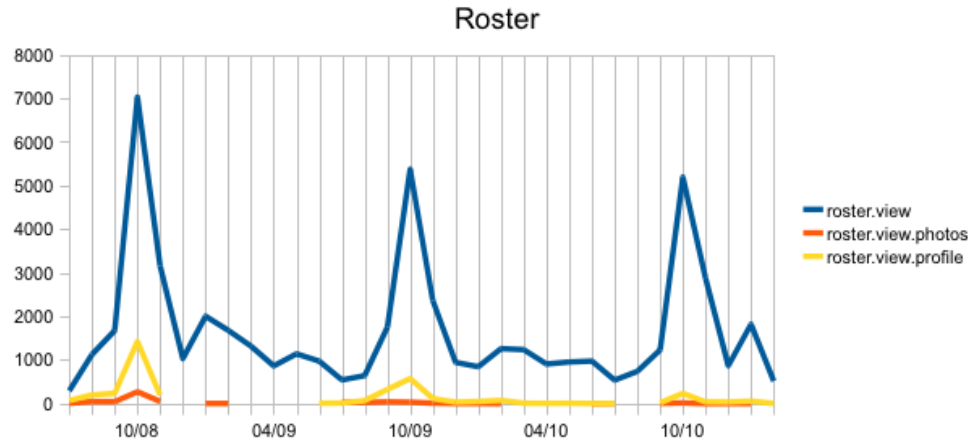
Permissions changes reflect people being added or updated on sites. Again peak activity is in October.



Forums again show a peak in October but their usage looks like it has tailed off.

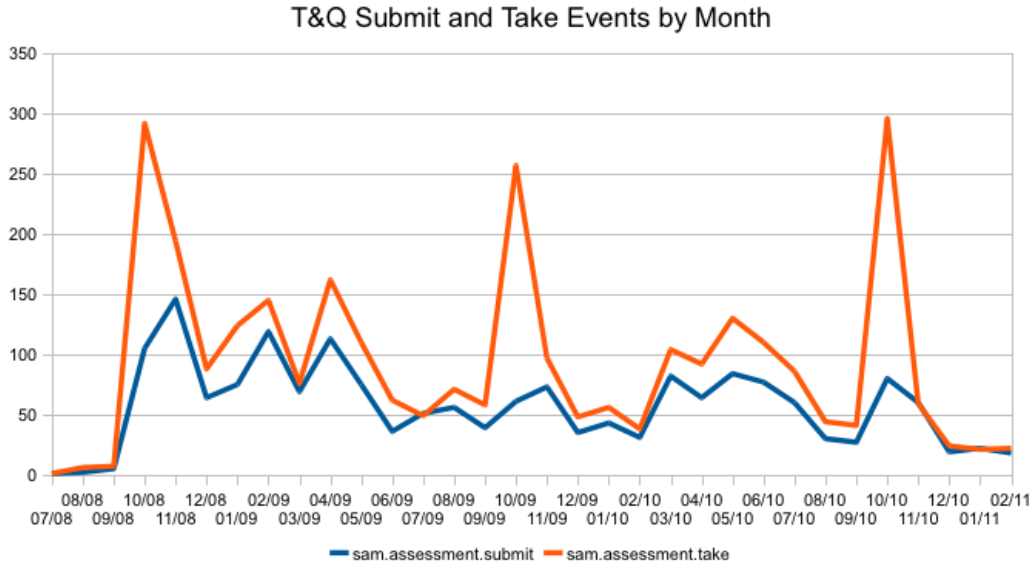


The chat tool also shows peaks in October and for the last year in November too.

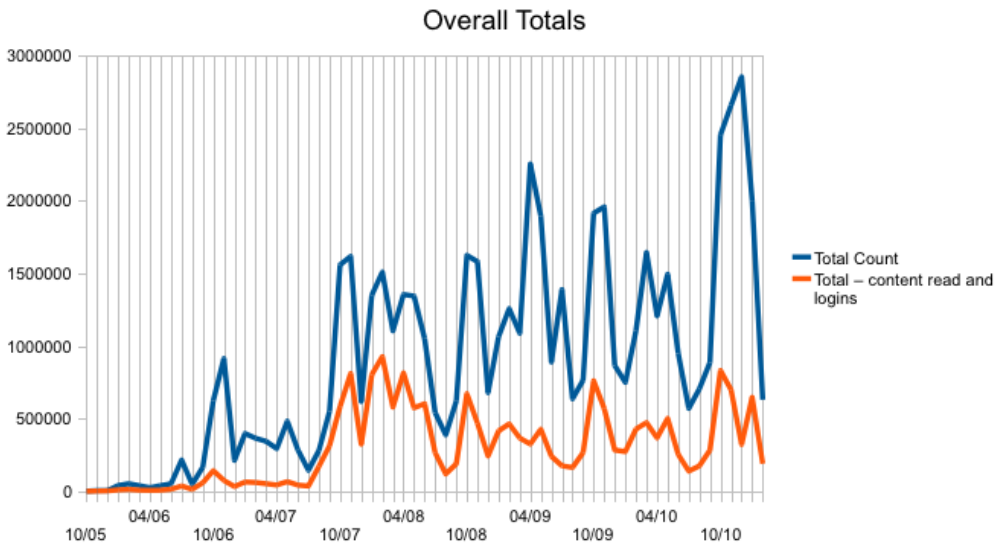


Again the roster tool, which is used to view other members of a site shows peak usage in October.

One area that I was particularly asked to look at was the Tests and Quizzes (T&Q) tool. The scale for this is again of a much smaller order of magnitude but again shows October peaks:



This final chart shows the overall totals for each month (blue) and the monthly totals excluding logins and content read (orange). This again shows a peak in October and other peaks corresponding to the terms. There is not much growth, in terms of event counts, over recent years in these other activities.



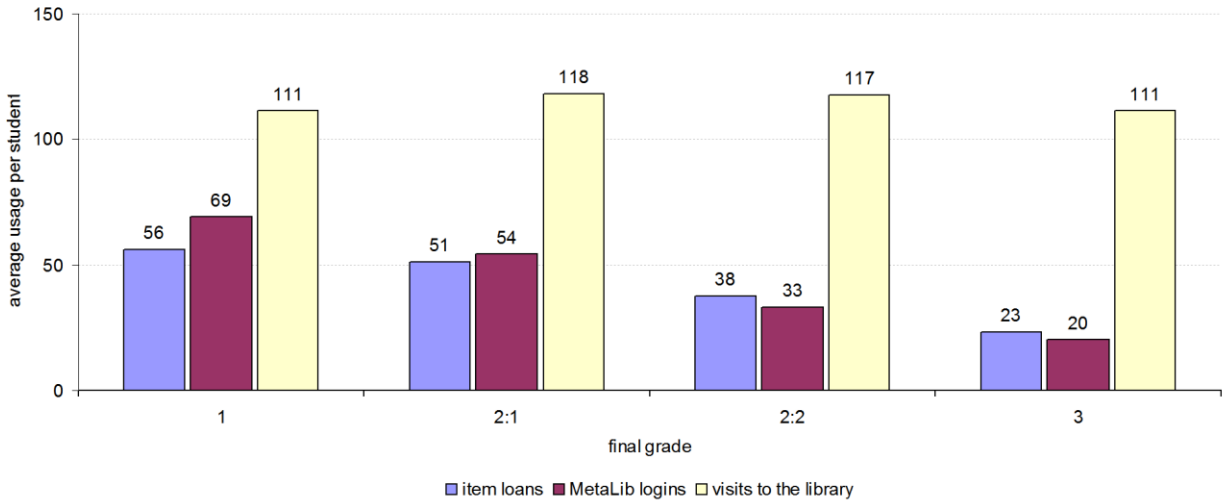
Conclusions

Many activities show a clear usage peak corresponding with the start of the Academic year in October. However reading content, which accounts for over 44% of the counts, has a higher peak in November than in October. It would be interesting to learn exactly what is causing this November peak.

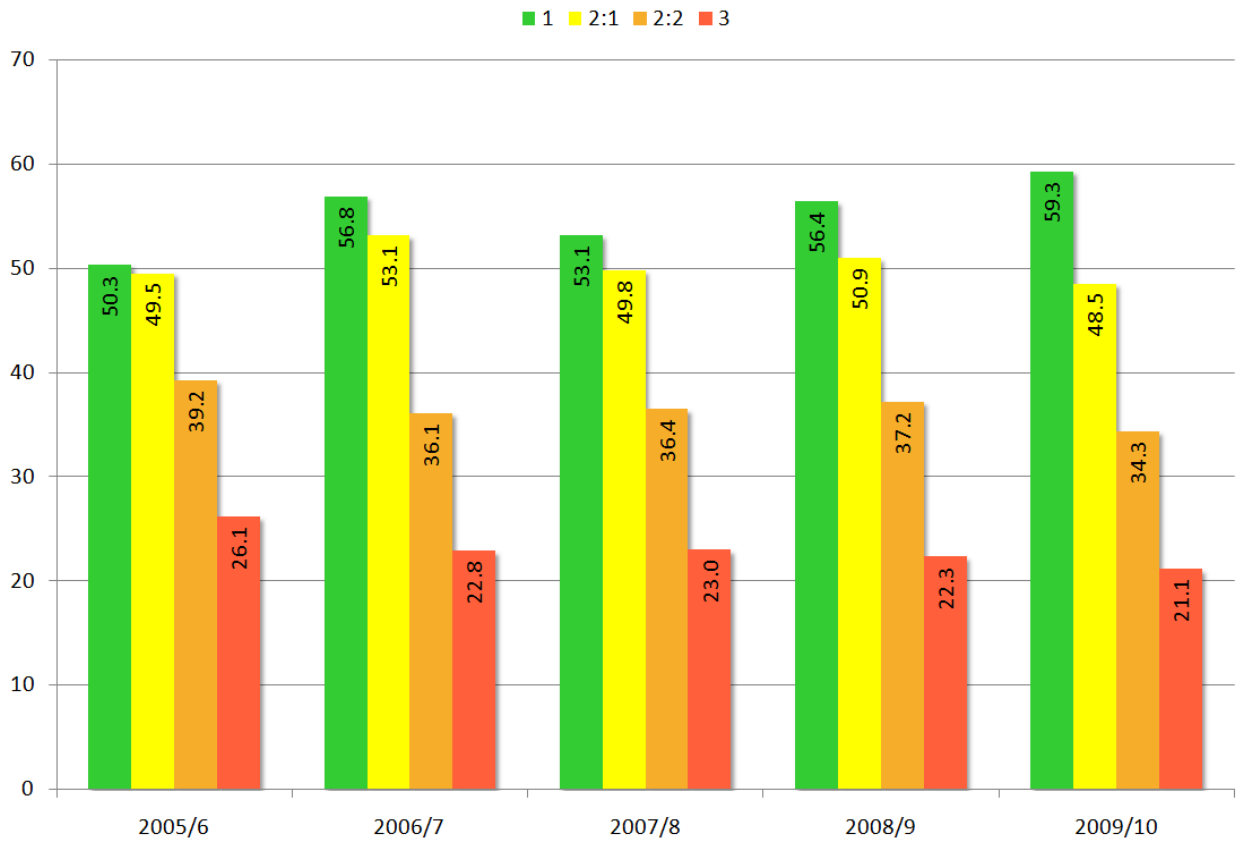
It also seems likely, given the different pattern in logouts, that the high spikes in login activity seen in the first chart are caused by spurious data and further investigation is needed here.

Library Impact Data Project (LIDP)

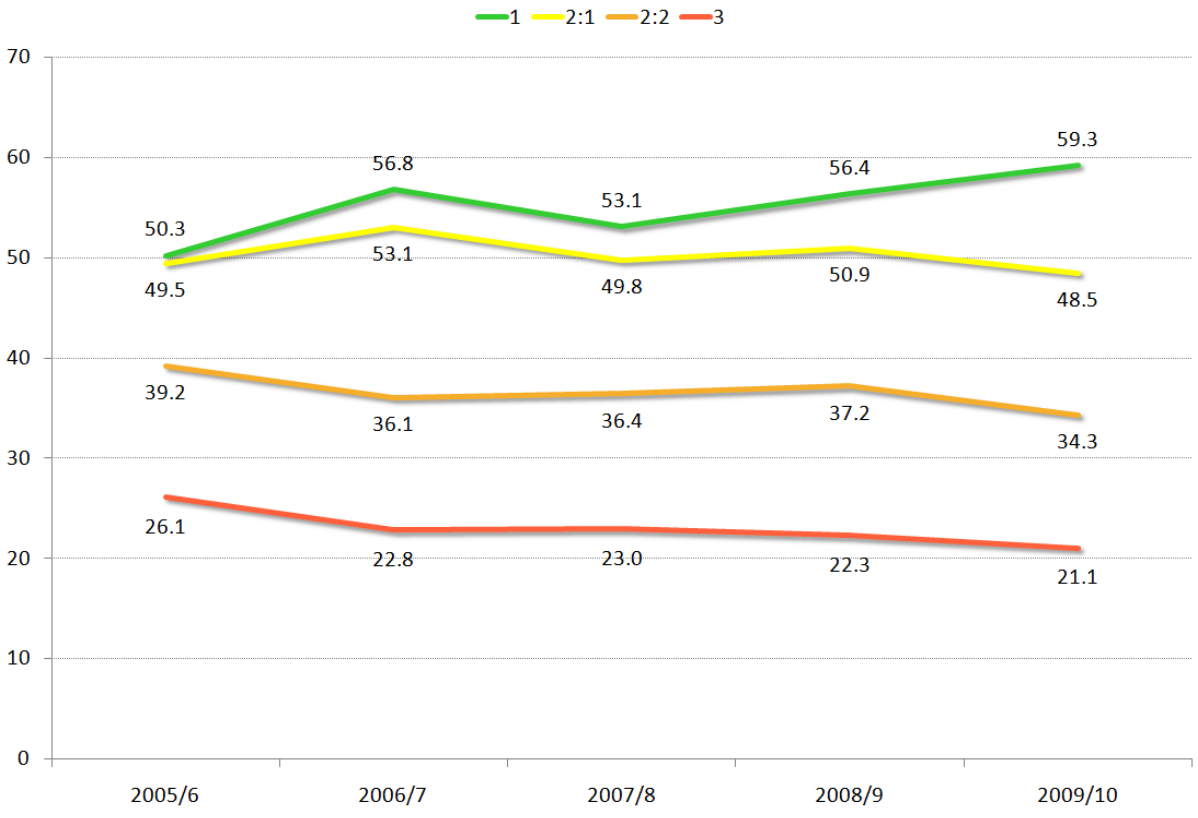
LIDP produced several kinds of graphs as visualisations showing how numbers of library loans differentiated between students who went on to obtain different class degrees.



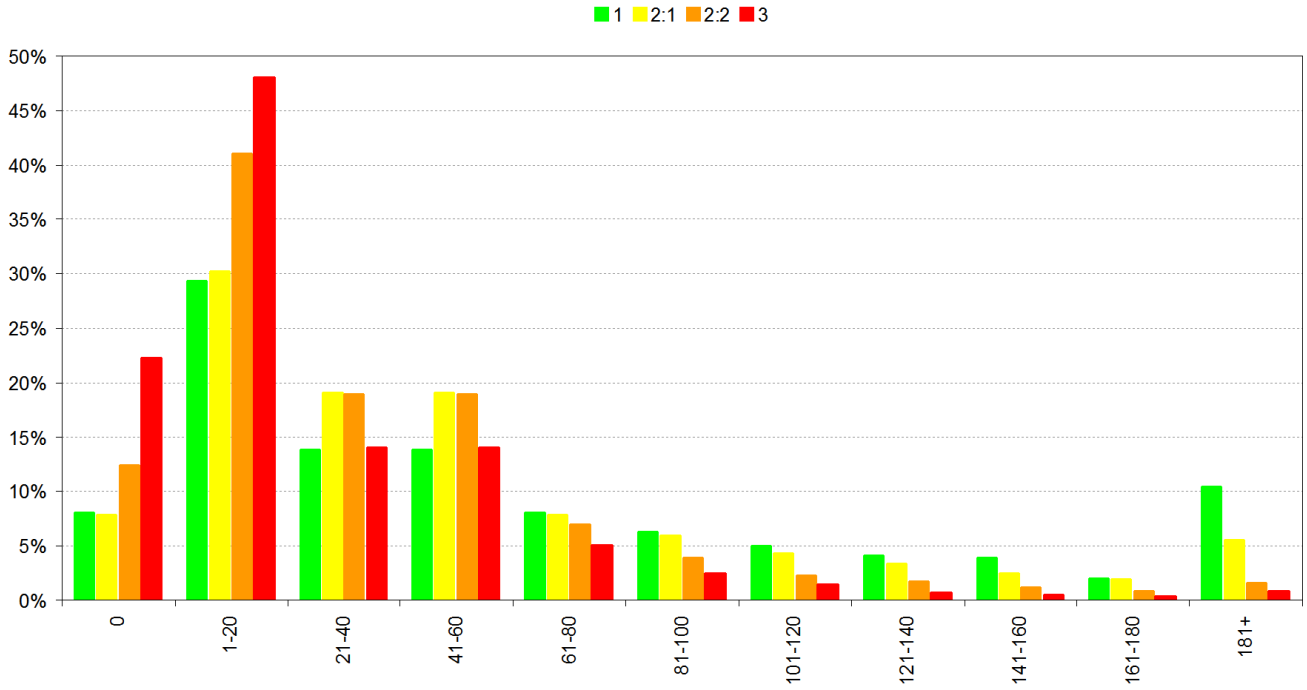
Usage of library resources against degree classification 2007/8 & 2008/9



Degree classification against number of books borrowed 2005/6 - 2009/10 (bar chart)



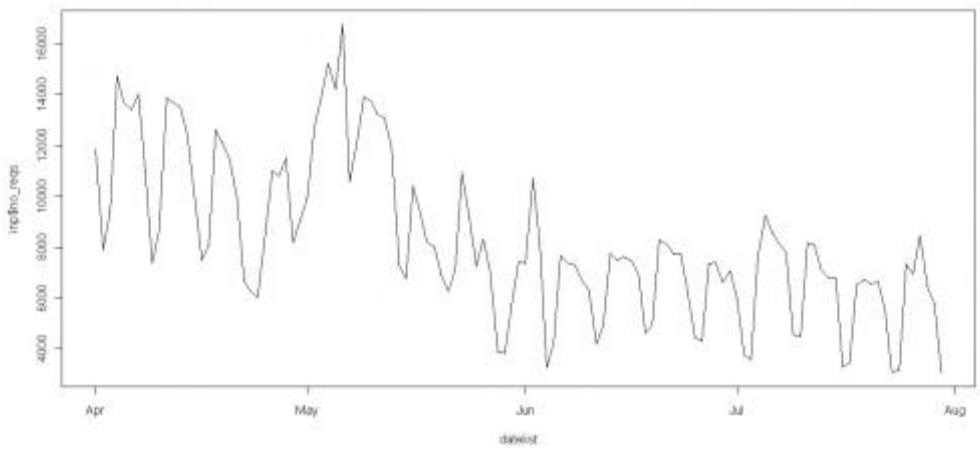
Degree classification against number of books borrowed 2005/6 - 2009/10 (line graph)



Percentage of students using amount of e-resources and degree classification

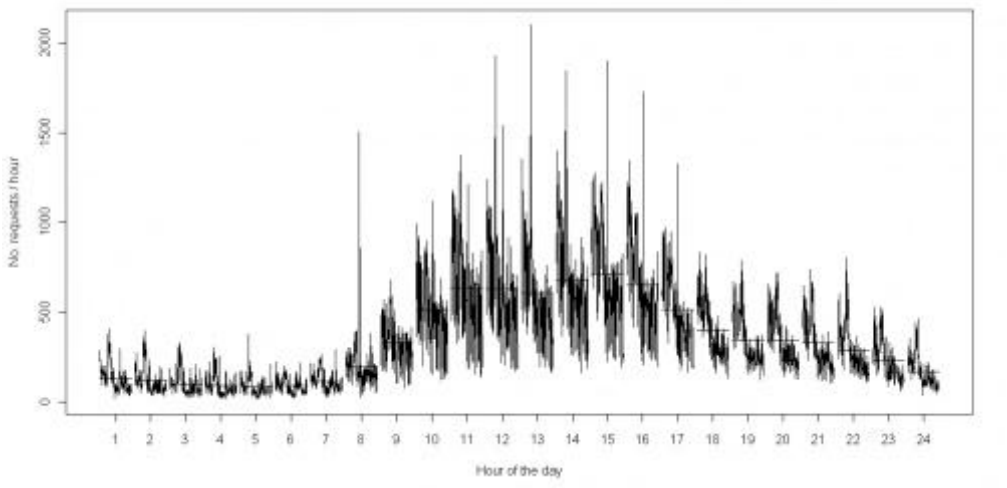
Using OpenURL Activity Data

The Using OpenURL Activity Data project led to a number of different ways of visualisation of the data. The project produced a number of summary graphs. such as:



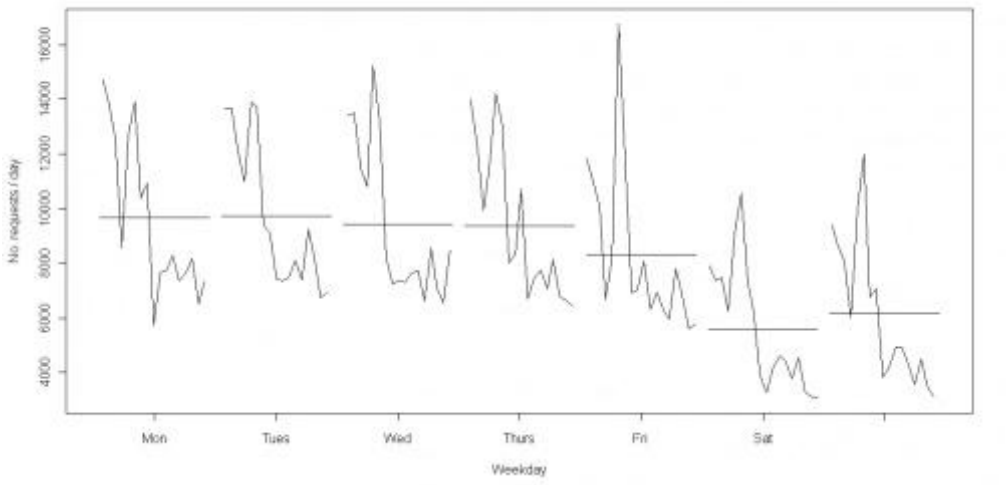
OpenURL Router Data: Total Requests by Date

From this it can be seen that there is a peak in usage at exam time and that usage drops off over the summer. The recipe for how to produce this is available, and the code and data for them all can be found at http://figshare.com/figures/index.php/OpenURL_Router_Data:_No_of_requests_by_date



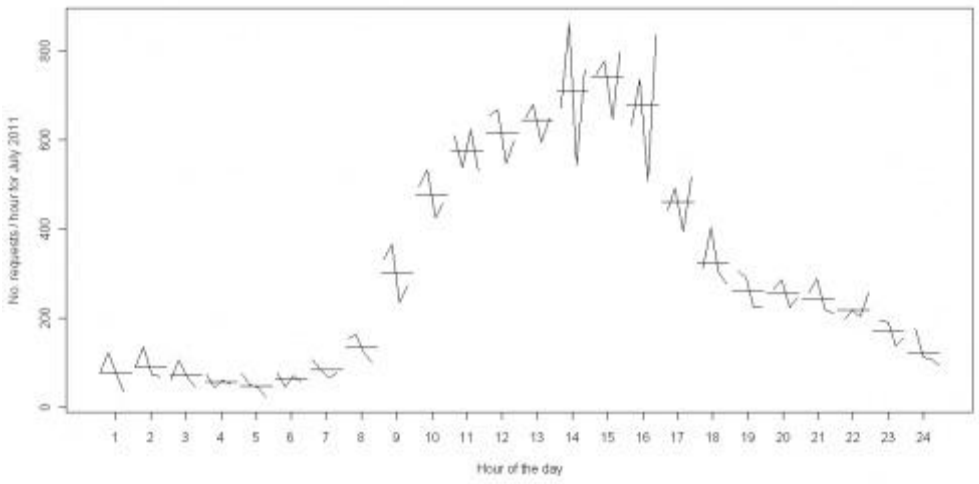
OpenURL Router Data: Total Requests by Hour

Note the sharp peak at the start of hours during the day - perhaps users are looking up things that they have just heard about in lectures? It would certainly be interesting to look further.

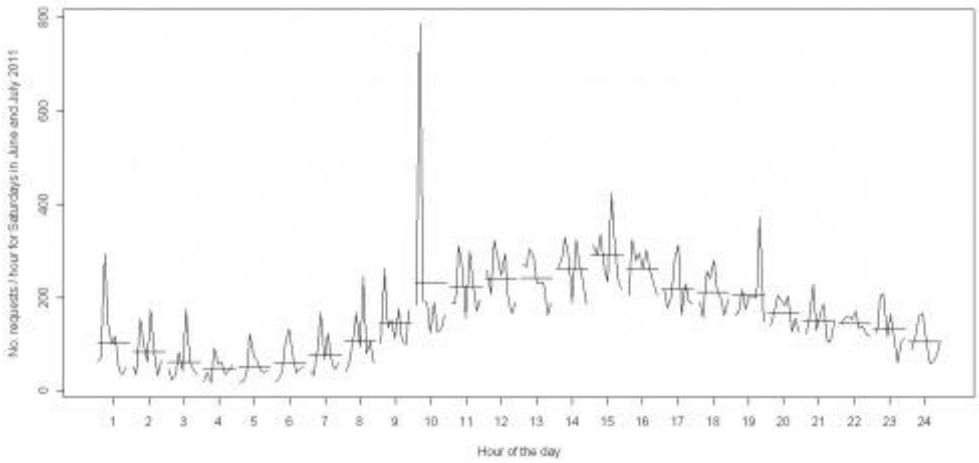


OpenURL Router Data: Total Requests by Weekday

There is a common pattern each day, with the greatest number of searches in the late morning, but why especially on Fridays?



OpenURL Router Data: Total Requests by Hour for Mondays in July 2011



OpenURL Router Data: Total requests per hour for Saturdays in June and July 2011

What did happen on the 9th?

User interface for recommender systems

Several of the projects have evaluated the usability of recommendation presentation. While sometimes usability evaluations do not reveal one correct answer, there are issues to consider in displaying recommendations:

- The number of recommendations to display (AEIOU found that five or six seems right)
- The most appropriate location on the search result(s) page for the recommendations?
- If there should be links to recommendations from the top of the page, if the recommendations appear below the fold (i.e. below the initially displayed area, for a particular sized screen)

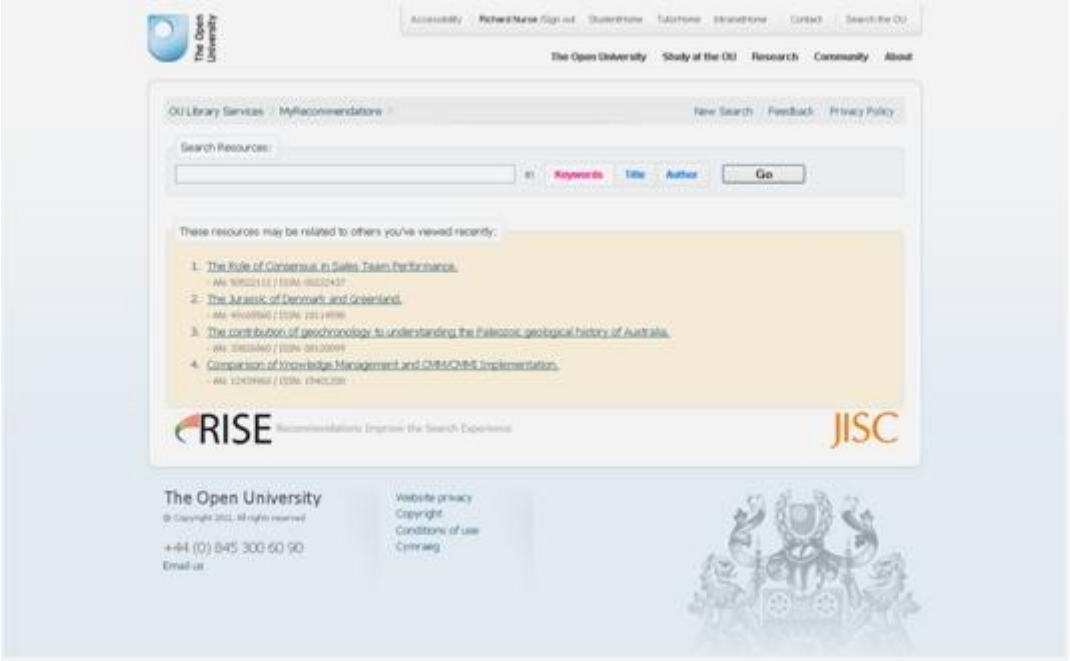
Projects

- RISE

Recommendations Improve the Search Experience (RISE)

A screenshot of the RISE front page is shown below.

This uses the standard Open University corporate style (with appropriate links to the RISE project and JISC) and provides a search box for the OU Library's EBSCO Discovery Solution.



Once searches have been initiated from this page, search results with recommendations are placed in subsequent pages. Interestingly users are prompted to rate individual results. See the MyRecommendations screencast for details.

Related work

There is much other work going on in the area of using activity data. Here we point to some of the other work, which we have divided up as follows:

- Other JISC funded work,
- Other work in the UK,
- Other work elsewhere in the world.

Related JISC funded programmes and projects

The following lists some of the other JISC programmes and projects that are undertaking related work:

- **JISC Business Intelligence programme.** which defines business intelligence in the following ways:
    - Business Intelligence (BI)
      Evidence-based decision-making and the processes that gather, present, and use that evidence base. It can extend from providing evidence to support potential students' decisions about whether to apply for a course, through evidence to support individual faculty/department and staff members, teams and departments, to evidence to support strategic decisions for the whole institution.
- Business Intelligence System (a BI system, a management dashboard)
  A system that compiles and presents key internal and external information in a concise, pictorial or graphical format, to support decision-making, planning and strategic thinking. It provides easy interactive access to reliable, current, good quality interdepartmental information, when needed. It allows senior management to be confident in the integrity and completeness of the information as they move between an overview and a detailed view. Advanced BI systems provide reliable, comprehensive information to all interested parties and include flexible user- defined views for senior managers and planning staff, and fixed views for public access and other users.

Projects in the programme JISC Business Intelligence programme:

- **BIRD - Business Intelligence Reporting Dashboard**: Using the JISC InfoNet BI Resource for guidance, this project will work with key stakeholders to re-define the processes that deliver the evidence base to the right users at the right time and will subsequently develop the BI system using Microsoft SharePoint to deliver the user interface (linked to appropriate data sets through the data warehouse). We will use this interface to simplify the process for requesting data/analysis and will provide personalisation facilities to enable individuals to create an interface that provides the data most appropriate to their needs. As part of the project we will review the JISC InfoNet BI resource, providing recommendations to support the evaluation of the product.
- Bolt-CAP: Using the requirements of HEFCE TRAC as the base model, the JISC Business Intelligence Infokit and an Enterprise Architecture approach, this project will consider the means by which effective data capture, accumulation, release and reuse can both meet the needs of decision support within the organisation and that of external agencies.
- **Bringing Corporate Data to Life**: The aim of the project is to make use of the significant advances in software tools that utilise in- memory technologies for the rapid development of three business intelligence applications (Student Lifecycle, Corporate Performance and Benchmarking). Information in each application will be presented using a range of fully interactive dashboards, scorecards and charts with filtering, search and drill- down and drill- up capabilities. Managers will be engaged throughout the project in terms of how information is presented, the design of dashboards, scorecards and reports and the identification of additional sources of data.
- **Business Intelligence for Learning About Our Students**: The goal of this project is develop a methodology which will allow the analysis of the data in an aggregate way, by integrating information in different archives and enabling users to query the resulting archive knowledge base from a single point of access. Moreover we aim to integrate the internal information with publicly available data on socio-economic indicators as provided by data.gov.uk. Our aims are to study, on a large scale, how student backgrounds impact their future academic achievements

and to help the University devise evidence informed policies, strategies and procedures targeted to their students.

- **Enabling Benchmarking Excellence**: This project proposes to gather a set of metadata from Higher Education institutions that will allow the current structures within national data sets to be mapped to department structures within each institution. The eventual aim is to make comparative analysis far more flexible and useful to all stakeholders within the HE community. This is the first instance where such a comprehensive use of meta-data to tie together disparate functional organisations has been utilised within the sector, making the project truly innovative.
- **Engage - Using Data about Research Clusters to Enhance Collaboration:** The Engage project will integrate, visualise and automate the production of information about research clusters at the University of Glasgow, thereby improving access to this data in support of strategic decision making, publicity, enhancing collaboration and interdisciplinary research, and research data reporting.

  Aims and Objectives

  - Field Test JISC InfoNet BI resource and provide feedback to the JISC InfoNet BI team.
  - Explore definitions of what a research cluster, theme or group is and the data requirements of research clusters.
  - Create a method for visualising and presenting data associated with research clusters.
  - Share our story including any issues encountered with the sector.
- **IN-GRiD**: The project addresses the process of collection, management and analysis of building profile data, building usage data, energy consumption data, room booking data, IT data and the corresponding financial data in order to improve the financial and environmental decision making processes of the University of Manchester through the use of business intelligence. The main motivation for the project is to support decision making activities of the senior management of the University of Manchester in the area of sustainability and carbon emissions management. The project will minimise the impact on the existing university infrastructure and will facilitate the take up of the project results by the wider community through the adoption of open source technologies and through active participation in dissemination and evaluation activities in collaboration with JISC.
- **Liverpool University Management Information System (LUMIS)**: The objectives of LUMIS are to design and implement an MI solution, combining technology with data integrity, business process improvement and change management to create a range of benefits including:
  - effective monitoring and analysis of performance against KPIs
  - timely provision of information to support strategic planning and operational management
  - provision of tools for managers to support devolution of responsibility
  - improved capability for evidence-based decision making
  - standards to support accurate reporting, including statutory returns
  - creation of single sources of data and improved data collection efficiency
  - enhanced capability for data analysis, assessment and interpretation
- **RETAIN: Retaining Students Through Intelligent Interventions**: The goal of the RETAIN project is to extend the existing Business Intelligence (BI)functionality that is currently in use at the Open University (OU).

  The focus will be on using BI to improve student retention. RETAIN will make it possible to:

  - include additional data sources with existing statistical methods
  - use predictive modelling to identify ‘at risk’ students.

Methods will be trialled with a view to longer term uptake and further extensions to BI functionality. The predicted benefits are improved retention and progression, leading to a financial cost saving benefit for the Open University and a better student experience.

- **Supporting institutional decision making with an intelligent student engagement tracking system**: This project aims to examine the extent to which the adoption of a student engagement tracking system can support and enhance institutional decision making with evidence in three business intelligence (BI) data subject categories: student data and information, performance measurement and management and strategic planning. More specifically, the project will assess the current and expected institutional BI maturity level in the three chosen BI data subject categories, work with SSP to develop an open source version of the BI decision support toolkit, identify key BI implementation issues and showcasing and promotion of the toolkit to the wider HE community.

- **Visualisation of Research Strength (VoRS)**: Many HEIs now maintain repositories containing their researchers‟ publications. They have the potential to provide much information about the research strength of an HEI, as publications are the main output of research. The project aims to merge internal information extracted from an institution‟s publications repository with external information (academic subject definitions, quality of outlets and publications), for input to a visualisation tool. The tool will assist research managers in making decisions which need to be based on an understanding of research strengths across subject areas, such as where to aim internal investment. In the event that the tool becomes a part of a BI resource, It could lead to institution vs institution comparisons and visual benchmarking for research.

- **Publisher and Institutional Repository Usage Statistics 2 (PIRUS2) project**. The original PIRUS project demonstrated that it is technically feasible to create, record and consolidate usage statistics for individual articles using data from repositories and publishers, despite the diversity of organizational and technical environments in which they operate. If this is to be translated into a new, implementable COUNTER standard and protocol, further research and development will be required, specifically in the following areas:

  - Technical: further tests, with a wider range of repositories and a larger volume of data, will be required to ensure that the proposed protocols and tracker codes are scalable/extensible and work in the major repository environments.

  - Organizational: the nature and mission of the central clearing house/houses proposed in the original project has to be developed, and candidate organizations identified and tested

  - Economic: we need to assess the costs for repositories and publishers of generating the required usage reports, as well as the costs of any central clearing house/houses; investigate how these costs could be allocated between stakeholders

  - Advocacy: the broad support of all the major stakeholder groups (repositories, publishers, authors) will be required. Intellectual property, privacy and financial issues will have to be addressed

  The objective of PIRUS2 is to address these issues and by doing so specify standards, protocols, an infrastructure and an economic model for the recording, reporting and consolidation of online usage of individual articles hosted by repositories, publishers and other entities.

- Journal Usage Statistics Portal (JUSP) project. Libraries spend millions of pounds on electronic journals each year, but gathering statistics about their use hasn't always been easy. Diminishing budgets must demonstrate value for money, and reliable data is key. Comparative usage statistics help evaluate the impact of e- resources and inform future purchasing decisions. The Journal Usage Statistics Portal (JUSP) provides a "one-stop shop" for libraries to view, download and analyse their usage reports from NESLi2 publishers. It responds to current financial challenges with time and cost saving benefits.

- **Making Our Shared Activity Information Count (MOSAIC):** MOSAIC investigated the technical feasibility, service value and issues around exploiting user activity data, primarily to assist resource discovery and evaluation in Higher Education. Such activity data might be combined from:

- The circulation module of Library Management Systems (the initial project focus)
- ERM system / Resolver journal article access
- VLE resource and learning object download
- In addition, reading lists (from a variety of institutional sources, without activity data) may provide key indicators

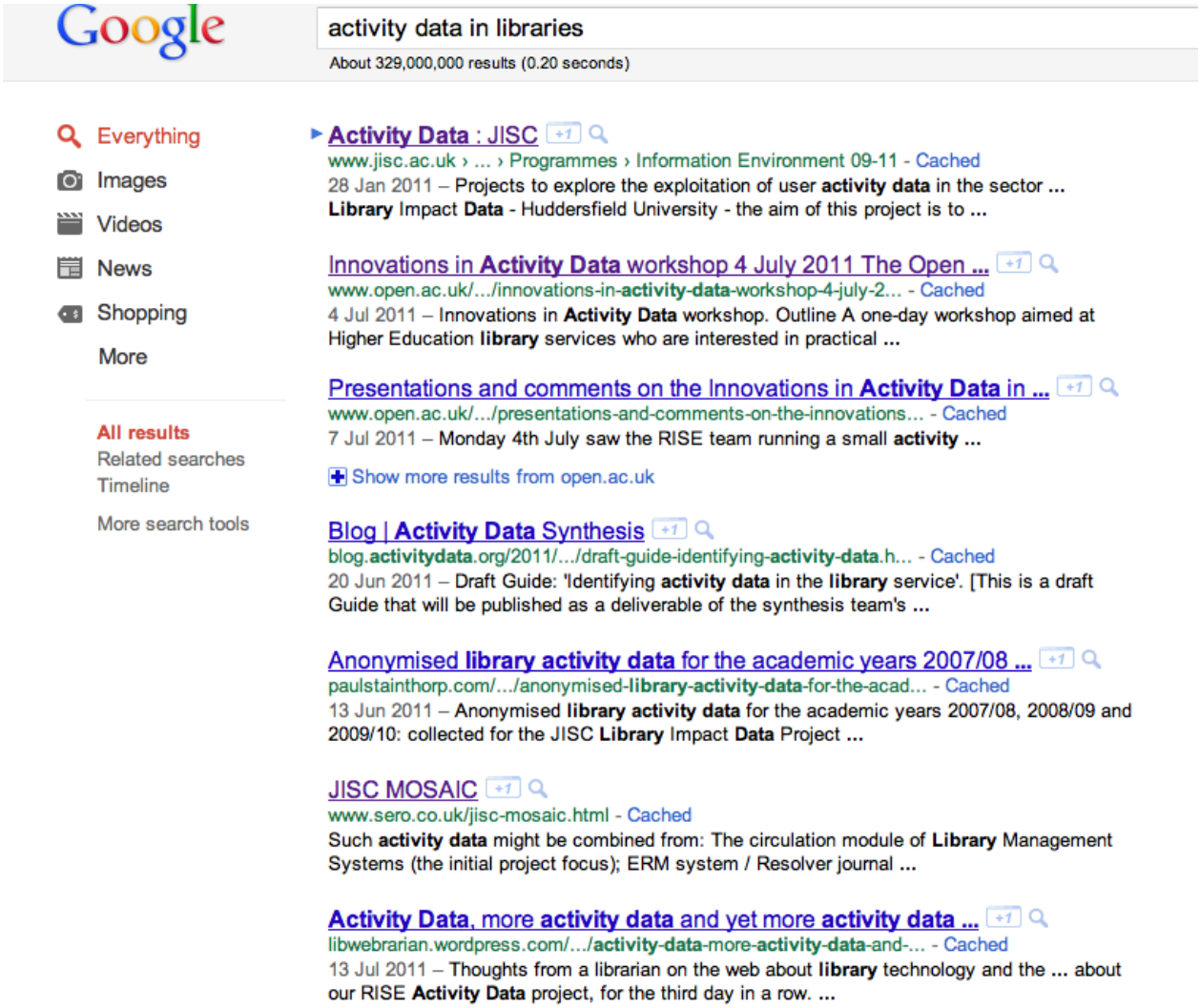
- SUSHI (the Standardized Usage Statistics Harvesting Initiative) Protocol has been developed by NISO to enable the automated harvesting of usage data, replacing the time- consuming user-mediated collection of online usage reports. The SUSHI Protocol is designed to be both generalised and extensible, meaning it can be used to retrieve a variety of usage reports. An extension is designed specifically to work with the COUNTER usage reports, which are by far the most widely retrieved usage reports. More information on SUSHI may be found on the NISO website (http://www.niso.org/workrooms/sushi/)
- **JISC Infonet** are developing a business intelligence guide which will be at http://www.jiscinfonet.ac.uk/bi

International

This section is not intended to be a comprehensive listing or review of activity data related work in UK or global Higher Education. It does however highlight work we have come across that is of particular relevance with reference to the JISC Activity Data programme.

Google searches indicate that the UK Higher Education sector to be playing a leading role in the field of activity data with reference to libraries. A search for 'activity data in libraries' (10- sep 2011) returned UK projects and events as hits 1-7. Whilst this may to some extent be true, the JISC programme reported here established valuable engagements with US projects engaged in the library field - notably LibraryCloud and Metridoc (see collaborative partners). In addition we should be aware of developments in the e-journals space, not least the MESUR project and the ExLibris bX product (see prior developments).

Furthermore, we should take semantics in to account when researching developments outside the UK and beyond the library domain. Much work elsewhere, notably but not exclusively in North America, is referenced as 'Analytics', whilst the expression 'User Tracks' has been used in Denmark. Sections Academic analytics resources from Educause and Other Publications and Online Resources note educational and learning analytics research, projects and tools that present important insights to inform UK developments.



Collaborative partners

The Activity Data programme was able to engage in ongoing dialogue with a number of North American projects, especially through the four online information sessions organised by the synthesis project. Whilst these collaborations were informal and restricted to sharing ideas rather than joint actions, they proved an important mutual asset, which we hope will continue and expand.

Harvard Library Cloud

The Library Cloud project is directed by David Weinberger and Kim Dullin of the Harvard Library Innovation Laboratory and the Harvard Law School.

In the words of co-director David Weinberger, the Library Cloud project aims "to make metadata from multiple libraries openly available in order to spur innovative applications. We're starting with circulation data and hope to move on to include other sorts of metadata as well. So far we've contacted local libraries and are just now getting our first data from them. Whilst we're quite early in the process, it is clear that we are facing issues you've been working on for a while."

LibraryCloud.local is focused on metadata about book use and reader contributions. LibraryCloud.global is intended to be an open resource built by a consortium of libraries and other knowledge- based institutions. It has no commercial aims.

What types of applications might LibraryCloud enable with reference to scholarly materials?

1. Browsing and discovery tools (such as ShelfLife)
2. Semantic-related webs
3. Recommendation engines

4. Social networks
5. Data browsing and analysis tools for librarians and those researching scholarly activities
6. Delivery systems based on profiles, recommendations, and availability

Learning Registry

Meanwhile the Learning Registry, a joint initiative between the US Departments of Defense and Education, is seeking to make federal learning resources easier to find, easier to access and easier to integrate *wherever* they are stored - and therefore more interconnected and personalized learning solutions.

The Learning Registry is creating the infrastructure that will enable projects to share their data in the public (or within a secure environment if required). This is based on a schema-free (or 'noSQL') database, which means that projects can donate their datasets without needing to transform them to meet the requirements of a specific database schema.

Steve Midgley, Deputy Director at the Office of Education Technology, is open to sharing data and exploring technical models for aggregation with any activity data projects, offering to give technical support to anyone who has data that they would like to put into the registry.

Metridoc

To aid the research library community in addressing a wide of range assessment requirements and in building infrastructure for decision support and planning, the Penn University Libraries is developing Metridoc (http://code.google.com/p/metridoc/).

Metridoc is a data integration / job framework to assist libraries with data ingestion and normalization, with the end result being a repository to help deal with usage reporting. Although the API focuses on libraries, it is also useful for writing job scripts to address database migrations, data loading and other system integration needs.

Support has come from Institute for Museum and Libraries Services (IMLS) for key portions of the project, including the development of demonstration project, a generalizable IT framework, and mechanism to disseminate code and documentation as part of an open source release of the software.

Joe Zucca (Project Lead) and Mike Winkler (University of Pennsylvania Director of Information Technologies and Digital Development) presented the project at our July workshop.

Prior developments

Appendix 1 of the MOSAIC project report (www.sero.co.uk/jisc- mosaic.html) introduces a number of seminal project and service developments, including

- California Digital Library - The Melvyl Recommender Project team explored two methods of generating recommendations. The first method used circulation data to determine linkages between items ("patrons who checked this out also checked out..."). A second, content- based, strategy used terms from the bibliographic records to develop queries for similar items ("more like this..."). The project showed strong evidence that library users are interested in receiving recommendations to support both academic and personal information needs. See http://www.dlib.org/dlib/december06/whitney/12whitney.html
- MESUR - MESUR stands for MEtrics from Scholarly Usage of Resources. The MESUR database contains **1 billion usage events** obtained from **6 significant publishers, 4 large institutional consortia and 4 significant aggregators**. MESUR produces large-scale, longitudinal maps of the scholarly community and a survey of more than 60 different metrics of scholarly impact. See http://www.mesur.org/MESUR.html
- Ex Libris bX - The bX service is the result of research into scholarly recommender systems conducted by the bX team and leading researchers Johan Bollen and Herbert Van de Sompel who were responsible for the MESUR project. Based on data captured through a large- scale aggregation of link-resolver usage logs, bX is an extension of the OpenURL framework. See http://www.exlibrisgroup.com/category/bXOverview

- BibTip - This recommender system is based on statistical evaluation of the usage data. All the data stored and processed are anonymous. See http://www.dlib.org/dlib/may08/monnich/05monnich.html#2
- The University of Minnesota Libraries - The University of Minnesota Libraries have created a 'MyLibrary' portal, with databases and e-journals targeted to users, based on their *affiliations*. For more on affinity strings see http://journal.code4lib.org/articles/501

Academic analytics resources from Educause

Educause (http://www.educause.edu/) is a strong source of material relating to the use of 'academic analytics' to support student success, which is available from their Academic Analytics page http://www.educause.edu/Resources/Browse/Academic%20Analytics/16930

What Educause calls academic analytics is very similar to what we are calling activity data, with emphasis on the tools, presentation and use. 'Academic Analytics: The Uses of Management Information and Technology in Higher Education' (below) locates academic analytics at "the intersection of technology, information, management culture, and the application of information to manage the academic enterprise."

Here we highlight five resources.

Seven Things You Should Know About Analytics

http://net.educause.edu/ir/library/pdf/ELI7059.pdf

Educause produces a series of '7 Things' reports. These are very brief and include a story / case study, definition and some of the key issues. They can be very useful introduction to those who do not already know about what you are doing, and come from an independent authoritative source.

2011 Horizon Report

http://www.educause.edu/Resources/2011HorizonReport/223122

The Educause Horizon report is produced annually and looks at technologies that are going to have an impact over the next year, two to three years and four to five years. Of course, not all the candidate technologies make it from important in four to five years to important now.

The report is part based on survey and part expert discussion and provides a very broad brush overview of the technology. The 2011 report picked out Learning Analytics for the four to five year time frame (pp28- 30). It provides a two page overview and some examples and further reading.

Signals: Applying Academic Analytics

http://www.educause.edu/EDUCAUSE+Quarterly/EDUCAUSEQuarterlyMagazineVolum/SignalsApplyingAcademicAnalyti/199385 or http://bit.ly/c5Z5Zu

This is a fascinating case study from Purdue University, indicating that the use of analytics has improved results, and led those in greatest danger of failing to switch courses earlier. They ran the trial using a control group (although they don't say how the outcome compares with the control group) and courses were sufficiently large for the results to be meaningful. Here is an extract:

"Over the succeeding weeks, 55% of the students in the red category moved into the moderate risk group (in this case, represented by a C), 24.4% actually moved from the red to the green group (in this case, an A or B), and 10.6% of the students initially placed in the red group remained there. In the yellow group, 69% rose to the green level, while 31% stayed in the yellow group"

Academic Analytics: The Uses of Management Information and Technology in Higher Education

http://www.educause.edu/ers0508

This book discusses analytics in HE and whilst dated 2005 it is still of interest. Among the things to note is the sources of data people were using in their analytics (N = 213). Note that there is no mention of Library systems of any type and the strong emphasis on administrative systems rather than academic systems.

Source	Percentage
Student information system	93.0%

Financial system	84.5%
Admissions	77.5%
HR system	73.7%
Advancement	36.2%
Course management system	29.5%
Ancillary systems (e.g., housing)	28.2%
Grants management	27.7%
Department-/school-specific system	22.5%
Comparative peer data	20.2%
Feeder institutions (high schools)	9.4%

How the ICCOC Uses Analytics to Increase Student Success

http://www.educause.edu/EDUCAUSE+Quarterly/EDUCAUSEQuarterlyMagazineVolum/HowtheICCOCUsesAnalyticstoIncr/219112

This Case Study reports on the use of analytics to improve student retention, raising it from 77% to 85%. However, it is not clear is how much of the improvement relates to the analytics and how much derives from other work to improve student success.

Other Publications and Online Resources

National Federation of Advanced Information Services

NFAIS has published Information Access and Usage Behavior in Today's Academic Environment. Papers included:

- Information Discovery on Campus: Serving Today's Users from University of Minnesota, which argues that "Users draw little distinction between discovery and delivery; systems, data, and information objects should be optimised for fulfilment." This has implications for recommender systems.
- The Netflix Effect: When Software Suggests Students' Courses discusses work at Austin Peay State University using activity data to recommend courses to students based on their previous work, and claims better grades and fewer dropouts.

Learning & Knowledge Analytics

The Learning & Knowledge Analytics blog (http://www.learninganalytics.net/?page_id=2) is devoted to this growing area of interest to educators, administrators, and entrepreneurs. Starting January 2011, this site proposed to serve as the starting point for an open course, offered to support the Learning and Knowledge Analytics 2011 Conference. The course is facilitated by George Siemens (TEKRI, Athabasca University), Jon Drown (SCIS, Athabasca University), Dave Cormier (University of Prince Edward Island), Tanya Elias (Athabasca University), and Sylvia Currie (BCcampus).

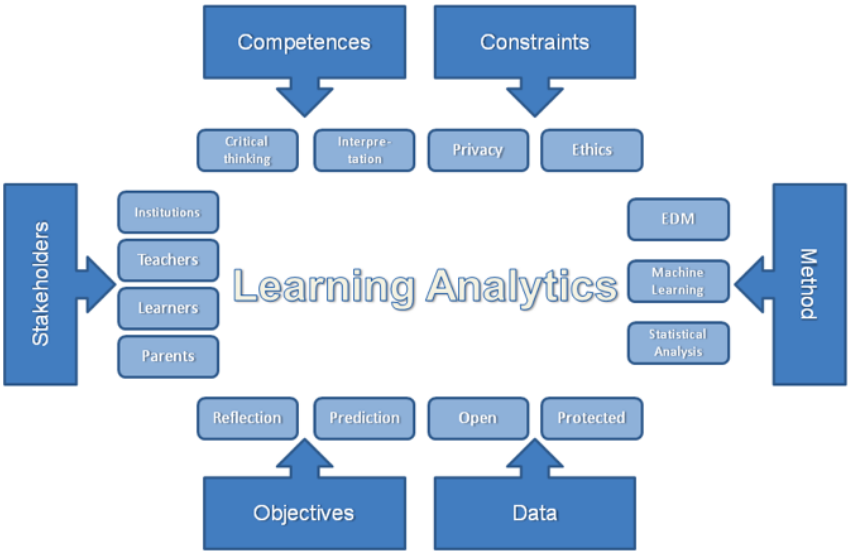
For example, George Siemens' post of 5-aug 2011 provides a useful framework for understanding activity data in the broader educational business context, mapping the space that includes 'educational data- mining', 'academic analytics' and 'learning analytics'

Type of Analytics		Level or Object of Analysis	Who Benefits?
Learning Analytics	**Educational data mining**	**Course-level:** social networks, conceptual development, discourse analysis, "intelligent curriculum"	Learners, faculty
		Departmental: predictive modeling, patterns of success/failure	Learners, faculty
Academic Analytics		**Institutional:** learner profiles, performance of academics, knowledge flow	Administrators, funders, marketing
		Regional (state/provincial): comparisons between systems	Funders, administrators
		National and International	National governments, education authorities

.

A Framework for Learning Analytics

Hendrik Drachsler and Wolfgang Greller of the Open University of the Netherlands (CELSTEC) have submitted 'Translating Learning into Numbers: Toward a Generic Framework for Learning Analytics' for publication in the Educational Technology and Society (JETS): Special Issue on Learning Analytics (ed. George Siemens). See http://www.slideshare.net/Drachsler/turning- learning-into- numbers-a- learning-analytics-framework for further insights in to their framework.



Learning Analytics framwork (Greller & Drachsler 2011)

The JISC activity data projects

- The JISC Activity data programme funded nine projects to investigate a variety of different areas of activity data. These were

Recommender systems

These are systems which offer alternative suggestions when you are searching or retrieving material.

- The AEIOU project is designed to increase the use of material in the Welsh Repository Network (a joint virtual open access repository for Welsh Universities). It is designed to encourage people to look for other material when they come in from a search rather than just retrieving a single article and moving on.
- The RISE project is designed to support learners to find a wider range of material through offering recommendations based on three possibilities (students on the same course viewed these resources, these resources may be related to the ones you recently looked at, and students using similar search terms used these resources).
- The SALT project is designed to help users to find rarely used material in libraries. While most recommender systems offer choices that are commonly associated with what the user is looking at the SALT project deliberately encourages users to consider more rarely used material.
- While the OpenURL Activity Data project did not directly develop recommender systems by exposing data from thousands of students and researchers it can form the basis of a recommender system. The data can also be used in many other ways.

Improving the student experience

These projects are concerned with supporting student success through early diagnosis of potential problems.

- The EVAD undertook early work in understanding what the activity data that is produced by their VLE means. This can be used in a number of ways including understanding how the VLE is being used, identifying which tools and facilities are being used, patterns of usage etc.
- The STAR-Trak: NG project developed a tool to help staff and students to measure and monitor their performance in order to help diagnose any issues early.
- In a rather different way the LIDP project demonstrated that there is a clear link between use of library resources and eventual degree classification.
- The RISE and SALT projects mentioned above where concerned with improving the student experience by supporting students to find material that they might otherwise not have found.

Resource management

With the current budgetary constraints it is especially important that resources are used as effectively as possible.

- The AGtivity project used usage patterns of the national videoconferencing system to help to provide better support to users by understanding where faults occurred and the patterns of usage of the various videoconferencing
- The EVAD project, mentioned above, was also concerned with determining where the VLE was being used effectively and what support the University might need to provide in order to make encourage more effective use of the VLE.
- Several of the other projects also provide data that can be used to enhance or improve the management of processes including RISE and SALT.
- A full alphabetic list of the projects is:
- Activity data to Enhance and Increase Open-access Usage (AEIOU)
- Exploiting Access Grid Activity Data (AGtivity)

- Exposing VLE activity data (EVAD)
- Library Impact Data Project (LIDP)
- Recommendations Improve the Search Experience (RISE)
- Surfacing the Academic Long Tail (SALT)
- Student Tracking And Retention (Next Generation) (STAR-Trak: NG)
- User-Centric Integration of Activity Data (UCIAD)
- Using OpenURL Activity Data

Activity data to Enhance and Increase Open-access Usage (AEIOU)

Lead institution: Aberystwyth University

Project home page: http://www.wrn.aber.ac.uk/aeiou/

Project hypothesis:

"The provision of a shared recommendation service will increase the visibility and usage of Welsh research outputs".

This will be demonstrated through quantitative and qualitative assessments:

- By a [significant] increase in attention and usage data for items held within the six core institutional repositories
- By establishing a user focus group to explore the potential of the recommendation service and its impact on repository users

The project evaluation concluded:

> "The AEIOU project has successfully developed and implemented a repository resource recommendation service. Due to considerable challenges around the project timescale it has not been possible to undertake a robust evaluation of the service to assess the extent to which its implementation would prove the project hypothesis. The positive feedback to date could be indicative that the recommendation service may encourage further use of repository resources; however, further data over a longer period would be needed to validate this.
>
> "More general feedback suggests that the service is seen as a positive enhancement to the repository and can provide benefits to users as well as exposing more of the repository content to users. "

Project description

Anecdotal evidence has suggested that the majority of user traffic into an institutional repository is generated through hits from Web search engines such as Google. Searchers will be directed towards a single item or record and have no additional incentive to look for further relevant information within the repository. The purpose of the proposed AEIOU Wales Project is to increase the visibility and usage of all academic research by aggregating Welsh institutional repository activity data to provide a “Frequently viewed together” recommendation service, such as those used by Amazon and many other e-commerce websites.

This project will utilise existing relationships and build on the national repository infrastructure established through the Welsh Repository Network (WRN), a collaborative venture between the twelve Higher Education Institutions (HEIs) in Wales. We propose to create a shared recommendation service by aggregating activity data from each WRN repository so that searchers accessing an item within an institutional repository will be presented with recommendations to further research outputs from across Wales. We will also explore the potential of promoting relationships between cross- institutional, collaborative research groups within Wales via the recommendation service, as well as enhancing the visibility and usage of all academic research by directing users to a Welsh Research Gateway.We therefore hypothesise that the creation of a shared recommendation service will increase the visibility and

usage of that research and will be demonstrated through a significant increase in attention and usage data at both an institutional and national level.

Benefits:

- Supporting student success
- System and service improvement

Other

- Project evaluation report

Exploiting Access Grid Activity Data (AGtivity)

Lead institution: University of Manchester

Project home page: http://grace.rcs.manchester.ac.uk/AGProjects/?cat=4

Project hypothesis:

The hypothesis is that by combining the usage data of these activities with other external sources, the UK Access Grid community will be able to evaluate their usage more accurately, in terms of the time nodes are used, audience sizes and environmental impact, and that they will see an overall improvement in Advanced Video Conferencing meetings through more targeted support by the AGSC staff of potentially failing nodes and meetings.

Project description

This project aims to solve a number of problems related to Advanced Video Conferencing use in the UK. These are that the users do not know when their nodes are being used and so cannot evaluate whether they need additional capacity; that administrators of groups offering teaching using these technologies cannot tell who is attending their courses; that users may not report that they are having problems with their nodes, resulting in a perception of poor quality of Advanced Video Conferencing meetings; and that users do not have access to the potential environmental savings of using these technologies over travelling to conferences.

The hypothesis is that by combining the usage data of these activities with other external sources, the UK Access Grid community will be able to evaluate their usage more accurately, in terms of the time nodes are used, audience sizes and environmental impact, and that they will see an overall improvement in Advanced Video Conferencing meetings through more targeted support by the AGSC staff of potentially failing nodes and meetings.

The use of Advanced Video Conferencing facilities as supported by the Access Grid Support Centre (AGSC) has modified over the last five years from an initial focus on large research management meetings to a more recent emphasis on multi-site distributed teaching and learning. Over the last couple of years detailed video conferencing activity data covering all these user types has been recorded. The aim of this project is to re-purpose this activity data and mash it together with external data sources to solve the problems highlighted above. This will result in both an improved service to the users, through the new ability to analyse their own service usage in terms of actual hours as well as audience sizes, and also in terms of environmental statistics from a series of diary reports; and also an improvement in the ability of the AGSC staff to recognise problems early and so improve the experience of the technologies for the entire community.

A key set of responsive users will be targeted in each of the above areas to exploit these resulting regular diary reports produced by this amalgamation. These users will be used to test the hypothesis set out above, through an evaluation performed using interviews and questionnaires, and also through further examination of the usage data captured.

The project had relatively small datasets, nevertheless, they found a wide variety of technologies and standards useful in their work. As with any such project they encountered a number of issues some of which they described as wins and fails.

Benefits:

- Service and system improvement

Recipes

Presenting data

- Plotting Calendar Data with GNUplot
- Producing PDF Reports Programmatically

Further work

The project suggests the following further work.

AGtivity: Technologies and Standards

Technologies

C++ using Standard Template Library- used to code the primary log parsers and produce mappings of hashed strings which allowed quick matching of a lot of textural information.

- C++ was used due to developer's familiarity and allow us to quickly get a first look at the data.
- The type of data available to the logging process meant that a number of hash tables are needed to correlate the different type of entries.
- Complicated by the lack of unique identifies for devices and sources.
- These could be rewritten in Python to ease maintenance.

Python + ReportLab Toolkit (PDF module) - looking for a simple and effective way to quickly produce PDF files with text, tables and images programmatically led us to look at Python as there were a number of PDF modules available.

- ReportLab appears to be the most widely used and allows production of reports with a few simple function calls.
- ReportLab requires the Python Image Library for image support.
- Look out for our PDF creation recipe!

Gnuplot - this portable and scriptable graphing package provides the features and flexibility to produce the graphs required for reporting.

- We are using the development version 4.5 as we had a found a bug in the 'with circle' rendering feature in the stable release.
- For data that clusters together in an otherwise sparse plot using transparency allows us to observe the density of plot points.
- Gnuplot can also plot vectors which meant we could show the when and how meetings were occurring in the form of a 'timetable'.
- Gnuplot scripts were created that are used to batch process data for each node and venue.

NetBeans IDE - supports a number of programming environments which has allowed it to be used for both C++ and Python developments.

- Especially useful for the integrated debugging tools and function/object completion tools when using new APIs.
- Main issue is the lack of support for Microsoft Visual Studio C++ compilers.

GoogleMaps API - easy to use but the restrictions on usage and number of requests made it unsuitable for calculating the distances between locations.

Though not intended as an outcome to this 6month project we also looked at web technologies that could provide interactive data presentation to allow users explore their own data. In addition to the popular

JQuery and JQuery UI toolkits we are very interested in providing graphs using Highcharts. Of several graphing toolkits this appears to be the most useful as it can handle dynamic data easily therefore affording the user greater control and flexibility.

Standards

PDF - for end user report production

CSV (Comma Separated Variable) - The log files are CSV and this format is also used as output after parsing and filtering since the data needs to be able to be processed by spreadsheet and graphing software. Though we could create yet another XML format we 'd still need to covert that to CSV for process by existing desktop software.

AGtivity: Wins and fails

Data corruption due to miss-typing: human data entries. The data sets we received had numerous human errors and inconsistency. People at times re- entered names but often slightly differently. Association tags had to be created in order to correlate data entries and create meaningful statistics.

Ontologies for cross-linking. A set of rules were used to associate meta- data. This did not always correlate and for example sets of virtual venue names were used for the same purpose and thus should have an ontology to describe their use as well as just their unique name.

Automate recording and log data as long as possible. Data has been logged for the last seven years, but is not complete and only over the last 10 months had detailed data logs been recorded. This data goes finer grain than currently needed involving individual streams of use. It is postulated as important to record as much as possible to aid in discovery.

Merging data from other sources is not trivial. A key useful component is to cross- merge data sets. This project used multiple sets which mostly correlated but again human miss-typing was an issue. Although these created the best serendipitous outcome they were the hardest to achieve, and also they are difficult to validate.

Minor issue was to have unified time stamps - we spent ages not realising some servers had moved to BST from GMT. Only an hour difference but caused lots of issues in calculating correlations and giving unusual results.

Holiday planning - expecting evaluations to occur within June and July is not necessarily the best of plan as carelessly users appear to want to go on holiday. Plans are to discuss outcomes beyond the project deadline and it is to be used in the future roles within JANET.

AGtivity: Numbers

Details of our size of data we have been handling. Not the biggest but quite rich.

- 10 months of detailed logging
- 122,000 entries
- 4,500 meetings
- 20,000 attendances by AG nodes
- <1sec our parser to read and process the whole 10months
- 85 virtual venues
- 400 AG nodes (munged from ~1000 badly typed names)
- <5secs for scripts to generate a PDF report with graphs and tables, for a node/venue, from the raw log file.

AGtivity: next steps

Areas that should be explored next;- likes and opportunities to be exploited, and issues to be addressed

- Integrate with other video conferencing sources; JVCS H.323, EVO, Skype.
- Continue to log at same level of detail; next stage should consider even smaller uses for example individual devices (camera feeds) could tell site involvement.
- Cross-correlate with student data attendance and completion rates: this was an unused task in the current time-table of work-plan. This should be a win-win and extra funds are being sought for this role.
- Tentative ideas on cross-linking with library and resource data downloads; mainly for course module lectures. This activity data from the lecture times would give a window of time to search for library or blackboard activity.
- Next level of statistics and visualization needs to consider trends, for example auto- correlations within longer time series data.

Exposing VLE activity data (EVAD)

Lead institution: University of Cambridge

Project home page: http://vledata.blogspot.com/

Project hypothesis:

We have four hypotheses we want to test:

- Senior stakeholders in our VLE would like richer information about VLE/ VRE usage, so that we can show growth potential, whether across the campus or in specific faculties or departments. We will test this by presenting the visualisations of our activity information to our Centre's management committee (or equivalent decision- making committees) and gathering their responses to the information, as well as obtaining their opinion on whether a case is made for a change in investment level.
- We aim to identify 'usage signatures' which indicate either skilled use of the VLE, or users who may be struggling, but who do not ask for help. In the former case, we'd like to share what they're doing; in the latter case, we'll look at the relation between access to our help documentation and our helpdesk tickets. We will test this by correlating a usage signature with the reported experiences of academics, gained via phone or email interview.
- We believe we can change our academics' attitudes towards the institutional VLE, by providing clear presentations including visualisations of activity information. We plan to test this by experimenting with different presentations of elements of our activity information to establish what the most effective and engaging presentational formats are. We will survey academics at the start and end of the project to measure their attitudes towards the VLE, which should allow us to measure our results.
- We think a comparison of VLE usage information across the universities of Cambridge, Oxford and Hull should prove valuable to the sector, as we may be able to identify similarities and differences in VLE usage which may inform future consideration of the transferability of VLE project results and concepts across institutions.

Project description

We will bring together activity and attention data for our institutional virtual learning environment (VLE) to create useful and informative management reporting including powerful visualisations. These reports will enable us to explore improvements to both the VLE software and to the institutional support services around it, including how new information can inform university valuation of VLEs and strategy in this area. We will also release anonymised datasets for use in research by others. We will reuse existing tools and methods for data analysis and visualisation as far as possible, and share our experiences and learning on a project blog in a highly accessible and reusable form. This will substantially assist other teams operating, supporting and managing VLEs in working with their own activity data, both as part of major institutional projects and within informal efforts to better understand how academics and students use VLEs.

See visualisation of the data.

Benefits:

- Supporting student success
- System and service improvement

Recipes

Processing data

- How to Pivot data in Open Office Spreadsheet

Presenting data

- How to work with Gephi to visualise a network of Sites connected by Users

Library Impact Data Project (LIDP)

Lead institution: Huddersfield University

Project home page: http://library.hud.ac.uk/blogs/projects/lidp/

Project hypothesis:

That there is a statistically significant correlation across a number of universities between library activity data and student attainment

See here for some comments from the project on the hypothesis, and further discussion of the implications of the evidence

There answer is a **YES!**

There is statistically significant relationship between both book loans and e- resources use and student attainment. And this is true across all of the universities in the study that provided data in these areas. In some cases this was more significant than in others, but our statistical testing shows that you can believe what you see when you look at our graphs and charts!

Where we didn't find a statistical significance was in entries to the library, although it looks like there is a difference between students with a 1st and 3rd, there is not an overall significance. This is not surprising as many of us have group study facilities, lecture theatres, cafes and student services in the library. Therefore a student is as just likely to be entering the library for the above reasons than for studying purposes.

We want to stress here again that we realise THIS IS NOT A CAUSAL RELATIONSHIP! Other factors make a difference to student achievement, and there are always exceptions to the rule, but we have been able to link use of library resources to academic achievement.

Project description

This project aims to prove a statistically significant correlation between library usage and student attainment. Using activity data from three separate systems and matching these against student records which are held in a fourth system, this project will build on in- house research previously undertaken at the University of Huddersfield. By identifying subject areas or courses which exhibit low usage of library resources, service improvements can be targeted. Those subject areas or courses which exhibit high usage of library resources can be used as models of good practice.The partner Universities represent a cross- section of size and mission and will provide a rich data set on which to work.

Benefits:

- Increased impact
- Student success

Recipes

Extracting data

- Extract authentication counts from the EZproxy Authentication and Access Software

- Extract circulation data from the Horizon Library Management System
- Extract entry statistics from the Sentry library gate entry system

Processing data

- Preparing attention data for processing by SPSS
- Stitching together library data with Excel

Publications

- A list of publications is available.

Further work

The project suggests the following further work.

LIDP: Thoughts on the hypothesis

Bryony Ramsden wrote about the hypothesis:

> "Since the project began, I've been thinking about all the issues surrounding our hypothesis, and the kind of things we'll need to consider as we go through our data collection and analysis.
>
> For anyone who doesn't know, the project hypothesis states that:
>
> “There is a statistically significant correlation across a number of universities between library activity data and student attainment”
>
> The first obvious thing here is that we realise there are other factors in attainment! We do know that the library is only one piece in the jigsaw that makes a difference to what kind of grades students achieve. However, we do feel we'll find a correlation in there somewhere (ideally a positive one!). Having thought about it beyond a basic level of “let’s find out”, the more I pondered, the more extra considerations leapt to mind!
>
> Do we need to look at module level or overall degree? There are all kinds of things that can happen that are module specific, so students may not be required to produce work that would link into library resources, but still need to submit something for marking. Some modules may be based purely on their own reflection or creativity. Would those be significant enough to need noting in overall results? Probably not, but some degrees may have more of these types of modules than others, so could be worth remembering.
>
> My next thought was how much library resource usage counts as supportive for attainment. Depending on the course, students may only need a small amount of material to achieve high grades. Students on health sciences/medicine courses at Huddersfield are asked to work a lot at evidence based assignments, which would mean a lot of searching through university subscribed electronic resources, whereas a student on a history course might prefer to find primary sources outside of our subscriptions.
>
> On top of these, there all kinds of confounding factors that may play with how we interpret our results:
>
> 1. What happens if a student transfers courses or universities, and we can't identify that?
> 2. What if teaching facilities in some buildings are poor and have an impact on student learning/grades?
> 3. Maybe a university has facilities other than the library through the library gates and so skews footfall statistics?
> 4. How much usage of the library facilities is for socialising rather than studying?
> 5. Certain groups of students may have an impact on data, such as distance learners and placement students, international students, or students with any personal specific needs.

For example some students may be more likely to use one specific kind of resource a lot out of necessity. Will they be of a large enough number to skew results?

6. Some student groups are paid to attend courses and may have more incentive to participate in information literacy related elements eg nurses, who have information literacy classes with lots of access to e- resources as a compulsory part of their studies.

A key thing emerging here is that lots of resource access doesn't always mean quality use of materials, critical thinking, good writing skills… And even after all this we need to think about sample sizes - our samples are self-selected, and involve varying sizes of universities with various access routes to resources. Will these differences between institutions be a factor as well?

All we can do for now is take note of these and remember them when we start getting data back, but for now I set to thinking about how I' d revise the hypothesis if we could do it again, with a what is admittedly a tiny percentage of these issues considered within it:

"There is a statistically significant correlation between library activity and student attainment at the point of final degree result"

So it considers library usage overall, degree result overall, and a lot of other factors to think about while we work on our data!"

LIDP: Implications of the evidence

As they wrote on their blog:

"We have been out and about disseminating the early findings of the LIDP project over the last few weeks. We have been delighted with the feedback we have received from conference delegates and a lot of the comments about possible future directions for research from the CILIPs, SCONUL and LIBER conferences have given us food for thought. Many of these comments will appear in the final project blog post before the end of July. However, we had the opportunity at the Business Librarians Association Conference at Sheffield (http://www.bbslg.org/2011Conference.aspx) of testing some of these thoughts. After our presentation (http://eprints.hud.ac.uk/10949/) we divided delegates up into a number of groups to discuss a variety of scenarios.

Scenario 1

If we assume a link between library usage and attainment, what does good practice look like? What are the students who gain a first doing differently to their colleagues who get lower grades? Do high achievers choose 'better' resources, or are they 'better' at choosing resources?

Two groups reported back on this scenario with the following recommendations:

- Talk to high achievers to find out what they are doing, eg

  Working

  Using data effectively

  Using the right resources

- Establish what good practice is, eg finding, using interpreting
- Consider the requirements of the subject, for example mathematics courses often require much less resource use than other subjects such as history
- Qualitative statistics need to be considered in addition to quantitative statistics
- Consider the impact of information literacy and support services
- Find out the student's own personal goals, eg why are they attending the course - as a work requirement etc.
- Look at which resources are being used, such as extended reading, not just how much

- Teach the students evaluation skills to help them find appropriate resources, not just 'better'

Scenario 2

If students are not using the library or the resources, what can we do to change their behaviour? Is non- use a resourcing issue or an academic/information skills issues? How could gender, culture and socio-economic background affect library usage and how could this be addressed? Are there scenarios where we should NOT try to increase library use?

Groups considered a number of factors that could be used to change behaviour:

- Incentives

  Attached to an assignment

  Work with and win over the academics

  Encourage student champions

  Make sure the resources are embedded and relevant to the subject

Regarding non-use, the groups thought that both issues were relevant. The skills issues required further training and the resources needed simplifying.

Gender, culture and socio-economic background were themes brought out at both the SCONUL and LIBER conferences. One group looked at international students where it was considered that they were too dependent on Google - does this means our resources are too difficult to understand? It was also considered that there is a focus on generalisations, eg international students, rather than looking at individuals. Another group considered that it was a cultural issue and that students were guided to the 'right answer' via reading lists, rather than reading around the subject.

Finally discussion turned to work-life balance and whether students should be logging in at 2am, and whether our culture of 24×7 access was a healthy one.

Scenario 3

Can we actually demonstrate that the library adds value? eg if a student enters university with average UCAS points and attains a first class degree having used library resources to a high level, does this prove the library has added value to the student achievement? Have we done anything? Do they need us?

The short answer to this scenario was yes!

We receive feedback, both internal and external and have provided learning spaces and essential resources at the very least. We can also show that we have promoted our services and embedded information literacy skills into the curriculum by working successfully with academic staff. It was thought that we add to the employability of students by teaching them research skills and giving certification, eg Bloomberg etc.

Scenario 4

If the hypothesis is proved to be correct, does cutting library budgets mean that attainment will fall? Is this something that can be used at director level to protect resource budgets/subject librarians? Should we be concerned about implications for publishers if the hypothesis is proven?

The group that looked at this scenario considered that further use of statistics were required to find out what students were reading. This would allow stock to be rationalised and the reduced budget could be used to better target appropriate resources.

In addition it was suggested that other services such as, inductions and information literacy training by audited and evaluated in order to provide more effective targeting.

It was also felt that there was an absolute minimum spend for resources, once this level was passed impact would be huge with insufficient resources to support courses.

The group felt that this could be used at Director level and that evidence would be required to support this.

Big deals came up in the final point from this scenario. Discussion centred on a standoff between the need for better products verses ongoing financial commitments

Many thanks to all the delegates for allowing us to blog about their comments and to the BLA for letting us loose at their conference. We'll be adding some of these comments to our final blog post."

LIDP: Articles and conference papers

Journal articles

Stone, Graham, Pattern, David and Ramsden, Bryony (2011) Does library use affect student attainment? A preliminary report on the Library Impact Data Project. LIBER quarterly (accepted)

Stone, Graham, Ramsden, Bryony and Pattern, David (2011) Looking for the link between library usage and student attainment. ARIADNE (in press)

Goodall, Deborah and Pattern, David (2011) Academic library non/low use and undergraduate student achievement: a preliminary report of research in progress. Library Management, 32 (3). ISSN 0143- 5124 http://eprints.hud.ac.uk/7940/

Goodall, Deborah, Pattern, David and Stone, Graham (2010) Making resources work harder. Library and Information Gazette . p. 5. ISSN 1741220X http://eprints.hud.ac.uk/8453/

White, Sue and Stone, Graham (2010) Maximising use of library resources at the University of Huddersfield. Serials, 23 (2). pp. 83-90. ISSN 0953-0460 http://eprints.hud.ac.uk/7811/

Conferences

Pattern, David (2011) If you want to get laid, go to college… In: Welsh Libraries, Archives and Museums Conference , 12 -13 May 2011, The Metropole, Llandrindod, Wales. http://eprints.hud.ac.uk/10506/

Adams, Philip (2011) Library Impact Data Project update. In #MashDMU, 17 May 2011, De Montfort University. http://www.slideshare.net/Libran/library- impact-data- project- update

Stone, Graham (2011) Looking for the link between library usage and student attainment. In CILIPS Annual Conference, Glasgow, 7 June 2011. http://eprints.hud.ac.uk/10655/

Stone, Graham (2011) There's gold in them there hills. In SCONUL Conference, Cardiff University, 8- 10 June 2011. http://eprints.hud.ac.uk/10654/

Pattern, David (2011) If you want to get laid, go to college… In Welsh Higher Education Library Forum colloquium, Gregynog Hall, Wales, 14 June 2011. http://eprints.hud.ac.uk/11003/

Stainthorp, Paul (2011) Making an impact: the JISC Library Impact Data Project (LIDP) In UC&R East Midlands Members' Day, Kimberlin Library, De Montfort University, Leicester, 28 June 2011. http://www.slideshare.net/pstainthorp/making-an- impact-the- jisc-library- impact- data- project-lidp

Stone, Graham, Pattern, David and Ramsden, Bryony (2011) Does library use affect student attainment? A preliminary report on the Library Impact Data Project. In: LIBER 40th Annual Conference, 29 June - 2 July 2011, Universitat Politècnica de Catalunya, Barcelona. http://eprints.hud.ac.uk/10208/

Stone, Graham, Pattern, David and Ramsden, Bryony (2011) Business Librarians Conference, University of Sheffield, July 2011. http://eprints.hud.ac.uk/10949/

Forthcoming conferences

Stone, Graham, Pattern, David and Ramsden, Bryony (2011) The Library Impact Data Project: hit miss or maybe. In: 9th Northumbria International Conference on Performance Measurement in Libraries and

Information Services: Proving value in challenging times, 22- 25 August 2011, University of York. http://eprints.hud.ac.uk/10210/

Stone Graham (2011) In National Acquisitions Group Annual Conference, Manchester, 7- 8 September 2011

Ramsden, Bryony and Pattern, Dave (2011) Internet Librarian International Conference, London, 27- 28 October

LIDP: Next steps

Although this project has had a finite goal in proving or disproving the hypothesis, we would now like to go back to the original project which provided the inspiration. This was to seek to engage low/non users of library resources and to raise student achievement by increasing the use of library resources.

This has certainly been a popular theme in questions at the SCONUL and LIBER conferences, so we feel there is a lot of interest in this in the library community. Some of these ideas have also been discussed at the recent Business Librarians Association Conference.

- There are a number of ways of doing this, some based on business intelligence and others based on targeting staffing resources. However, we firmly believe that although there is a business intelligence string to what we would like to take forward, the real benefits will be achieved by actively engaging with the students to improve their experience. We think this could be covered in a number of ways.
- Gender and socio-economic background? This came out in questions from library directors at SCONUL and LIBER. We need to re-visit the data to see whether there are any effects of gender, nationality (UK, other European and international could certainly be investigated) and socio-economic background in use and attainment.
- We need to look into what types of data are needed by library directors, eg for the scenario 'if budget cuts result in less resources, does attainment fall'? The Balanced Scorecard approach could be used for this?
    - We are keen to see if we add value as a library through better use of resources and we have thought of a number of possible scenarios in which we would like to investigate further:
    - Does a student who comes in with high grades leave with high grades? If so why? What do they use that makes them so successful?
    - What if a student comes in with lower grades but achieves a higher grade on graduation after using library resources? What did they do to show this improvement?
    - Quite often students who look to be heading for a 2nd drop to a 3rd in the final part of their course, why is this so?
    - What about high achievers that don't use our resources? What are they doing in order to be successful and should we be adopting what they do in our resources/literacy skills sessions?
- We have not investigated VLE use, and it would be interesting to see if this had an effect
- We have set up meetings with the University of Wollongong (Australia) and Mary Ellen Davis (executive director of ACRL) to discuss the project further. In addition we have had interest from the Netherlands and Denmark for future work surrounding the improvement of student attainment through increased use of resources

In respect to targeting non/low users we would like to achieve the following:

- Find out what students on selected 'non-low use' courses think to understand why students do not engage
- To check the amount and type of contact subject teams have had with the specific courses to compare library hours to attainment (poor attainment does not reflect negatively on the library support!)

- Use data already available to see if there is correlation across all years of the courses. We have some interesting data on course year, some courses have no correlation in year one with final grade, but others do. By delving deeper into this we could target our staffing resources more effectively to help students at the point of demand.
    - To target staffing resources
- Begin profiling by looking at reading lists
    - To target resource allocation
    - Does use of resources + wider reading lead to better attainment - indeed, is this what high achievers actually do?
- To flesh out themes from the focus groups to identify areas for improvement
    - To target promotion
    - Tutor awareness
    - Inductions etc.
- Look for a connection between selected courses and internal survey results/NSS results
- Create a baseline questionnaire or exercise for new students to establish level of info literacy skills
    - Net Generation students tend to overestimate their own skills and then demonstrate poor critical analysis once they get onto resources.
    - Use to inform use of web 2.0 technologies on different cohorts, eg health vs. computing
- Set up new longitudinal focus groups or re-interview groups from last year to check progress of project
- Use data collected to make informed decisions on stock relocation and use of space
- Refine data collected and impact of targeted help
- Use this information to create a toolkit which will offer best practice to a given profile
    - eg scenario based

Ultimately our goal will be to help increase student engagement with the library and its resources, which as we can now prove, leads to better attainment. This work would also have an impact on library resources, by helping to target our precious staff resources in the right place at the right time and to make sure that we are spending limited funds on the resources most needed to help improve student attainment.

How can others benefit?

There has been a lot of interest from other universities throughout the project. Some universities may want to take our research as proof in itself and just look at their own data; we have provided instructions on how to do this at http://library.hud.ac.uk/blogs/files/lidp/Documentation/DataRequirements.pdf. We will also make available the recipes written with the Synthesis project in the documentation area of the blog, we will be adding specific recipes for different library management systems in the coming weeks: http://library.hud.ac.uk/blogs/projects/lidp/documentation/

For those libraries that want to do their own statistical analysis, this was a was a complex issue for the project, particularly given the nature of the data we could obtain vs. the nature of the data required to specifically find correlations. As a result, we used the Kruskal Wallis (KW) test, designed to measure whether there are differences between groups of non- normally distributed data. To confirm non-normal

distribution, a Kolmogorov- Smirnov test was run. KW unfortunately does not tell us where differences are, the Mann Whitney test was used on specific couplings of degree results, selected based on visual data represented in boxplot graphs. The number of Mann Whitney tests have to be limited as the more tests conducted, the higher the significance value required, so we limited them to three (at a required significance value of 0.0167 (5% divided by 3)). Once Mann Whitney tests had been conducted, effect size of the difference was calculated. All tests other than effect size were run in PASW 18; effect size was calculated manually. It should be noted that we are aware the size of the samples we are dealing with could have indicated relationships where they do not exist, but we feel our visual data demonstrates relationships that are confirmed by the analytics, and thus that we have a stable conclusion in our discarding of the null hypothesis that there is no relationship between library use and degree result.

Full instructions of how the tests were run will first be made available to partner institutions and disseminated publicly through a toolkit in July/August

Recommendations Improve the Search Experience (RISE)

Lead institution: Open University

Project home page: http://www.open.ac.uk/blogs/RISE/

Project hypothesis:

That recommender systems can enhance the student experience in new generation e- resource discovery services.

This hypothesis was chosen quite carefully for a number of reasons. We've only recently implemented our Ebsco Discovery Solution aggregated search system so we are still in an evaluation stage and are really still assessing how students at the OU will get the best out of the new system. We have a particular perspective at the Open University in that the use that students make of our search systems varies widely from course to course. So we will particularly want to look at whether there is variation between the levels of students in their reaction to recommendations.

There is also a discussion of the method for evaluating the hypothesis.

It has shown that recommendations can be made from EZProxy data, that users like recommendations and that overall there is a value in showing recommendations to users of new generation discovery solutions.

Project description

As a distance-learning institution, students, researchers and academics at the Open University mainly access the rich collection of library resources electronically. Although the systems used track attention data this data isn"t used to help users search. While library loans data was made available through the JISC MOSAIC project there is no comparable data available openly of e- resource use. RISE aims to exploit the unique scale of the OU (with over 100,000 annual unique users of e- resources) by using attention data recorded by EZProxy to provide recommendations to users of the EBSCO Discovery search solution. RISE will then aim to release that data openly so it can be used by the community. The project will also test the feasibility of deploying recommendations in the OU Google Apps environment and potentially into the institutional VLE by building a library search and recommendation Google Gadget.

Project Objectives

The overall objectives of the RISE project are to:

- Establish a process to collect and analyse attention data about the use of electronic resources from the EZProxy proxy referral system.
- Create a recommender service using attention data to answer questions such as *'people on my course are looking at these resources'*
- Identify metrics to detect changes in user behaviour as a result of service use.
- RISE will create a personal recommendations service, **MyRecommendations** for OU users of the EBSCO Discovery Solution (EDS).

- It will explore issues (of anonymity, privacy, licensing and data format/ standards) around making this data available openly and will aim to release it openly so it can be re-used by the wider community in innovative ways.
- RISE will use the EDS API to create a Google Gadget for the OU Google Apps environment and will aim to test in the OU Moodle Virtual Learning Environment (VLE) using features developed by the JISC DOULS project.
- RISE will evaluate the pros and cons of providing recommender data to students of an e-resource discovery service.
- Overall RISE will provide the wider community with an example of the benefits to users of discovery solutions of using e-resource activity data, will aim to make that data available to the wider community, and will provide a tool that can be adapted and reused.

Benefits:

- Student success

Recipes

Extracting data

- Extract user ID and course code from CIRCE MI

Processing data

- Process EZProxy log
- Provide course based recommendations
- Provide relationship based recommendations
- Search term based recommendations

Further work

The project suggests the following further work.

RISE: How do we plan to evaluate the hypothesis?

We are planning to approach the evaluation in three ways.

- By establishing some website metrics to allow us to assess how user behaviour is affected by the recommendations. We expect to build two versions of the search interface, one with recommendations and one without. This will allow us to A/B test the interfaces, so we can track the impact that different options make on users behaviour. Using Google Analytics we will track where users click and where they go.
- We will track the use that users make of the rating feature to see whether there is evidence that it is actively being used.
- We will actively encourage user feedback on the tools, carry out surveys with students and run a short series of focus groups to test the hypothesis.

As part of our evaluation work we will be looking to assess whether there are variations that can be ascribed to course or course level in how useful students find the recommendations to be, and whether there are circumstances where they are not useful. We will also be testing a variety of different types of recommendations and will aim to assess which are found to be most useful.

The evaluation report will detail the activities and results of the work to test the hypothesis and we will look at using Quora to record evidence.

To talk in a bit more detail about the A/B testing. It is possible to handle this within Google's tools by coding the elements of the page so they are displayed 'randomly' to users and the different options are

shown the same amount of times. So typically we would show different types of recommendations. So one user might be shown recommendations such as people on your course are looking for this, another might see 'you might like these similar articles'. This page on Google explains a bit more http://www.google.com/support/websiteoptimizer/bin/static.py?page=guide.cs&guide=29619&topic=29621

What we would be looking at is to see how user behaviour changes. Which types of recommendations are looked at most often, how often do users go for the recommendations rather than the basic search results.

We intend to back this up with some more focused testing with groups of users, where we know more about the courses they are studying. What we also have is the recommendations database that we are building and that will give us some evidence of user behaviour which we will be looking at.

RISE: Next steps

A major barrier to the open release of data from RISE has been the lack on open article level metadata. One future area of work could be to encourage providers to open up their article level metadata to allow others to build services that use it. It isn't always clear what can be done with data that can be obtained through APIs and web services and it would be helpful to have a resource that recorded what different data is out there, how it can be accessed (i.e. what record keys do you need), what data can be retrieved, and what you can or cannot do with the data.

Opening up activity data to an extent that it could be aggregated is clearly in the very early stages. It would be useful if some standards and formats were established and agreed that could be used and then applied systems. Part way through the RISE project the OpenURL project at EDINA released the first batch of OpenURL data and we did a comparison between the data stored by RISE and the EDINA OpenURL data. If you start with EZProxy data then there is very little cross- over with the OpenURL standard.

Surfacing the Academic Long Tail (SALT)

Lead institution: MIMAS, University of Manchester

Project home page: http://salt11.wordpress.com

Project hypothesis:

Library circulation activity data can be used to support humanities research by surfacing underused "long tail" library materials through search.

Overwhelmingly, the groups found the recommender useful. They were keen that their comments be fed back to developers and that work should continue on the recommender to get the results right as they were keen to use it and hoped it would be available soon. Further details can be found in the evaluation.

Project description

SALT will test the hypothesis that Library circulation activity data can be used to support humanities research by surfacing underused 'long tail' library materials through search. We will investigate how issues of relevance and frequency of borrowing might shift within the particular use case of humanities research. An API onto JRUL's ten years of circulation data will be made available to the HE/FE community, and the project will pay specific attention to the sustainability of an API service as a national shared service for HE/FE that both supports users and drives institutional efficiencies (eg collections management).

Working with ten years+ of aggregated and anonymised circulation data amassed by JRUL. Our approach will be to develop an API onto that data, which in turn we'll use to develop the recommender functionality in both services.

Our overall aim is that by working collaboratively with other institutions and Research Libraries UK, the SALT project will advance our knowledge and understanding of how best to support research in the 21st century. Libraries are a rich source of valuable information, but sometimes the sheer volume of materials they hold can be overwhelming even to the most experienced researcher — and we know that researchers' expectation on how to discover content is shifting in an increasingly personalised digital world. We know that library users — particularly those researching niche or specialist subjects — are

often seeking content based on a recommendation from a contemporary, a peer, colleagues or academic tutors. The SALT Project aims to provide libraries with the ability to provide users with that information. Similar to Amazons, 'customers who bought this item also bought….' the recommenders on this system will appear on a local library catalogue and on COPAC and will be based on circulation data which has been gathered over the past 10 years at The University of Manchester's internationally renowned research library.

Benefits:

- Service and system improvement

Recipes

Extracting data

- Extract anonymised loan data from Talis Alto

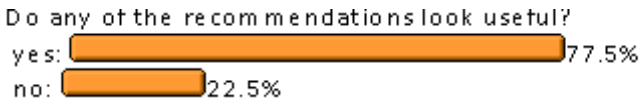
Further work

The project suggests the following further work.

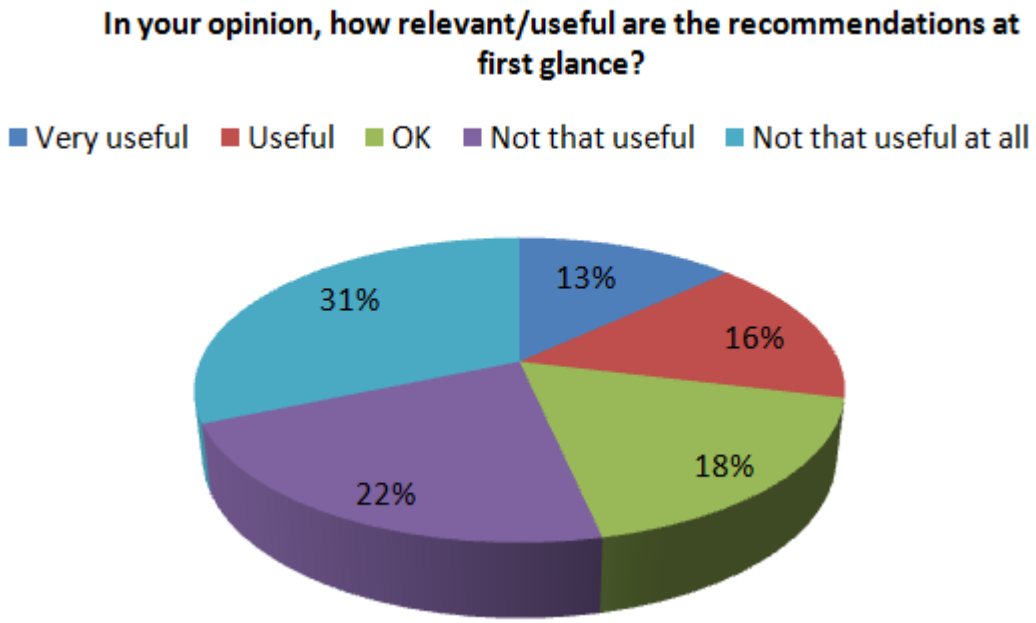
SALT: Evaluation

The project evaluated the hypothesis, writing:

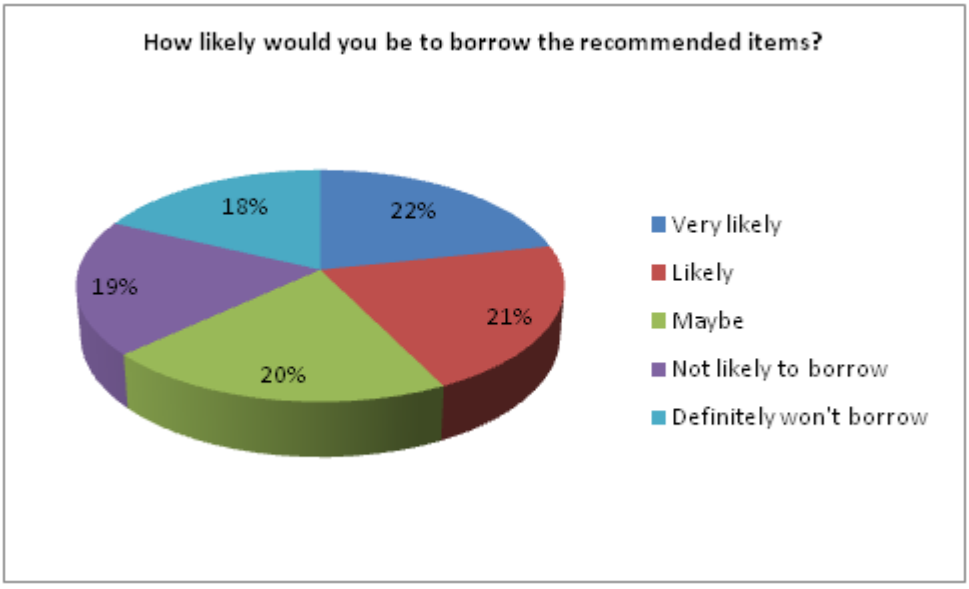
> Following internal in house testing the recommender was open to the users. In the last week of July 18 humanities postgraduates passed through the SALT testing labs, (11 PhD students, 3 taught Masters students and 4 research students). Lisa and I held three focus groups and grilled our potential users about the SALT recommender. The research methods used were designed to answer our objectives, with an informal discussion to begin with to find out how postgraduate students approach library research and to gauge the potential support for the book recommender. Following the discussion we began testing the actual recommender to answer our other research objectives which were:
>
> - Does SALT give you recommendations which are logical and useful?
> - Does it make you borrow more library books?
> - Does it suggest to you books and materials you may not have known about but are useful and interesting?
>
> As a team we agreed to set the threshold of the SALT recommender deliberately low, with a view to increasing this and testing again if results were not good. **As our hypothesis is based on discovering the hidden long tail of library research we wanted the recommender to return results that were unexpected - research gems that were treasured and worthy items but had somehow been lost and only borrowed a few times.**
>
> 42 searches in total were done on the SALT recommender and of those 42, **77.5% returned at least one recommendation, (usually many more) that participants said would be useful.** (As an aside, one of the focus groups participants found something so relevant she immediately went to borrow it after the group has finished!)



> However the deliberately low threshold may have caused some illogical returns. The groups were asked to comment on the relevance of the first 5 recommendations, but **quite often it was the books further down the list that were of more relevance and interest.** One respondent referred to it as a 'Curate's egg' however, assured me this was in reference to some good and some bad. His first five were of little relevance, 'only tangentially linked', his 6th, 7th, 8th, 9th, 11th and even 17th recommendations were all 'very relevant'. Unfortunately this gave disappointing results when the first 5 suggested texts were rated for relevance, as demonstrated in the pie chart below.



However the likelihood of borrowing these items gave slightly more encouraging results;



Clearly we've been keen on the threshold count. Lessons need to be learnt about the threshold number and this perhaps is a reflection of our initial hypothesis. We think that there would be much merit in increasing the threshold number and retesting.

On a positive note, initial discussions with the researchers (and just a reminder these are seasoned researchers, experts in their chosen fields familiar and long term users of the John Ryland's University Research Library) told us that the recommender would be a welcome addition to COPAC and the library catalogue. **99% of the researchers in the groups had used and were familiar with Amazons recommender function and 100% would welcome a similar function on the catalogues based on circulation records.**

Another very pertinent point, and I cannot stress this strongly enough, was the reactions expressed in regards to privacy and collection and subsequent use of this data. The groups were slightly bemused by questions regarding privacy. **No one expressed any concern about the collection of activity data and its use in the recommender.** In fact most assumed this data was collected anyway and encouraged us to use it in this way, as ultimately it is being used to develop a tool which helps them to research more effectively and efficiently.

Overwhelmingly, the groups found the recommender useful. They were keen that their comments be fed back to developers and that work should continue on the recommender to get the results right as they were keen to use it and hoped it would be available soon.

SALT: next steps

There are a number of steps that can be taken as a result of this project - some imminent 'quick wins' which we plan to take on after the official end, and then others that are 'bigger' than this project.

What we plan to do next anyway:

- Adjust the threshold to a higher level (using the 'usefulness' benchmark given to us as users as a basis) so as to suppress some of the more off- base recommendations our users were bemused by.
- Implement the recommender in the JRUL library search interface.
- Once the threshold has been reset, consider implementing the recommender as an option feature in the new COPAC interface. We 'd really like to, but we 'd need to assess if the results are too JRUL- centric.
- Work with JRUL to determine most appropriate mechanisms for hosting the data and supporting the API in the longer term (decisions here are dependent on how, if at all, we continue with this work from a Shared Services perspective)
- Work with JRUL to assess the impact of this in the longer term (on user satisfaction, and on borrowing behaviour)

The Big Picture (what else we 'd like to see happen):

1. **Aggregate more data**. Combine the normalised data from JRUL with processed data from additional libraries that represent a wider range of institutions, including learning and teaching. Our hunch is that only a few more would make the critical difference in ironing out some of the skewed results we get from focusing on one data set (i.e. results skewed to JRUL course listings)

2. **Assess longer term impact**. Longer-term analysis of the impact of the recommender functionality on JRUL user satisfaction and borrowing behaviour. Is there, as with Huddersfield, more borrowing from 'across the shelf'? Is our original hypothesis borne out?

3. **Requirements and costs gathering for a shared service**. Establish the requirements and potential costs for a shared service to support processing, aggregation, and sharing of activity data via an API. Based on this project, we have a fair idea of what those requirements might be, but our experience with JRUL indicates that such provision need to adequately support the handling and processing of large quantities of data. How much FTE, processing power, and storage would we need if we scaled to handling more libraries? Part of this requirements gathering exercise would involve identifying additional contributing libraries, and the size of their data.

4. **Experiment with different UI designs and algorithm thresholds to support different use cases.** For example, undergraduate users vs 'advanced' researcher users might benefit from the thresholds being set differently; in addition, there are users who want to see items held elsewhere and how to get them vs those who don't. Some libraries will be keen to manage user expectations if they are 'finding' stock that's not held at the home institution.

5. **Establish more recipes** to simplify data extraction from the more common LMS's beyond Talis (Horizon, ExLibris Voyager, and Innovative).

6. **Investigate how local activity data can help collections managers** identify collection strengths and recognise items that should be retained because of association with valued collections. We thought about this as a form of "stock management by association." Librarians might treat some long- tail items (eg items with limited borrowing) with caution if they were aware of links/associations to other collections (although there is also the caveat that this wouldn't be possible with local activity data reports in isolation)

7. **More ambitiously, investigate how nationally aggregated activity data could support activities such as stock weeding by revealing collection strengths or gaps** and allowing librarians to cross check against other collections nationally. This could also inform the number of copies a library should buy, and which books from reading lists are required in multiple copies.

8. **Learning and teaching support.** Explore the relationship between recommended lists and reading lists, and how it can be used as a tool to support academic teaching staff.

9. **Communicate the benefits to decision-makers.** If work were to continue along these lines, then a recommendation that has come out strongly from our collaborators is the need to accompany any development activity with a targeted communications plan, which continually articulates the benefits of utilising activity data to support search to decision- makers within libraries. While within our community a significant amount of momentum is building in this area, our meetings with librarians indicates that the 'why should I care?' and more to the point 'why should I make this a priority?' questions are not adequately answered. In a nutshell, 'leveraging activity data' can easily fall down or off the priority lists of most library managers. It would be particularly useful to tie these benefits to the strategic aims and objectives of University libraries as a means to get such work embedded in annual operational planning.

Student Tracking And Retention (Next Generation): STAR-Trak: NG

Lead institution: Leeds metropolitan University

Project home page: http://leedsmetstartrak.wordpress.com/

Project hypothesis:

That by harvesting user activity (usage and attention) data that already resides in existing institutional systems, combining it with demographic information and presenting it in a single portal- type application, we can improve our support services by revealing new information, providing students, tutors and student support officers with a broader picture of a student's engagement with the University at both an academic and social level.

Furthermore, being able to predict students at risk of dropping out, based on lack of engagement, will enable us to develop targeted personalised interventions appropriate to the type and level of non-engagement, and do so in a more joined-up and timely manner.

Project description

This project will provide an application (STAR-Trak:NG) to highlight and manage interventions with students who are at risk of dropping out, identified primarily by mining student activity data held in corporate systems. STAR- Trak:NG represents a synthesis and extension of the outputs of two JISC projects: our current STAR- Trak (Student Tracking And Retention) proof-of- concept application, and MCMS (Mining Course Management. System) developed by TVU. Whilst far more sophisticated than STAR- Trak, MCMS is currently tightly bound to TVUs IT architecture, source systems for activity data, business processes and culture, and hence is not yet reusable by other institutions. In summary, our project will:

- Take the best elements of MCMS into Star-Trak as "STAR-Trak:NG" to work with user activity data, business processes, structures and cultures at Leeds Metropolitan
- Provide a stepping-stone, supported by TVU and Coventry Universities, to creating an open source application that can, with future development and support through JISC and OSS- Watch, be shared across the HE and FE sectors

A fuller illustrated description can be found here.

Benefits:

- Student success

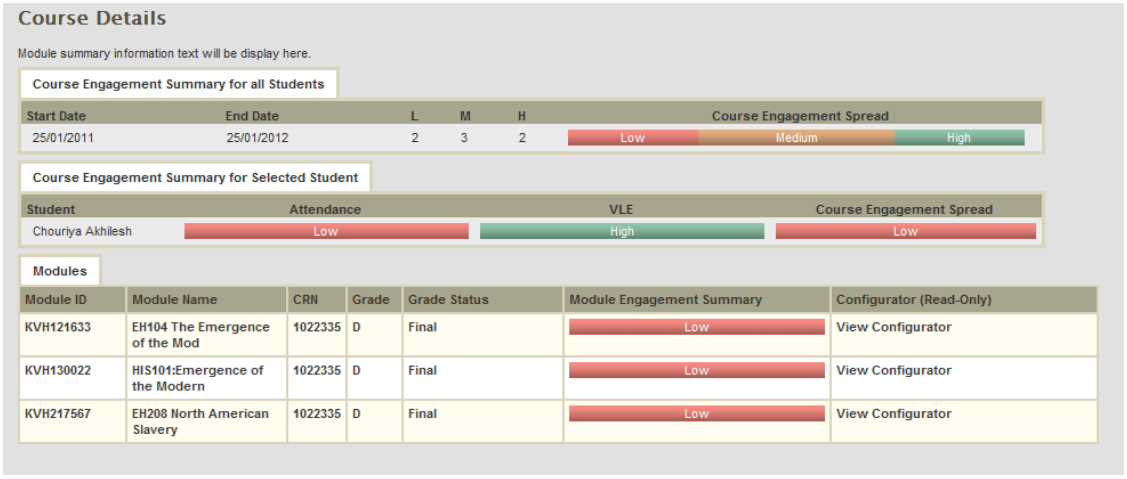
Further work

The project will be continuing. See Next steps.

STAR-Trak: NG

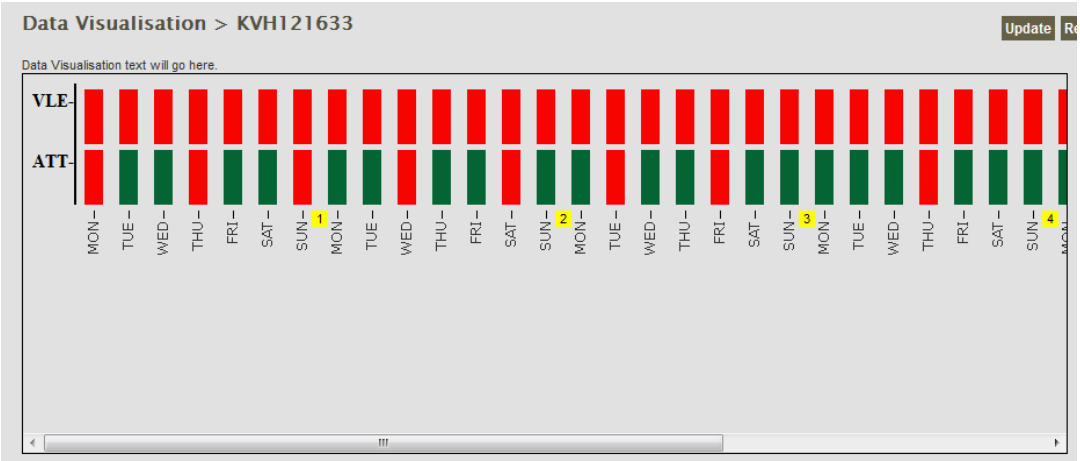
STAR-Trak: NG developed a dashboard, which amongst many other things, allows staff and students to see how their actions compare to other members of the same cohort and correlations between those actions and later outcomes (module and degree results). The aim is to provide an early warning to both staff and students and to enable appropriate interventions that may boost students' performance. To do

this they look at each student's performance and compare it against those of the whole cohort using a set of criteria to weigh the different aspects of this. Students can both view the results on their personal dashboard as shown below and, if they want, receive email warnings when their status changes together with weekly reminders of their status for each module.

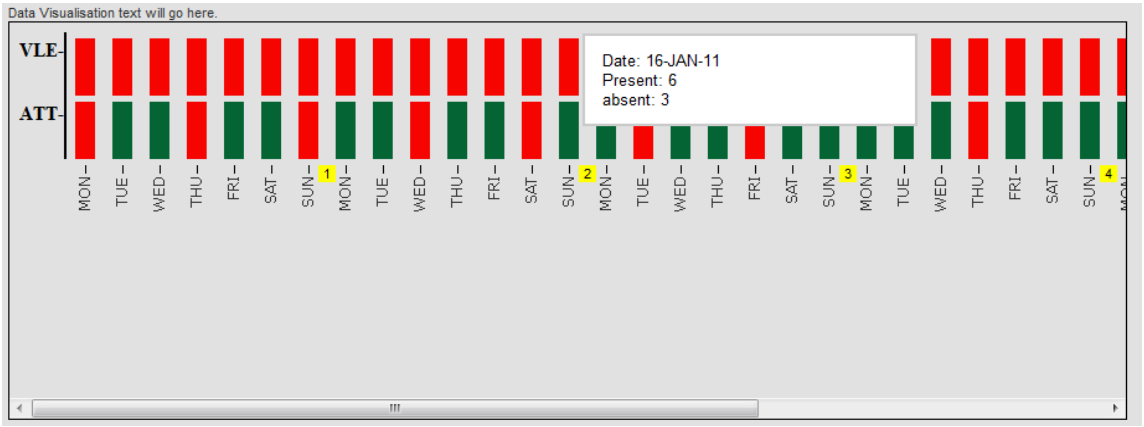


Personal dashboard from STAR-Trek: NG

Tutors are also able to look at the overall performance of students on particular activities, and then drill down and see who has, and has not been active.

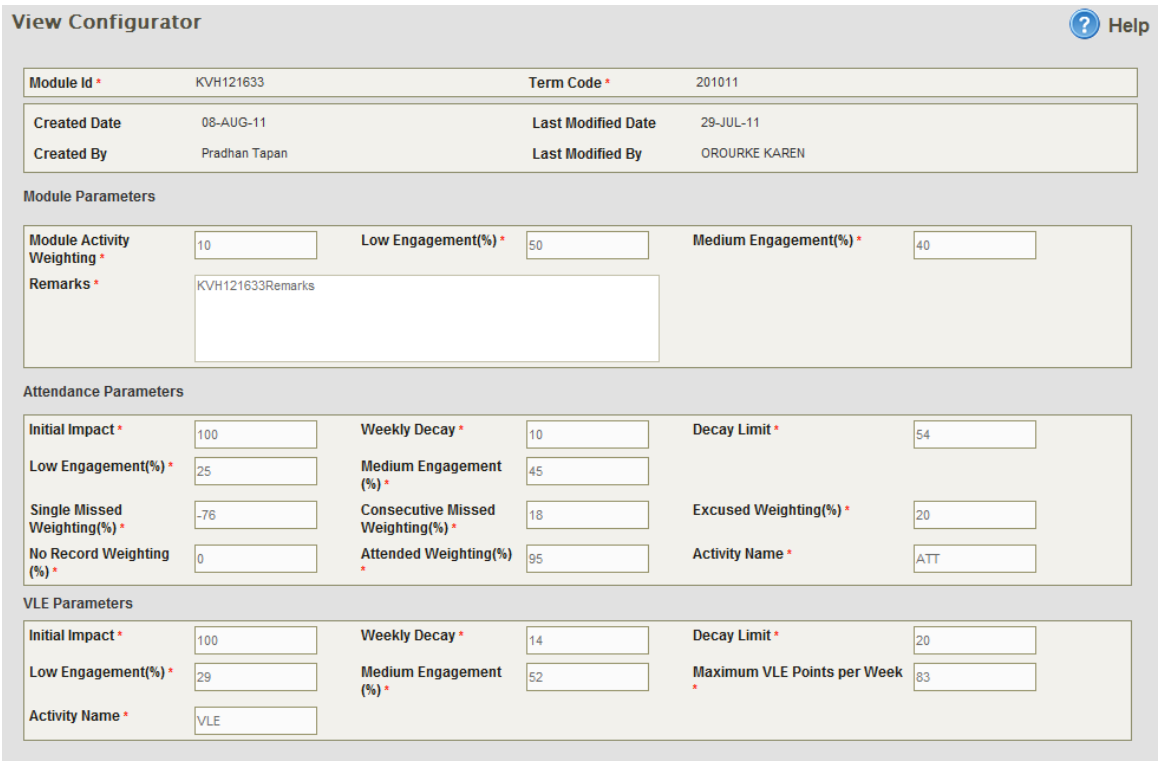


Engagement in studies: use of VLE and attendance at lectures



Engagement in studies: use of VLE and attendance at lectures showing attendance figures for Sunday Week 2.

In order to determine students' status it is necessary to configure the system to understand how important various attributes of performance are. For instance, missing a single lecture might not be very significant, whilst missing two consecutive ones might be a much more significant indicator of concern. Similarly, missing a recent lecture might be significant, but four or five weeks later this might be of only marginal interest. The criteria will vary between courses. For instance in distance courses the use of the VLE may be more significant than for campus based courses. They have therefore developed a tool that allows lecturers to specify the importance of each type of learning activity and how that importance decays with time. It is likely that there will be a need for a set of standardised templates for different types of course and different subject areas and that significant training will be required to correctly configure the tool. Equally, it may be possible to use historical data to arrive at appropriate criteria where modules have been run sufficiently long with enough students to ensure that the data is meaningful.



STAR-Trak: NG configurator tool

STAR-Trak: NG next steps

The JISC Activity Data project may be completed, but STAR-Trak lives on! We have also been involved in another JISC programme - Flexible Service Delivery (FSD) and have just completed our SOAP Opera 2 project. SOAP Opera 2 has enabled us, amongst other things, to develop our understanding of what enterprise architecture can mean for Leeds Metropolitan and we plan to bring STAR-Trak into a wider "innovation programme" that will enable us to develop our skills in a number of related areas including:

- Data warehouse/ business intelligence
- Extract, transform and load (ETL) tools
- Enterprise Service Bus (ESB)

BizTalk will be used as our ESB and this will form a mediation layer between the source data systems (student record system, VLE, time tabling etc) and the STAR-Trak application tables and the data warehouse. As an intermediate step the staging tables, which currently provide temporary storage for

incoming data feeds, will remain, but eventually these will be subsumed within BizTalk. Of course we don't really need BizTalk to manage the data extracts, but this is an ideal project to have a close look at how it might become a key component of our enterprise architecture in the long term. (By the way, if you have a Microsoft Campus Agreement you can purchase BizTalk at an incredibly low price !)

Talend will be used to manage the transform and load stages. We like the look of this suite of programs and have hopes that it may form part our master data management (MDM) architecture in the future.

Our business intelligence layer will be provided by the BIRT tools. Currently we don't have a corporate data warehouse or BI solution, although we use Business Objects for corporate reporting and have purchased their Excelsius product. There is no point in heavy investment until the business is able to fully articulate its requirements: BIRT gives IT and the business the opportunity to learn together at minimal cost, before making a long-term investment decision.

Getting this loosely-coupled end-to-end architecture up and running will provide a wonderful learning environment for the business and IT (I hope to post up an amended architecture diagram soon). From an architectural perspective it will help us develop our understanding of how to attain that elusive goal of "business agility" where IT can at last claim the right to be an enabler of change. Getting back to the core objective of STAR-Trak, it will also provide us with valuable tools for student, staff and corporate planners to aid retention and encourage engagement in all aspects of student life.

User-Centric Integration of Activity Data (UCIAD)

Lead institution: Open University

Project home page: http://uciad.info/ub/

Project hypothesis:

Hypothesis 1: Taking a user-centric point of view can enable different types of analysis of activity data, which are valuable to the organisation and the user.

Hypothesis 2: Ontologies and ontology-based reasoning can support the integration, consolidation and interpretation of activity data from multiple sources.

Discussion of the hypotheses can be found here, and a discussion of using, refining, customising and reusing ontologies can be found here.

Project description

UCIAD is addressing the integration of activity data spread across various systems in an organization, and exploring how this integration can both benefit users and improve transparency in an organization.

Both research and commercial developments in the area of user activity data analysis have until now mostly focused on logging user visits to specific websites and systems, primarily in order to support recommendation, or to gather feedback data from users. However, data concerning a single user are generally fragmented across many different systems and logs, from website access logs to search data in different departments and as a result organizations typically are not able to maintain an integrated overview of the various activities of a given user, thus affecting their ability to provide optimal service to their users. Hence, a key tenet of the UCIAD project is that developing a coherent picture of the interactions between the user and the organization would be beneficial both to an organization and to its users.

Specifically, the objective of UCIAD is to provide the conceptual and computational foundations to support user- centric analyses of activity data, with the aim of producing results which can be customized for and deployed in different organizations. Ontologies represent semantic models of a particular domain, and can be used to annotate and integrate data from heterogeneous sources. The project will therefore investigate ontological models for the integration of user activity data, how such models can be used as a basis for a pluggable data framework aggregating user activity data, and how such an infrastructure can be used for the benefit of the users, providing meaningful (and exportable) overviews of their interaction with the organization.

Recipes

Extracting data

- Extract trace data from log files

Processing

- Exploitable artefact: UCIAD Ontology

See also

- position paper for the W3C Workshop on Web Tracking and User Privacy: Self-tracking on the Web.

Further work

The project will be continuing. See Next steps.

UCIAD: Hypotheses

Hypothesis 1: Taking a user-centric point of view can enable different types of analysis of activity data, which are valuable to the organisation and the user.

In order to test this hypothesis, one actually needs to achieve such user- centric analysis of activity data. This implies a number of technical and technological challenges, namely, the need to integrate activity data across a variety of web sites managed by an organisation, to consolidate this data beyond the "number of visits", and to interpret them in terms of user activities.

Ontologies are formal, machine processable conceptual models of a domain. Ontology technologies, especially associated with technologies from the semantic web, have proven useful in situations where a meaningful integration of large amounts of heterogeneous data need to be realised, and to a certain extent, reasoned upon in a qualitative way, for interpretation and analysis. Our goal here is to investigate how ontologies and semantic technologies can support the user- centric analysis of activity data. In other words, our second hypothesis is

Hypothesis 2: Ontologies and ontology-based reasoning can support the integration, consolidation and interpretation of activity data from multiple sources.

To test this the first task was to build an ontology capable of flexibly describing the traces of activities across multiple web sites, the users of these web sites and the connections between them. The idea is to use this ontology (or rather, this set of ontologies) as a basis for a pluggable software framework, capable of integrating data from heterogeneous logs, and to interpret such data as traces of high-level activities.

UCIAD: Using, refining, customising and reusing ontologies

Ontology modelling is a never ending task. Elements constantly need to be corrected and added to cover more and more cases in a way as generic as possible. It is even more the case in UCIAD as the approach is to create the ontology depending on the data we need to treat. Therefore, as we will progressively be adding more data from different sources, including server logs from different types of web sites, activity logs from systems such as VLEs or video players, the ontologies will evolve to include these cases.

Going a step further, what we want to investigate is the user-centric analysis of activity data. The ontologies will be used to provide users with views and analysis mechanisms for the data that concern their own activities. It therefore seems a natural next step to make it possible for the users to extend the ontologies, to customize them, therefore creating their own view on their own data.

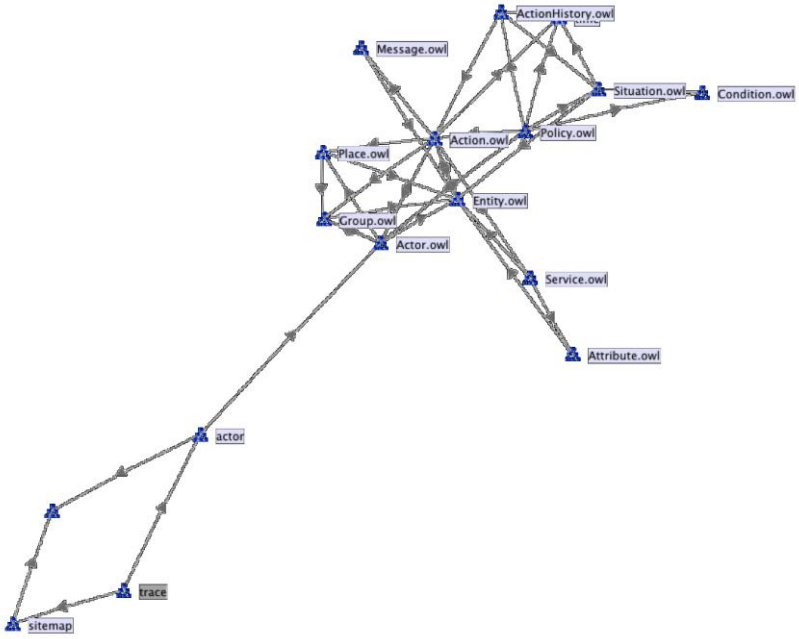
Reusing existing ontology

When dealing with data and ontologies, reuse is generally seen as a good practice. Apart from saving time from not having to remodel things that have already been described elsewhere, it also helps anticipating on future needs for interoperability by choosing well established ontologies that are likely to have been employed elsewhere. We therefore investigated existing ontologies that could help us define the notions mentioned above. Here are the ontology we reused:

- The Friend of a Friend (FOAF) ontology is commonly used to describe people, their connections with other people, but also their connections with documents. We use FOAF in the Actor Ontology for human users, and on the Sitemap Ontology for Web pages (as Documents).

- The Time Ontology is a common ontology for representing time and temporal intervals. We use it in the Trace Ontology.
- The Action ontology defines different types of actions in a broad sense, and can be used as a basis for representing elements of the requests in the Trace Ontology, but also as a base typology for the Activity ontology. It itself imports a number of other ontologies, including its own notion of actors.

The graph representing the dependencies between the UCIAD ontologies and others is represented below.



While not currently used in the base ontologies, other ontologies can be considered at a later stage, for example to model specific types of activities. These include the Online Presence Ontology (OPO), as well as the Semantically- Interlinked Online Communities ontology (SIOC).

Identifying concepts and their relations

The first step in building our ontology is to identify the key concepts, i.e., the key notions, that need to be tackled, bearing in mind that the ultimate goal is to understand activities. The main concepts to be considered are therefore the ones that support the concept of activity. Activities relate to users, but not only. The system relies extensively on website logs as sources of activity data. In these cases, it is possible to investigate requests both from human users and from robots automatically retrieving and crawling information from the websites. The server logs in question represent collections can be seen as traces of activities that these users/robots are realising on websites. therefore it is necessary to model these other aspects, which correspond to actions that are realised by actors on particular resources. These are the three kinds of objects that, in the context of Web platforms, the system to models, so that they can be interpreted and classified in terms of activities. The project therefore proposes four ontologies to be used as the basis of the work in UCIAD:

- The Actor Ontology is an ontology representing different types of actors (human users vs robots), as well as the technical setting through which they realise online activities (computer and user agent).
- The Sitemap Ontology is an ontology to represent the organisation of web pages in collections and websites, and which is extensible to represent different types of web pages and websites.
- The Trace Ontology is an ontology to represent traces of activities, realised by particular agents on particular web pages. As the current focus is on HTTP server logs, this ontology contain specific sections related to traces as HTTP requests (e.g., methods as actions and HTTP response code). It is however extensible to other types of traces, such as specific logs for VLEs or search systems.

- The Activity Ontology is intended to define particular classes of activities into which traces can be classified, depending on their particular parameters (including actors and web pages). The type of activities to consider highly depends on the systems considered and to a certain extent on the user. The idea here is that specific versions of the ontology will be built that fit the specific needs of particular systems. It is the possible to extract the generic and globally reusable part of these ontologies to provide a base for an overarching activity ontology. Ultimately, the idea in UCIAD is that individual users will be able to manipulate this ontology to include their specific view on their own activities.

UCIAD: next steps

There are a lot of things to mention here, some of them we have already mentioned several times. An obvious one is the finalisation, distribution and deployment of the UCIAD platform. A particular element we want to get done at a short term is to investigate the use of the UCIAD platform with various users, to see what kind of extensions of the ontologies would be commonly useful, and generally to get some insight into the reaction of users when being exposed to their own data.

More generally, we think that there is a lot more work to do on both the aspects of user-centric activity data and on the use of ontologies for the analysis of such data, as described in particular in our Wins and Fails post. These includes aspects around the licensing, distribution and generally management of user-centric data (as mentioned in our post on licensing). Indeed, while "giving back data to the users" is already technically difficult, there is a lot of fuzziness currently around the issues of ownership of activity data. This also forces us to look a lot more carefully at the privacy challenges that such data can generate, that didn't exist when these data were held and stayed on server logs.

Using OpenURL Activity Data

Lead institution: EDINA, University of Edinburgh

Project home page: http://edina.ac.uk/projects/Using_OpenURL_Activity_data_summary.html

Project hypothesis:

OpenURL Router data can be used to support recommendations

Project description

This is an invited proposal by JISC which takes forward the recommendations made in scoping activity related to collection and use of OpenURL data that might be available from institutional OpenURL resolvers and the national OpenURL router shared service which was funded between December 2008 - April 2009 by JISC. The work will be done in two stages: an initial stage exploring the steps required to make the data available openly, followed by making the data available and implementation of prototype service(s) using the data.

Recipes

Processing data

- Manipulating activity data using Linux commands
- Manipulating activity data using Ruby

Presenting data

- OpenURL Router Data: Total Requests by Date
- Visualising OpenURL Referrals Using Gource

Further work

The project will be continuing. See Next steps.

Project final report

A final report is available.

OpenURL: next steps

During the course of the project many interesting areas of research surfaced, which we were not able to pursue due to time constraints.

Making Further Data Reliably Available

- The information in the Level 2 data made available relates only to 'redirect to resolver' requests. There is further information in the Level 0 logs that may be of use to others. We noted a Level 1 data set that would include as much of the data as possible while not identifying institutions. A brief analysis and a small amount of development would be needed to make this available.
- Several institutions indicated that they would be willing to share data from their resolvers. Further legal work would be required to enable this as well as a solution to the challenge of de- duplicating the data that would appear both in the OpenURL Router and institutional resolver logs. Making this expanded data set available in the same way as the files released during this project would likely be very useful to others.
- Manual steps are currently involved in producing the data files. It would be valuable to invest effort to fully automate monthly production of the data files and include this as part of the OpenURL Router service so that others could rely upon regular production of the data.
- Not all UK Higher Education Institutions (HEIs) that have an OpenURL resolver are registered with the OpenURL Router service. Working with these institutions to facilitate their registration and encouraging them to participate in the data aggregation would expand the data set available.

Develop Services Based on the Data

- Based on the data from 2) above develop a reporting service for institutions to give them statistics about what journals are sought via the OpenURL Router for their institution compared with a) all UK institutions b) some grouping eg Russell Group. This depends on whether they pass all information via the OpenURL Router, but for those that do it will enable them to compare publisher data on statistics with an independent source, and to determine the value for money of the packages that are sold by the publishers.
- Develop a method for an institution to take the data set and integrate it with existing recommenders (possibly working with the RISE Project). This is likely to involve further work with vendors.

Develop the Prototype Recommender

- Develop the prototype recommender into a service producing relevant recommendations via machine interface that could be integrated into any of the existing library systems.
- One of the OpenURL resolver vendors indicated that they would be willing to share a large volume of data; inclusion of this in the data set for the prototype recommender would enable a significantly greater number of links to be made, thus improving the recommendations for end users.
- Develop a journal-level recommender for undergraduates, ideally with a visual interface to allow them to explore the network of journals that may interest them.
- Explore the use of content filtering in combination with collaborative filtering (or adding session identifiers) to enable use of the whole data set, and not just the proxy free data set.

Investigate the Data Files Made Available During this Project

- Determine whether the data held by the OpenURL Router is just the same as citation data, or whether it holds links that are not obvious from citation data.
- Enable visualisation of the data set, such as with a Force Directed Graph for displaying links between the articles based on the recommendation data, or adapting technology from the MOSAIC Book Galaxy project

Explore Potential Uses for the Data

- As indicated elsewhere we expect use of the data that we cannot anticipate. Nevertheless, some ideas have emerged within the community which are listed here:
    - Analysis to identify patterns of requests for journals or articles;
    - Use by students learning to analyse large data sets;
    - Use by researchers as the basis for a thesis;
    - Use by publishers to compare their listings of journals and articles with those sought by users;
    - Linking with other data sets for analysis of wider activity.

Technical recipes

A recipe documents a process performed by a project to solve a particular problem. There are three uses for recipes:

- Where a recipe is an exact match for something someone needs to do the recipe is simply a set of steps to follow.
- As a source for ideas as to what you might do to solve a problem in an activity data project that you are engaged in.
- To gain a flavour of what activity data projects have done, to better understand what the activity data project you are contemplating might do.

Recipes appear under three main headings

- Extracting activity data
- Processing activity data
- Presenting activity data

There is also a further section, that contains relevant recipes from elsewhere on the web:

- Other recipes

Perforce, all recipes are about techniques applied to particular systems. However, some recipes are more system-specific than others. For example, all the recipes in the *extracting activity data* section are very specific, with most so being *extract entry statistics from the Sentry library gate entry system*.

While still system-specific, the processing activity data section contains some more recipes that describe technology that can be used more widely in processing activity data. There are five such recipes. An example is manipulating activity data using Ruby, where the OpenURL Router log data is used as an exemplar activity data set that is processed using Ruby.

Extracting data

The recipes for extracting data are:

- Extract authentication counts from the EZproxy Authentication and Access Software
- Extract circulation data from the Horizon Library Management System
- Extract entry statistics from the Sentry library gate entry system
- Extract user ID and course code from CIRCE MI
- Extract anonymised loan data from Talis Alto - SALT
- Extract trace data from log files - UCIAD

Extract authentication counts from the EZproxy Authentication and Access Software

Originators/Authors

Graham Stone and Dave Pattern

LIDP, University of Huddersfield

Purpose

To extract the number of times authentication to e-resources were performed as an indicator for the number of e- resources actually accessed. Specifically, the number of times the student authenticated through EZProxy.

Background

To investigate whether there is a statistically significant correlation across a number of universities between library activity data and student attainment.

Ingredients

- EZProxy log files
- Perl

Assumptions

User level logging has been enabled. Note that the default is anonymous logging.

The Perl script is run in the same directory as the log files (or a local copy of them).

Warnings

This is a crude but common measure of actual e-resource usage.

Method

- Run Perl script
- Set start date and end date (eg an academic year).

Individual steps

- Run the Perl script shown in Appendix B
- When prompted enter the start and end dates in dd/mm/yyyy format

Output data

Information will be placed in a csv file called EZProxyUsageyyyymmdd- yyyymmdd.csv

File format

UserID	Count of authentications
Xxx	36
Xxy	3

Appendix A: Sample output

UserID, Count of authentications

Xxx, 36

Xxy, 3

Appendix B: Scripts

Please email d.c.pattern at huddersfield.ac.uk for the script

Extract circulation data from the Horizon Library Management System

Originators/Authors

Graham Stone and Dave Pattern

LIDP, University of Huddersfield

Purpose

To extract circulation data from the Horizon Library Management System. Extracts the number of items each registered library user borrowed from the library.

Background

To investigate whether there is a statistically significant correlation across a number of universities between library activity data and student attainment.

Ingredients

- Horizon Library Management System
- Perl
- Perl ODBC module to access the Horizon database

Assumptions

Select access to the Horizon database holding the circulation data.

ODBC connection to the Horizon database has already been configured.

Warnings

Has the potential to impact on performance of the Horizon system.

For security reasons it is advisable to set up a read only account to use the ODBC connection.

Method

- Run Perl script
- Set start date and end date (eg an academic year).

Individual steps

- Run the Perl script
- When prompted enter the start and end dates in dd/mm/yyyy format

Output data

Information will be placed in a csv file called HorizonUsageyyyymmdd- yyyymmdd.csv

File format

UserID	Count of items borrowed
Xxx	17
Xxy	3

Appendix A: Sample output

UserID, Count of items borrowed

Xxx, 17

Xxy, 3

Appendix B: Scripts

Please email d.c.pattern at huddersfield.ac.uk for the script

Extract entry statistics from the Sentry library gate entry system

Originators/Authors

Graham Stone and Dave Pattern

LIDP, University of Huddersfield

Purpose

To extract entry statistics from the Sentry library gate entry system. Showing the number of times each person entered the library, e.g. via a turnstile system that requires identity card access.

Background

To investigate whether there is a statistically significant correlation across a number of universities between library activity data and student attainment.

Ingredients

- Sentry Access Control Software System
- Perl
- Perl MySQL module to access the Sentry database

Assumptions

Select access to the MySQL database holding the Sentry data.

Warnings

Has the potential to impact on performance of the Sentry system.

Method

- Run Perl script
- Set start date and end date (eg an academic year).

Individual steps

- Run the Perl script
- When prompted enter the start and end dates in dd/mm/yyyy format

Output data

Information will be placed in a csv file called SentryUsageyyyymmdd- yyyymmdd.csv

File format

UserID	Count of times entered library
Xxx	77
Xxy	43

Appendix A: Sample output

UserID, Count of times entered library

Xxx, 77

Xxy, 43

Appendix B: Scripts

Please email d.c.pattern at huddersfield.ac.uk for the script

Extract user ID and course code from CIRCE MI

Authors

Paul Grand

RISE, Open University

Purpose

Extract user ID and course code from CIRCE MI

Background

To investigate the hypothesis that recommender systems can enhance the student experience in new generation e-resource discovery services.

Ingredients

- CIRCE MI regular SQL dump file, stored in a Samba share
- PHPMyAdmin for the import

Assumptions

- Column names in the SQL dump stay the same
- Students will change course modules over time
- Only one course module per student is specified at the moment

Warnings

Students will in the future have all their course modules listed against their real student IDs (OU Computer User Name – generally up to four initials and them numbers) in the SQL dump.

Method

Run an import using PHPMyAdmin

Output data

Users and Users Courses tables in the RISE MySQL database will be updated with current registration information while maintaining historic data.

users table

User id	oucu
3	XY4569

users_courses table

User id	Course_id
3	A200

Extract anonymised loan data from Talis Alto

Originators/Authors

Andy Land and Steve Campbell

SALT, University of Manchester

Purpose

To extract loan transactions from Talis Alto.

For each transaction (the borrowing of a physical item but not its renewal or return)the script extracts data showing the item id, basic bibliographic details, borrower id and transaction date.

Background

The loan transactions were needed to provide data for use in the SALT book recommender service.

Ingredients

- Talis Alto Library Management System
- Perl

Assumptions

Select access to the Sybase database holding the loan data.

Warnings

Has the potential to impact performance of the Talis system. The routine is best run on an alternative instance of the database if available and/or scheduled at a time when performance is less crucial.

Method

- Run Perl script
- Set start date and end date (eg an academic year), or use a range of loan_id numbers

Individual steps

- Run the Perl script shown in Appendix B
- Put the start and finish LOAN_ID numbers into the relevant sql line and into the output file name.

Output data

Information will be placed in a csv file called salt_data_xxxxxxx-yyyyyyy.out

Appendix A: Sample output

This is a single line as might appear in the CSV output file

'Feb 29 2000','4000000','155567','39106','0416181902','Carsten, F. L., Francis Ludwig',. - The rise of fascism','1970',' '

Appendix B: Scripts

```
salt_data_4000000-4250000.pl

#! /usr/local/bin/perl

$BLCMP_HOME=$ENV{"BLCMP_HOME"};

$TALIS_HOME=$ENV{"TALIS_HOME"};

$MIS_HOME="$TALIS_HOME/mis";

$LOCAL_MIS_HOME=$ENV{"LOCAL_MIS_HOME"};

require "sybperl.pl";

require "$TALIS_HOME/perl_tools/std_utils.pl";

require "$TALIS_HOME/perl_tools/mis_utils.pl";

$Database = "prod_talis";

&Std_open_db();

open (LOG, "> salt_data_4000000-4250000.out");

 ($result) = &sql($d,"

 select getdate()
```

```
 ");

(@result) = &sql($d,"

SELECT substring(L.CREATE_DATE,1,11), L.LOAN_ID, L.BORROWER_ID,  W.WORK_ID,

W.CONTROL_NUMBER, W.AUTHOR_DISPLAY, W.TITLE_DISPLAY, W.PUB_DATE,
W.EDITION_MAIN

from WORKS W, LOAN L, ITEM I, BORROWER B

WHERE L.LOAN_ID between 4000000 and 4250000 and L.STATE=0 and
L.ITEM_ID=I.ITEM_ID

AND I.WORK_ID=W.WORK_ID

and L.BORROWER_ID=B.BORROWER_ID AND B.TYPE_ID not in (7,17)

");

foreach $result (@result)

{

 ($t1,$t2,$t3,$t4,$t5,$t6,$t7,$t8,$t9)=split('~',$result);

 print LOG "'$t1','$t2','$t3','$t4','$t5','$t6','$t7','$t8','$t9'\n";

}
```

Extract trace data from log files

Originators/Authors

Salman Elahi and Mathieu d'Aquin

UCIAD, Open University

Purpose

To extract activity data from log files in a flexible way in order to deal with a multiplicity of log file formats.

Background

To integrate log file data from diverse systems.

Ingredients

- Log files
- UCIAD Parderer

Assumptions

- Log files relate to online resources
- Parderer is parameterised with a parameter for the specific parser class applicable to the log file format
- A parser class can have specific parameters, notably a regular expression describing log file entries
- Output from Parderer is in RDF

Warnings

A specific Parderer for a particular log file format needs to be deployed on the server where the log file (or a copy) is stored

Method

- Edit configuration file to specify the parameter class for Parderer and information about the server, eg where the log files are stored, the URI to describe the server in the RDF
- Run Parderer via cron jobs, currently always daily

Individual steps

- By default Parderer runs automatically (thanks to cron), but it can be run manually for specific dates via command line parameters

Output data

Information will be placed in a RDF output file as specified in the configuration files

File contents are in RDF XML format and comply with the ontology at available at http://github.com/uciad - see http://uciad.info/ub/2011/03/uciad-ontologies-a- starting-point/ for initial documentation.

Appendix A: Sample output

RDF for the trace

GET request to the URI http://data.open.ac.uk/resource/person/ext-718a372e10788bb58d562a8bf6fb864e

<rdf:RDF>

<rdf:Description rdf:about="http://uciad.info/trace/kmi- web13/ede2ab38da27695eec1e0b375f9b20da">

<rdf:type rdf:resource="http://uciad.info/ontology/trace/Trace"/>

<hasAction rdf:resource="http://uciad.info/action/GET"/>

<hasPageInvolved rdf:resource="http://uciad.info/page/0b9abc62fcf90afc53797b938af435dd"/>

<hasResponse rdf:resource="http://uciad.info/response/ea95add1414aba134ff9e0482b921a33"/>

<hasSetting rdf:resource="http://uciad.info/actorsetting/119696ec92c5acec29397dc7ef98817f"/>

<hasTime rdf:datatype="http://www.w3.org/2001/XMLSchema#string">13/Jun/2011:01:37:23+0100</hasTime>

</rdf:Description>

</rdf:RDF>

<rdf:RDF>

<rdf:Description rdf:about="http://uciad.info/page/0b9abc62fcf90afc53797b938af435dd">

<rdf:type rdf:resource="http://uciad.info/ontology/sitemap/WebPage"/>

<isPartOf rdf:resource="http://uciad.info/ontology/test1/dataopenacuk"/>

<onServer rdf:resource="http://kmi-web13.open.ac.uk"/>

<url rdf:datatype="http://www.w3.org/2001/XMLSchema#string">

/resource/person/ext-718a372e10788bb58d562a8bf6fb864e

</url>

</rdf:Description>

<rdf:RDF>
<rdf:Description rdf:about="http://uciad.info/ontology/test1/dataopenacuk">
<rdf:type rdf:resource="http://uciad.info/ontology/sitemap/Website"/>
<rdf:type rdf:resource="http://uciad.info/ontology/test1/LinkedDataPlatform"/>
<onServer rdf:resource="http://kmi-web13.open.ac.uk"/>
<urlPattern rdf:datatype="http://www.w3.org/2001/XMLSchema#string">/*</urlPattern>

</rdf:Description>
<rdf:Description rdf:about="http://uciad.info/response/ea95add1414aba134ff9e0482b921a33">
<rdf:type rdf:resource="http://uciad.info/ontology/trace/HTTPResponse"/>
<hasResponseCode rdf:resource="http://uciad.info/ontology/trace/200"/>
<hasSizeInBytes rdf:datatype="http://www.w3.org/2001/XMLSchema#int">1085</hasSizeInBytes>
</rdf:Description>

<rdf:Description rdf:about="http://uciad.info/actorsetting/119696ec92c5acec29397dc7ef98817f">
<rdf:type rdf:resource="http://uciad.info/ontology/actor/ActorSetting"/>
<fromComputer rdf:resource="http://uciad.info/computer/7587772edef21a8461f6af0efaf150fc"/>
<hasAgent rdf:resource="http://uciad.info/actor/ceec6f92bcc8167cbb665f7e51b2a6b3"/>
</rdf:Description>

<rdf:Description rdf:about="http://uciad.info/computer/7587772edef21a8461f6af0efaf150fc">
<rdf:type rdf:resource="http://uciad.info/ontology/actor/Computer"/>
<hasIPAddress
rdf:datatype="http://www.w3.org/2001/XMLSchema#string">129.13.186.4</hasIPAddress>
</rdf:Description>

<rdf:Description rdf:about="http://uciad.info/actor/ceec6f92bcc8167cbb665f7e51b2a6b3">
<rdf:type rdf:resource="http://uciad.info/ontology/actor/ActorAgent"/>
<agentId rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
ldspider (BTC 2011 crawl, harth@kit.edu, http://code.google.com/p/ldspider/wiki/Robots)
</agentId>
</rdf:Description>

Appendix B: Documentation to describe Parderer configuration

Here is an example Parderer configuration with comments

```
# the regular expression for a line of log when parsing apache logs
logPattern = ^([\\d.]+) (\\S+) (\\S+) \\[([\\w:/]+\\s[+\\-]\\d{4})\\] \"(.+?)\" (\\d{3}) (\\S+) \"(.*)\"  \"([^\"]+)\"
# the URI of the server which logs are being considered
serverURI = http://kmi-web13.open.ac.uk
# name of the server as it will appear in the data
serverName = kmi-web13
# local path to the log file
logFileDirectory = /web/logs/lucero.open.ac.uk
# pattern of the log file name
logFileNamePattern = access_log_%Y-%m.log
# number of variables in the regular expression
numberOfFields = 9
# URI of the SESAME triple store where the data should be uploaded
repURI = http://kmi-web01.open.ac.uk:8080/openrdf-sesame
# name of the repository in the SESAME triple store
repID = UCIADAll
# directory where the data dumps (in zipped RDF) should be placed
zippedFileOutPutDir = /web/lucero.open.ac.uk/applications/parsedLogs
```

Processing data

The recipes for processing data are:

- Stitching together library data with Excel
- Preparing attention data for processing by SPSS
- Manipulating activity data using Linux commands
- Manipulating activity data using Ruby
- Provide course based recommendations
- Process EZProxy log
- Provide relationship based recommendations
- Search term based recommendations
- How to Pivot data in Open Office Spreadsheet
- Exploitable artefact: UCIAD Ontology

Stitching together library data with Excel

Author

Philip Adams,

LIDP, De Montfort University

Introduction

Part of my contribution to the Library Impact Data Project has been compiling the data from different sources into a single spreadsheet. We have been able to gather data on items borrowed in a given year, visits to the Kimberlin library building, logins to the network in the libraries and use of an Athens account to access electronic resources. Some of the techniques used were new to me, and may be worth knowing about (or remembering if I have to run the same jobs again next year).

Data merging scenario

You have two spreadsheets, one with data about items loaned from the library and one from the Access Control system showing who entered the building. Both cover the same date range and are organised by some kind of user ID.

Loans file

p123456	17
p098765	12
p456098	6

Visits file

123456	45
098765	20
674523	23

To match the two spreadsheets, we need a common field. The first column in each nearly does this, but we need to fix the missing 'p' from the start of each ID number. There is a merge columns function in Excel , so you could create a new first column, copy 'p's into each cell and then merge the two columns into one in Column C. This is a bit tedious if you have thousands of lines of data. Another way of achieving the same results would be to adapt another function, using concatenate to add a prefix letter to each cell, e.g. =CONCATENATE("p",A1). Copy this formula into the cells in the inserted B column to populate with the correct ID data that we can use to match with the other spreadsheet.

Column format

At this point you might complain that nothing of the sort happens, that all you get is a repeated view of the formula itself and not its result. Most likely this is because the column you want to use to create the merged data has been set to be in text format. Quite rightly, at least in its own terms, Excel it treating what you type in a text, rather than instruction. Change the format to 'General' and the formula will perform its magic.

There may be a good reason why you set the column to text though. Excel generally expects to be working with numbers, rather than strings. Unless the column is set to 'text' it is likely to quietly amend an ID like '098765' to '98765'.

When importing .csv files this kind of change is carried out automatically, so a way round this is to change the file format to .txt and then when you import it, you get more control over how each column is to be handled.

Merging spreadsheets with VLOOKUP

If you don't already have sensible labels for your columns, here is a good moment to put them in.

With both speadsheets open you can begin to merge the data. In the first cell of the column where the data is to be inserted, click on the fx button. A popup list of functions will appear. Search through the list until you see 'VLOOKUP'. Select that and a new menu appears which enables you to set how the process is to work. The four values required are for: Lookup_value; Table- array; Col_index_num; Range_lookup. There is more on these fields on the Microsoft Excel site , but basically:

- Lookup_value is the identifier common to both spreadsheets. In the spreadsheet to which you want to bring the new data, click on the first relevant cell (A2).
- In the spreadsheet that you want to collect the new data from, click on the top left and bottom right cells to highlight the relevant data.
- The data that you want to import is in the second, right hand column, so type '2' into the Col_index_num field
- Some of the rows will not produce matches. Someone could regularly visit the library without borrowing anything. In that case we only want the exact matches, so 'false' is the correct answer for this field.

Apply this with 'OK' and you should get either a value inserted into your cell, or an 'N/A' if there is no match. Copying and dragging this formula into the rest of the cells for the column should bring in the remaining data. This can be the most tedious part of the process if you have thousands of rows to copy into.

So now you should have:

Loans file

Identifier	Loans for year	Visits for year
p123456	17	45
p098765	12	20
p456098	6	N/A

Checking your merges

For the sake of data integrity and general peace of mind you will want to check that the data has been correctly carried across. You can use VLOOKUP again to check on the merge, but this time reverse the spreadsheets so that you are copying back data from the original destination sheet.

Preparing attention data for processing by SPSS

Originators/Authors

Graham Stone and Dave Pattern

LIDP, University of Huddersfield

Purpose

To merge the data extracted from the following systems:

- Student record system
- Library management system
- Authentication system
- Turnstile entry system

In order to produce a single file that can be processed by SPSS or an equivalent system.

Background

To investigate whether there is a statistically significant correlation across a number of universities between library activity data and student attainment.

Ingredients

File formats for the following are shown in Appendix C.

- Extract of Student Record System data
- Extract of data from Library Management System
- Extract of data from Authentication system
- Extract of data from Turnstile entry system

Assumptions

Data has previously been extracted from the student record system and one or more of the other systems.

Student ID is of identical format in each of the files to be merged.

Warnings

All files should be for the same period to ensure statistical validity.

Method

- Run Perl script
- Follow prompts

Individual steps

- Run the Perl script
- When prompted select CSV files as appropriate

Output data

Information will be placed in a CSV file called LibraryImpactData.csv

File format

UserID	academic year of graduation	course title	length of course in years	type of course	Full time/part time	grade achieved	school/academic department	Count of items borrowed	Count of authentications	Count of times entered library
Xxx	2010	Nursing	3	UG	FT	2.1	Nursing	17	36	77
Xxy	2010	Sociology	3	UG	FT	1	Social sciences	3	3	43

Appendix A: Sample output

UserID, academic year of graduation , course title , length of course in years , type of course, Full time/part time, grade achieved, school/academic department, Count of items borrowed, Count of authentications, Count of times entered library

Xxx, 2010, Nursing, 3, UG, FT, 2.1, Nursing, 17, 36, 77

Xxy, 2010, Sociology, 3, UG, FT, 1, Social sciences, 3, 3, 43

Appendix B: Scripts

Please email d.c.pattern at huddersfield.ac.uk for the script

Appendix C: input file formats

Student Record System data

> UserID, academic year of graduation , course title , length of course in years , type of course, Full time/part time, grade achieved, school/academic department
>
> Xxx, 2010, Nursing, 3, UG, FT, 2.1, Nursing
>
> Xxy, 2010, Sociology, 3, UG, FT, 1, Social sciences

Library management system data

> UserID, Count of items borrowed
>
> Xxx, 17
>
> Xxy, 3

Authentication data

> UserID, Count of authentications
>
> Xxx, 36
>
> Xxy, 3

Turnstile entry data

> UserID, Count of times entered library
>
> Xxx, 77
>
> Xxy, 43

Manipulating activity data using Linux commands

Author

Tony Hirst writing at his blog

Purpose

To demonstrate how Linux commands can be used to manipulate a activity data (in this case OpenURL log file contents

Description

{ ColorSyncTransformCreator = "Apple CMM"; ColorSyncTransformDestinationSpace = "RGB "; ColorSyncTransformSourceSpace = "RGB "; "com.apple.cmm.TransformType" = NULLTransform; "com.apple.cmm.cmmstorage" = <a7abaa32>; }You know those files that are too large for you to work with, or even open? Maybe they ’re not….

> Picking up on Postcards from a Text Processing Excursion where I started dabbling with Unix command line text processing tools (it sounds scarier than it is… err… maybe?!;-), I thought it would make sense to have a quick play with them in the context of some “real” data files.
>
> The files I've picked are intended to be intimidating (maybe?) at first glance because of their size: in this post I'll look at a set of OpenURL activity data from EDINA (24.6MB download, unpacking to 76MB), and for a future post, I thought it might to interesting to see whether this approach would

work with a dump of local council spending data from OpenlyLocal (73.2MB download, unzipping to 1,011.9MB).

To start with, let's have a quick play with the OpenURL data: you can download it from here: OpenURL activity data (April 2011)

What I intend to do in this post is track my own preliminary exploration of the file using what I learned in the "Postcards" post. I may also need to pick up a few new tricks along the way… One thing I think I want to look for as I start this exploration is an idea of how many referrals are coming in from particular institutions and particular sources…

Let's start at the beginning though by seeing how many lines/rows there are in the file, which I downloaded as *L2_2011-04.csv* :

```
wc -l L2_2011-04.csv
```

I get the result *289,691* ; older versions of Excel used to only support 65,536 rows per sheet, though I believe more recent versions (Excel 2007, and Excel 2010) can support over a million; Google Apps currently limits sheet sizes to up to 200,000 cells (max 256 columns), so even if the file was only one column wide, it would still be too big to upload into a single Google spreadsheet. Google Fusion Tables can accept CSV files up to 100MB, so that would work (if we could actually get the file to upload… Spanish accent characters seemed to break things when I tried… the workaround I found was to split the original file, then separately upload and re-save the parts using Google Refine, before uploading the files to Google Fusion tables (upload one to a new table, then import and append the other files into the same table).

..which is to say: the almost 300,00 rows in the downloaded CSV file are probably too many for many people to know what to do with, unless they know how to drive a database… which is why I thought it might be interesting to see how far we can get with just the Unix command line text processing tools.

To see what's in the file, let's see what's in there (we might also look to the documentation):

```
head L2_2011-04.csv
```

Column 40 looks interesting to me: *sid* (service ID); in the data, there's a reference in there to *mendeley* , as well as some other providers I recognise (EBSCO, Elsevier and so on), so I *think* this refers to the source of the referral to the EDINA OpenURL resolver (@ostephens and @lucask suggested they thought so too. Also, @lucask suggested "OpenUrl from Endnote has ISIWOS as default SID too!", so we may find that some sources either mask their true origin to hide low referral numbers (maybe very few people ever go from endnote to the EDINA OpenURL resolver?), or to inflate other numbers (Endnote inflating apparent referrals from ISIWOS.)

Rather than attack the rather large original file, let's start by creating a smaller sample file with a couple of hundred rows that we can use as a test file for our text processing attempts:

```
head -n 200 L2_2011-04.csv > samp_L2_2011-04.csv
```

Let's pull out column 40, sort, and then look for unique entries in the sample file we created:

```
cut -f 40 samp_L2_2011-04.csv | sort | uniq -c
```

I get a response that starts:

```
12

 1 CAS:MEDLINE

 1 EBSCO:Academic Search Premier

 7 EBSCO:Business Source Premier

 1 EBSCO:CINAHL

...
```

so in the sample file there were 12 blank entries, 1 from *CAS:MEDLINE* , 7 from *BSCO:Business Source Premier* and so on, so this appears to work okay. Let's try it on the big file (it may take a few seconds…) and save the result into a file (*samp_uniqueSID.csv* :

```
cut -f 40 L2_2011-04.csv | sort | uniq -c > uniqueSID.csv
```

This results of the count will be in arbitrary order, so it's possible to add a sort into the pipeline in order to sort the entries according to the number of entries. The column we want to sort on is column 1 (so we set the sort -k key to 1; and because *sort* sorts into increasing order by default, we can reverse the order (-r) to get the most referenced entries at the top (the following is NOT RECOMMENDED… read on to see why…):

```
cut -f 40 L2_2011-04.csv | sort | uniq -c | sort -k 1 -r >  uniqueSID.csv
```

We can now view the *uniqueSD.csv* file using the more command (more *uniqueSD.csv*), r look at the top 5 rows using the *head* command:

```
head -n 5 uniqueSID.csv
```

Here's what I get as the result (treat this with suspicion…):

```
9181 OVID:medline
 9006 Elsevier:Scopus
 7929 EBSCO:CINAHL
 74740 www.isinet.com:WoK:UA
 6720 EBSCO:jlh
```

If we look through the file, we actually see:

```
1817 OVID:embase
 1720 EBSCO:CINAHL with Full Text
 17119 mendeley.com:mendeley
 16885 mendeley.com/mendeley:
 1529 EBSCO:cmedm
 1505 OVID:ovftdb
```

I actually was alerted to this oops when looking to see how many referrals were from mendeley, by using grep on the counts file (if grep complains about a "Binary file", just use the -a switch…) :

```
grep mendeley uniqueSID.csv
```

```
17119 mendeley.com:mendeley
 16885 mendeley.com/mendeley :
```

17119 beat the "top count" 9181 from OVID:medline – obviously I' d done something wrong!

Specifically, the sort had sorted *by character* **not** by numerical value… (17119 and 16885 are numerically grater than 1720, but 171 and 168 are less (in string sorting terms) than 172. The reasoning is the same as why we 'd index aardman before aardvark).

To force sort to sort using numerical values, rather than string values, we need to use th - n switch (so now I know!):

```
cut -f 40 L2_2011-04.csv | sort | uniq -c | sort -k 1 -r - n >
uniqueSID.csv
```

Here's what we get now:

```
74740 www.isinet.com:WoK:UA

 34186

 20900 www.isinet.com:WoK:WOS

 17119 mendeley.com:mendeley

 16885 mendeley.com/mendeley:

 9181 OVID:medline

 9006 Elsevier:Scopus

 7929 EBSCO:CINAHL

 6720 EBSCO:jlh

...
```

To compare the referrals from the actual sources (e.g. the aggregated EBSCO sources, rather than EBSCO:CINAHL, EBSCO:jlh and so on), we can split on the ":" character, to create a two columns from one: the first containing the bit before the ':', the second column containing the bit after:

```
sed s/:/'ctrl-v<TAB>'/ uniqueSD.csv | sort -k > uniquerootSID.csv
```

(Some versions of sed may let you identify the tab character as \t; I had to explicitly put in a tab by using ctrl-V then tab.)

What this does is retain the number of lines, but sort the file so all the EBSCO referrals are next to each other, all the Elsevier referrals are next to each other, and so on.

Via an answer on Stack Overflow , I found this bit of voodoo that would then sum the contributions from the same root referrers:

```
cat uniquerootSID.csv | awk '{a[$2]+=$1}END{for(i in a ) print  i,a[i] }'
| sort - k 2 -r -n > uniquerootsumSID.csv
```

Using data from the file *uniquerootSID.csv* , the *awk* command sets up an array (a) that has indices corresponding to the different sources (EBSCO, Elsevier, and so on). It then runs an accumulator that sums the contributions from each unique source. After processing all the rows (END), the routine then loops through all the unique sources in the a array, and emits the source and the total. The sort command then sorts the output by total for each source and puts the list into the file *uniquerootsumSID.csv* .

Here are the top 15:

```
www.isinet.com 99453

 EBSCO 44870

 OVID 27545

 mendeley.com 17119

 mendeley.com/mendeley 16885

 Elsevier 9446

 CSA 6938

 EI 6180

 Ovid 4353
```

```
 wiley.com 3399

 jstor 2558

 mimas.ac.uk 2553

 summon.serialssolutions.com 2175

 Dialog 2070

 Refworks 1034
```

If we add the two Mendeley referral counts that gives ~34,000 referrals. How much are the referrals from commercial databases costing, I wonder, by comparison? Of course, it may be that the distribution of referrals from different institutions is different. Some institutions may see all their traffic through EBSCO, or Ovid, or indeed Mendeley… If nothing else though, this report suggests that Mendeley is generating a fair amount of EDINA OpenURL traffic…

Let's use the *cut* command again to see how much traffic is coming from each unique institution (not that I know how to decode these identifiers…); column 4 is the one we want (remember, we use the *uniq* command to count the occurrences of each identifier):

```
cut -f 4 L2_2011-04.csv | sort | uniq -c | sort -k 1 -r -n >
uniqueInstID.csv
```

Here are the top 10 referrer institutions (columns are: no. of referrals, institution ID):

```
41268 553329

 31999 592498

 31168 687369

 29442 117143

 24144 290257

 23645 502487

 18912 305037

 18450 570035

 11138 446861

 10318 400091
```

How about column 5, the *routerRedirectIdentifier* :

```
195499 athens

 39381 wayf

 29904 ukfed

 24766 no identifier

 140 ip
```

How about the publication year of requests (column 17):

```
45867

 26400 2010

 16284 2009

 13425 2011

 13134 2008

 10731 2007
```

```
 8922 2006

 8088 2005

 7288 2004
```

It seems to roughly follow year?!

How about unique journal title (column 15):

```
258740

 277 Journal of World Business

 263 Journal of Financial Economics

 263 Annual Review of Nuclear and Particle Science

 252 Communications in Computer and Information Science

 212 Journal of the Medical Association of Thailand Chotmaihet
thangphaet

 208 Scandinavian Journal of Medicine & Science in Sports

 204 Paleomammalia

 194 Astronomy & Astrophysics

 193 American Family Physician
```

How about books (column 29 gives ISBN):

```
278817

 1695 9780470669303

 750 9780470102497

 151 0761901515

 102 9781874400394
```

And so it goes..

What's maybe worth remembering is that I haven't had to use any tools other than command line tools to start exploring this data, notwithstanding the fact that the source file may be too large to open in some everyday applications…

The quick investigation I was able to carry out on the EDINA OpenURL data also built directly on what I' d learned in doing the Postcards post (except for the voodoo awk script to sum similarly headed rows, and the sort switches to reverse the order of the sort, and force a numerical rather than string based sort). Also bear in mind that three days ago, I didn't know how to do any of this…

…but what I do suspect is that it's the sort of thing that Unix sys admins play around with all the time, e.g. in the context of log file hacking…

PS so what else can we do…? It strikes me that by using the date and timestamp, as well as the institutional ID and referrer ID, we can probably identify searches that are taking place: a) within a particular session, b) maybe by the same person over several days (e.g. in the case of someone coming in from the same place within a short window of time (1-2 hours), or around about the same time on the same day of the week, from the same IDs and searching around a similar topic).

See also Playing with OpenURL Router Data in Ruby

Manipulating activity data using Ruby

Author

Mark van Harmelen writing at Hedtek's blog.

Purpose

To demonstrate how scripts written in Ruby can be used to manipulate a activity data (in this case OpenURL log file contents)

Description

This is an alternative method to that used by Tony Hirst in Playing With Large (ish) CSV Files, and Using Them as a Database from the Command Line: EDINA OpenURL Logs while experimenting with processing a reasonably large data set with *NIX command line tools. The data set is the recently published OpenURL Router Data . Inspired by this post I wondered what I could hack up in Ruby to process the same data, and if I could do this processing without a database. The answer is that it is pretty simple to process.

First what is the OpenURL Router, and what is its data? What we need to know here is that the Router effectively enables library services to find the URLs for online bibliographic resources (more detail). A simplification is that the Router supply a translation from bibliographic data to the URL in question. The OpenURL router is funded by JISC and administered by EDINA in association with UKOLN .

Suitably anonymised OpenURL Router Data has been published by Using OpenURL Activity Data Project. This project is participating in JISC's Activity Data Programme where Hedtek is collaborating in the synthesis of the outputs of the projects participating in this programme. Hence my interest in the data and what can be done with it.

My initial interest was in who has what proportion referrals. Tony computed this, and I wanted to replicate his results. In the end I had a slightly different set of results.

Downloading and decompressing this CSV data was pretty easy, as was honing in on one field of interest, the source of the data being referred to. Tony's post and the OpenURL Router Data documentation made it pretty easy to hone in on the 40th field in each line of this CSV formatted data.

My first attempts were to use a ruby gem, CSV, from the ruby interpreter irb. This went well enough but I soon discovered that CSV wouldn't handle fields with a double quote in them. Resorting to the my OS X command line

```
tr \" \' < L2_2011.csv > nice.csv
```

soon sorted that out.

It soon emerged that I needed to write a method, so I flipped to the excellent RubyMine , and soon hacked up a little script. Interestingly, I found that the representation of the site with the requested resource often had a major component and a minor component, separated by a colon, thus

```
EBSCO:edswah

EBSCO:CINAHL with Full Text
```

etc

Having been excited by previous mention of Mendeley by Tony and wanting to find out the percentage of references to Mendeley's data for another piece of work I am doing, I stripped out the minor component, and came up with the following code

```ruby
require 'csv'

def create_totals
  records = CSV.read("nice.csv", {:col_sep=>"\t"})
  totals = Hash.new(0)
  records.each do |rec|
    unless rec[39]
      index='undefined'
    else
      index = rec[39].gsub(/:.*/,'')
    end
    totals[index] += 1
  end
  totals.each do |t|
    puts "#{t[1]} \t#{t[0]}"
  end
end

create_totals
```

While its open to a good refactoring, it did the job well enough, producing an unsorted list of results. A quick refactor resulted in the following, which also coalesced both mendeley.com and mendeley.com/mendeley into one result.

```ruby
require 'csv'

SOURCE_COL = 39

def make_index (record)
  record[SOURCE_COL] ? record[SOURCE_COL].gsub(/(:|\/).*/,'') : 'undefined'
end

def create_totals
  records = CSV.read("nice.csv", {:col_sep=>"\t"})
  totals = Hash.new(0)
  records.each do |rec|
    totals[ make_index(rec) ] += 1
  end
  totals.each do |t|
    puts "#{t[1]} \t#{t[0]}"
  end
end

create_totals
```

To sort the output I used a command line sort after the script invocation

```
ruby totals.rb | sort -nr
```

and obtained the following, here only listing those sites with more than 1000 references

```
99453 www.isinet.com

44870 EBSCO

34186 undefined

34004 mendeley.com

27545 OVID

9446 Elsevier

6938 CSA
```

6180 EI

4353 Ovid

3399 wiley.com

2558 jstor

2553 mimas.ac.uk

2175 summon.serialssolutions.com

2070 Dialog

1034 Refworks

The rest, working out percentages, is easy thanks to Excel, see the middle column

99453	34.3	www.isinet.com
44870	15.5	EBSCO
34186	11.8	undefined
34004	11.7	mendeley.com
27545	9.5	OVID
9446	3.3	Elsevier
6938	2.4	CSA
6180	2.1	EI
4353	1.5	Ovid
3399	1.2	wiley.com
2558	0.9	jstor
2553	0.9	mimas.ac.uk
2175	0.8	summon.serialssolutions.com
2070	0.7	Dialog
1034	0.4	Refworks

Provide course based recommendations

Authors

Paul Grand

RISE, Open University

Purpose

To provide, given a searcher's course, recommendations as to serial articles or other e- resources that might be indexed in EBSCO EDS

Background

To investigate the hypothesis that recommender systems can enhance the student experience in new generation e-resource discovery services.

Ingredients

- RISE database
- MyRecommendations web service
- Logged in user with OUCU (OU computer username)that can be associated with the user's course code

Assumptions

- Can obtain user's course code from OUCU
- At the moment only one course code is associated with a user, later this will change

Warnings

If the RISE data base does not contain a course code (because the searcher is a new student and CIRCE MI import is not done yet, or if the searcher is a member of staff) then the recommendation gathering process will not be performed and no course base recommendations will be shown via the user agent.

Method

MyRecommendations does the following:

- Uses OUCU to obtain the user's course (see warning, if no OUCU exists further processing here is not done)
- Checks resource views table to see which resource IDs appear most frequently for the given course ID, with a limit of 4
- Return these via the user agent as a list of titles with hypertext links

Output data

Appears in every page, including the search results page, showing up to the top four items most accessed by people on that course (now and historically)

Title	Sample Title
URL	XY4569
EBSCO accession number	01234567
ISSN	12345678

Appendix A: MyRecomendations web service source code

See http://code.google.com/p/rise-project/

Appendix B: ER diagram for RISE database

http://www.open.ac.uk/blogs/RISE/wp- content/uploads/2011/04/ERD1404111.png

Process EZProxy log

Author

Paul Grand

RISE, Open University

Purpose

Process EZProxy log, to extract historic usage data

Background

To investigate the hypothesis that recommender systems can enhance the student experience in new generation e-resource discovery services.

Ingredients

- EZProxy log, a zipped full format log file (as opposed to the abbreviated log file)

- Parsing script in PHP

Assumptions

- EZProxy log files can be customised; this recipe relies on the OU's version of the log files; these are modified to contain the OUCU (OU computer username)
- Script can be run nightly on the day's logs
- Log files are in a known directory that is on the same server that is running the script

Warnings

The log files are generated daily, the script will process log files for which there is no entry in the RISE database (in the table log_files)

Method

Run as a cron job

Provide relationship based recommendations

Author

Paul Grand

RISE, Open University

Purpose

To provide, given a searcher's last two viewed e-resources (via the EZProxy logs and/or MyRecommendations), recommendations as to serial articles or other e-resources that might be indexed in EBSCO EDS.

Background

To investigate the hypothesis that recommender systems can enhance the student experience in new generation e-resource discovery services.

Ingredients

- RISE database
- MyRecommendations web service
- Logged in user with OUCU (OU computer username to yield integer user ID in the RISE database

Assumptions

- The user has previously viewed resources that are stored in the RISE database

Warnings

If there are no previously viewed resources by the user then there will be no recommendations made.

Method

MyRecommendations does the following

- Uses OUCU to obtain the user's numeric ID (done earlier before this particular processing)
- Uses the numeric user ID to find the user's last viewed resources, limit 2, returns resource IDs
- Users view resources in order. Let's imagine that some users have looked at resource A then resource B immediately afterwards. This will be stored in the RISE database. Now imagine our current user looks at resource A. We would like to retrieve resource B as a recommendation.

- We have a numeric id for resource A, so we interrogate the RISE database to see which resources were viewed (by any user) after resource A. With simple numeric counts derived from database entries, we can determine the most frequently viewed resources following the viewing of resource A. These resources form the basis of the recommendations.

  In actuality the top four recommendations (ie most viewed/rated, see the numeric index information below) yielded from the user's last two viewings are used as the recommendations. Also, a value field is used in forming recommendations, where this is set by users rating the quality of recommendations (+2 for useful, +1 for viewed, -1 for not useful).

Output data

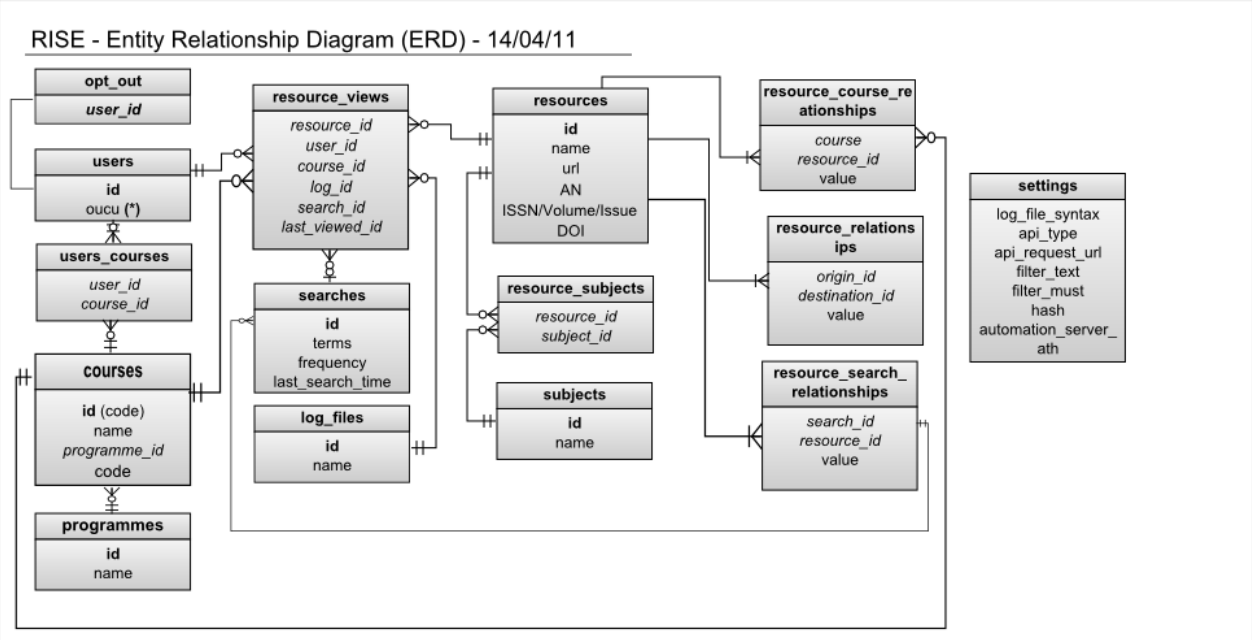
Appears in every page, including the search results page. Shows up to the top four items most accessed/rated by people who viewed the same items (now and historically)

Title	Some Sample Title
URL	http://.../.../.../
ISSN	12345678

Appendix A: MyRecomendations web service source code

See http://code.google.com/p/rise-project/

Appendix B: ER diagram for RISE database



Source: http://www.open.ac.uk/blogs/RISE/wp- content/uploads/2011/04/ERD1404111.png

Search term based recommendations

Authors

Paul Grand

RISE, Open University

Purpose

To provide, given a searcher's current search term, recommendations as to serial articles or other e-resources that might be indexed in EBSCO EDS

Background

To investigate the hypothesis that recommender systems can enhance the student experience in new generation e-resource discovery services.

Ingredients

- RISE database
- MyRecommendations web service
- User's current search term

Assumptions

- The user has entered a search with a non-null search term
  One or more users has used the search term and subsequently viewed at least one resource from the returned result set

Warnings

If there are no records of the same search term with at least one resource having been access from the search term results, then no recommendations will be returned.

This is entirely reliant on the MyRecommedations service, and does not draw on historic EZProxy data.

Method

MyRecommendations makes the following database queries

- Search the searches table for the terms used to get a numeric search ID, if there is no search ID that already exists, then do not look for search related recommendations
- Select from the resource_views table according to a match on the search ID, order by the most commonly viewed resources, limit 4.

  In actuality also a value field is used informing recommendations, where this set by users making recommendations (+2 for useful, +1 for viewed, -1 for not useful).

Output data

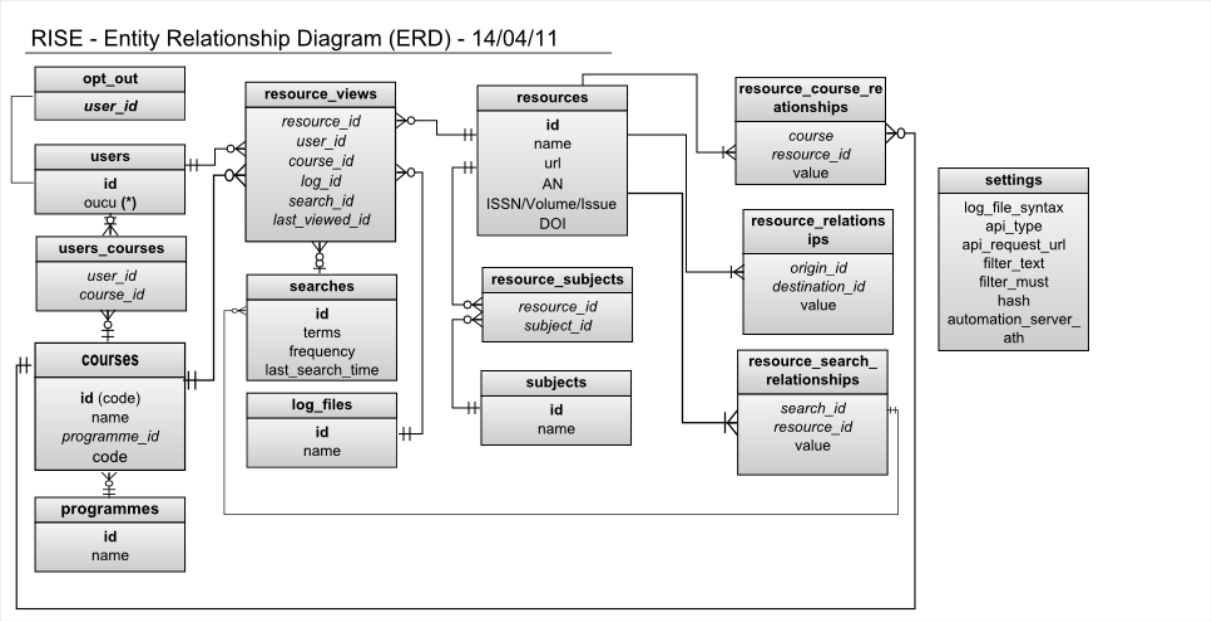
Appears in every page, including search results page, up to the top the top four items most accessed/rated by people who viewed the same items (now and historically)

Title	alphanumeric
URL	http://.../.../.../
ISSN	12345678

Appendix A: MyRecomendations web service source code

See location http://code.google.com/p/rise-project/

Appendix B: ER diagram for RISE database



source: http://www.open.ac.uk/blogs/RISE/wp- content/uploads/2011/04/ERD1404111.png

How to Pivot data in Open Office Spreadsheet

Originators/Authors

Anne Clarke

Exposing VLE activity data, University of Cambridge

Purpose

Instructions on how to work with output stats data in an OpenOffice spreadsheet to create a Pivot Table.

Background

We have developed tools that can be run to extract various counts (unique logins, number of each type of event) on a granular scale; for instance per month or per week. This allows investigation of the fluctuation of these interactions.

However to see aggregated totals per Event Type or per Institution we can use the Pivot functionality within a spreadsheet. This document gives instructions for OpenOffice.

Ingredients

- CSV file containing granular results of running Perl Stats routines on the
- database eg total_events_by_week.csv, unique_instids_by_month.csv
- This file should have appropriate column headings in the first row OpenOffice application

Assumptions

The user has basic spreadsheet knowledge and some familiarity with OpenOffice.

Method

You will first use OpenOffice to open the CSV file. Then you select the data, enter grouping information and create a Pivot table giving an aggregated view of the data.

Individual steps

- Open your results CSV file in OpenOffice:

	A	B	C	D	F	G	H
1	Year	Month	Academic Ye	Event Type	Count		
2	2005	10	2005_06	alias.add	12		
3	2005	10	2005_06	alias.del	3		
4	2005	10	2005_06	annc.delete.own	1		
5	2005	10	2005_06	annc.new	45		
6	2005	10	2005_06	annc.revise.own	1		
7	2005	10	2005_06	calendar.new	2		
8	2005	10	2005_06	calendar.revise	2		
9	2005	10	2005_06	chat.delete.any	0		
10	2005	10	2005_06	chat.delete.own	3		
11	2005	10	2005_06	chat.new	177		
12	2005	10	2005_06	content.delete	1		
13	2005	10	2005_06	content.new	61		
14	2005	10	2005_06	content.read	600		
15	2005	10	2005_06	content.revise	2		
16	2005	10	2005_06	disc.new	9		
17	2005	10	2005_06	disc.null	5		
18	2005	10	2005_06	mail.delete.any	4		
19	2005	10	2005_06	mail.new	18		
20	2005	10	2005_06	mail.revise.any	0		
21	2005	10	2005_06	memory.reset	0		
22	2005	10	2005_06	prefs.add	0		
23	2005	10	2005_06	prefs.del	4		
24	2005	10	2005_06	realm.add	18		
25	2005	10	2005_06	realm.upd	74		
26	2005	10	2005_06	realm.upd.own	17		
27	2005	10	2005_06	site.add	13		
28	2005	10	2005_06	site.upd	63		
29	2005	10	2005_06	user.add	4		
30	2005	10	2005_06	user.del	1		
31	2005	10	2005_06	user.login	106		
32	2005	10	2005_06	user.logout	296		
33	2005	10	2005_06	user.upd.any	1		
34	2005	10	2005_06	user.upd.own	1		
35	2005	11	2005_06	alias.add	14		
36	2005	11	2005_06	alias.del	10		
37	2005	11	2005_06	annc.delete.own	2		
38	2005	11	2005_06	annc.new	22		
39	2005	11	2005_06	annc.revise.own	15		
40	2005	11	2005_06	asn.new.assignment	1		
41	2005	11	2005_06	asn.new.assignment	0		
42	2005	11	2005_06	calendar.new	12		
43	2005	11	2005_06	calendar.revise	3		
44	2005	11	2005_06	chat.delete.own	3		
45	2005	11	2005_06	chat.new	96		
46	2005	11	2005_06	content.delete	34		

Illustration 1: Sample granular data

- Use short cut keys to select all the data (make sure you select just the data not the whole spreadsheet).
- Data->DataPilot->Start
- Ok to using current selection
- Wait patiently for some time – you will see the Data Pilot window:
- Move the field names to the Rows and Columns as appropriate. Normally you will be putting a numeric value in the data field.
- The example below users Academic Year, Year, Month as the column headings for the pivoted table and event type to aggregate values over. The sum of the counts form the data fields.
- Click on the More button
- Choose to display “Results to” in a new sheet (if you don't do this, the results will appear below the current sheet. Click Ok.

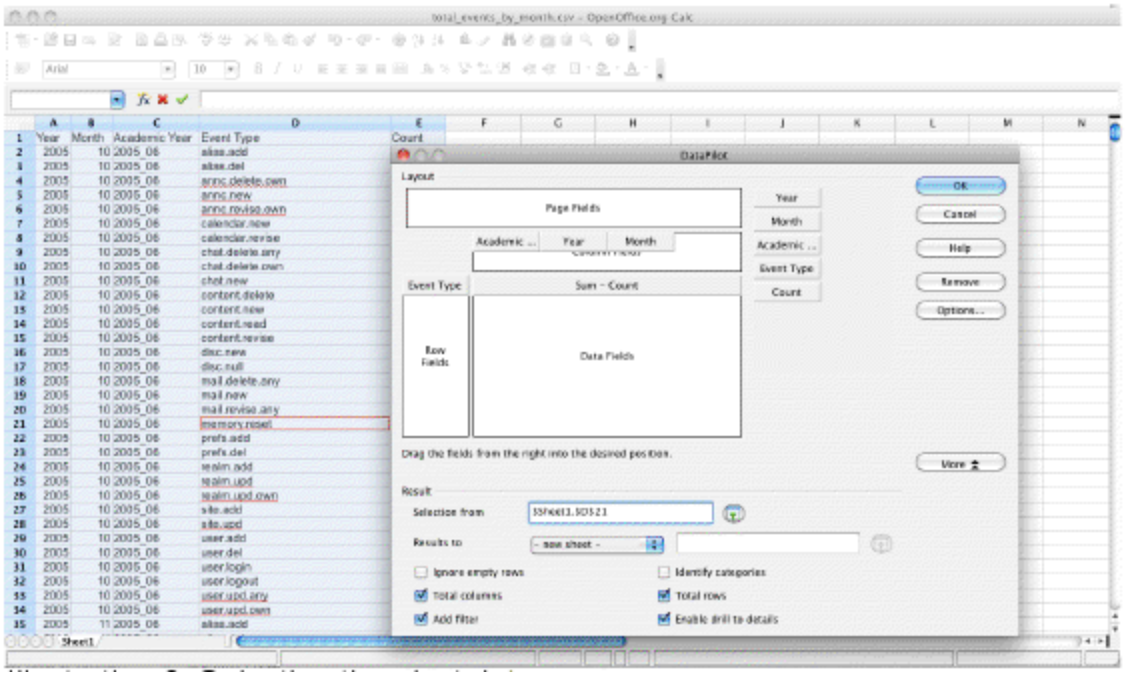


Illustration 2: Selecting the pivot data

- Now the pivot table is there.
- Format columns etc to beautify. Select all the columns then
- Format>Column>Width

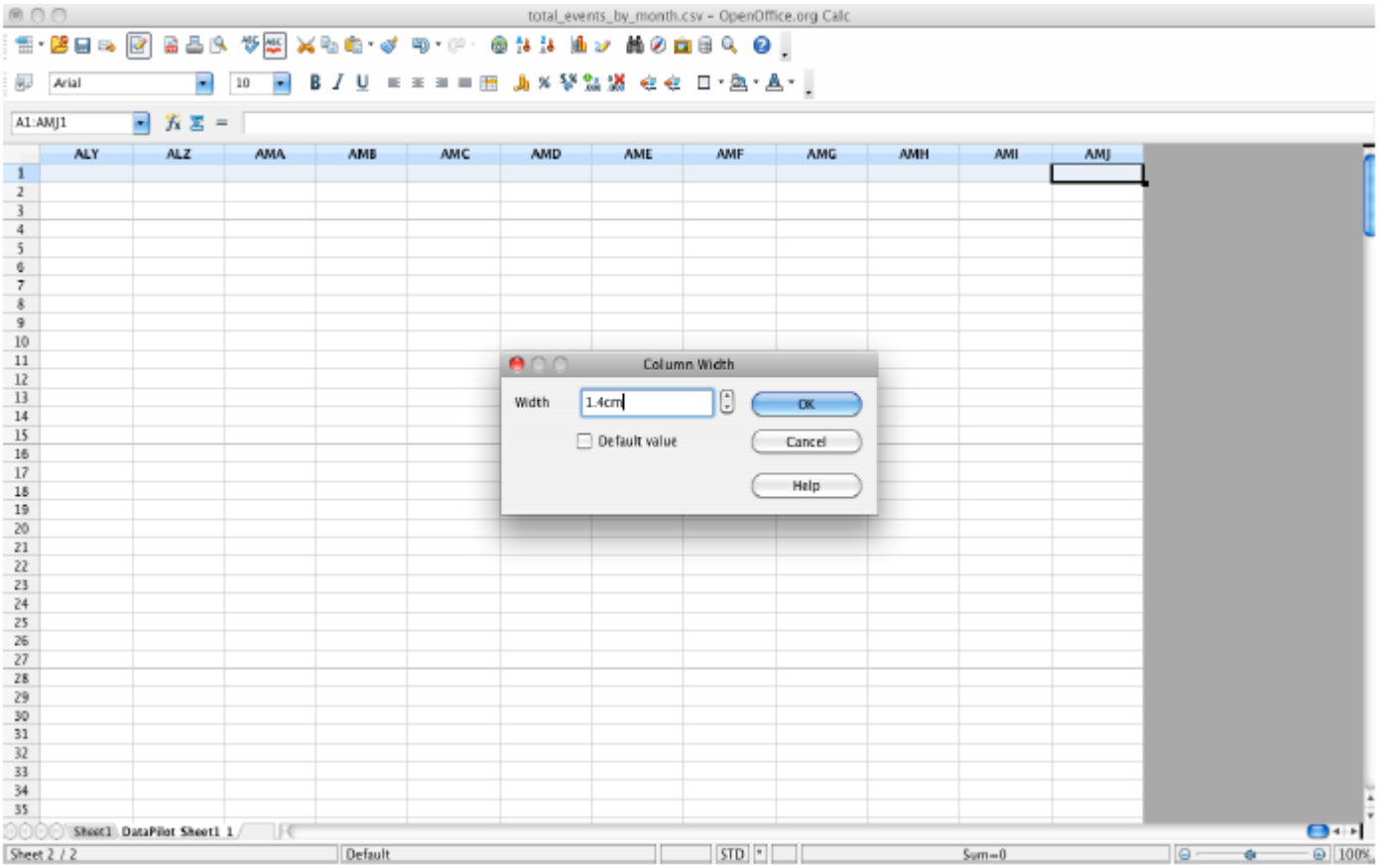


Illustration 3: Setting the Column Width

- Format the left and rightmost columns … and here is the result:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Filter																
2																	
3	Sum - Count	Academic Year	Year	Month													
4		2005_06										2006_07		Total Result			
5		2005			2006							2006					
6	Event Type	10	11	12	1	2	3	4	5	6	7	8	9				
7	Added new section					1	2	2						5			
8	alias.add	10	7	7	26	3	2	18	24	33	68	30	148	377			
9	alias.del	2	12	2	7	13	13	10	6	2	41	14	46	167			
10	ann.delete.any				1	2	1	2	3	1	0	7	6	21			
11	ann.delete.own	1	2	0	1	1	24	13	24	13	1	2	22	103			
12	ann.new	83	51	13	102	104	120	25	17	107	226	153	237	1238			
13	ann.revise.any			0	1	0	1	4	1	0	2	30	8	47			
14	ann.revise.own	5	17	3	18	10	21	1	42	7	5	5	56	190			
15	asn.delete.assignment										0	0	3	3			
16	asn.new.assignment		0				2		0	0	2	3	22	30			
17	asn.new.assignmentcontent		1				2		2	0	3	4	22	34			
18	asn.revise.assignment						2		2	1	1	1	7	14			
19	asn.revise.assignmentcontent						1		3	2	0	2	1	9			
20	asn.submit.submission					2				1		1	1	4			
21	calendar.new	56	75	9	213	548	116	172	438	131	99	51	188	2096			
22	calendar.revise	9	8	1	47	85	153	39	120	132	104	323	1303	2324			
23	chat.delete.any	0			1		0	5	8	6	15		1	35			
24	chat.delete.own	5	2		1	0	13	3	5	2	2	1	3	36			
25	chat.new	83	117	3	2	39	568	215	59	134	219	265	101	1786			
26	content.delete	0	429	921	6	853	310	149	611	1163	7026	38	5760	17291			
27	content.new	270	1543	1468	2969	1481	426	960	1924	2790	17724	4020	5394	41069			
28	content.read	464	1920	1396	18150	6696	7229	5875	2822	14533	42400	1559	3075	106420			
29	content.revise	6	20	19	194	376	231	8	123	101	126	855	64	1922			
30	digest.del				0				2		0		2	4			
31	disc.delete.any			0	1			2	3	0	1		6	11			
32	disc.delete.own				0	1	3	7	3	4	0		1	19			
33	disc.new	6	3	1	1	39	66	10	1	53	52	18	11	262			
34	disc.new.category											14	33	48			
35	disc.null	4	4	2	6	7	45	26	16	11	5		6	133			
36	disc.revise.any				4									4			
37	disc.revise.own				1	3		0	1	1	0		3	9			
38	help.access											79	670	749			
39	help.search											24	107	131			
40	mail.delete.any	4							1			1	4	9			
41	mail.new	6	16	3	11	8	17	1	2	73	81	184	10	411			
42	mail.revise.any	1	0		2	5	6	0	3	8	8	0	11	43			
43	memory.reset	2	0											2			
44	org.theospi.glossary.updateAdd						0							0			
45	prefs.add	1	1		3	1	3	4	1	3	8	2	18	44			
46	prefs.del	3	2	1	5	0	1	1	6	4	14	3	41	82			
47	prefs.upd		3		1	5	6	0	13	0	5	1	4	32			
48	realm.add	16	17	8	95	32	5	106	119	13	302	37	160	929			
49	realm.del		0		3	6	30	5	5	16	3	24	37	130			
50	realm.upd	133	79	27	19	1386	913	3	352	1417	634	384	2382	7730			
51	realm.upd.own	6	9	7	46	40	36	5	60	10	23	39	16	303			
52	Redefined section memberships to				0	2	1							3			
53	Removed section						1							1			
54	section.add									3	1	2	6	12			
55	section.disband									1		3	7	10			
56	section.join											3		3			
57	section.members.reset												2	2			
58	section.student.reg=true									0	0	1		1			
59	section.student.switch											3		3			
60	section.student.switch=false											0		0			
61	section.student.switch=true									1	1			1			
62	section.update									1		7	8	16			
63	site.add	45	11	12	50	185	85	94	43	148	77	95	47	893			
64	site.del		0		1	1					1	2	7	12			
65	site.upd	39	56	9	19	12	227	361	257	462	39	66	1619	2964			
66	site.upd.grp.mbrshp				0	0	0						1	1			
67	syllabus.delete.posted		1		3		0	2	8			2		13			
68	syllabus.post.change		1		4	7	11	0	23	2	2	4	7	60			
69	syllabus.post.new		1		2	3	2	2	1	1	1	7	0	20			
70	Updating site: allow student reg					1								1			
71	Updating site: allow student swi					1								1			
72	user.add	2	2		19	11	78	32	8		5	1	20	180			
73	user.del	3				1	4		1		30	0		39			
74	user.login	261	389	320	670	994	2509	4365	944	7465	5425	5972	14507	46836			
75	user.logout	295	309	154	677	1912	2190	678	1059	644	1901	2147	1666	13562			
76	user.upd.any	1	1		1	1	4	2	1	1	4		0	15			
77	user.upd.own	2	1		3	2	2	1	3	0		2	1	16			
78	wiki.revise								105	2199	2691	841	1192	7027			
79	Total Result	1823	5124	4404	23389	15692	15466	13225	9367	31690	92374	17131	38862	257933			
80																	
81																	

Illustration 4: Sample showing data pivoted for Academic Year 2005-06

Exploitable artefact: UCIAD Ontology

Originators/Authors

Salman Elahi and Mathieu d'Aquin

UCIAD, Open University

Purpose

To describe activity data from log files in a flexible way being able to deal with a multiplicity of different sources, and to enable reasoning over activity data.

Background

To provide flexible and extensible views over activity data from a user-centric point-of-view

Ingredients

- UCIAD Ontology expressed in OWL

Assumptions

- Dealing with online activity data
- Ontology complies with OWL2RL Profile
- Relies on FOAF, W3C Time, and Action Ontologies

Warnings

Under continuous development

Specification

See http://uciad.info/ub/2011/03/uciad-ontologies-a-starting-point/ - more complete and formal specification will be made available when version 1.0 will be realised)

Availability

Source code and the ontology are available at http://github.com/uciad (direct link for ontology https://github.com/uciad/UCIAD-Ontologies)

Presenting data

The recipes for presenting data are:

- Producing PDF Reports Programmatically
- Plotting Calendar Data with GNUplot
- How to work with Gephi to visualise a network of Sites connected by Users
- Visualising OpenURL Referrals Using Gource
- OpenURL Router Data: Total Requests by Date

Producing PDF Reports Programmatically

Originators/Authors

James S Perrin

AGtivity, University of Manchester

Purpose

To automatically produce a report for end users in charge of a physical Access Grid Node or a Virtual Venue, showing usage statistics and patterns by combining images and data in PDF document. A complete report based on usage data should be produced without manual intervention.

Background

Programs and scripts had been developed that would parse the usage logs from the Access Grid servers and generate graphs from the processed and filtered output. This works fine for the AG support staff, but there was a need to combine the data, graphs and other information into a report for the end- user so they could make their own decisions from how their node or venue was being utilised. The scripts can be run automatically for all nodes (> 400) and virtual venues (>50) as manual compilation of these reports would be laborious.

We describe the basic elements needed to produce a PDF document using Python. There are PDF libraries for other scripting languages such as PERL but Python was chosen as it considered easier to learn and understand.

Ingredients

- Python
- Reportlab Toolkit (PDF module)
- Python Image Library (to load
- Data sources e.g.
    - Graphing software (GNUplot)
    - Log file and log parser (C++ executable in this case)

Assumptions

Some programming knowledge. Python specific knowledge is a bonus, however this was my first Python project.

Method

Install Python, Reportlab Toolkit and Python Image Library. Write a Python script that can call your log parser, generate graphs and then combine the results into a PDF.

Run the Python script with the name of the node (or venue) of interest. Optionally filter for a range of dates. Email the PDF to the end user.

Individual Steps

Installation

If not already available install Python http://www.python.org

Install the following modules (check the versions are compatible with the version of Python installed e.g.

2.7)

- Reportlab ToolKit http://www.reportlab.com/software/opensource/rl- toolkit/
- Python Image Library http://www.pythonware.com/products/pil/

Reading Script Arguments

We need to import the sys module so we can reading values passed to the script

```
# add this at the top of the script
import sys
# get the venue the report will be produced for
venue = sys.argv[1]
```

Executing External Commands

If you want to run external commands such as your log parser or graphing tool you need the following

code:

```
# add this at the top of the script
import os
# then you can execute a command like this
os.system('mylogparser –option foo.log > bar.csv')
```

Reading CSV Files

Python has support for reading CVS files as a core module:

```
# add this at the top of the script
import cvs
# open a file
csvfile = open(venue+‘.csv', 'rb')
reader = csv.reader(csvfile)
# read in rows
total = 0
for row in reader:
    # skip header
    if(reader.line_num > 1):
        total += float(row[2]) # sum the 3 rd column
```

Setting up a PDF

There is very little to set up, it is just a matter of importing all the right modules. The following allows a PDF with text, images and tables to be created:

```
from reportlab.lib.styles import getSampleStyleSheet
from reportlab.lib.pagesizes import A4
from reportlab.lib.units import inch
from reportlab.platypus import SimpleDocTemplate, Paragraph, Image,
Spacer, PageBreak, Table, TableStyle
```

Some different text style would be nice. We can grab these from SampleStyleSheet that Reportlab provides. We can use these as is or modify them:

```
styles = getSampleStyleSheet()
styleNormal = styles['Normal']
styleHeading = styles['Heading1']
styleHeading.alignment = 1 # centre text (TA_CENTRE)
```

Creating a PDF ‘Story’

Reportlab has a few different ways to construct a PDF, the easiest is to create a story. This is simply a list of objects that will be rendered into a PDF:

```
story = []
# Text is added using the Paragraph class
story.append(Paragraph(‘AGtivity Report’, styleHeading))
story.append(Paragraph(‘This report describes the usage of the AG venue
’+venue, styleNormal))
# Images just need a filename and dimensions it should be printed at
story.append(Image(‘graph-‘+venue+’.png’, 6*inch, 4*inch))
# Spacers and pagebreaks help with presentation e.g.
```

```
story.append(Spacer(inch, .25*inch))
…
story.append(PageBreak())
# Data can be best presented in a table. A list of lists needs to
declared first
tableData = [ [‘Value 1’, ‘Value 2’, ‘Sum’],
[34, 78, 112],
[67,56, 123]
[75,23, 98]]
story.append(Table(tableData))
```

Tables have their own TableStyle class which can set fonts, alignment and boxes. The Reportlab documentation covers this in detail.

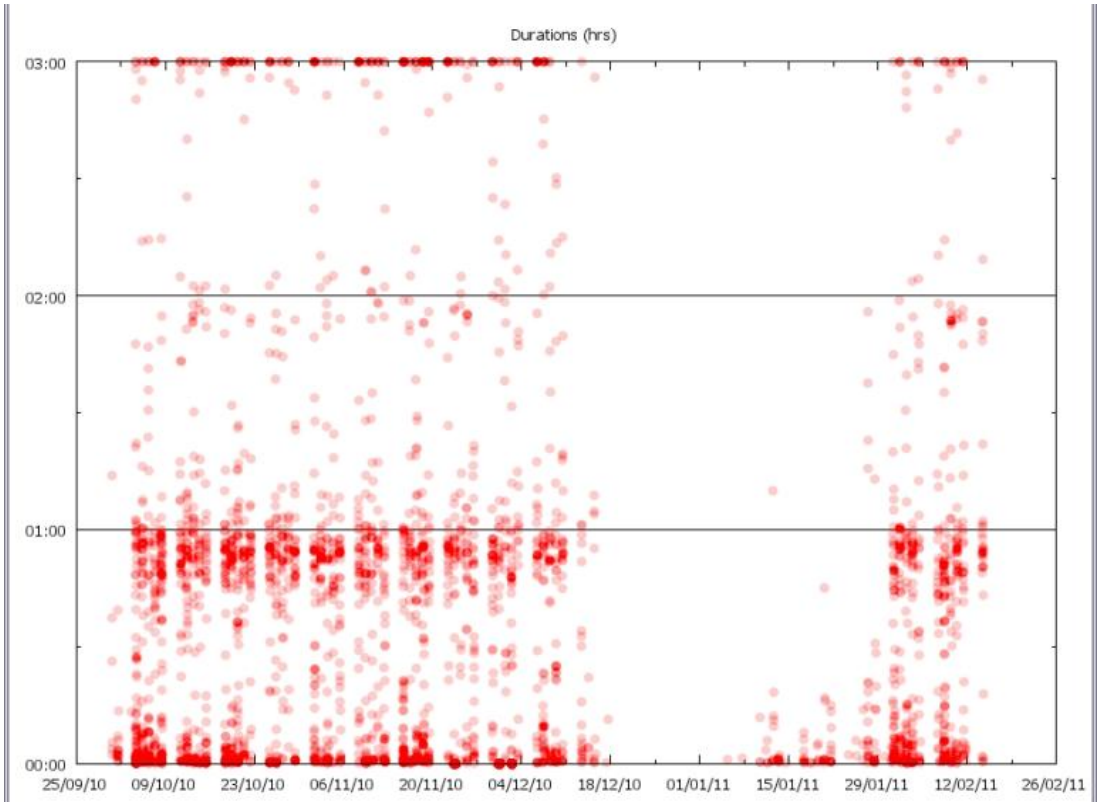
Creating the PDF File

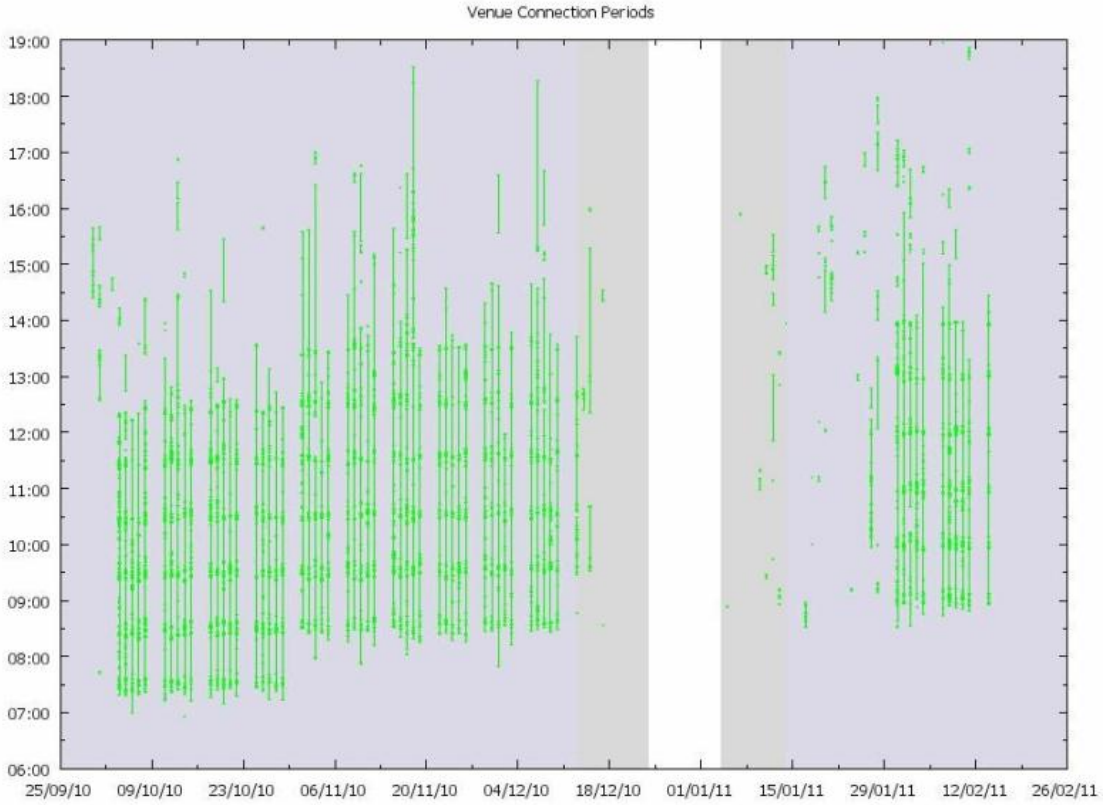
The SimpleDocTemplate class takes care of most of the work for a straight forward document.

```
doc = SimpleDocTemplate(‘report.pdf', pagesize = A4, title = "Access
Grid Venue Usage Report “+venue, author = "AGSC")
doc.build(story)
```

Output Data

A PDF report is produced with graphs of meeting durations and timetable. For venues, occupancy is also plotted. A table of statistics for venues attended by the node, or nodes attending the venue is generated, as well as supporting data on average, minimum/maximum meeting durations.




Venue Connection Periods
19:00
18:00
17:00
16:00
15:00
14:00
13:00
12:00
11:00
10:00
09:00
08:00
07:00
06:00
25/09/10
09/10/10
23/10/10
06/11/10
20/11/10
04/12/10
18/12/10
01/01/11
15/01/11
29/01/11
12/02/11
26/02/11

AG Node	No.of Visits	Total	Average	Median	Min	Max
ALL	1459	2459:29:46	1:41:08	0:49:07	0:00:00	95:15:39
Cardiff - Maths	70	901:48:10	12:52:58	3:03:18	0:00:14	91:35:09
Leeds Magic	97	338:15:06	3:29:13	0:58:53	0:00:12	77:09:15
manchester@magic	157	232:21:34	1:28:47	0:58:13	0:00:11	6:32:08
Loughborough Magic	158	180:51:36	1:08:40	0:57:53	0:00:06	4:59:49
Birmingham Magic	145	155:24:00	1:04:18	0:50:29	0:00:00	19:38:43
agn2.mss.surrey.ac.uk	43	111:14:34	2:35:13	0:13:23	0:00:10	95:15:39
York MAGIC	128	92:00:58	0:43:07	0:46:38	0:00:08	2:52:20
Newcastle Magic	63	82:02:32	1:18:08	0:49:41	0:00:09	22:37:44
Sheffield MAGIC	70	75:46:44	1:04:57	0:55:30	0:00:12	6:20:24
ivmagic	102	55:10:40	0:32:27	0:34:34	0:00:12	1:55:59
Southampton MAGIC	78	51:38:53	0:39:43	0:48:44	0:00:16	1:54:10
MTH Mobile Magic	84	45:26:54	0:32:27	0:42:37	0:00:16	1:52:52
Nottingham Magic	56	36:07:09	0:38:41	0:48:51	0:01:02	1:53:36
GWR exeter@hrs	40	21:58:24	0:32:57	0:11:37	0:00:09	3:29:32
Durham MAGIC	26	21:35:20	0:46:15	0:48:46	0:02:40	1:52:08
Leicster Magic	24	14:52:16	0:37:10	0:48:18	0:00:05	0:55:25
Lancaster MAGIC	12	9:43:27	0:48:37	0:53:57	0:01:08	1:52:28
U of East Anglia	17	7:48:21	0:27:33	0:06:56	0:00:14	1:19:45
magic	11	6:07:07	0:33:22	0:40:36	0:01:06	0:58:26
pbd.nerc-essc.ac.uk	17	5:10:44	0:18:16	0:08:50	0:00:52	0:52:43
ACET-ag	8	3:25:59	0:25:44	0:28:00	0:00:18	0:54:57
Tom Murray	11	2:49:34	0:15:24	0:05:16	0:00:36	0:58:39
Christian Bick	7	1:29:21	0:12:45	0:09:36	0:00:56	0:39:00
AVS UCF Peter Lanyon Building	2	1:15:20	0:37:40	0:37:40	0:02:23	1:12:57
GWR University	3	1:09:32	0:23:10	0:17:42	0:00:14	0:51:36
Ian Dennet visitnet	2	1:09:01	0:34:30	0:34:30	0:03:30	1:05:31
Jitesh_gajjar@magic	1	0:56:22	0:56:22	0:56:22	0:56:22	0:56:22
Jitesh Gajjar	4	0:54:13	0:13:33	0:14:10	0:01:25	0:24:28
Keele MAGIC	10	0:27:21	0:02:44	0:01:07	0:00:08	0:15:24
anon	4	0:13:18	0:03:19	0:03:07	0:00:53	0:06:11
Cardiff - Magic	1	0:07:07	0:07:07	0:07:07	0:07:07	0:07:07
Sam Elliott - Leeds	1	0:05:39	0:05:39	0:05:39	0:05:39	0:05:39
Surrey Computing	4	0:01:51	0:00:27	0:00:24	0:00:12	0:00:50
GWR University of Exeter	1	0:00:39	0:00:39	0:00:39	0:00:39	0:00:39

Table I: Statistics for duration of visits for each attending AG nodes (H:M:S). 34 AG nodes in total.

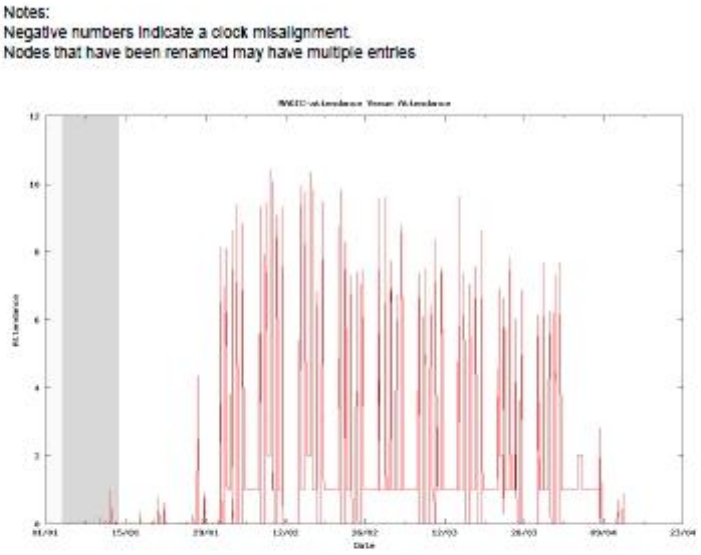


Figure 3: Number of attending AG nodes

Plotting Calendar Data with GNUplot

Originators/Authors

James S Perrin

AGtivity: University of Manchester,

Purpose

To display multiple discrete events which have a start and end time, such as meetings, in a diary/calendar like plot.

Background

The Access Grid activity data quite literally describes what a group of people were doing at discrete periods of time i.e. have an Access Grid meetings. It is therefore of benefit to present this data in a

calendar or diary format to describe when the activities at virtual venue or physical node were occurring. We could look at room booking data but this only shows what was proposed to happen, meetings occur ad- hoc or cancelled without updating the system. The diary format allows trends and patterns in actual room usage to be observed. We could go further and use cyclic plots to compare activity on a week by week basis say.

Ingredients

- GNUplot
- Some data for events with start and end times

Assumptions

We assume that each event occurs within one day ie starts after 00:00 and finishes before 24:00 the same day (we can clamp the data if necessary). The timestamps should be give the time and date, UNIX timestamps (seconds since UNIX Epoch) are most easy to handle.

The user is already familiar with basic GNUplot commands and usage.

Warnings

GNUplot has its own epoch value and Excel using a floating point number for timestamps.

Method

- Define some new flat ended arrow styles in GNUplot
- Set up the data formats and axes to handle the time stamps correctly
- Plot data ‘with vectors’

Individual Steps

Set Arrow Styles

We want to use flat headed arrows to show the extent of each event.

- unset style arrow
- set style arrow 1 heads back nofilled linetype 2 linewidth 2.000 size screen 0.001,90.000,90.000

Set Time Format and Axes

Tell GNUplot that time will be plotted on both X and Y axes:

- set xdata time
- set ydata time

and let it know that the value will be in UNIX time:

- set timefmt x “%s”
- set timefmt y “%s”

Then tell GNUplot how to format the axes, the date along the X and time along the Y axis:

- set format x "% g"
- set format y "%H:%M"

of course we should label the axes too:

- set xlabel “Date”
- set ylabel “Time”

Manipulating and Plotting the Data

Let's assume the start and end times are in columns 1 and 2.

Multiple events occurring on the same day should be vertically aligned ie plotted at whole days rather than at a day plus seven hours and fourteen minutes. The x values are therefore rounded down using the following expression:

(floor($1)/86400)*86400)

Where 86400 is the number of seconds in day. The floor() function is used as a belt and braces approach to convert the (by default) float value into an integer. Along the Y axis we just need the time of day, using the modulus operator and the seconds per day we get the remainder:

(floor($1)%86400))

GNUplot vectors use a start point and a relative offset (x,y,dx,dy) so all we need to compute the duration of the event is:

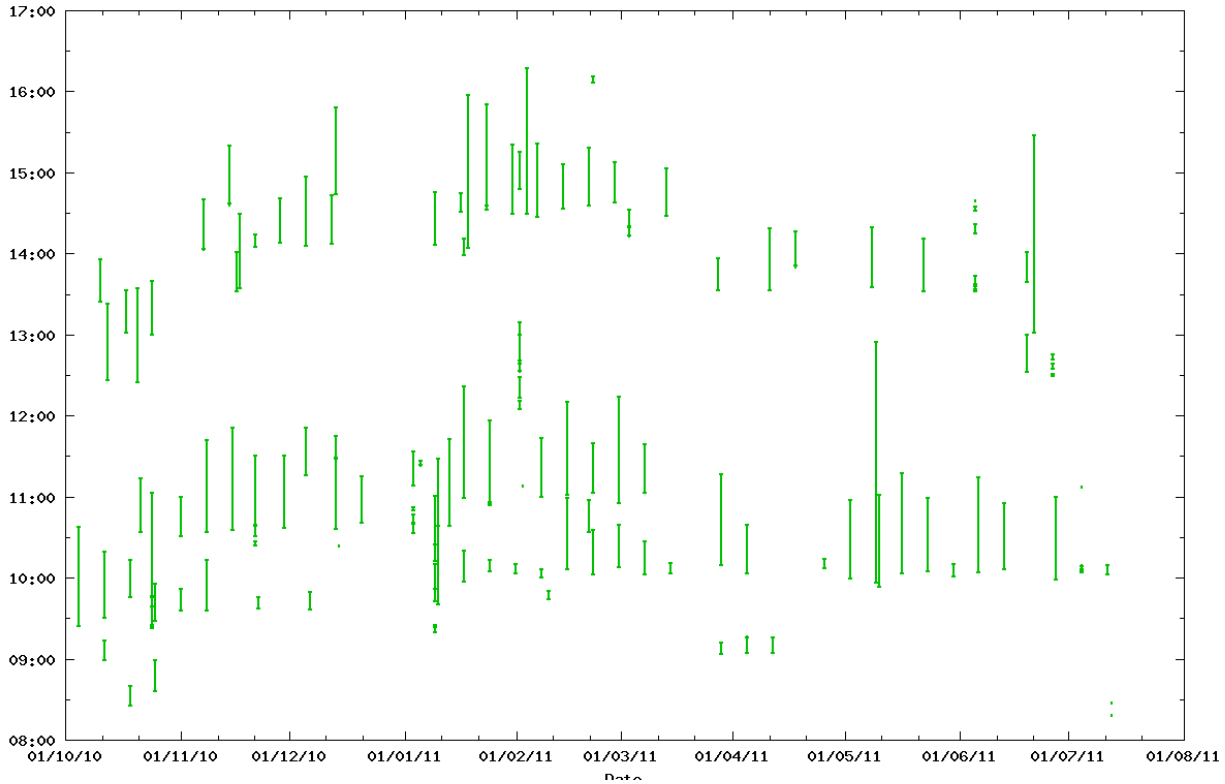
($2-$1)

Putting these expressions altogether in the plot command as follows:

plot mydata.csv using ((floor($1)/86400)*86400):(floor($1)%86400):(0):($2-$1) with vectors as 1

The data is plotted with vectors using the arrowstyle (as) 1 that was defined earlier.

Output Data



Resources

The batch_timetable.plt GNUplot script is available as part of our log parsing package:

http://wiki.rcs.manchester.ac.uk/community/AGProjects?action=AttachFile&do=get&target=aglp.zip

How to work with Gephi to visualise a network of Sites connected by Users

Originators/Authors

Anne Clarke

EVAD, University of Cambridge

Purpose

Instructions on how to import data from a list of sites and their active users and visualise the network using Gephi.

Background

We have developed tools that can be run to extract a given list of sites and their active users from the VLE logging data. This gives us users that have actually visited the site as opposed to those who have been given access to the site.

Ingredients

- Perl statistics tool developed for this project (this may need customising to fit your version of the VLE logging data).
- Gephi application
- Basic knowledge of speadsheets – OpenOffice was used here.

Assumptions

The user has downloaded the current version of the Gephi application and worked through at least the basic on-line tutorial.

Method

You will first prepare the data for import into Gephi. After importing it, you will be able to visually explore the network using Gephi

Individual steps

Prepare Data for Gephi:

- Create a list of sites and active users (eg all_active_users_for_sites.csv):
  This can be done by inputting a list of sites to the perl stats.pl script and running this to produce a list of active users for each site. Eg active_users_for_.csv

Join these files together: *cat act* > all_active_users_for_sites.csv*

- Use spreadsheet sort to sort in order of Site and remove all the heading lines (Site, User etc).

Check and correct the user ids at this point – I found some (eg thu21) which had been mangled by the spreadsheet into date format.

- Run gephi perl script against this to produce list of nodes and edges for import into Gephi.

eg *perl gephi.pl -f all_active_users_for_sites.csv -w*

This determines the number of active users for each site and the number of shared users between any 2 sites. It will produce 'nodes.csv' and 'edges.csv' in the input file location.

Importing Data into Gephi:

- Open up Gephi and select File > New Project
- Click on the Data Laboratory Tab
- Click on Import Spreadsheet and browse to the location of the nodes file that you created above.
- Make sure 'As Table' is set to Nodes Table
- Click on Next
- Set the Total Users field to Integer – choose from the drop down list (you may need to go up though the list of options)
- Click on Finish
- Nodes table should now be imported

- Repeat for Edges table
- Go to Overview tab
- You should now see a network !

Working with Gephi:

(Please follow the Gephi Getting Started tutorial to learn the details for these steps)

Setting the node size to Number of Users:

- Select the Ranking tab at the top right of the screen
- Choose Total Users as the Ranking Parameter
- Click on the colour wheel button
- Set the Colour range
- Click on Apply
- The nodes will now have colour intensity linked to the number of users
- Click on the diamond icon
- Set the Min Size (eg 10) and the Max Size (eg 100). The range should already be set to the max and min no of users.
- Click on Apply
- The node sizes should now reflect the number of users.

Changing the Network Layout:

- From the layout window on the bottom left of the page select Force Atlas
- Set Repulsion strength 10000, tick Adjust by sizes.
- Click on Run
- Click Stop when it has stopped changing

Showing the Labels

- Click on the black 'T' icon at the bottom of the Graph display
- You should see the node labels
- You can use the rightmost slider at the bottom of the the Graph display to change the text size.

To stop the Labels overwriting each other:

- From the Layout tab – bottom of left hand side of page:
- Choose Label Adjust
- Click on Run

To highlight clusters of nodes:

- Click on the Statistics tab on the left hand panel.
- Click to run Modularity – this looks for clusters of nodes that are tightly associated with each other and more loosely linked to other nodes.

- Click on the Partition tab on the right hand panel.
- Choose Modularity Class and click on Run
- You should now see tightly associated nodes have the same colours.

Hiding Nodes

- Go to Filters menu
- Click on Topology then drag 'Degree range' down to the Query section.
- Use the slider to change the Degree range settings
- Press filter
- You should see that the less connected nodes have been hidden.

Interaction with Gephi

- If you click on one of the nodes in Overview mode you can see the linked nodes highlighted.

Producing an output view

- Click on the Preview tab.
- Click on Refresh
- If you don't see anything try clicking on the Reset Zoom button
- Use the Default preset with the following settings:
    - Show labels ticked
    - Increase label font if needed
    - Undirected Curved set
    - Rescale weight ticked (this makes sure the connecting lines are not really thick)
    - Thickness – maybe set to 4 – this thickens up the lines reduced above
    - Proportional Label size un-ticked
    - Click on Refresh
- You can use the Preview ratio slider to filter out less active nodes.
- Export to SVG/PDF to save a copy

Output

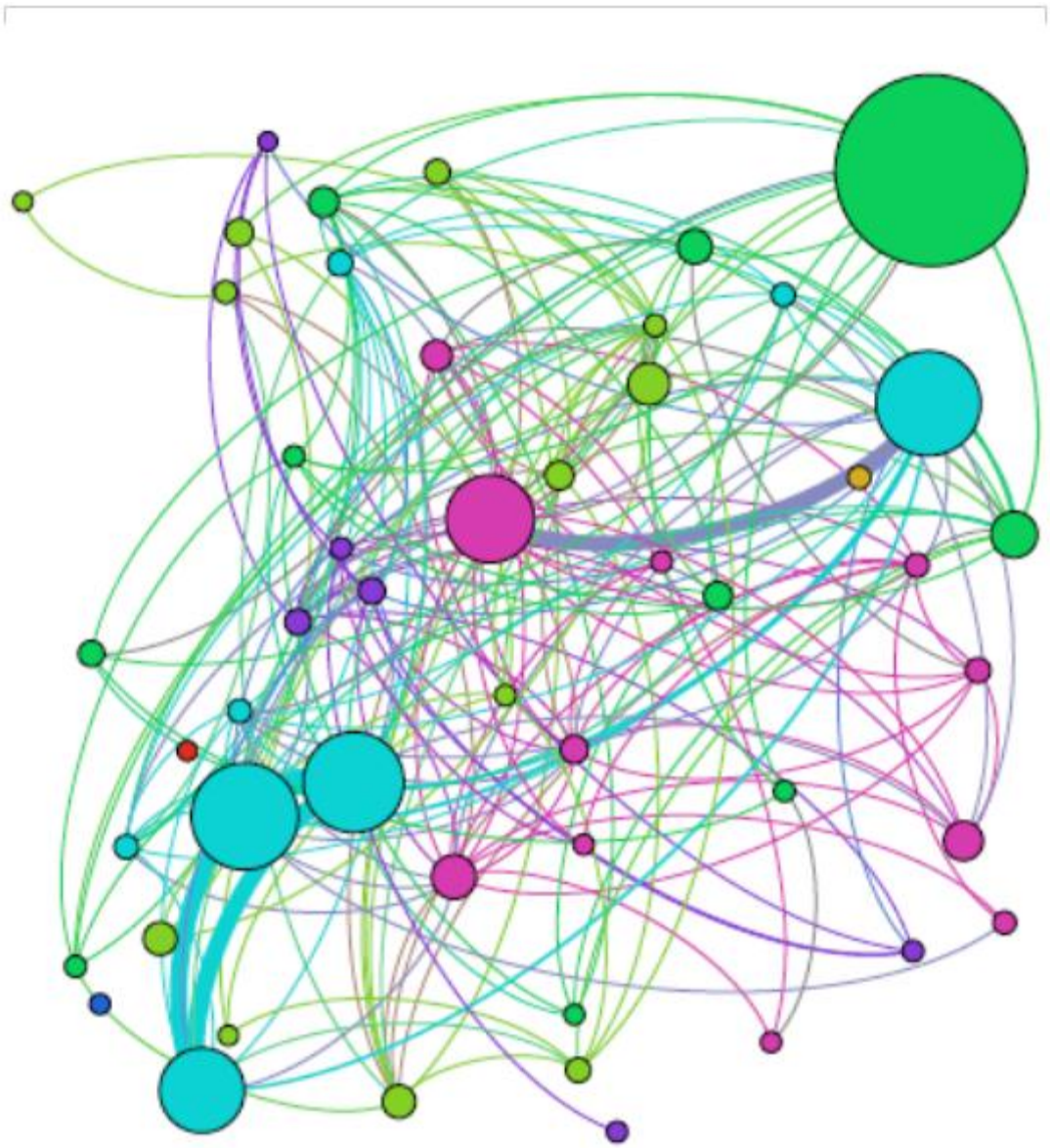


Illustration 1: Gephi network connecting the Top50 Sites from one of our Departments

Visualising OpenURL Referrals Using Gource

Tony Hirst writing at his blog on visualising OpenURL Logs shows how this can be done using Gource:

> Picking up on the OpenURL referrer data that I played with here, here's a demo of how to visualise it using Gource [video]:
>
> If you haven't come across it before, Gource is a repository visualiser (Code Swarm is another one) that lets you visualise who has been checking documents into and out of a code repository. As the documentation describes it, "software projects are displayed by Gource as an animated tree with the root directory of the project at its centre. Directories appear as branches with files as leaves. Developers can be seen working on the tree at the times they contributed to the project."
>
> One of the nice things about Gource is that it accepts a simple custom log format that can be used to visualise anything you can represent as a series of actors, doing things to something that lives down a path, over time… (So for example, PyEvolve which visualises Google Analytics data to show website usage.)

In the case of the EDINA OpenURL resolver, I mapped referring services onto the "flower"/file nodes, and institutional IDs onto the people. (If someone could clarify what the institutional IDs - column 4 of the log - actually refer to, I' d be really grateful?)

To generate the Gource log file - which needs to look like this:

> timestamp - A Unix timestamp of when the update occurred.
>
> username - The name of the user who made the update.
>
> type - initial for the update type - (A)dded, (M)odified or (D)eleted.
>
> file - Path of the file updated.

That is: 1275543595|andrew|A|src/main.cpp

I used a command line trick and a Python trick:

```
cut -f 1,2,3,4,40 L2_2011-04.csv > openurlgource.csv

head -n 100 openurlgource.csv > openurlgource100.csv
```

(Taking the head of the file containing just columns 1,2,3,4 and 40 of the log data meant I could try out my test script on a small file to start with…)

```
01 import csv

02 from time import *

03 f=open('openurlgource.csv', 'rb')

04

05 reader = csv.reader(f, delimiter='\t')

06 writer =
csv.writer(open('openurlgource.txt','wb'),delimiter='|')

07 headerline = reader.next()

08 for row in reader:

09 if row[4].strip() !='':

10 t=int(mktime(strptime(row[0]+" "+row[1], "%Y-%m- %d
%H:%M:%S")))

11
writer.writerow([t,row[3],'A',row[4].rstrip(':').replace(':','/')])
```

(Thanks to @quentinsf for the Python time handling crib:-)

This gives me log data of the required form:

```
1301612404|687369|A|www.isinet.com/WoK/UA

 1301612413|305037|A|www.isinet.com/WoK/WOS

 1301612414|117143|A|OVID/Ovid MEDLINE(R)

 1301612436|822542|A|mendeley.com/mendeley
```

Running Gource uses commands of the form:

```
gource -s 1 --hide usernames --start-position 0.5 -- stop-
position 0.51 openurlgource.txt
```

The video was generated using ffmpeg with a piped command of the form:

```
gource -s 1 --hide usernames --start-position 0.5 -- stop-
position 0.51 -o - openurlgource.txt | ffmpeg - y -b 3000K -  r 60
```

```
-f image2pipe -vcodec ppm -i - - vcodec libx264 - vpre  slow -
threads 0 gource.mp4
```

Note that I had to compile ffmpeg myself, which required hunting down a variety of libraries (e.g. Lame, the WebM encoder, and the x264 encoder library), compiling them as shared resources (./configure -- enable-shared) and then adding them into the build (in the end, on my Macbook Pro, I used *./configure - enable- libmp3lame -enable- shared - enable-libvpx - enable- libx264 -enable- gpl -disable- mmx - arch=x86_64* followed by the usual *make* and then *sudo make install*).

Getting *ffmpeg* and its dependencies configured and compiled was the main hurdle (I had an older version installed for transforming video between formats, as described in ffmpeg - Handy Hints, but needed the update), but now it's in place, it's yet another toy in the toybox that can do magical things when given data in the right format: gource:-)

OpenURL Router Data: Total Requests by Date

Originators/Authors

Sheila Fraser

OpenURL, EDINA, University of Edinburgh

Purpose

To display a graph of the OpenURL Router data total requests by date as one example of how the data can be used

Background

All OpenURL requests made by end users via the Router at openurl.ac.uk are logged, and (subject to the metadata included by the referring service) provide a record of the article that user was attempting to find via their local resolver. A subset of the log has recently been released as open data and this can be analysed in a variety of ways. This example gives the count of the total number of requests on each date, which shows a weekly usage pattern.

Other related data and visualisations, such as a weekly cycle chart, can be found together with the scripts to generate them from http://figshare.com/figures/index.php/Openurl

Ingredients

- OpenURL Router data (http://openurl.ac.uk/doc/data/data.html)
- R (free software from http://cran.r-project.org/)

Assumptions

- Availability of OpenURL Router data
- Knowledge of how to run Perl scripts on the downloaded file

Warnings

The data is being regularly published, but may be a little out of date following the end of the project.

Individual steps

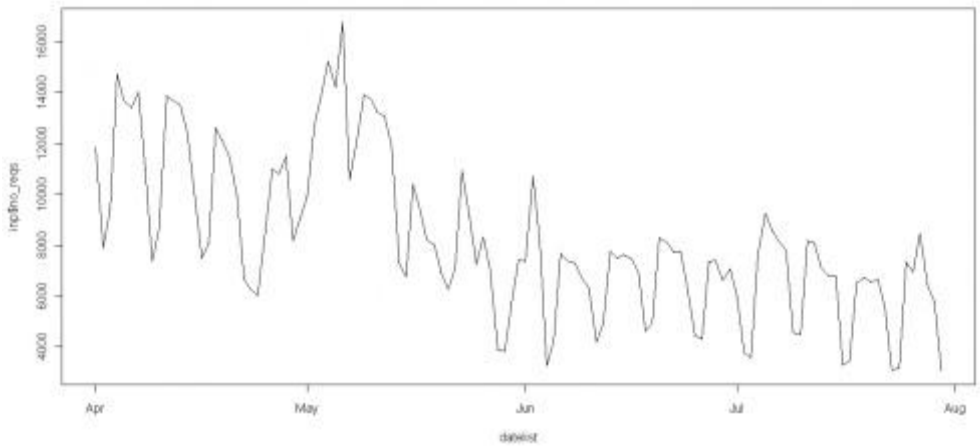
- Download the OpenURL Router data
- Run the script to pre-process the data
- Run the R script to generate the graph

Intermediate data

date	no_reqs	day
01/04/11	11836	Friday

02/04/11	7894	Saturday
03/04/11	9399	Sunday
04/04/11	14719	Monday

Appendix A: Sample output



Appendix B: Pre-processing script

```
#!/usr/local/bin/perl

use strict;

use Data::Dumper;


my $logDir = "../logs/2011_raw/rotated/"; # change to the Level 0 log file name to read from my  $fileName = "2011-08-01_Router.log";


open LOGFILE, $logDir."$fileName" or die "cannot open file for reading\n"; my $line; my %dates;


while (my $line = <LOGFILE>){
 chomp($line);


 # if it's a request to a resolver then count it in
 if($line =~ "redirect.resolver"){
 my $dateTime = (split(/\t/,$line))[0];
 (my $date, my $time) = split(/T/, $dateTime);
 my $hour = (split(/:/, $time))[0];
 $dates{"$date"}++;
 }
}
```

```
# print the counts on the terminal
foreach my $key (sort keys %dates) {
 print "$key\t".$dates{$key}."\n";
```

Appendix C: R script to generate the graph

Using R to generate this example graph, the script is:

```
filepath <- system.file("ourldata", "no_of_reqs_csv_dates_Apr_Jul.csv", package="datasets")
inp <- read.csv(filepath, header=TRUE)
inp$date <- factor(inp$date)
datelist <- as.Date(inp$date, "%d/%m/%Y")
plot(datelist, inp$no_reqs, pch=".")
lines(datelist, inp$no_reqs)
```

Other

Some open data recipes which offers the step by step recipes for working with open data, including:

- An introduction to sourcing open data
- Foraging for data with Scraper Wiki
- Preparing your data
- Sprinkled Statistics
- Fusion Cooking with Foraged Data
- Visualising and exploring survey data with IBM Many Eyes
- A dataset mixture with Yahoo Pipes

Synthesis method

The programme reported here sought to develop robust evidence of practice and of business case as well as technical foundations to establish positively that the HE-specific exploitation of activity data has merits for both the institution and the user. The role of the Synthesis Project has therefore been to support, cohere and leverage this work in the form of documentation to fulfil and sustain the objectives of the call, notably:

- Identifying approaches or technical **solutions** that can be rolled out more widely.
- Supporting the development of **knowledge and skills**.
- Consolidating evidence of valid **business cases** for institutions to engage.
- Taking account of other projects, in the UK (eg JISC Business Intelligence) and elsewhere.

The project ran for 8 months, therefore requiring most work to be undertaken as the funded projects developed. To this end, the Synthesis project used a range of complementary activities, working in partnership with the projects, that would lead to a set of cross- referenced artefacts, accessible though a single reference site (http://activitydata.org/).

- **Live Synthesis** - This involved near-realtime interaction with the projects, refining the synthesis approach developed for the JISC LMS programme; i.e. projects were tasked to make blog posts at key points; these posts plus #jiscad tagged tweets were amalgamated in to a project feed; this informedinformed the synthesis blog and populated a (typically) weekly digital tabloid. available from http://blog.activitydata.org/search/label/digest
- **Activity Project Visits** - Each project was visited at least once. The initial visits took place in the early stages after the project plan had been finalised and the project staff are in place.
- **Activity Strand Events** - We supported the JISC Programme Manager in running two face- to-face events (March & July 2011), bringing together all the projects to share approaches, themes, business cases and sustainability challenges and to refine forward objectives.
- **Dissemination -** Wider dissemination was limited in the programme timeframe (8 months) and was therefore focused on specific communities of interest, notably involving the UKOLN Web Manager's Conference, learning technologists at ALT-C 2011 and comparable research libraries in Europe at the Aarhus workshop.
- **Online Tech Exchange Events** - Drawing on emerging themes, we held 4 online events open to the projects and the wider international community, notably involving US colleagues. These events included project and guest presentations and discussions on such as data models, aggregation, privacy and visualisation. We used Elluminate, Go to Meeting and Webex - finding Elluminate most successful and free of technical glitches. Two of these are available at: http://blog.activitydata.org/search/label/virtualevent
- **Recipes -** We defined a recommended format and then worked with the projects to document technical 'Recipes' encapsulating their approaches to such as data schemas, extraction, standardisation and manipulation. Written by the projects with our editorial input, the number of contributions exceeded expectation.
- **Guides** - We developed a format concise for 'how to' / 'read and do' Guides to provide non-technical advice on key issues, such as rights and business cases; as planned, 12 guides were produced.
- **Final Synthesis** - The synthesis website is modelled as a mind map, which links together all public project and synthesis deliverables. This is an extensive resource and therefore readers are recommended to enter via the themed menu. In addition, a summary report for library directors and institutional managers focused on business cases.
- **Recommendations** - We developed a recommendations report for JISC evaluation and planning purposes.

Recommendations for further work

This is an informal report outlining the likely recommendations from the Activity Data projects to help JISC to determine future work in the area. This is not intended as a public document, but rather to stimulate discussion and lead to a more formal document at a later stage.

There are two things to note at this stage:

- The greatest impact (and ROI) will be in the support of student success. This is one of the wide varieties of end uses of activity data within UK HE;, as exemplified by the range of projects in this programme. We strongly recommend this as one of the components of any future funding.
- We suggest that the next call explicitly funds other universities to pick up of the techniques and / or software systems that have been developed in this programme. This to if they are useful beyond the initial institution, and in this process, discover what the issues may be to make effective use of the techniques and / or systems in a broader context. However, this may not be in accordance with JISC's standard practice and is not an essential part of the recommendations.

The recommendations appear under the following topic areas:

- Student success
- User experience
- Service and system improvement
- Other

Projects in the current Programme addressed the first three of these topics thus:

Project	Student success	User experience	Service and system improvement.
AEIOU		X	
AGtivity			X
EVAD	X		X
LIDP	X		
OpenURL		X	X
RISE		X	
SALT		X	X
STAR Trak	X		
UCIAD		X	

Student success

"It is a truth universally acknowledged that" early identification of students at risk and timely intervention must lead to greater success. One of the successes of the AD Programme is that it has demonstrated, through LIDP's results, that activity data can identify students at risk from patterns in activity data. These students could be supported by early intervention. It has also been demonstrated in work in the US that it can help students in the middle to improve their grades.

Projects working in this area

- LIDP
- EVAD

- STAR-Trak:NG

Recommendations

In year 2, JISC should fund research into what is needed to build effective student success dashboards

Work is needed at least in the following areas:

- Determination of the most useful sources of data that can underpin the analytics
- Identification of effective and sub-optimal study patterns that can be found from the above data.
- Design and development of appropriate algorithms to extract this data. We advise that this should include statisticians with experience in relevant areas such as recommender systems.
- Watching what others are doing including in the areas of learning analytics, including Blackboard and Sakai developments. This can also draw on the work of STAR-Trak:NG.
- Development of a common vocabulary

At this stage it is not clear what the most appropriate solutions are likely to be; therefore, it is recommended that this is an area where we need to "let a thousand flowers bloom". However, it also means that it is essential that projects collaborate in order to ensure that projects, and the wider community, learn any lessons.

Pilot Systems

In year 2 or 3, JISC should pilot the following systems developed in the current programme:

- **LIDP** - further refinement of the algorithms used for instance to look at the effect of different patterns of activity as well as the overall level of activity.
- **EVAD -** trial the EVAD approach and (part) codebase elsewhere at the other Sakai implementations in the UK and/or apply the approach to other VLEs (Blackboard / Moodle).
- **STAR-Trak:NG** - Further development with trial(s) elsewhere. This could include developing a generic framework to support identification of students at risk

User experience

This area is primarily concerned with using recommender systems to help students and (junior) researchers locate useful material that they might not otherwise find, or would find much harder to discover.

Projects working in this area

- RISE
- SALT
- AEIOU
- OpenURL
- UCIAD

Recommendations

It is recommended that in year 2, JISC fund additional work in the area of recommender systems for resource discovery.

In particular work is needed in the following areas:

- Investigation and implementation of appropriate algorithms. This should look at existing algorithms in use and their broader applicability. We advise that this should include statisticians with experience in areas such as pattern analysis and recommender systems.

- Investigation of the issues and tradeoffs inherent in developing institutional versus shared service recommender systems. For instance, there are likely to be at least some problems associated with recommending resources that are not available locally.
- Investigating and trialling the combination of activity data with rating data. In doing this there need to be acknowledgement that users are very frequently disinclined to provide ratings, and that ways to reduce barriers to participation and increase engagement with rating processes need to be discovered in the context of the system under development and its potential users.

Pilot systems

- **RISE -** try elsewhere using either the software, or if with a different VLE and LMS, then the methods and algorithms.
- **SALT -** try elsewhere, either discipline based or at other institutions. SALT may also help to enhance the impact of COPAC. A similar approach could also be tried as a shared service, using activity data from a representative sample of university libraries.
- **OpenURL -** Run a trial to use OpenURL data to enhance existing or build new recommender systems within one or more institutions.

Service and system improvement

Activity data provides information on what is actually being used / accessed. The opportunity exists to use data on and how and where resources are being used at a much finer level of granularity than is currently available. Activity data can therefore be used to help inform collection management.

Note that this is an area where shared or open data may be particularly valuable in helping to identify important gaps in a collection.

Projects working in this area:

- AGtivity
- EVAD
- SALT
- OpenURL

Recommendations

It is recommended that in the coming year JISC should fund work to investigate how activity data can support collection management.

In particular work is needed to

- Consider how activity data can supplement data that libraries are already obtaining from publishers, through projects such as JUSP and the UK Research Reserve.
- Assess the potential to include Open Access journals in this work.
- Pilot work based on SALT and OpenURL to see if the data that they are using is helpful in this area.

It is recommended that JISC fund projects to use activity data from VLEs and related tools to understand how these actually being used.

- Consider building on the work of EVAD to use activity to understand how VLEs and related tools are actually being used.

Other

The following are important areas that JISC should pursue.

It is recommended that JISC should fund work to develop (or reuse and integrate existing open source tools) to visualise activity data

Tools of this kind could support at least

- The exploration activity data before designing for its use
- The display the outputs of analysis of activity data, eg to display learning analytics

JISC might do this by encouraging and funding projects that

- Incorporate useful reusable visualisation components
- Develop general purpose activity data visualisation toolkits (and demonstrate their use)

It is recommended that JISC continue work on open activity data

In particular fund projects that

- Develop one or more repositories for open activity data, including APIs to store and access the activity data. These repositories should be capable of accommodating the diversity of activity data formats.
  Suitable noSQL databases could be used for this purpose.

With repository use, the project(s) should

- Identify common and variant activity data in different kinds of activity data
- If possible, formulate common formats that might become standards

It is recommended that one or more projects in year 2 should investigate the value of a mixed activity data approach in connection with noSQL databases in order to maximise flexibility in the accumulation, aggregation and analysis of activity data and supporting data sets; the US Learning Registry project may be relevant.

This is distinct in aim from noSQL's potential use in Recommendation 6, there the intent is storage and retrieval, here it is leveraging greater effect from using multiple datasets.

It is recommended that JISC should continue to fund work on of the discovery and use of ontologies in the analysis and interpretation of activity data.

It is recommended that JISC ask appropriate experts (such as Naomi Korn / Charles Oppenheim or JISC Legal) to provide advice on the legal aspects such as privacy and data sharing, similar to *Licensing Open Data: A Practical Guide* (written for the Discovery project).

Acknowledgements

The Activity Data Synthesis team would like to express their sincere thanks to all the following who have contributed to the development of this site.

Firstly we would like to thank JISC for funding this work, and the projects whose work we are bringing together on this site, and without which we would not have been able to undertake this work.

Secondly we would like to thank all the project staff who have been universally extremely helpful in giving their time and information and allowing us to re-use the words that they have written. We hope that they have found the experience beneficial.

Most of the material for this web site has been drawn from material produced and written by the activity data projects funded by the JISC, and in particular from their blog postings, which are linked to from projects pages. As we have made such extensive use of their words and this is not an academic paper we have decided for purposes of readability not to cite their words where they are used. We have also altered them in order to provide greater stylistic cohesion.

The authors are not cited each time that there work has been used as this would make the text unreadable, and much of their work has been edited to give it a common style.

We would like to thank the following project staff:

Activity data to Enhance and Increase Open-access Usage (AEIOU), Aberystwyth University

- Antony Corfield
- Jo Spikes

Exploiting Access Grid Activity Data (AGtivity), University of Manchester

- Martin Turner
- James Perrin

Exposing VLE Activity Data (EVAD), CARET, University of Cambridge

- John Norman
- Tony Stevenson
- Verity Allan
- Dan Sheppard

Library Impact Data Project (LIDP), University of Huddersfield

- Graham Stone
- Dave Pattern
- Bryony Ramsden
- Leo Appleton, Liverpool John Moores University

Recommendations Improve the Search Experience (RISE), Open University

- Richard Nurse

Surfacing the Academic Long Tail (SALT), MIMAS, University of Manchester

- Joy Palmer
- Janine Rigby
- David Chaplin
- Andy Land

- Lisa Charnock

STAR-Trak: NG (Next Generation)

- Robert Moores

User-Centric Integration of Activity Data (UCIAD), Open University

- Mathieu d'Aquin

Using OpenURL Activity Data, Edina, University of Edinburgh

- Sheila Fraser

Synthesis project, University of Manchester

- David Kay, Sero Ltd
- Helen Harrop, Sero Ltd
- Mark Van Harmelen, HedTek Ltd
- Tom Franklin, Franklin Consulting

Other people who have contributed include:

- Andy McGregor, JISC
- Tony Hirst, Open University

The project also wishes to thank international colleagues and projects for their engagement in developing a mutual understanding of the potential of Activity Data and the enabling technologies. special mentions go to:

- Learning Registry at the US DOE / DOD - Steve Midgley & Dan Rehac
- Library Cloud at Harvard University - David Weinberger & Kim Dulin
- Metridoc at the University of Pennsylvania - Michael Winkler & Joe Zucca
- Jens Hofman Hansen of the Statsbiblioteket, Arhus and the 2011 On Tracks conference
- David Massart from EUN
- Susan VanGundy from the US National Science Digital Library

licence