

Automatic Metadata Generation – Use Cases

David Kay, Sero Consulting Ltd on behalf of the JISC MOSAIC project

Contents

1. Scope	1
2. Why is this important?	3
3. Scenarios	6
4. Tools.....	7
5. Standards	7
6. Key Issues	8
7. Recommendations	9
Appendix - National recommender service.....	10

1. Scope

The 'Automatic Metadata Generation – Use Cases' (AMG-UC) project is gathering information about a variety of metadata types from a wide and informed community. To this end the project has commissioned this report on usage-related metadata.

In order to scope the context for the report, it is important to define what might be classified as 'usage-related metadata'. In so doing, three categories are proposed:

- Content usage records
- Proxies for usage records
- User generated metadata

The issue of aggregation is also introduced.

1.1 Content Usage Records

Content activity data, sometimes known as 'attention data'¹, covers more than just 'use' in the range of web activity; it may also cover searching (as in Google search hint stats), browsing, bookmarking (such as 'saved for later' or user defined lists) and purchasing.

Examples of content usage records in Higher Education are as follows:

- A library use record might contain the information that *'a first year Chemistry undergraduate at the University of Hull borrowed a book "Organic Chemistry: A Primer" in academic year 07/08'*.
- Use records could be generated from more sources than library circulation systems, notably from Open URL Resolvers and from VLEs; for example *'a first year Chemistry undergraduate at the University of Hull downloaded the presentation for a lecture on organic chemistry in academic year 07/08'*.
- In addition there is potential for use records to be generated by other content systems, such as institutional and Open Education Resource repositories.

Overall, it is suggested that such usage records have greatest value when linked to the user's circumstance, most significantly their status (staff, student, etc), their course and /or module affiliation and their level of study (Undergraduate, Masters, etc).

¹ One blogger has produced a handy introduction, saying that 'Content is what you think it is, metadata is information about the content, and attention data is qualitative information about the content as deduced from an audience's interaction around the content.'
[<http://walkerfenton.blogspot.com/2007/08/content-metadata-attention-data.html>]

1.2 Proxies for Usage Records

Some records can be considered as providing proxy information with some equivalence to or reinforcement of actual usage data.

- Course or module reading lists represent the pinnacle of 'likely use', especially where an item can be found on multiple reading lists within the UK HE system
- Short loan collection status represents higher likelihood of use than might be assumed for general catalogue records
- Volume of physical holdings represents some indication of popularity, though this may be historic in many cases

It is suggested that such information is best used in tandem with evidence of actual usage, thus providing additional indicators; for example: *'25 first year and 40 second year Chemistry undergraduates at the University of Hull borrowed a book "Organic Chemistry: A Primer" in academic year 07/08, which is on 2 reading lists as follows ...'*

Librarians may also to use this proxy information in isolation to inform management decisions (e.g. relating to stock), especially if comparable data is available from other institutions.

1.3 User Generated Metadata

It may be argued that User Generated Content (UGC) such as reviews and comments should not be classified as metadata and that they are far from automated. Notwithstanding such philosophical distinctions, user contributions linked to course and module affiliations can certainly add value to the selection and ranking of content based on usage activity. Such contributions fall into four categories:

- Ratings – as used by such as Amazon and eMusic, typically scoring a book on a visible scale (e.g. up to 5 stars)
- Reviews – as used by such as Amazon and eMusic, providing personal opinions on content
- Annotations (including Tags, Comments) – applying to the metadata or to the content itself, such contributions seem least popular amongst users
- Lists – as used by such as Amazon and eMusic, lists provide a popular vehicle for making links at all levels, from personal interest to deep specialism; reading lists could be seen as a special form of user generated list.

Again, it is suggested that such information is best used in tandem with evidence of actual usage, providing additional indicators; for example: *'25 first year and 40 second year Chemistry undergraduates at the University of Hull borrowed a book "Organic Chemistry: A Primer" in academic year 07/08 ... it was ranked 3 stars by 15 first years and 3.5 stars 12 by second years ... it is on the 2 reading lists and 4 personal lists ... and has been reviewed as follows.'*

1.4 Aggregation

The capacity for aggregating all three categories of usage records to provide a consortium, national or web-scale perspective should be considered:

- The usage records for one item can be aggregated across a range of universities, user levels and courses
- The associated proxy data and user generated content can be integrated in the same space
- The unit of aggregation is important – for example, at the work or the edition level
- The level of storage is significant – either store all usage records and aggregate in response to queries or store as aggregations (e.g. work / institution / course / year –

involving possibly one order of magnitude less records)

2. Why is this important?

2.1 Historical Context

2.1.1 The Amazon Factor

Amazon has set standards and styles followed and exceeded by a range of similarly configured services, leading to a common recognition and expectation of recommendation devices based on

- overall popularity
- the purchasing patterns of others
- user bookmarks and lists
- the idea of users like me

Network effect is a vital to the interest inspired by such information – contributing both volume and specialism

There is therefore some onus on library and other educational content services to present themselves in a similar manner.

2.1.2 UK HE Community

The JISC & SCONUL LMS Study (2007-8) emphasized the 'concentration' of users and data in web-scale services as the key to generating network effect. The subsequent JISC TILE project (2008-9) indicated that services which capture and concentrate activity are likely to provide critical mass for value added services, such as recommendations and other forms of User Generated Content (UGC).

A range of institution level services have the potential to collect data that tracks user activity. Some of those services already capture such activity (e.g. LMS, ERM, VLE) whilst others have the potential to do so (e.g. Repository, OER services). The activity data sets from all such local systems can be logically linked through student registration data (year, course, module options) without compromising personal data. They also offer opportunity for aggregation of client behaviours at consortium, national or wider levels

The work of Dave Pattern at the University of Huddersfield has demonstrated these possibilities in a local application which wraps the library catalogue in the user context ('Users like you ...'). Huddersfield has also liberated its activity data, linked explicitly to course names, for wider re-use, opening the possibility of aggregation (<http://library.hud.ac.uk/data/usagedata/readme.html>). The power is perceived to lie more in the concentration of data rather than in the local interface.

2.1.3 Large Scale Library Services

Recent shifts in perception amongst libraries regarding the management, location and scale of services may be setting the scene for the sort of services within which usage data may play an important part.

The survey conducted on behalf of SCONUL for the HEFCE Shared Services Study (May 2009) indicated strong and widespread interest in shared services from the 83 respondent institutions (c.45% of the sector). Over 60% of the respondents indicated that they are currently involved in or planning some form of shared services activity. Leveraging larger (web) scale services was seen by 73% as a high/medium 3-5 year priority but only an immediate priority for 46%. Shared library services involving individual user data attracted a lower level of interest (e.g. User Recommendations - 15%), though this may be motivated by uncertainties regarding security and DPA obligations as well as competitiveness.

The potential for large scale usage data is illustrated by projects and services in wider community. In the United States, the MESUR project (MEtrics from Scholarly Usage of Resources – Los Alamos National Laboratory - www.mesur.org) ran from October 2006 to October 2008. Its major objective was to enrich the toolkit used for the assessment of the impact of scholarly communication items with metrics that derive from usage data. The 'bX' recommender service was developed by ExLibris in collaboration with Los Alamos. It is focused solely on the scholarly domain and recommendations are based on usage of articles based on Link Resolver data. In addition to global usage, recommendations can be generated based on usage at their own institution, their consortium, and peer institutions.

Elsewhere, the statewide Ohiolink platform brings together catalogue data from 89 colleges and universities, including over 47 million books. It has yielded the largest and most diverse set of academic usage data for books ever collected, containing both circulation and renewal statistics by institution. From a statistical perspective, the results are limited to the Ohio academic libraries. However, a recent report concludes that 'because of the number and diversity of the OhioLINK libraries, most of the findings, certainly the general trends, are expected to apply to most academic libraries.'

2.2 Benefits

For Higher Education, the business case for applications exploiting usage data is unproven, unless we take broadly focused and commercial services such as Amazon as exemplars. Key challenges around issues of motivation, scale and exclusivity need to be addressed; for example

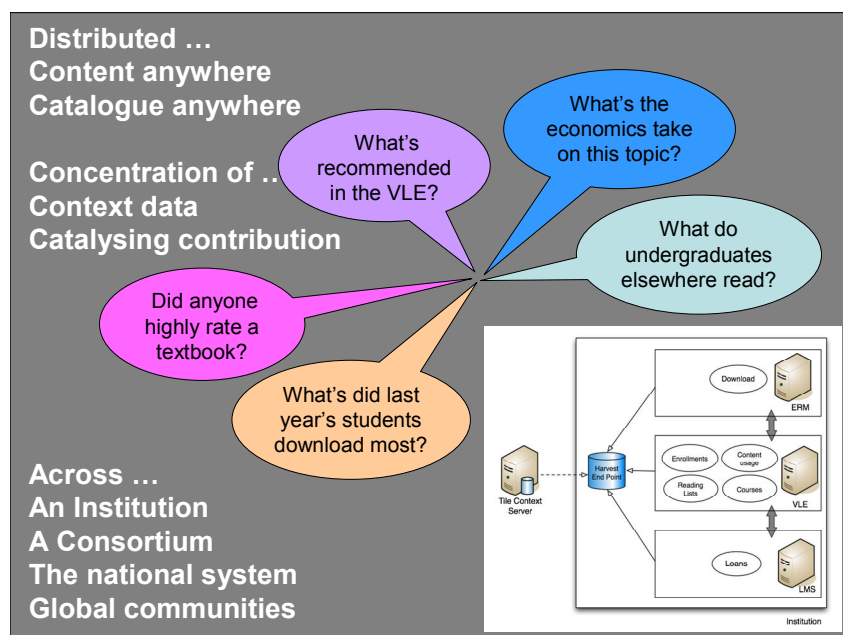
- How will such capabilities address the motivations of students, lecturers or researchers?
- Will they serve the objectives of Higher Education, not least the promotion of individual enquiry?
- Will intelligence drawn from activity data enhance the positioning library services and resources amongst the user strategies for accessing the best content at the right time?
- Should activity data enriched with course affiliations be seen as a differentiator for HE specific discovery services or is such data best deployed as a value add to web-scale services such as Google Scholar?
- Does activity and recommendation data work at the level of the local institution (as suggested by the Huddersfield student response) or is it better considered as an aggregated service, faceted according to such as institution and course?

It is however arguable that a Web 2.0 business mind would initially assess the data (bad data and low volumes will not engage users) and then assess user opinion on the back of a rapidly prototyped implementation (good data will almost certainly generate novel applications, empowering innovative workflows and learnflows).

The JISC MOSAIC project is tasked with investigating benefits for students, for support librarians and for the libraries themselves. As illustrated here and in the scenarios that follow, the benefits are currently perceived as follows

- **Students** – fast tracking content selection, getting ahead in areas of interest
 - What's been used (borrowed, downloaded) by previous students?
 - What's being used by current students?
 - What's used by students elsewhere or at other levels?
- **Postgraduate Researchers** – accessing the latest scholarly literature, identifying the long tail, engaging with the wider community
 - What's out there that my institution does not hold
 - What are the related papers
- **Teaching Staff** – monitoring cohorts, improving recommendations, enhancing courses

- What are my students accessing?
 - What's being used on my reading list?
 - What's on reading lists elsewhere?
- **Libraries** – advising students, ensuring supply, tracking short loan requirements, pruning stocks and subscriptions
 - How popular is this book now?
 - Who else is holding this title?
 - Do I need more copies?
 - Do I need this at all?



Benefits derived from concentrated intelligence

2.3 Services

2.3.1 New Opportunities

Usage data opens up the possibility of providing services that would otherwise be unavailable.

The JISC TILE project recommended concentrating the intelligence (patterns of user activity that exist in institutional and national HE systems) in new web-scale services of value to undergraduates, lecturers, researchers and institutions. Such services potentially contain a critical mass of activity data from 'Day 1' on account of the historic data held in many (though not all) local library circulation systems. This 'running start' would bring to life the opportunity to engage and curate user contribution (e.g. ratings, reviews).

The critical service success factors are:

- The DNA of user 'academic affiliation' – HE institutions have an advantage over commercial services on account of the knowledge available about the user's context (subject, level and place of study) which can run through all activity records. This data can be leveraged without any real threat to anonymity ('a named student read this'), though it might be perceived to undermine institutional advantage and integrity ('students on a specified university course used and even rated this').
- The benefits of scale – whilst some subject areas will deliver scale within a single institution, a consortium or national scale services will deliver unique benefits in terms of similarity (comparative activity elsewhere, other reading lists, wider feedback) and scarcity (identifying the long tail of subject interests).

2.3.2 Source Systems

The richness of such services is likely to increase in line with the range of systems that are able to contribute auto-generated usage data. The key examples are

- Circulation & Inter-lending modules of Library Management systems (LMS)
- Virtual Learning Environments (VLE)
- OpenURL resolver / ERM derived information
- Institutional and Open Education Resource Repositories

3. Scenarios

3.1 Scenario 1

Title	Access to content activity data – an undergraduate scenario
Basis	This scenario is based on experimental technologies (including large scale aggregation) which are currently being evaluated (e.g. JISC MOSAIC project, MESUR, OhioLink), drawing on data which exists in some Library Management Systems but is not yet being exploited.
Narrative	<p>Eleanor is starting her undergraduate studies in History. The VLE contains reading lists for her initial modules and she has also gathered recommendations from several lectures, as well as observing her classmates hogging the short loan collection. She quickly realizes that there is too much choice, restricted availability and too little time and that she currently lacks the background to navigate the options.</p> <p>The university's 'Recommender' service proves invaluable. It combines library, VLE and other Open Resources, providing indicators based on circulation, download and links to reading lists. It not only weighs the options but also opens up new possibilities, offering usage data from on similar courses and modules, in previous years, at other levels and in related subjects. Whilst not essential and somewhat sparse, the ratings and reviews of other students are interesting – but the real bonus is the lists provided by previous students for specific essay topics.</p>

3.2 Scenario 2

Title	Interaction with a community recommender service – a postgraduate scenario
Basis	This scenario is based on experimental technologies (including large scale aggregation) which are currently being evaluated (e.g. JISC MOSAIC project, MESUR, OhioLink), drawing on data which exists in some Library Management Systems but is not yet being exploited.
Narrative	<p>Francois is successfully progressing his Masters work on the literary style of Virginia Woolf. In so doing he has become increasingly aware of the need to provide additional commentary on related writers, such as Vita Sackville-West. The challenge is where to begin in the time available Using Amazon he can make the connections but finds it impossible to value the material, not least because any reviews and ratings lack validation – exactly who is E. Grossberger and what does he or she know?</p> <p>Thankfully the national recommender service has some critical mass in this subject area. An initial search across reading lists, especially those of cited academics, clarifies the importance of specific authors (Mustn't forget James Joyce! Perhaps Leonard Woolf was more than just the man around the house?) and related academics and journals. Then navigating from reading lists to usage data helps reject texts used only by undergraduates on generic courses. Down to a short list, the gold dust could be the reviews. There aren't many but one review comment on a Vita book from a German researcher reveals the key to the door, in the form of Herr Grossberger's ongoing PhD thesis – 'A comparative study of style in the extended Bloomsbury circle'. Time for an email!</p>

3.3 Scenario 3

Title	Usage informed decision support – a librarian scenario
Basis	This scenario is based on experimental technologies (including large scale aggregation) which are currently being evaluated (e.g. JISC MOSAIC project, MESUR, OhioLink), drawing on data which exists in some Library Management Systems but is not yet being exploited.
Narrative	<p>The university library needs to free space for working areas and has decided to focus initially on the monographs collection. A starting point has been identified in the reading lists of modules that have been discontinued for 2 years or more. Jim, the subject librarian for Engineering, is able to take advantage of M25 consortium usage data to inform his decisions.</p> <p>For each book on the agreed reading lists, he examines the volumes of local circulations and renewals on an annual basis over 5 years. He also investigates levels of stock and annual usage of the same books at other M25 universities, thinking in terms of inter-lending as well as differences in course emphasis. The exercise indicates very different patterns at Imperial and so Jim decides to investigate what else is on their comparable reading and short loans lists. The outcome is a reasoned recommendation to faculty to remove 2600 books from the shelves. Ironically, Jim also raised the issue that the three of the most popular books at Imperial were not in the local library, leading to a rapid reading list revision in one module and the purchase of 12 new books. Such is the nature of intelligent pruning!</p>

4. Tools

In the current library and VLE systems environment, it will be valuable for vendors and systems librarians to develop publicly available tools which can support work with the datasets described here; for example

- Extracting Data from systems – e.g. techniques for different LMS systems
- Merging data sets – e.g. joining circulation with registry data
- Anonymising records – e.g. introducing GUIDs and loan sequences to records, removing singleton records

Institutions collaborating in the JISC MOSAIC project will be in a position to share experience and code (e.g. Dundee, Huddersfield, Lincoln, Sussex, Wolverhampton) and Dave Pattern has established a wiki for this purpose within the project.

5. Standards

There is currently no standard for the definition or representation of usage data of the type described in this paper. It should be noted that only three of the four major vendors of LMS systems to UK HE libraries are currently understood to support the automated generation of usage data (e.g. typically within circulation modules).

The JISC MOSAIC project is building on the work of Dave Pattern in drafting guidelines for the definition and representation of LMS derived usage records and reading list records. These guidelines have been used to extract data from ExLibris, SirsiDynix and Talis systems. The guidelines can be downloaded from www.sero.co.uk/jisc-mosaic-documents.html.

The MOSAIC guidelines recognise the importance of the following standards, specifications and approaches in the further development of standards and shared formats for the exchange of library and other content usage data.

5.1 Metadata definition

- Dublin Core (DC – www.dublincore.org) - to assist in standardisation of attributes across

media types (e.g. Author becomes generic DC Creator)

- Functional Requirements for Bibliographic Records (FRBR – www.frbr.org) – to assist in determination of level of usage aggregation and UGC attachment (e.g. Work or Edition)
- Academic Institution Internal Structure Ontology (AIISO -<http://vocab.org/aiiso/>) - may provide a means of formalising course and module descriptions
- IMS Global Learning Consortium (IMS – www.imsglobal.org) – It is not clear that IMS specifications support specific aspects of this work; however a dialogue may be of value as possibilities emerge
- Attention Profile Mark-up Language (APML - www.apml.org) – will be of interest to developers building applications around usage data; APML allows sharing of personal Attention Profiles like OPML allows the exchange of reading lists between News Readers. The idea is to compress all forms of Attention Data into a portable file format containing a description of your ranked interests.

5.2 Data format & representation

- XML – it is highly desirable that a portable data format is adopted; whilst it should be expected that a community definition (DTD) will be developed as the field matures, the JISC MOSAIC project is working with an exploratory format building on the work of Dave Pattern at the University of Huddersfield
- RDF – Use of RDF to expose activity data as a service will be important part of developing the possibilities of activity data in terms of aggregations and mash ups.

6. Key Issues

Issues that could critically affect the success of your scenarios

IPR	IPR is not an issue in the case of activity metadata, assuming it only contains the essential parts of catalogue records, including reliable links to full records and thereby to any digital content. In the case of user generated content, it will be important for users to agree that their contributions are open to re-use upon subscribing to such a service.
Privacy	<p>Privacy of activity data has been identified a major concern under the terms of the Data Protection Act. It is therefore essential that automatically generated activity records cannot be traced back to individual users when they appear in services.</p> <p>DPA Provision 6 allows for processing when '<i>necessary in order to pursue the legitimate interests of the data controller or third parties (unless it could unjustifiably prejudice the interests of the individual).</i>'</p> <p>There is no issue for activity data that is aggregated so long as singleton cases are removed (and possibly even double- and thrice-only mentioned records).</p> <p>A more significant challenge arises from what MOSAIC calls Level 2 data, which contains GUIDs and activity sequence numbers to derive information such as 'users who borrowed this also borrowed / previously borrowed / next borrowed'. This data cannot by definition be aggregated so a problem arises in the case of very small courses where a set of individuals might be easily 'known'. However, having removed at least singleton records, it is highly unlikely that one could easily (or at all) extract prejudicial data.</p> <p>In the foreseeable future, fair processing notices (agreements students and staff sign) will emerge as a legally-inspired belt and</p>

	braces approach to cover this kind of library and resource use data as a belt and braces approach.
Competitiveness	Universities and individual academic staff may perceive that opening up usage data, reading lists and UGC such as ratings beyond institutional boundaries to be <ul style="list-style-type: none"> • Surrendering their competitive advantage – for example, by sharing specialized reading lists • Compromising their reputation – for example, if reading lists are not well used or if UGC is of poor quality or even negative
Scale & Aggregation	Scale above and beyond the institution is generally regarded as a key asset in reaping value from usage data. National scale would seem to represent a valuable systemic level for UK HE usage data. As illustrated in Scenarios 2 & 3, scale enables comparison and opens up the possibilities of 'the long tail'. It is currently suggested that scale is likely to be achieved by aggregation of usage data (auto-generated and user generated) in to a central system as opposed to a federated solution. As indicated under Aggregation (see 1.4), the potential volumes of usage records are very significant. Whilst aggregation might reduce volumes by an order of magnitude, great care should be taken not to discard what might later be regarded as key indicators (as a minimum institution, unit of study & year should be preserved).
Integration	It is argued that the link between user context (subject affiliation - whether course or thesis) and activity is the essential DNA that would enable the HE community to mount a uniquely valuable service. This is data that is not readily available to services outside the sector (such as Google Scholar). However, the creation of such datasets will typically be reliant on links between resource management systems (LMS, VLE, VRE, Repository, OER) and student registry and HR systems. This implies a degree of integration and synchronization that may not yet have been practicably achieved by many HEIs. Underlying this requirement is the deeper challenge of identifying a suitably granular context with which to associate a usage occurrence. This would ideally be a module as well as a course – though the latter would be much more easily achieved.

7. Recommendations

Recommendations to JISC that would significantly advance the approaches described here.

7.1 Recommendation – User Interest

Drawn from the TILE Architectural Proposals paper

Undertake research to ascertain whether the range of library users would be inclined to pay attention to recommendations arising from activity and other usage data (both local and national) and to use the services that might be constructed around such data. The JISC MOSAIC project is tasked to undertake initial focus group investigations in autumn 2009. Further research may be required.

7.2 Recommendation – Data Standards

Drawn from the TILE Architectural Proposals paper

Work with the community to establish a robust **format** for the extraction and exchange of the usage datasets, with the potential to cover resources other than LMS content (e.g. VLE, Repository, OER). The JISC MOSAIC project is tasked to develop initial guidelines (see above) that are 'good enough' to support early experiments. This should provide the basis for community developments which must involve systems vendors.

7.3 Recommendation – National Recommender Service

Evaluate the business case (including technical feasibility) for a national recommender service. A vision for such a service, which would address the three scenarios set out in this paper, is appended. The MOSAIC project is tasked to provide business case and technical information and forward options in its final report to JISC in November 2009.

Appendix - National recommender service

Proposal

This scenario is based on a national sector-wide approach to the aggregation of explicit and implicit recommendation data. Explicit data, also known as User Generated Content (UGC) will come in the form of ratings, reviews and shared lists. Implicit data will be aggregated from national services, institutional library and learning systems – drawing on records of such as circulation, content downloads and reading lists.

Benefits

National aggregation has the potential for critical mass in most subject areas. Inclusion of implicit recommendations derived from activity data and reading lists means data could be visible at Day 1 regardless of UGC levels. Handling at national level relieves the requirement for each LMS and VLE vendor to address this issue and for each HEI to implement a solution.

Implications

For Corporations – willingness to share activity data; use of common Digital Object Identifiers for all object types; data protection issues of sharing data that may be tracked back to an individual (in cases of 'singleton' records); risk that local reading lists and collections will become open to criticism

For Channels – someone needs to develop and provide the service, within the HE community or a third party; the service would need to de-duplicate linked assets in an appropriate manner

For Clients – success will be dependent on motivation and visible payback; factors may include the interface, integration with social networks and bookmarking services, openness of the data for mash ups and lack of restrictions on participants

Other Possibilities

If successful, the recommender service could provide a platform for all the scenarios examined in this paper, supporting such as

- recommendations based user habits, not just activity volumes (e.g. Users who did this also did that)
- further interactions with the 'catalogue' such as tagging
- qualification of recommenders according to status
- a variety of personal list functions

The service could also benefit libraries in guiding acquisition and pruning decisions.