CENTRE FOR DIGITAL
LIBRARY RESEARCH

cdlr.strath.ac.uk

Online Catalogue and Repository Interoperability Study (OCRIS)

Appendice 2 to OCRIS Final Report

Case Studies in 2 Higher Education Institution Libraries

Duncan Birrell, Gordon Dunsire and Kathleen Menzies

OCRIS
Online Catalogue and Repository
Interoperability Study

JISC

Centre for Digital Library Research, University of Strathclyde
November 2009

Project Acronym: OCRIS
Document: Appendice 2 to OCRIS Final Report
Version: 1
Contact: Kathleen Menzies or Gordon Dunsire
Date: November 2009

Case Studies in 2 Higher Education Institution Libraries

Introduction

The case studies carried out within two research intensive institutions - Cambridge University and the University of Glasgow - as part of Workpackage 3 of the OCRIS project, presented an opportunity to discuss in depth many of the issues addressed by the project and which confront the professionals who work, as a matter of course, with LMSs, OPACs and IRs as they attempt to meet the challenges facing academic libraries when providing an integrated e-infrastructure for research.

A key goal of the Case Studies was to discern how the systems under examination relate to the wider agendas and priorities of the institutions to which they belong, and what were their actual or perceived roles in terms of institutional mission, research strategy, administrative requirements, and the curation/preservation of university publications and other research output. Further consideration was also given as to how such issues impact on library resources, workflows and staff development.

In accordance with the stated terms of the OCRIS Project, the following was discussed with case study participants:

- Technical issues involved in the implementation and running of library systems both discretely and with regard to integration and interoperability.
- Metadata schemes and standards used
- Views of staff (and, by proxy, users) with regard to service and service development.
- Workflows and procedures currently governing development and use of IRs and OPACs and how these might be refined or improved.
- The socio-political and economic factors and imperatives informing the strategic aims and objectives of HEIs; how these affect collection policy, development policy, work practices and the attitudes of staff and other stakeholders.

Both HEIs are research-intensive and members of the Russell Group; their University Libraries are members of Research Libraries UK (RLUK) and both are involved in the HEFCE REF Pilot Exercise. The structural and operational organisation of the two Universities and their respective libraries is, however, significantly different.

It is unsurprising that within both Cambridge University Library (CUL) and the University of Glasgow Library (UGL), staff demonstrated a highly informed expertise and showed ample awareness of the issues under consideration by OCRIS: duplication of content (at both item and record level), the compatibility or otherwise of format and content standards and, crucially, the need for systems that share information as part of a broad, interoperable "ecosystem" across not only library services but institutions as a whole.

That the workflows and practices involved in addressing these issues are complicated and at times hindered by factors outwith the control of library staff – economic, political and cultural factors; research assessment obligations; the technical capabilities of the systems with which they work; the quality and availability of externally sourced bibliographic data – encourages debate and collaboration. Thus, while there are many pressures upon staff there is also much innovation and continuous, iterative assessment and development.

Project Acronym: OCRIS
Document: Appendice 2 to OCRIS Final Report
Version: 1
Contact: Kathleen Menzies or Gordon Dunsire
Date: November 2009

Cambridge University Library

Staff participating in Case Study:
Barbara Bültmann: DSpace@Cambridge Support and Liaison Officer (Day1)
Edmund Chamberlain: Systems Librarian (Day 2)
Patricia Killiard: Head of Electronic Services and Systems (Day 2)
Elin Stangeland: DSpace@Cambridge Repository Manager (Day1&2)
Hugh Taylor: Head, Collection Development and Description (Day 1&2)

Overview

Cambridge University

Cambridge, 800 years old this year, is arranged as a federated collegiate University, meaning that its structure is quite complex – under its Statutes and Ordinances each of the 31 Colleges are self-governing, with only minimal "top down" central administration throughout the institution (primarily of Schools, Faculties and Departments in conjunction with which the Colleges carry out teaching of their students). Centralised systems are accordingly limited and include identity management, institutional email and authentication services.

The University Offices provide support to the Schools, Faculties and Departments (administrative groupings of related subjects) of the University. Within these Offices sit the administrative departments most relevant to OCRIS: Finance, Human Resources, the Research Services Division and Student Administration and Records.

The Resource Allocation Model applies to all 6 Schools equally; research and teaching is organised by Faculty, with Faculty Boards responsible to the General Board. However, "the Faculties have different organisational sub-structures which partly reflect their history and partly their operational needs."[1]

Cambridge University Library

Cambridge University Library (CUL) is a legal deposit library and holds over 7 million items. It states in its current strategic plan that it:

"*Will continue to offer a 'hybrid library' service, combining paper and electronic resources as appropriate in each subject area. The growth of electronic resources will not be matched by an equivalent decline in the publication or use of traditional paper-based library resources, except possibly in some of the sciences. The Library will have to meet the challenge of providing an integrated infrastructure for research, by expanding the digital library and at the same time safeguarding the print collections that still underpin research in many subjects.*"

Further, it acknowledges that:

"*Developments in ICT will allow more sophisticated services to be offered, and the growth in computer literacy and widespread use of computers by all groups of user will lead to expectations for more library services to be delivered directly to the desktop.*"

The CUL comprises the main University Library building (in West Road) and four dependant libraries – the Squire Law Library, the Medical Library, the Central Science Library and the Betty and Gordon Moore Library. Each has its own Strategic Plan, aligned with that of the main library. A co-ordinated rather than integrated approach is taken to the provision of university wide library services:

[1] http://www.cam.ac.uk/deptdirectory/moreinfo.html [Last accessed 30th July, 2009].

"*During the period covered by* [the strategic] *plan, there will be increasing co-ordination between the libraries of the University, but a fully integrated 'University Library Service', with a Director of Library Services, will not be introduced; if it is, a fundamental revision of this plan will be required. In the absence of any such development, the University Library will continue to implement the Library Syndicate's policy of operating as a single co-ordinated service and collection on five sites."* [2]

In fact, there has recently been internal discussion of movement towards just such an integrated *'University Library Service'*, under the stewardship of a Director. This is partly in response to economic pressures and would, if taken forward, be done incrementally and in close consultation with faculties and departments.

Each of the Colleges has its own independent libraries (often one for Students and one for Fellows) - so too, of course, do many faculties and departments. These libraries are under no obligation to contribute records of publications to the main Library and do so only on a voluntary basis. There is, therefore, no centralised system in place for the complete aggregation of catalogue records.

The federal arrangement of Cambridge University therefore impacts significantly upon the operation of the library culturally, administratively and economically, in ways which are somewhat unique and which are of direct consequence to the "libraries@cambridge" project (which includes the activities of what was formerly known as the Union Catalogue project) as well as the operation of the central IR (DSpace@Cambridge).

At the same time, the centralised aspects exert their own forces. For example, the Medical Library's *Strategic Plan 2006-11* states that "The [new] Resource Allocation Model will make it increasingly difficult to secure adequate UEF [University Education Fund] funding for central services such as the Medical Library" – further, any "Reduction in UEF funding without an increase in external funding will" negatively impact on their ability "to provide core services, including adequate provision of books, journals and electronic resources." This, as they seek to "continually liaise with other libraries to ensure optimum co-ordination of resources".[3]

The Library Management System and the OPAC

System

The Library Management System in use at Cambridge University Library is Voyager, with WebVoyage acting as the OPAC. This was installed in 2002 and was supplied by Endeavor Information Systems Inc. (acquired in 2006 by Ex Libris Ltd.).

"Newton" (named for Sir Isaac Newton, a Trinity College Graduate) is the major online catalogue for the libraries of the University, including the CUL and its dependants as well as most faculty, departmental and college libraries. To quote:

"Due to the size and complexity of the libraries in Cambridge, the catalogue is currently divided into several smaller catalogues, each holding information on a specific set of libraries or library collections. A test Universal Catalogue interface provides a means of searching all of the smaller catalogues."[4]

This is explored in more detail below.

[2] Fox, P. (2006). *Cambridge University Library. Strategic Plan 2006/7 – 2010/11*. Available online at: http://www.lib.cam.ac.uk/strategic_plan.html [Accessed 30th July, 2009].
[3] Morgan, P. (2006). *Cambridge University Library. Medical Library, Strategic Plan, 2006-2011*. Available online at: http://www.lib.cam.ac.uk/strategic_plan.html, pp.1-2.
[4] University of Cambridge. (2007). *About the Newton Catalogue*. Online at: http://ul-newton.lib.cam.ac.uk/help/about.htm. [Accessed 21st August, 2009].

Configuration of the LMS and OPAC

The underlying arrangement of Newton is rather complex – it consists of separate catalogues held within 9 separate Voyager databases, configured as separate modules. These are ORACLE Relational Databases, the core technology used by Voyager being ORACLE's RDMS (Relational Database Management System). 3 of these databases are for the catalogues of Departmental and Faculty Libraries, 2 are for College Libraries, 1 is for CUL Manuscripts and Theses, 1 is for the holdings of the CUL and 1 is for affiliated member institutions or 'other' libraries. The last is a non-public database holding information on Legal Deposit receipts, for which there is no data in the system at the point of arrival.

A separate database for the now defunct Cambridge Union List of Serials (CULoS) is no longer maintained and cataloguing is incorporated within whichever database relates to the relevant holding library.

Each contributing library creates and maintains its own records within the system but minimum standards for cataloguing have been agreed upon for all of the Union Catalogue libraries, (i.e. members of the former Union Catalogue project), local practice being taken into account within these. The sharing of bibliographic records is however possible *within* a single database.

Barriers to the creation of a University-wide Union Catalogue

A Union Catalogue which exisited previously at Cambridge included the records of Faculty, Department and College libraries as well as the *dependants* of the CUL but not those of the CUL itself. Clearly it was desirable that a Cambridge-wide Union Catalogue be established.

However, when it came to doing so, certain barriers to interoperability became clear, hence the modular structure described above. Although financial resourcing was to some extent as issue, the problem primarily lay with the Endeavor system itself – it did not have the system capacity required to reflect the complex acquisition, borrowing and circulation control rules in operation across Cambridge's colleges, faculties and departments within one database. The representation of the circulation rules was achieved, with considerable effort from the supplier, by the development of a unique 'clustering system' for the University.

Staff had assumed Union catalogue products would help them provide a single point of access to holdings information, bringing together (from the user's perspective) their fragmented databases and catalogues. However, this did not work out as intended.

Such lack of scalability within the federated system, demonstrates the difficulties inherent in attempting to develop a truly interoperable, cross-searchable LMS - working with other institutional systems (whether external or internal) at machine to machine level - were the current Voyager installation to remain in place.

The Universal Catalogue allows only the records of the smaller contributing libraries to be searched as a group.

Libraries Contributing

In total, there are 104 libraries within Cambridge (108 if we include the 4 CUL dependents). Approximately 89 libraries have collections which are described within Newton; Newton’s records however, can often relate to libraries which no longer exist, which have changed name or have been absorbed into/merged with other University libraries; the number of contributing libraries must therefore remain provisional.

15 of these 89 contribute records only for Serials holdings.

Although open to all University libraries, matters of historical legacy, cultural or economic factors affect whether or not a library chooses to contribute to the CUL catalogues. Anecdotal evidence suggests that, in some cases, Colleges seem to prefer to limit knowledge of which collections they possess, perhaps viewing themselves as 'closed communities'; whether by accident or design, there have been instances of records for Rare Books being suppressed.

In an example of the anomalies perhaps inherent in any federal system, the 'Working Library' catalogue of the Gonville and Caius Library is searchable via Newton; however, the catalogue of its 'Old Library' (the College's large collection of manuscripts, early printed books, and other rare material) is not. The Department of Engineering maintains its own OPACs – one for its Serials holdings, another for its book stock; and these also are not included in Newton.

A few of the 89 libraries are not in fact part of the University of Cambridge but were added under previous, more generous inclusion policies around 10 years ago.

Some libraries are not included because they do not wish to be – the federated system allows them this privilege – while others plan to join at a later date. The Rosemary Murray Library at New Hall (Murray Edwards College) will soon be adding records.

Institutions listed by the Cambridge University directory, and with whom Cambridge staff may be associated, might not be in scope for Newton. For example, the Babraham Institute has the status of a postgraduate department within the University of Cambridge and trains PhD students who are registered with the University's Faculty of Biology. Possessing an extensive collection on Developmental Biology, Neuroscience, Immunology and Molecular Signalling, and with extensive publications lists for staff, it is out of scope for the OPAC and the IR as it is not a part of Cambridge University proper, its stock being controlled by the BBRSC and various Research Councils.

The MRC Collaborative Centre for Human Nutrition Research Library was previously more closely aligned with the Medical Library than it is currently and has ceased contributing to the central catalogue. They maintain their own publications database.

Other Catalogues

The pre-1978 Supplementary Catalogue contains records for items that were not considered to be of academic importance at the time of receipt (for example, novels). It is in two sequences: a sheaf catalogue on slips for works published from 1800 to 1905 and a card catalogue for publications from 1906 to 1977. These are searchable within Newton. Re-cataloguing of the sheaf catalogue is currently being undertaken as part of the "Tower Project" (http://www.lib.cam.ac.uk/deptserv/towerproject/).

Users are able to search the National Library of Scotland's catalogue, the Library of Congress' catalogue and the University of Oxford's OLIS OPAC alongside the Newton Catalogues. This is achieved via a virtual union catalogue which uses the z3950 protocol.

Cambridge's JANUS project makes catalogues for archives and manuscripts available via a set of OPACs which are quite distinct from Newton. To quote:

"*JANUS is a self-funded project, established in October 2002 to provide a single point of networked access to catalogues of archives and manuscript collections held throughout Cambridge. The number and range of participating* repositories - *both University and non-University* - *continues to widen, promising in due course the near comprehensive coverage of archives in the city and surrounding area.*"[5]

[5] http://janus.lib.cam.ac.uk/db/node.xsp?id=Webpages%2FPublic%2Fabout [Accessed 1st August 2009].

Its catalogues conform to ISAD (G) with NCA rules for the construction of personal, family and place names used for authority control and access points as well as the UNESCO thesaurus and the Getty geographical thesaurus.

Automated validation checks are used to highlight possible duplication of index terms.

32 institutions within or associated with the University contribute to JANUS, of which the following 6 fall within the remit of the CUL Syndicate: British and Foreign Bible Society's Library, Cambridge University Archives, Manuscripts, Royal Commonwealth Society Collections, Royal Greenwich Observatory Archives and the Squire Law Library.

Items in Scope

Data content within Newton is broken into 14 item types. These are:

1. Book
2. Serial
3. Electronic journal
4. Electronic resource
5. Disc (CD/DVD)
6. Music Score
7. Map
8. Non-musical Recording
9. Musical Recording
10. Archive/Manuscript
11. Kit
12. Mixed Material/Collection
13. Mixed Material
14. Visual Material

These item types derive from MARC21 coded data fields and fields held in Voyager's ORACLE databases and which are themselves derived from MARC21 encoding. However, anything which can be described using AACR2 and MARC21 is in scope for the catalogue.

Standards

MARC21 is used as the format standard for bibliographic records. The content standard supported is AACR2 with Library of Congress and in-house Subject and Classification schemes used for purposes of Authority control. An unusual, '3-figure classification' scheme developed by the library in the early 20th century[6] is also used for some holdings.

There is no single class scheme used across the University and even within a single library various schemes may sometimes be used. One example is the Cambridge University Engineering Department Library which has developed its own ‘CUED Library Classification' system, based on LCC.

Deduplication, Authority Control and Metadata Issues

The Electronic Services and Systems teams are heavily involved in cleaning up over 4 million records, reducing duplication by getting rid of the lowest quality records while enhancing the higher quality ones. Much of this work relates to the cleaning-up of bibliographic records that results from the weekly updating of authority files (i.e. the "downstream work" that results when Heading A is changed to Heading B). CUL would, ideally, like its records to be usable as authoritative metadata sources, as 'clean' as the OCLC versions, and this is what they work to achieve.

[6] See *Classification Scheme.* At: http://www.lib.cam.ac.uk/class/index.html [Accessed 04 August 2009].

There are fundamental differences with the bibliographic records contained in each of the catalogue databases due to the varying quality of records and the local practices in operation across the various libraries. Managerial and administrative changes and the attendant changes in workflow can make the enforcement of consistent standards complex, as indeed can questions of granularity (some catalogues and some types of user may prefer a 'richer' set of metadata than others).

It was formerly the case that all libraries owned their own bibliographic as well as holdings records. This inevitably led to duplication and complicated the de-duplication process, with the best solution being that another library simply add their holdings information to any duplicate record.

With the current Voyager system records can – in the case of the 3 'Faculty' and 'Departmental' databases - be fully de-duplicated and merged as the relevant libraries contributing to Newton now own only their *holdings records*. This is because the need to de-duplicate has become more of a priority, not least in terms of efficiency saving. However, the bibilographic records held in the 'Colleges' and 'Others' databases are still "owned" by the contributing libraries and they do not subscribe to this more efficient model.

Partly this is because some of technical difficulties: some of these libraries simply upload files of records to Newton and do not use Voyager to do their cataloguing work; therefore it would be difficult to "roll out" this model to them.

The Universal Catalogue contains the bibliographic records from all of the individual public Voyager databases. In addition, 'stub' holdings records are created which include the information necessary to link back to the source databases in order to provide the user with current holdings and item data.

Software allows the records of the separate databases to be de-duplicated based on 'match point' detection to present a single record for a single item to users – however as some older records in the database are brief and do not contain the information used as a 'match point' duplicate bibliographic records on occasion remain visible.

To further complicate the workflow, the de-duplication options offered by Voyager are limited; useful secondary checks such as ISSN, date or country of publication (taken from the 008 field) aren't possible because of the size of the database, with a limit on the number of records against which you can 'test' each element separately. Sometimes *both* duplicates are mistakenly thrown out or the same records may appear in the pool of potential duplicates more than once. De-duplication is done in order of preference and the system re-build once this has been carried out is very slow.

Questions therefore remain – at what point can you be sure that your data is reliable enough to inform/justify de-duplication, or to share? This is especially important if you are wishing to make it available openly (e.g. as part of the 'Semantic Web').

It is difficult even to discern the authoritative source for the official names of schools and departments – for example, which departments are allowed to make use of the title 'Cambridge University'? Is it correct that the Department of Engineering (also sometimes referred to as 'Cambridge University Department of Engineering') is, at times, referred to as 'Cambridge University Engineering Department', or is it the case (as was tradition) that only the CUL and the Cambridge University Press (CUP) can officially use this title? Even experienced staff remain unclear on such distinctions. It is therefore difficult to ensure conformity or create reliable authority controlled lists and rules for use across systems.

Manuscripts and Theses

It was decided approximately 15 years ago that Manuscripts and theses should be made more visible to users. Records held in the database, however, may have been created many years ago. The card catalogue was updated to make it partly conform to MARC standards. Yet, in a number of cases, holdings were only described at the 'high' levels of Collection and Series, rather than at Item level, clearly making them difficult to include in item level search results. Limited to a restricted set of indices, access to such materials would therefore remain limited in comparison with that of other records.

While a few hundred music manuscripts were catalogued to Item level, the Manuscripts Department changed from MARC to adopt EAD as their bibliographic standard.

The University Library Manuscripts and Theses catalogue is therefore effectively now 'live' only for Theses records and a small number of new music manuscript records. As Theses are effectively 'book like items' it was suggested that it may be more appropriate to move them into the main database where book records are held; these are perhaps conceptual, rather than technical questions but they are relevant ones.

Development of OPAC Tools and Interfaces

Electronic Services and Systems staff have been attempting to improve users' search experience in terms of look, feel and design, as well as introducing 'Web 2' features and tools. This is in direct response to comments from Undergraduate users via Department and Faculty libraries, or more formally through focus groups. There have been several recent developments:

CUL staff installed a pilot Voyager 7 interface for the UL and Dependents catalogue at the end of 2008 (http://ul-newton.lib.cam.ac.uk:7708/vwebv/searchBasic?sk=en_US) which is now 'live'. The layout and user interface are different, with scroll menus replaced by drop-down ones and the option of viewing your search history. Otherwise, the indices being searched and the way in which you can search them is more or less the same. Differences are that 'any of these', 'all of these' or 'as a phrase' can be used as limiters with search terms. The status of a journal 'currently published', 'ceased publication' or 'unknown' can also now be selected. The number of records displayed on a page can also be limited. The main change is the way in which results lists are presented – it is more obvious to users what they can 'do' with the items matching their search through the addition of a simpler, less crowded array of 'print', 'export', 'email' and 'bookmark' buttons, the provision of a permanent URL so that you can 'link to this item' and more visible filtering options. The new interface is also easier to configure.

Systems staff are interested in the idea of providing different, customised interfaces for different types of user - for example, library staff, whose non-public Resource Files at present use an interface nearly identical to the public one.

Another departure from tradition is the 'Library Toolbox' (http://www.lib.cam.ac.uk/toolbox/) which offers OpenSearch "plug-ins", an iGoogle search gadget and Reference Management tools (the Zotero citation collection and management suite which works with the Firefox web browser and can export Newton results lists to RefWorks and EndNote). RSS feeds are set up to alert users to new electronic resources and science resources being subscribed to and any electronic resources trials being run. The web browser toolbar allows multiple field searching of any Newton catalogue without having to go to the web interface. This is based on the Virginia Tech University Libraries' LibX Open Source framework.

Not all of the tools listed above are managed by the library.

Links to other Institutional Systems

Duplicate patron records do not need to be created by Library and HR staff as the barcodes on student and staff cards, which are fed into the LMS and used for authentication and for purposes of patron account management, are derived from (and also utilised by) the HR System. Local copies of patron records are therefore stored by the system across various institutional databases. Users registered with more than one library are thereby presented with just a single patron record aggregated from details held across the systems.

The University's 'Identity Management Group' is investigating how records from the Management Office relating to staff and students might be better shared across systems, with possibilities including a portable ID that could track an individual throughout their time at Cambridge. This ID would be persistent, even were you to graduate from Cambridge and return at some point as a conference delegate or visiting professor.

There are multiple 'data owners' within Cambridge and establishing means of data sharing would necessarily be a challenge. For example, devising a comprehensive name authority list or ID from the Cambridge HR system would be complicated by the high number of staff and patrons who are not in fact affiliated with Cambridge. Many individuals who do not qualify for an institutional email address are still allowed access to its libraries. There is no common login scheme in place across Cambridge, and often initials rather than full names are used.

The issue is complicated further in the case of Continuing Education students who might be attending only a day school or summer school or learning within one of the University's many regional study centres. Similarly, at record or item level there may not be any direct link between an individual and the institution. They may have contributed to a paper or symposium but no name authority would exist for them anywhere at Cambridge.

Virtual Authority Files and similar initiatives could in future help with this, but at local level this remains problematic.

Workflow inefficiencies will inevitably be encountered when changing from a system of local control to one which relies on the flow of data to-and-from a centralised, "external" system. Such a solution, implying, as it would, the removal of local control and the ability to directly and effectively manipulate information at the federal level, may bring its own additional problems. For instance, staff could lose certainty over how recent and up-to-date the records being pushed or pulled into their systems were.

Additionally, the LMS at Cambridge has its own customised system for handling financial information and the management and interoperability of this service would also have to be considered.

Attitudes, Problems and Future Issues

Electronic Services and Systems staff, with a wealth of experience, are more than aware of the problems and issues concerning duplication of scope and content, the use and imposition of standards and the need for systems that interact and interoperate as part of a broad, interoperable 'ecosystem'. Their decision to install a Resource Discovery Platform (see page 15 for more information) is motivated not only by the attitudes of users but by the need to include and expose Repository content and other types of learning, research and teaching content within the OPAC (or at least, the library's "shop front").

However, the workflows and practices involved in this are complicated and at times hindered by factors outwith their control – economic, political and cultural factors within Cambridge, the technical capabilities of the systems with which they work and the quality and availability of externally sourced records and data among them. There are many pressures upon staff.

Some members of ESS feel that, in general terms, there is not a sufficient framework in place allowing practitioners' concerns to influence developments, or be translated into change, within the University. Many library staff working at Departmental level, for example, are not empowered enough within the hierarchical and federal system to influence decisions, their libraries remaining servants of the faculties - which of course have their own priorities. Power is confined to too high a level within the institution, remote from the concerns of those directly experiencing the situation 'on the ground'.

Sourcing of bibliographic records

At a practical, day-to-day level, issues surrounding the quality and availability of bibliographic records, digital content, and the regulations governing these, throw up some pressing questions for the Electronic Services and Systems team, who must always keep an eye on the future. For example:

Regulations are currently going through Parliament which will determine the formats required for the electronic legal deposit items. What issues might arise for CUL Systems when they begin receiving the metadata for these? It is likely that the CUL will not receive all of these records (excepting MARC records for eBooks) and they may consider sourcing them from OCLC or Nielsen BookData (for example) to populate a database which they could then edit for use within their own system. The British Library's Resource Discovery strategy will also be influential.

But nothing can be presumed about electronic legal deposit records at present beyond the fact that they are likely to constitute a considerable problem for which effective solutions will need to be in place; cost will need to be balanced against the timely need to add the new records.

Acquiring bibliographic records and associated metadata from external sources would save both time and resources although something has to be done to ensure that the quality of these records is guaranteed. This is in keeping with the recommendations of the recent RIN (2009) report *Creating catalogues: bibliographic records in a networked world*.[7]

Exposing Data and Re-imagining the OPAC

Some Systems staff feel that the OPAC should be re-envisaged as a Repository in its own right; it might include a mechanism for publishing and exposing data via XML/Google, separating format from content and making it available for use elsewhere. CUL services sitting 'between' the 3rd party services being subscribed to would allow for the personalisation of services to reach a particular audience - something exemplified by the 'Science@Cambridge'[8] portal which makes heavy use of RSS feeds.

At present, curating web material in such a way that it could be used in an Open Environment is easier with the IR than with the LMS where the data is 'locked in' and where closed standards are used. Open Source library software has great potential here and might be integrated with the OPAC – for example the Xtensible Catalog project[9] has already released some OAI and metadata manipulation toolkits.

Externally sourced software and systems may not always provide the most appropriate or efficient means of linking data for users. OpenURL links in subscription citation databases such as Web of Science and CSA are added to every result, regardless of the CUL's actual holdings (i.e. whether the library in fact possesses a copy of the item being cited). Google Scholar, on the other hand, has harvested the CUL's holdings and only displays links where appropriate. Staff feel that the latter is a much better option and wonder why more vendors have not adopted this approach.

The forging of stronger links with related stakeholders (for example, the Centre for Applied Research in Educational Technologies (CARET) or DSpace@Cambridge staff) and shared strategies and workflows could also allow the vision of a multi-purpose, Repository-like OPAC to become a reality.

[7] Research Information Network. (2009). *Creating catalogues: bibliographic records in a networked world. A Research Information Network report. June 2009.* Research Information Network: London.
[8] See: http://www.lib.cam.ac.uk/scienceportal/ [Last accessed 04 August 2009].
[9] The Extensible Catalog (XC) project is run by the University of Rochester. See: http://www.extensiblecatalog.org/

These are only ideas and represent an additional layer above and beyond core library services - as with everything taking them forward at either Departmental or Institutional level would be dependent on resourcing. They do however demonstrate the forward-thinking nature of the Cambridge Systems team and their interest in open data and data sharing and 'stakeholder' engagement.

There are no visible tensions between staff working with the IR and those working with the LMS (except that DSpace staff would like to have more bibliographic services staff made available to them). The ESS team feel that digital materials and the systems which make them available to students help position the CUL as more central than it once was to the student experience. Previously it would have been that needs would be met largely by the College, Departmental or Faculty libraries (depending on the course being studied). Cambridge students and staff are perhaps now more aware of the role played by the CUL because of the range of Electronic Services and Systems being provided by them.

Institutional Repositories

Systems

The main Institutional Repository at Cambridge is DSpace@Cambridge, which seeks to gather content from across the University. It was established in 2003 and is based on DSpace version 1.45.1.

There are three other IRs at Cambridge administered at Departmental level. These are the Computer Laboratory Technical Reports Repository - which uses a custom-built database and supports Simple DC – and the Cambridge University Engineering Department Publications Database (CUED), an EPrints installation holding bibliographic data only. The Department of Earth Sciences has a similar EPrints set-up.

The Chemistry Department are establishing a repository under the JISC-funded 'Chemical Laboratory Repository In/Organic Notebooks' (CLARION) project, which will determine how best to store, process, preserve, enhance and make available to others, crystallographic, spectroscopy and chemical syntheses data. This builds on the work of the 'Submission, Preservation and Exposure of Chemistry Teaching and Research Data' (SPECTRA) and 'Submission, Preservation and Exposure of Chemistry Teaching and Research Data from Theses' (SPECTRA-t) projects, which were a joint venture between the libraries and Chemistry Departments at Cambridge and Imperial College London. The Chemistry Department does deposit some of its data with DSpace@Cambridge, however it clearly has special requirements (not least because the file formats and the IPR issues relating to the ownership and reuse of scientific data are complex[10]).

These additional subject based/departmental IRs were **not** directly discussed as part of the OCRIS Case Study as they are not associated with CUL - although they may be referred to.

Configuration

DSpace@Cambridge runs on several Linux machines. Its database "runs on PostgreSQL and the web application on Tomcat. Several Solaris-based servers provide the assetstore, with a capacity of about 100TB. Backups are stored on disk, on off-site servers.

To keep data secure, an off-site 'mirror' is in place, keeping a history of file changes. Servers are located at two different locations within Cambridge with the backups being stored off-site on large Sun servers with a ZFS filestore."[11]

Standards

[10] Morgan P., and Tonge, A. (2007). *Project SPECTRA. JISC Final Report*. Available online at: http://www.lib.cam.ac.uk/spectra/FinalReport.html [Last accessed 03 August 2009].

[11] See: http://www.lib.cam.ac.uk/repository/Technical_details/ [Last accessed 03 August 2009].

DSpace@Cambridge supports Simple DC as a format standard. From September 2009 onwards, the Electronic Theses Online Service (EThoS) metadata content standard, UKETD DC, will be supported. The Scholarly Works Application Profile (SWAP) may be supported at some point within DSpace (possibly in the Version 2 release) but this depends on the activities of the external development community.

DSpace was created to support any file format, for a range of media types.

Staff would like to start supporting data formats other than Simple Dublin Core, as it does not offer sufficient granularity. It would probably not, however, be MARC or any of its variants that would be chosen. Staff wish, in the long-term, to take a more flexible approach and to possibly hold metadata in multiple standards for items in the repository (for example, DC for basic dissemination activities, PREservation Metadata: Implementation Strategies **(**PREMIS) for preservation, and possibly content specific metadata such as Encoded Archival Description (EAD) for manuscripts and archival material or discipline-specific standards for research data etc.

METS may be used to wrap data expressed in XML in order to improve interoperability.

Contributors

The Repository is open to all members (staff and students) of Cambridge University. Others who wish to deposit, such as those working with institutions which are not technically part of Cambridge but associated with it in some way – may contact staff seeking permission to make deposits. However, the inclusion of data from these affiliated associations is limited. Theses yield from the MRC Collaborative Centre for Human Nutrition Research Unit, for example, are eligible but not other data or research output; similarly with the outputs of the Needham Research Institute.

Staff from within Cambridge Colleges are allowed to deposit scholarly works free of charge but will be charged the curation and preservation costs of other data types. Anyone depositing a large collection of items will also be charged, based on disk space occupied and management costs.

Position within Institution

Although DSpace@Cambridge is an established service, it is funded on a temporary basis. The current funding runs out in 2012 and a strong business case will be promoted in order to justify its continued existence, making it clear that the value is worth more than the cost. This partly fed into their decision to make Theses a major collection within the IR (although these have always been in scope) as not only are Theses valuable and easy to source, but authors are generally very keen to make these visible to the wider community.

Items in Scope

DSpace@Cambridge does not seek to restrict the types of content (or formats) in scope choosing instead to engage with and respond to the needs of researchers from all disciplines and planning developments accordingly. Nevertheless, there are 21 types of item (dc.type values) listed as being in scope:

To use the terms employed by the OCRIS Questionnaire, the 15 item types in scope are:

- Book
- Book item
- Book review
- Conference item
- Conference paper
- Journal article
- Journal item

- Learning object
- News item
- Other
- Report
- Research dataset
- Scholarly text
- Thesis or dissertation
- Working or discussion paper

Staff have created a dc.type.version field with the following defined values:

- draft
- submitted version
- accepted version
- published version
- updated version

DSpace@Cambridge currently holds 202,826 items, of which 170,000 are chemistry datasets describing molecular structures bulk uploaded from the Chemistry Department's Unilever Centre for Molecular Informatics.

University staff are not mandated to deposit publications with DSpace@Cambridge - deposit relies on the decisions and needs of an individual rather than being overseen or mediated at Departmental or Faculty level.

De-duplication, Authority Control and Metadata Issues

DSpace@Cambridge employs absolutely no authority control. There are no browsable subject headings and no name authorities are used. This can lead to confusion as keywords supplied largely by those depositing are implemented as supposed Subject index terms for browsing - complete with spelling errors, author names, mathematical and chemical formulae and unexplained acronyms presented as index terms.

It is felt that those authority lists supplied by default with Repository software (Norwegian and Swedish Science Indexes) are not appropriate to or sufficiently granular for use by the specialised, UK-based community of end users. It would be feasible to link to databases such as PubMed, whose MeSH headings could be utilised, although the vocabularies and ontologies required for many other subjects would still remain a problem.

As an interesting aside (based on the Repository Manager's previous experience at the University of Bergen), we might consider Norway, where much more is done centrally in terms of library records via BIBSYS (the supplier of library and information systems for all the Norwegian university Libraries, the National Library of Norway, college libraries, and a number of research libraries and institutions). The Norwegian Science Index was used within the University of Bergen's Open Research Archive (BORA - https://bora.uib.no/) and was implemented as a browsable menu. However, comparable lists are not yet freely available in the UK.

Many end users access DSpace@Cambridge items via the Google search engine, where keyword searching on the full-text provides entry points. In light of this, it could be argued that the issue of Authority Control (although clearly an ideal) is not as pressing a concern as it once was – but many cataloguers still contest this point.

At the same time, Repository staff do not necessarily feel that duplication of items or item descriptions between the OPAC and the IR is problematic – partly because of the different user communities accessing them, and partly because actual duplication of items is fairly small.

The "splash page" or "jump-off page" which users may reach in advance of an item they wish to locate is used to provide information on versioning or other important changes to a document (for example, illustrations included in the print version of a document may not be available in the electronic copy for reasons relating to IPR or it may be used for descriptions of non-text data). The issue of metadata leading to yet more metadata, associated with the "splash page", is not considered to be a problem in this instance.

Also helpful in terms of metadata editing will be developments already underway in the DSpace community to allow for the batch editing of records. The University of Auckland have built and are prototyping a system at present that works by:

"*...Exporting metadata (for an item, a collection, a community, or the whole repository) to a CSV file. That file can then be edited in a spreadsheet application. Changes can be made to metadata, new values added, and other values removed. New metadata-only records can also be added, and existing items can be moved between collections. You then import the changed CSV file. Changes are highlighted, and if you confirm the changes are OK, the changes will be processed.*"[12]

IPR, Open and Closed Access and Harvesting Issues

DSpace allows the assemblage of item Sets (Communities and Collections) via the creation of "base URLs" which can be implemented so that exposure to OAI-PMH harvesters can be restricted (or disallowed altogether) via the modification at local level of a standard DSpace "handle" or identifier. This means that not all items in the Open Access repository are in fact openly accessible and is required for a variety of reasons, notably where content is being stored to fulfil a preservation rather than an access function (and which in some cases are not intended for publication).

For example, preservation copies of the Scott Polar Research Institution's "Freeze Frame" collection[13] are stored within DSpace but they prefer to make them accessible via their own website. Low resolution versions of these will be released and available via DSpace@Cambridge later this year.

There are also issues with non-institutional outputs for which certain signatures are required before Cambridge can make them Open Access. Items added to the repository may be subject to 'embargos' while copyright issues are addressed or may be added to a 'dark' archive (only accessible with the right authentication privileges).

As the SPECTRA-t Final Report highlighted:

"*...discussions with chemistry research leaders had indicated the need for a restricted embargo procedure as a necessary requirement for the deposition of unpublished or commercially-sensitive material. Therefore, when data is deposited using SPECTRA deposition tools, the user is actively required to indicate both the length of the embargo period (0-3 years) and the default status (review or release) to be adopted at the end of the requested period. This information is held in metadata and immediately controls public access to the data held in the repository.*"

Widgets for publishing, editing or searching DSpace@Cambridge are reliant on open data standards and the need to be careful about who is able to access and re-use bibliographic data and content in situations such as the above would need to be thoroughly legislated for before such applications could be installed.

With an eye on the future, staff must consider the possible utility of retaining full, un-edited copies of documents from which content have been removed for IPR reasons, for when copyright expires thus removing the need for redaction.

[12] See: http://wiki.dspace.org/index.php/Batch_Metadata_Editing_Prototype [Last accessed 05 August 2009].
[13] See: http://www.freezeframe.ac.uk/explore/search [Last accessed 04 August 2009].

The base URL and Set system of DSpace means that in theory publications relevant to or submission being made for the RAE, REF and similar exercises could be established as Collections (with whatever level of access was deemed appropriate) as is the case with the existing 'Communities' and 'Collections' within the Repository. These could be instantiated as XML files or sets allowing interoperability.

Development of Repository, Tools and Interfaces

SFX link resolvers are used within DSpace@Cambridge, with an SFX link to "eresources@cambridge" being pulled up at the bottom of the page for searching Electronic Journals and Databases. There are problems with the utility of this feature; often it is rendered irrelevant because the resolver doesn't 'know' whether the Library's subscription package holds any related items or information.

There are options for harvesting data from DSpace into the 'eresources' system to increase the accuracy of these link resolver services, but this is not currently viewed as a priority. In the majority of cases there are no other versions of the DSpace item available and if they did exist, the SFX resolver and its Knowledge Base would not be aware of them. One exception would be for theses containing 3rd party copyright material. If permissions cannot be granted to release this content electronically, the SFX button would be a useful tool to locate the complete print copy.

The software may be changed at some point (although it apparently would not be EPrints that would be used instead of DSpace). It is suggested that waiting for developments within the DSpace and wider development and research communities can make it difficult to plan CUL service enhancements with any degree of accuracy. The promise, for example, that SWAP will be implemented within DSpace may not bear fruit, or may be too far into the future to allow for the construction of strategies and workflows.

Theses Workflow

A workflow for the cataloguing and uploading of theses will be integrated within the current DSpace workflow and will involve a member of staff from the Manuscripts Department who currently works with the print theses collection.

Part of the workflow involves a dialogue with the author of the thesis, designed to clarify information about rights status, format requirements and any other issues. IPR issues are mainly the concern of the Board of Graduate Studies (BOGS) and a member of BOGS staff might also be involved in such a discussion if resources permit. It is envisaged however that as much responsibility as possible will be pushed to the student in terms of submission preparation and upload.

The student will submit a print copy of their thesis, approved and checked by the Exam Board and BOGS, to the Library. Simultaneously they will be given access to DSpace to deposit a digital copy of their thesis. The theses and dissertations librarian will receive an email alert informing them that a thesis has been deposited and will log into DSpace to check (and correct) metadata. If there are no exceptions the thesis will be included in the public collections of the repository. The next stage is export of the thesis' metadata the CUL's Voyager system, where records for the print theses are held. The Dublin Core record for the electronic copy of the thesis will be converted to MARC using various scripts[14]. The workflow by necessity originates within DSpace as Voyager cannot handle the added submission (i.e. the deposit of an electronic item).

[14] This will be based on the methodology described by Boock, M. and Kunda, S. (2009). 'Electronic Thesis and Dissertation Metadata Workflow at Oregon State University Libraries'. In *Cataloging & Classification Quarterly*, Vol. 47, No. 3-4, 2009, pp. 297-30. Abstract online at: http://catalogingandclassificationquarterly.com/ccq47nr3-4.html [Last accessed 13th August 2009].

This workflow will improve the quality of the metadata associated with the DSpace theses collection as more fields will be used than are at present.

Advocacy and Support

There is a dedicated Support and Liaison officer within the DSpace@Cambridge team. They attend conferences and research events to promote and discuss DSpace and its relevance to the academic community. The officer will speak in depth with scholars and departments wishing to make use of the repository service in order to ascertain user needs, e.g. what work might need to be undertaken before ingest and what level of processing will be required.

Use Cases are made available on the DSpace@Cambridge website to help make the service seem less abstract and highlight its practical relevance to the scholarly community. However, deposit remains a matter for the individual, with no mediation, as yet, at Departmental level.

A pragmatic approach is taken to which Heads of Department get involved in DSpace (some may be targeted over others depending on whatever the current priorities and development plans of the team are, or who has expressed an interest). Departments may set up "communities" within DSpace but with researchers not currently obligated to upload content these may not necessarily be heavily used - as, for example, has been the case with the Department of Physics and its Cavendish Laboratory.

Links to other institutional systems

DSpace staff would like (in the long term) to pull in data from other institutional systems. They seek to demonstrate the relevance and utility of the repository to the academic community by encouraging deposit, whilst at the same time strengthen the perceived and actual value of its service to the university.

There may be opportunities for integration with the CamTools[15] service (a customised version of the Open Source Sakai Collaboration and Lesson Environment – a software platform for developing sets of collaborative learning teaching and research tools) which is not centralised but which is offered throughout Cambridge as a platform for VLE and VRE development by CARET. A widget has been created to be used within CamTools for deposit into DSpace although this has not been rolled out yet.

DSpace staff would like to form a stronger connection with CARET and CamTools - indeed, there have been recommendations made within the University that CARET become integrated with the library, or that the two are formally aligned.

The 3 year Arcadia@cambridge programme, run from within the library, carries out many research projects exploring the role of academic libraries in a digital age. At present one project is investigating the "development of a standard notation for the description of lists and resources, to include contextual as well as descriptive information" and its aims are worth listing as they clearly pertain to interoperability, service development and information sharing across systems:

- *investigation of existing solutions, technologies and standards*
- *consultation with academics, librarians and students*
- *development of a standard notation for the description of lists and resources, to include contextual as well as descriptive information*
- *based on this notation, development of a consistent way to build links from resources described on lists to library (and other) resources*
- *construction of a tool to harvest data from the lists into a central database for re-use*

[15] https://camtools.cam.ac.uk/access/content/public/help.html [Last accessed 13th August 2009].

- *design and implementation of interfaces and tools to allow interested parties to use this data to improve their services and working practice*

- http://arcadiaproject.lib.cam.ac.uk/readinglist/index.html

The project is considering how an XML mark-up system might link Departmental and Faculty book lists and undergraduate bibliographies with records held in the LMS and items held in the IR. Consideration is also being given to sharing Exam Papers within Sakai; these currently tend to be available in print, from University departments, and from the CUL and are not held by the OPAC or IR.

The JISC-funded Virtual Research Environment (VRE) Programme is funding CARET to create a joint publishing system combining DSpace and Sakai. This would allow publication and simultaneous deposit of a paper into DSpace and other research infrastructure systems including a REF publications/citations database, incorporating editing and text mining tools.

The University is currently running a pilot project using the Symplectic Publications system mainly to test how it can perform in an REF context. If a decision is made to move this forward (which DSpace@Cambridge staff believe is likely) they will start working with them to integrate DSpace with Symplectic. The goal is to allow researchers who are managing their publications in Symplectic to also have the option to upload the full text for deposit into DSpace. As metadata for publications in Symplectic would be available for re-use by end users, upload to DSpace would therefore be streamlined and appear much simpler than a reUGLar DSpace deposit. Technically this is achieved by linking the two systems together using web services based on Atom and the SWORD protocol.[16]

Attitudes, Problems and Future Issues

There is no evidence that the development activities occurring within the IR are in any way intended to impact upon or alleviate any difficulties with the LMS or OPAC and their methods of gathering, storing, preserving and making accessible information. In fact, staff feel that IRs may end up with the same fragmentation. For example, not all Cambridge College Archivists contribute to the JANUS catalogue but some contribute (or are likely in future to contribute) to DSpace.

Given the charges involved in depositing large collections with DSpace, it is tempting to speculate that, with the required resources, colleges, faculties or departments within Cambridge might begin to establish their own Repositories. It may be that they would then prefer to set up and maintain their own software systems or work with subject specific standards.

The uncertain future of DSpace@Cambridge beyond 2012 may affect the future of the Repository, with staff having to find time to devote to making their Business Case, however at present they focus on "business as usual," hence, future work. Resources and staffing are also issues in terms of cataloguing, consistency and standards, as evidenced by the fact that one of the Theses and Dissertations librarians has to be "re-purposed" in order to assist with the improvement of the Thesis deposit workflow.

Linking the LMS and IR – Resource Discovery Platforms and Federated Search

As noted earlier, staff at Cambridge do not necessarily feel that duplication of items or item descriptions between the OPAC and the IR is problematic.

Staff at CUL believe that "RDPs present a unique opportunity to search across the OPAC and the IR," solving a number of problems - "not least the multiple catalogue silo situation". Issuing an ITT for an RDP in April of this year, they were impressed by the resulting product demonstrations and are in the process of selecting the successful tender, whose product is required to:

[16] http://www.swordapp.org/ [Last accessed 06 August 2009].

"*...act as a single point of discovery for University collections.*

The system would harvest data from the current Ex-Libris Voyager Library Management System to create a central bibliographic metadata repository. Upon this, the system would provide a rich web interface, acting as an alternative to the existing library catalogues and replacing the existing Universal Catalogue service.

It is anticipated that it would also include library materials from digital collections within Cambridge and subscribed electronic material, via a Federated Search service, and by harvesting data from non-bibliographic sources, such as the DSpace institutional repository."

The RDP is also strongly viewed as an attempt to respond to the opinions of undergraduates – of which staff are aware via researchers and personal supervisors – that a "search engine style" system would be more appealing and familiar to them.

The platform would include, among other things, "a means to control de-duplication of data from incoming streams" and "full harvesting and storage of bibliographic and other forms of metadata from a variety of sources, including the […] Library Management System". Staff could choose to implement only limited de-duplication and could implement an "FRBR-ised" display; with no risk to the underlying data itself (content and carrier are separated) it would be the presentation and arrangement of data that would change. The installation of an RDP is therefore seen as low-risk.

The tender document asks that the entity-relationship model of the Functional Requirements for Bibliographic Records (FRBR) is to be supported if possible. The platform must be OAI-PMH compliant, being both harvester and harvestable. Integration with the SerialsSolutions Webfeat Express (federated search) application, installed in 2008, is also a necessity. Webfeat is used to search for electronic resources and journals within over 300 subscription databases and catalogues (including DSpace). OpenURL Links are to be generated where appropriate and passed to the relevant CUL system.

CUL also wish the RDP to "if possible make use of existing control authority data for authors" for browsing purposes and to codify results. Export mechanisms for data and identifiers from the RDP to WorldCat or other external systems are only "preferred" (as opposed to mandatory). Interestingly, it is required that staff be able to enter bibliographic data directly into the RDP, using Unicode MARC or MARCXML – this suggests that it is at least considered as a theoretical *potential* replacement for (or could job-share with) one or more of the library catalogues. Clearly this would have implications for the current cataloguing practices in operation at Cambridge.

Depending on which indexes are available, the extent to which additional ones could be added and the extent to which data elements included can be customised, the metadata exposed by an RDP might not reflect in sufficient depth the richness of the catalogue records being searched; it may not be able to offer certain search and browse facilities – for example, it may search across all fields in response to a keyword search or offer author, title and subject delimiters but not language or format. Search quality depends upon which indices are available and which features are offered by advanced search facilities. Inefficiencies in metadata, however, may also be exposed when certain fields are used as the basis for indices.

Systems staff have provided very specific details about how de-duplication should be carried out by any RDP to ensure that groups of duplicates will be adequately identified and edited/merged to create "best records" to be visible to users and "non-best" records which will be hidden from the view of end users (but retained and linked to the best record).

We might observe that while for *end users* the location of records becomes irrelevant, it being immaterial, for example, whether Manuscript descriptions are held in the Universal Catalogue, a separate catalogue or an IR, a potential problem presents itself to staff: they now have to work with and learn their way around multiple systems, altering workflows and practices accordingly.

Further speculation would be inappropriate, however, as Cambridge have not yet made public their preferred choice of platform.

University of Glasgow Library[17]

Staff participating in Case Study
Marie Cairney: Principal Library Assistant (Day 2)
William Nixon: Deputy Head of Library Information Systems (Day 1&2)
Rosemary Stenson: Head of Cataloguing (Day 1&2)
Karen Stevenson: Library Systems Administrator (Day 1)

Overview

Glasgow University

Glasgow University was founded in 1451 and is the fourth oldest University in the UK. Its main campus (and several other associated buildings such as those within the Yorkhill Academic Campus) are located in Glasgow although it has a number of buildings elsewhere including a facility at Loch Lomond, the University Marine Biological Station Millport and the Crichton Campus in Dumfries (which it jointly administers with the University of the West of Scotland, Dumfries and Galloway College and the Open University).

The structure of Glasgow University is in keeping with the terms of the Ancient Universities governance system as decreed in the *Universities (Scotland) Acts*, with University authorities comprising the University Court (responsible for Finances and Administration), the General Council (and advisory body) and the Academic Senate (who regulate the teaching and discipline of the University, and administer its property and revenues).

There are 9 faculties within the University, each overseeing a number of departments. For example, the Law, Business and Social Sciences Faculty oversee the School of Law while the Medicine Faculty oversees the Dental School, the Medical School and the School of Nursing.

University of Glasgow Library

University of Glasgow Library (UGL) holds over 2 million physical items.

There are no independent libraries within Glasgow University; all are dependants or branches of the University Library.

It is stated in the Annual Review (2006-07) that:

"*A major focus for development has been the Library catalogue (WebPAC). With different requirements and expectations from a new generation of users, the Library's system supplier (III) has introduced facilities to allow libraries a greater degree of customisation of their catalogues (WebPac Pro enhancement bundle) as well as an entirely new interface (Encore).*"[18]

The Library Management System and the OPAC

System

The Library Management System in use at the University of Glasgow Library is Millennium, with WebPAC Pro acting as the OPAC. This was installed in 1995 and was supplied by Innovative Interfaces Inc. (http://www.iii.com/).

[17] While direct quotations from our discussion are given throughout this report they are not attributed to any particular individual. They are not to be taken as the official viewpoints of the University and are provided for illustrative purposes.
[18] University of Glasgow Library. (2006). *Library Annual Review.* Available online at: http://www.gla.ac.uk/media/media_100619_en.pdf [Last accessed 07 August 2009].

The University Archives Catalogue (Cheshire) uses Cheshire for Archives (Cheshire 3) on top of the Millennium LMS.

Standards

MARC is used as the format standard for bibliographic records. Content standards supported are AACR, ISAD (G) and ISBD. This is identical to the standards used by the Resource Discovery Platform (RDP) (which searches the same database) except that the RDP also uses qualified DC as a format standard. Both use OCLC, Library of Congress and in-house name authority files.

Cheshire uses ISAD (G) as its content standard, Encoded Archival Description (EAD) as its format standard and ISAAR (CPF) for the construction of name authorities.

Libraries Contributing

Branches and other University of Glasgow Libraries contributing to the main Library are as follows:

Adam Smith Library (for students taking courses in Social Sciences or Psychology)
Chemistry Branch Library (for students taking courses in Chemistry)
James Herriot Library (for students taking courses in Veterinary Science)
James Ireland Library (for students taking courses in Dentistry)
Law Workshop - for students taking courses in Law
Language Centre Library (supports language learning and teaching within the University as well as a select number of smaller faculty/departmental libraries).

All materials (or, in the case of the Law Workshop, the *majority* of materials) from these libraries are catalogued in the GUL catalogue.

Other Catalogues

There are 6 other catalogues maintained by the Library. 5 of these relate to Special Collections materials and one to Archival items. They are:

- Cheshire [Archive Services Catalogue].
- Manuscripts Catalogue
- Scottish Theatre Archive Catalogue
- Nineteenth Century Novels Collection
- Scottish Chapbooks Catalogue
- Hill and Adamson [the Hill & Adamson Collection of early photographs]

Although some items from some of these collections are recorded in the main catalogue, each has its own separate online catalogue.

Items in Scope

The 12 item types listed as in scope for the main catalogue are:

- Books - (print & electronic)
- Journals - (print & electronic)
- Audiovisual materials - (video recordings, sound recordings, kits or artworks)
- Course materials - (selected materials that lecturers have asked the Library to place in Short Loan or acquire electronically for a specific course)
- Databases
- Electronic resources - (library material available via WWW, including eBooks, eJournals, eMaps, exam papers, but not online course materials)
- Exam papers - (University of Glasgow degree exam papers)
- Maps - (maps and atlases, print & electronic)
- Music scores - (music scores & sheet music)
- Rare books - (includes all printed items published before 1851)

- Sound recordings - (CDs, LPs or tapes)
- Theses - (print & electronic)

Printed Special Collections (SCs) and pre-1852 items held outwith the main library are also recorded.

The existence of other SCs catalogues is explained by the fact that, being neither monographs nor serials but ephemera or other types of Special Collection material, different domain-specific content standards that couldn't be accommodated by MARC such as ISAD (G) were used.

Archives have always been kept completely separate from the other library catalogues yet it is the point of view of the library that, much like electronic subscriptions should be made available in the Enlighten IR, so archival materials should also be visible in the catalogue - along with all the other specialist searchable databases.

There has already been a previous, unsuccessful project attempting to make these resources discoverable through the main catalogue with a "metadata builder" package. The intention was to harvest the archival records and create corresponding MARC records to describe items; trying to cross-walk EAD into MARC (when there is not a direct correlation between the two standards) proved prohibitive and the major reason for the failure of this project was due to problems with different forms of authority control.

De-duplication, Authority Control and Metadata Issues

"There is no singular, joined up way of thinking about classifications."

In a wide ranging discussion, staff stated that inadequate authority control presents "a huge, insurmountable problem" for effective cataloguing, information sharing, and as a result, interoperability. It was argued that cataloguers have always, by the very fact of their profession, understood intimately the need for authority control, have been advocating the adoption of clear standards from the very inception of Institutional Repositories, and are arguing for still greater convergence in cataloguing workflows across library services. Problems are multiple.

For instance, if one is to consider making use of existing records to improve efficiency, this is complicated by the fact that the authorities used by major systems such as the Library of Congress database do not necessarily match those being used by any given library thus do not accommodate local requirements - certainly this is true in the case of the GUL.

Mapping authorities or, in another example, being able to make use of unique, internationally recognised identifiers is something that might be helped by major projects such as the Virtual International Authority File or Resource Description and Access (which highlight such problems far more than do the Second Edition of the Anglo-American Cataloguing Rules (AACRII)). However, the sheer number of ongoing projects in the area of authority control makes the eventual outcome of such initiatives difficult to predict or plan for.

It is admitted that there are problems stemming from the configuration of the OPAC itself (whether at Glasgow or elsewhere), with students preferring to search by their own keywords; on the other hand, end user generated metadata attached to some records is not authoritative and does not conform to a recognised standard for retrieval; while it may be acknowledged that often faculty/departmental staff can place more comprehensive information on a record relating to subject/keywords for extraction than can the library, they cannot be expected to learn the intricacies of cataloguing.

Such problems are only increased when we move beyond undergraduate level to consider the needs of higher level academics. While those studying at a less advanced level may appreciate a "free text" field for searching or "user tagging" options, delivering effective solutions for the variety of scholars at Glasgow requires a stricter approach. Many faculties would certainly require controlled vocabularies, with sufficient granularity built-in, for the effective description and retrieval of items relevant to their specific subject area, within the catalogue.

Further, there is the problem of name authorities: there are multiple links within and between faculties and, even within a single department it is always possible that there may be more than one member of staff with the same name. When searching across the catalogue for the output of one researcher (discussed in detail in the previous section of this report) one encounters issues both of inconsistency and ambiguity. This is something which could impact on the proposed introduction of a "Glasgow author" field, to be used with both OPAC and IR records, for the identification of institutional staff.

Much work is being done in this area by the Bibliographic Services team. Infrequently, staff working with the main catalogue receive queries about authority control from users regarding the application of surnames or subject headings to their own publications; these are addressed accordingly.

The GUL receives *frequent* requests to catalogue so-called "hidden" collections (specialist collections held at departmental level). Even with the often inadequate data available, such "hidden" collections should, ideally, be represented in the GUL catalogue. Yet resources are insufficient to attain that ideal and while, for instance, resources have recently been made available from the Chancellors' Fund to catalogue materials held by the Department of Celtic and Gaelic such piecemeal work can only go so far.

Journals

Cataloguing modules and tools are sophisticated in comparison to those offered by most Institutional Repositories (IRs) – which initially didn't provide many of the features to improve workflow that are taken for granted in most modern LMSs (for example, de-duplication algorithms, auto-complete functions or bibliographic and authority editing). However, repositories are starting to play catch up.

Journals provide a good example of implicit authority control as well as an example of where LMSs and IRs might interoperate to the benefit of the OPAC. Traditionally, a journal title effectively becomes an inbuilt authority file – cataloguers and end users can search by a variety of fields (title, ISSN, publisher etc.). Enlighten uses the EPrints auto-complete feature to offer journal title, publisher and ISSN data which is pulled from records already deposited and made live in the service. Work is currently underway to add two additional fields ISSN (Online) and Journal Abbreviations to the service. These will become part of the Journal auto-complete feature so that as a journal title is entered, the related ISSN and abbreviations are provided in the record. There is a "Browse by Journals"[19] listing provided that displays all of the journal titles (and the number or articles available). This listing is also very useful for library staff in building greater consistency for *all* journal records.

Duplicate records are move into a separate "Duplicates" work area which removes them from the Editorial Review and helps to both manage the workflow and to assess how many duplicate records are likely to arise in future.

Development of the LMS and OPAC

> *"[There is a need] to make the material much more visible and available through standard library facilities. They should be available through one search facility – the LMS."*

The new "Encore" Resource Discovery Platform (RDP) installed in 2008 offers a lot of potential in terms of modern, improved service, as its development has been and can be influenced by the way young people now search; users "are happy to get a large hit of information and then refine it down".

At present Encore (the service stemming from this is known at Glasgow as "Quicksearch") searches the same catalogue as does the main interface (the "Classic Catalogue"), in essence enabling keyword searching of the same database and the same records with the application of different relevancy rankings. Currently resources are catalogued and made equally available to both systems. Exactly the same results are retrieved using both interfaces, with the exception of course materials.

[19] Available at: http://eprints.gla.ac.uk/view/journal_volume/ [Accessed 3rd September 2009].

While it is worth considering whether the adoption and development of the Encore RDP *may* complicate future cataloguing workflows with regard to duplication between it and the OPAC, this shouldn't be over-stated. One member of the Bibliographic team commented:

"There may be a question about what we catalogue in the main catalogue if, say, theses are already recorded in another database which Encore has the capability of searching, but this debatable."

Complicating any review of cataloguing practice are external requirements such as the UGL's commitment to export its records to WorldCat. As a member of Research Libraries UK (RLUK), it is also obligated to export records from its catalogue to the RLUK database.[20]

The future development of Encore at GUL therefore undoubtedly raises issues but also possibilities. Potential exists to include holdings of the IR (a harvester is available with Encore and the GUL team are looking at harvesting the records from the EPrints repository for use within Encore) but Encore was not specifically procured with the idea of interoperability with the IR at the forefront.

Whatever tensions may exist between the "Classic Catalogue" and "Quicksearch", the cataloguing process cannot be future proofed. It cannot be supposed that the user needs which "Quicksearch" caters for today, will bear any relation to the user needs of tomorrow.

Links to other Institutional Systems

"There is a need to push information out of the patron record to finance and other services."

Influenced by new financial directives and fresh priorities being formulated at the highest level of the University, GUL staff wish to see the LMS develop its relations to other institutional systems.

Much work is going on across the University at present regarding its Identity Management infrastructure (primarily using Management Information Systems within IT Services). There is now more convergence in terms of both strategy and the co-ordination of systems and services, with the infrastructure steadily becoming "much less nebulous" partly through the application of consistent, institution-wide "digital identities" known as Glasgow Unique Identifiers (or, GUIDs).[21]

The GUID is used by staff for a variety of services including access to HR and Research Systems and the submission of Time Allocation Schedules (TAS) as part of mandatory HEFCE reporting activities. Within the library, from the 1st of August 2009 the GUID replaced Athens as the method by which to access e-books, e-journals and subscription databases such as EDINA or MIMAS.

One member of staff stated that: "*Departments like their autonomy; technology is available to aid their integration but it relies in part on web services being introduced which push and pull data across distributed services rather than on being held centrally.*"

The decision to move to web services has come from University Management, who have become very aware of economic and efficiency issues pertaining to duplication and information "silos". Services must therefore be centralised or streamlined, with data being re-used across institutional systems as appropriate.

These imperatives can at times squeeze resources as staff have to learn new procedures and workflows; there may also be concerns about the advisability and purpose of widening access to certain types of personal information. At present, for example, the LMS does not "push" any patron data (regarding fines etc.) out; *"the LMS is the only place which knows a student has accumulated fines in the library."* Might conflicts arise with other departments or offices seeking monies owed them if they are able to see that it is also owed elsewhere?

[20] http://www.rluk.ac.uk/database [Accessed 3rd September 2009].

[21] For more information see: http://www.gla.ac.uk/services/it/projects/identitymanagement/
and
http://www.gla.ac.uk/services/library/howtofindinformation/accessingeresources/glasgowuniqueidentifierguid/
[Accessed 19th August, 2009].

Links already exist between the LMS and the HR System: the LMS "pulls" data in from HR as the basis for its Staff Patron Records and derives data from the Student Records System (SRS) for student Patron Records. This is undertaken by batch loading records on an overnight basis thus it does not constitute 'real-time' interoperability. Again, web services might be utilised to enable this information sharing and processing to be done more quickly were a need for increased speed deemed beneficial. At present staff are satisfied with the existing arrangement.

No interoperability of any sort occurs at present between Research Administration Systems and the LMS.

Exposing Data and Re-imagining the OPAC

> *"We have rules, but what should be focused on is how the end users use these rules."*

If the LMS/OPAC is to become a "junction box" - the 'share point' or 'portal' - for catalogue records, patron records, email and other re-usable data, as staff hope it might, a cultural as well as a technological shift will be required. Such an evolution in service delivery would require the integration of new workflows into standard cataloguing procedures, with the necessity to further improve the skills base of cataloguers. This would sit well alongside current plans to expand the functionality and role of Enlighten (the main Institutional Repository, discussed further below), with some integration and interoperability between the two systems.

One key area being investigated at present is the provision of links between the IR and the OPAC using technologies already established within library services, such as Open URL resolvers, particularly for books and book chapters.

To improve resource discovery and search quality for the end user, a "find more" option might be presented alongside retrieved records (e.g. for journal articles in the catalogue or an "article freely available from" link to the published media, bibliographic service or full text version). This is in some ways similar to the concept encapsulated by link resolvers such as SFX but would be implemented more reliably and enacted on a grander scale. Indeed, a nascent version of such a service has recently been set up within the Enlighten repository's test service, where links to OPAC holding records are revealed to users[22].

Links point both internally to other library catalogues and systems and externally to those of other organisations. Developing this further would combat the perception that for end users, searching the OPAC results in "digital dead ends". Some of this would depend on Open Standards and Open Data and in this way the LMS and IR could mutually develop.

It is beginning to be the case that online course materials (usually managed within the LMS) are being deposited in Enlighten as well as being held in the OPAC. Here, the workflow and the correspondence between the IR record and the existing OPAC record is not straightforward, nor does it necessarily result in interoperability since in many cases the original entry is a static file. However, information is being shared and made visible to a wider range of users as a result and the sharing of records should only be increased, even if this to some extent might be viewed as 'duplication'.

There is a strong feeling that library services cannot simply undertake to re-evaluate workflows and services based on the use by end users of the latest online platforms (such as Flickr or Twitter) – and that academic libraries should not be reduced to chasing users across the internet. Catalogues need to operate by recognised and agreed international standards. Taking advantage of major systems already sharing records such as OCLC WorldCat, to improve efficiency and value for money must be continued and as many items as possible must be catalogued and visible in the main library catalogue.

Institutional Repositories

[22] See: http://testservice-eprints.gla.ac.uk/4303/ for an example.

> *"It has been an iterative and evolutionary journey to get where we are."*

Systems

There are 3 Institutional Repositories at Glasgow University. These are:

1. ***Enlighten*** – for peer-reviewed articles, published papers and books. This runs on EPrints Version 3.1.1.
2. ***Glasgow Theses Service*** – for theses produced by staff and students of Glasgow University. This is also an EPrints installation (version 2.1).
3. ***Glasgow DSpace*** – for working papers, course materials, grey literature and technical reports. This runs on DSpace version 1.1.

All of the UGL's repositories use Open Source software as there were (and are) no equivalent commercial systems available. The administration of these systems is shared between Computing Services and Library staff.

Items in Scope

Enlighten (formerly known as the "Glasgow EPrints Service") is the University of Glasgow's institutional repository service for published research material including:

- peer-reviewed journal articles
- published conference papers
- books and book chapters

Item types in scope, using EPrints' terminology, are:

1. Article
2. Book Section
3. Conference Proceedings
4. Book
5. Patent
6. Artefact
7. Show/Exhibition
8. Composition
9. Performance

In line with the University’s Publication Policy it is intended that, where copyright permits, all peer reviewed journal articles and published conference papers published from September 2008 onwards will be held in Enlighten. It currently holds records for nearly 6,000 items with a target of 25,000 set for March 2010.

The **Glasgow Theses Service** is a collection of full text Higher Degree theses successfully defended at the University of Glasgow. The service does not contain *all* theses defended at the University of Glasgow but rather, around 600 of them (and rising).

DSpace is for "other research material" ('grey literature') such as working papers and technical reports and consists of various communities which map to University faculties and departments which have provided content. Within these communities, various sub-communities and collections have been created depending on the nature or range of material provided.

Background to the installations

For the UGL, the term "Institutional Repository" is synonymous with Enlighten, regardless of their other repository installations.

Each repository, differentiated by content type, was initially established as part of the "Data providers for Academic E-content and the Disclosure of Assets for Learning, Understanding and Scholarship"

(DAEDALUS) project, funded by JISC between 2002 and 2005. Since then they have moved through what staff describe as 3 phases:

1. "**DAEDALUS - Glasgow ePrints Service and DSpace**[23]

At the end of DAEDALUS in 2005 we were committed to maintaining the repository service. After DAEDALUS we did move to *only* accepting full text material rather than the mix of bibliographic and full text which we had been doing as the JISC project. We did anticipate maintaining both DSpace and EPrints but ultimately we decided to focus on the ongoing development of the Glasgow EPrints Service which would become Enlighten. Our local expertise in Perl and MySQL and the development path which the EPrints team demo'ed was closely aligned with the direction in which we wanted to take the repository.

2. **Post-DAEDALUS**

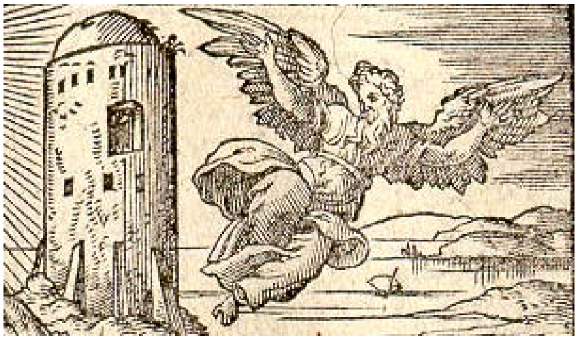
Enlighten was chosen name of the umbrella service and the Library's experiences with the RAE and subsequently as a REF Pilot site become part of the evolution of the IR Service.

3. **Publications Database**

Post-RAE Enlighten became synomous with the Glasgow ePrints Service and was positioned in the role of University Publications Database. Enlighten became an integral element of the University's new Publications Policy introduced in June 2008."

Here, we consider briefly all 3 IRs and all 3 phases, as all are relevant to OCRIS.

Daedalus from "Ovid: Metamorphoses", illustrated by Virgil Solis.



***F**rankfurt 1569 Glasgow University Library, Special Collections*

Each repository serves a different set of institutional requirements and is differentiated accordingly, including at the level of the underlying software. For example, a pragmatic decision was taken by Glasgow as part of DAEDALUS to test DSpace as the storage system for pre-prints and technical reports, with the view that "each piece of software is best suited for the particular needs of the content types being served."[24]

However, there have been changes in both attitude and strategic requirement since the project finished, notably the decision to focus on EPrints over DSpace. As well as the comments given above, staff state (in somewhat conceptual as well as practical terms) that they "opted for EPrints because Glasgow is not a pure repository but a hybrid breed."

Further:

"We are repositioning the Repository as the University's publications database. It is therefore not just about bibliographic data and Open Access materials."

[23] Theses were included in DSpace during this phase.

[24] http://eprints.gla.ac.uk/3718/1/Enlighten_oclc_article_.pdf [Accessed 3rd September 2009].

Glasgow has not got the technical resources to support two different platforms long term and ultimately EPrints was a better fit in terms of publications and the technical abilities of staff, than was DSpace.

Therefore EPrints has been nurtured and advanced while the original DSpace install has not been developed and, because of the old version in place is now considered to be, deprecated. Some of its content (such as 150 electronic dental electives) will be migrated from it to new services using EPrints. The UGL now only accepts material for DSpace in very exceptional circumstances, ensuring when they do that the depositor is aware of the status of the service (i.e. that it is being maintained on its existing server but that there are no plans to further develop or enhance it).

The aim of DAEDALUS was always to bring to light such issues and to provide a basis for comparison. Staff state therefore that "DSpace actually helped the cataloguing workflows." They are now satisfied with the configuration of their systems and by the improvements represented by each new EPrints release. Version 3 for example brought with it more fields and improved metadata format compatibility as well as an open URL resolver which aids the creation and maintenance of links between the IR and OPAC.

With their scopes (or "cores") now clearly defined, deciding which services and tools to implement on top of each Repository appears less of a challenge conceptually and internal development is very much ongoing.

To summarise: Enlighten acts as a comprehensive publications database with as much content being made freely available as possible (and in such a way that as much of that data can be re-used in as many different places as possible). The Theses service has a clear and self-evidently narrow scope, whilst DSpace (originally intended to make 'grey' and other non-peer reviewed outputs visible) is no longer actively accepting content.

Generally, it is felt that as long as there are staff with Perl and MySQL skills within the Computing Services team, modifying and customising the source code is relatively straightforward. A "Test Repository" – to all intents and purposes identical in look and feel to the live version – is installed on the same server as Enlighten and is used as a sandbox for testing new features (for example, the creation of new views, additional fields and services such as the OpenURL resolver).

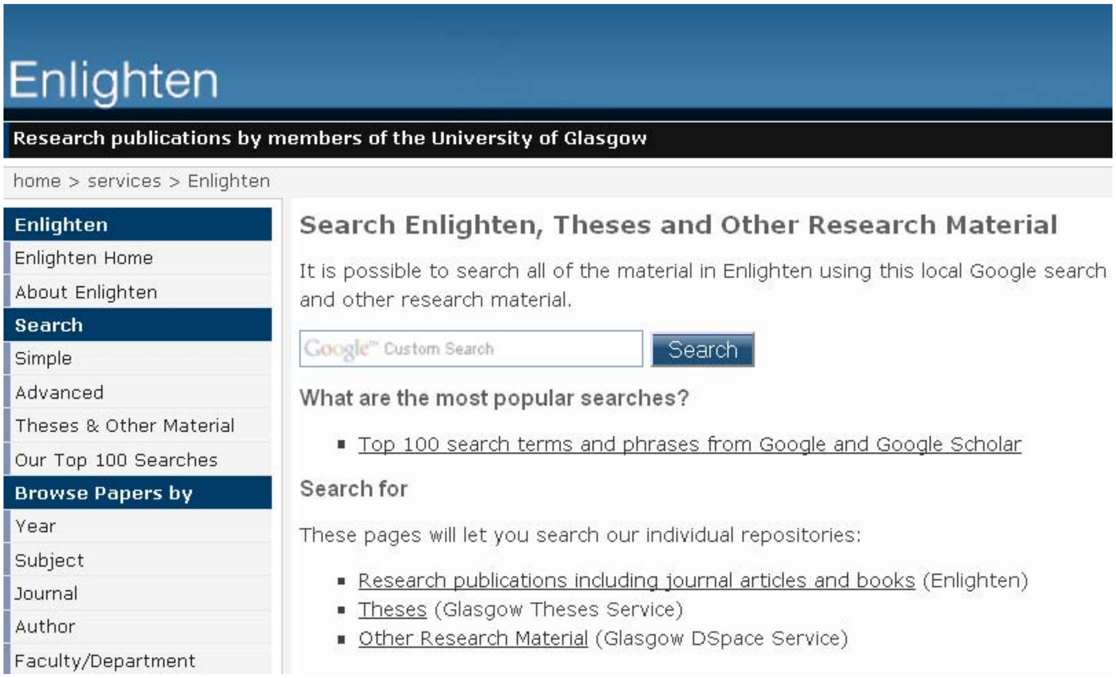
Presentation Layer

Remnants of the IRs' histories can still be glimpsed in some aspects of Enlighten's User Interface (UI) as it continues its journey towards becoming a central publications system.

For users accessing the homepage, it may appear that the Theses Service functions as a sub-set or specialised partition of Enlighten (Theses being referred to in its search options menu as visible in Figure 1) or that the three IRs are inter-linked. Indeed, during "Phase 2" when Enlighten was still the "umbrella title" for the entire suite of Institutional Repository services, this was the case.

Re-branding in early 2009 brought about the separation of the 3 services and although there are web links provided on Enlighten, allowing users to navigate and select easily between the 2 other services (thus conveying a seamless "one stop shop"), the underlying services are technically quite separate.

The Google Search box that can be seen in Figure 1 below is a custom service which was set-up to create a cross-repository search service, returning results from Enlighten, the Theses Service and DSpace while providing the familiar "look and feel" of Enlighten for both the searches and the results. It is configured to return only records and full-text PDFs. Browse lists by author and department etc. are excluded from the results.



Search Enlighten, Theses and Other Research Material

It is possible to search all of the material in Enlighten using this local Google search and other research material.

What are the most popular searches?

- Top 100 search terms and phrases from Google and Google Scholar

Search for

These pages will let you search our individual repositories:

- Research publications including journal articles and books (Enlighten)
- Theses (Glasgow Theses Service)
- Other Research Material (Glasgow DSpace Service)

Figure 1. The custom Google search page of Enlighten, Glasgow's core institutional repository.

Standards

Enlighten supports Simple DC, DIDL, METS, UKETDDC and Context Object metadata format standards. In-house authority files are used for names with LCSH used for Subject control.

The Glasgow Theses Service supports Simple DC and the UKETDDC standards. LCSH is used for Subject control.

DSpace supports qualified DC. LCSH is used to assign keywords.

Contributors

Enlighten is open to members of the University of Glasgow - this can include both students and members of staff.

The University's Publications Policy[25] (which came into effect in September 2008) requires staff to deposit copies of all peer-reviewed journal articles and conference proceedings into Enlighten and applies to material published from this date onwards.

The Theses Service is open to all those defending higher level theses (Masters and Doctorates) at the University of Glasgow. As of session 2007/2008 students have been required to submit one printed and one electronic copy of their thesis rather than two bound copies.

DSpace is open to all members of the Glasgow University community however it has also provided more flexibility than the other services. DSpace has also been used a platform to host material from conferences run by members of the University of Glasgow where some content has not been authored by members of the University. An example of this is the material from the 13th Annual Conference of Slavists [26] which was created for the Department of Slavonic Studies.

Position within Institution

[25] http://www.lib.gla.ac.uk/enlighten/publicationspolicy/ [Accessed 05 August 2009].
[26] https://dspace.gla.ac.uk/handle/1905/21 [Accessed 3rd September 2009].

"*An over-night success after 8 years.*"

Enlighten and the Glasgow Theses Service enjoy an important status within the University, with the success of library staff in building up a "critical mass" of deposits, establishing a positive reputation and gaining Senate approval for the establishment of Mandates (or, to use the GUL terminology, 'Publications Policies'), demonstrating the successful iterative and evolutionary journey that the IRs have taken.

Initially, "while project staff were successful in persuading academics from a range of Departments to deposit content" and while statements such as The Scottish Declaration on Open Archives[27] encouraged the deposit of material into Enlighten, "it was very clear that significant amounts of content were never likely to be deposited unless a mandate was in place and unless deposit was built into working practices" (Greig, 2009)[28].

Intensive advocacy and outreach work was therefore undertaken and directly influenced the decision to implement University-wide policies requiring electronic deposit. The "support of the University's Vice Principal for Research was vital in" gaining approval for the 2007 mandate for the deposit of electronic theses[29], followed, in 2008, by a policy (Mandate) for all research publications within Enlighten, following the presentation of a paper by the VP to the Senate.

Delivering on Enlighten is now one of the Key Performance Indicators (KPIs) for the University in the assessment of how well it has met the aims of its research strategy.

Institutional "buy-in" has been heavy, with "internal high level lobbying"[30] of the Principal, the University Management Group and the University Research Committee by the Library's strategic and operational planning team, from the time of the DAEDALUS project on. The proposed bibliometric aspect of the Research Excellence Framework (REF) has been a key factor in discussing how to centrally accommodate University publications and associated metadata and how to ensure that Enlighten is fit for this role.

In a review of the RAE 2008 and the University's work with the REF Pilot[31], the need for a university wide publications was identified as a key issue, as was HR not providing sufficient information about staff to other services (something revealed by the pilot to be common across the HE sector). The need to effectively link publications data with other university systems is now recognised as a critical activity, also across the sector. Library staff suggest that "three Rs" have been the driving forces behind aligning the repositories with institutional strategy and gaining the support and goodwill of academics:

15 RAE
16 REF
17 Relationships

Staff have "triangulated" the repository so it actually addresses those matters which the institution views as priorities. This has taken a great deal of patience, persistence, and outreach work. Working with the University's Research Office (RO) was a significant component of preparing publications

[27] http://scurl.ac.uk/WG/OATS/declaration.htm [Accessed 19th August, 2009].
[28] Greig, M. (2009) Achieving an 'enlightened' publications policy at the University of Glasgow. Serials, 22 (1). pp. 7-11. ISSN 0953-0460. Available online at: http://eprints.gla.ac.uk/5119/ [Accessed 19th August, 2009].
[29] Greig, M. (2009) Achieving an 'enlightened' publications policy at the University of Glasgow. Serials, 22 (1). pp. 7-11. ISSN 0953-0460. Available online at: http://eprints.gla.ac.uk/5119/ [Accessed 19th August, 2009].
[30] Ashworth, S. (2004). *Glasgow University institutional repositories: a case study.* Presentation available online at: www.lmba.lt/ppt/SusanAshworth.ppt [Accessed 06 August 2009].
[31] Glasgow University were one of 22 institutions involved in the HEFCE REF Pilot Exercise in the construction of bibliometric indicators of research quality. See http://www.hefce.ac.uk/Research/ref/Biblio/ for further details. [Accessed 19th August, 2009].

data for the RAE. Although Enlighten was not used per se, this experience helped staff to forge relationships with the RO as well as with other Departments across the University, aiding the gathering of crucial, up-to-date contact data and providing a driver for gathering data from the Science and Medical Faculties and the Accounting and Finance Department (on top of those Faculties with which they had stronger pre-existing connections such as Arts and Social Sciences). This has enabled the publications store to be expanded.

The "tripartite" publications policy at Glasgow encompasses publication details, open access and the citation of the Institution. This is viewed as a vital "sea-change" for the visibility of institutional publications. Glasgow has now, it appears moved to the next-generation of IR advocacy (where nobody now asks "why" they should deposit any more, only "how").

A mutually beneficial convergence between economic, research and teaching priorities might be seen to be crystallised within the efficiently functioning IR.

To take an example: money can be saved and workflows improved if online course materials can be found in Enlighten as well as in the OPAC given that the supply of such materials is, at present, relatively primitive via the LMS and is mainly limited to Arts based subjects and book sections etc. For Bibliographic Services, cataloguing these materials is not particularly straightforward due to, for instance, a reliance on unstable course-lists and course-codes; the IR enables faculties and departments to build up electronic study packs more readily.

Advocacy and Outreach

> "*Departmental administrators are acting as intermediaries with the IR, taking over workflows between the author and department and IR deposit"*.

Training workshops have been delivered to a wide range of Departmental Administrators within Glasgow. Library staff supply additional documentation for these workshops, and the 2 hour session covers the mechanics of adding records as well as addressing more intuitive questions such as "what's this about" and "why should we bother?" Workshops and presentations provides context, opportunity for the IR Advocacy Manager to speak to university staff face to face, to get administrators to register for the system and to provide use cases and examples.

The cultivation of faculty contacts and direct links with academics might be viewed as examples of 'Best Practice' in outreach.

In terms of mediated deposit, the preferred option for many scholars is "proxy deposit" by Research Assistants and Departmental Secretaries, with a small number of departments which already have a culture of self-archiving choosing to self deposit.

Enlighten staff are keen to work with Departmental Administrators as they are not losing sight of the necessity of getting hold of full text materials. They are working towards a scenario whereby mediators within a department would email relevant academics with reminders to deposit full text versions of their articles in the IR. This builds on the workflows developed in liaison with such staff during the RAE and they continually communicate updates, offering training and guidance. It is noted that when such electronically delivered services are fronted by real people, scholars can more readily perceive their value.

Duplication, Authority Control and Metadata Issues

Some authors or mediators are better than others at providing accurate, comprehensive and reliable metadata but while standard checks are carried out on author completed fields by Bibliographic Services staff this is not generally found to be prohibitively time consuming.

One solution to the problems of authority control is to work with reliable, consistent data from other, administrative systems, for example, the use of staff names, IDs and e-mail addresses from the university’s "Identity Vault". The authoritative version of person's name (as well as their staff number and department) would be taken from the Identity Vault and linked to the records of all publications associated with that individual. Here, the retrospective conversion process would be extensive as there are at present some 3,000 records to sort for staff numbers. For publications by staff from the Department of English Literature there are also separate Staff ID numbers to consider while there are 400 authors for Physics papers alone (due to the commonality of multiple co-authors). Here one also encounters the problem of authors not affiliated with Glasgow.

De-duplication workflows can become complex when dealing with multi-authorship, for example when author W from department X co-authors with author Y from department Z, and where each author/department may have deposited the document separately as part of their own workflows and output – which is the case with departments such as Statistics and Computing:

"The Department of Statistics are adding their papers, but co-authored materials may already be in the IR, having been deposited earlier by the collaborating authors from other departments".

Prior to removing a duplicate record, staff would check each version to see which record is more complete, (e.g. which has the most appropriate keywords attached) and will then simply strip mine the duplicate record. EPrints is flexible enough to permit one record to be attributable to more than one author, subject heading and department.

Names are notoriously unstable (for example, they change because of marriage) and are a high cause of duplication; this is yet another justification for using a separate unique identifier field. Publications linked to individuals recorded, for example, under their maiden names, would be pulled out and cross-matched against their staff numbers (regardless of the name an author is cited by). A "Researcher ID" tag derived from Web of Knowledge, for instance, might also be used.

For members of staff who publish prolifically and on services such as Web of Knowledge, there will still be a clash with the Identity Vault's authority records – again, Name Authorities in Library of Congress records will often also be different.

However, when an author moves on from the university, their staff number is not re-used, these remain unique; something which is seen by staff as extremely helpful – the University will keep a record once an individual member of staff moves on and their published material need not be modified or removed from the repository.

There are additional granularity issues for IRs (and OPACs). The UGL could potentially utilise top level Library of Congress Subject Headings (LCSH) alongside LCC classes for article level cataloguing, to be applied across the two systems. However, LCC as utilised in the IR may not best serve the very specific nature of some of the research output at Glasgow.

Developing and improving metadata

Developments addressing metadata issues within Enlighten continue apace with a clear focus on enriching and linking data to increase efficiency and improve utility for both core users and staff.

Bibliographic "cover-sheets" have been added to items in Enlighten, carrying all the relevant information needed for the published citation; these also carry the University crest, the unique identifier in Enlighten, the URL for the full text version, and a link back to the Enlighten search or home page. This addresses the concerns of researchers and funders regarding provenance issues in search services such as Google Scholar, which link directly to a PDF ("avoiding a beautifully crafted record") without any indication of source location.

There are plans also to include citation of the address of the University of Glasgow as the place where the data was produced. The inclusion of a field in EPrints to indicate institutional provenance would make research materials retrievable "on batch", for example, on Web of Knowledge or within UK-wide theses repositories. This would also enable IR staff (and other stakeholders) to extract such data for sharing and re-use.

Developers are experimenting with the use of the EPrints "auto-complete" function which informs you if a record for the title you are entering is already there. This cannot be entirely automated as checks will still need to be carried out to discern whether the same title is being used for different papers. In addition they hope to use auto-complete to match an author's name to his or her HR derived staff ID.

They are also considering the creation of an additional field called"Glasgow Authors" will allow newly established local name authorities to be detailed while leaving original citation details unaffected.

The planned authors field will include University of Glasgow e-mail addresses which will enable Enlighten to tie publications to staff and to display the staff name used in the University's A to Z service and held by Human Resources, regardless of the citation used by the author. A separate Glasgow Authors field will be displayed in the record and an additional browse view (and search option) of Glasgow only authors will be made available. This view will provide a range of groupings such as date, item type and keywords (as a tag cloud) and could be included in staff web pages.

IPR, Version Control and Open Access

Theses

There are many reasons why access to a thesis might be restricted – for example, if the work was commercially sponsored, if portions are due for publication in monograph form, or if sensitive information is to be exempted from FoI requests. In such cases it is up to the student to speak to his or her Supervisor or Faculty Freedom of Information Scotland Act Coordinators. Requests are then considered by the relevant Faculty Graduate School and if accepted, the content will be placed under "embargo" for a period of 3 years.

Students are also responsible for seeking copyright clearance allowing the publication of 3rd party content, although Enlighten staff will provide help and advice if contacted. Where clearance has not been granted, the student has the option of providing a version of their thesis without this material which will be made available. If no version is available the thesis is not added to the service and message "Due to Embargo and/or Third Party Copyright restrictions, this thesis is not available in this service" is displayed in the record.

Journal articles and working papers

There has been a shift in recent years, with "versioning" having become a major concern **-** deposit of an author's "final version" in an Open Access repository is often a condition of the research grant allocation. While some funders are happy to accept the "author final" version as being the one which appears in the IR, others prefer the final, post-print copy as it appeared in its official place of publication, or a copy as close to the final as possible, and certainly post-peer review.

With regards to working papers for research, for instance, advance online publications are often ineligible for the research evaluation depending on the publication date. Pre-prints are therefore *not* permissible for deposit into Enlighten.

Many publishers such as *Nature* are happy for the author's final version to be used within the IR, provided they are credited accordingly and indeed, publishers may themselves attempt to deposit a final version in the author's "home" repository; there are opportunities emerging for the SWORD plug-in to be used to at the point at which you deposit in *Nature* that also deposits a version into an academic's local institutional repository.

For all item types, there are copyright concerns for materials deposited from subject areas such as History of Art (for obvious reasons such as the reproduction of artworks) and from Archaeology (where illustrations from Ordinance Survey maps may be used). Clearance for art images may have been granted for *journal* publication but not for versions being held in a repository. Yet academics can often be unhappy about "not for publication" versions being made available on Open Access.

IR staff check copyright for any uploaded journal articles. Publishers tend not to be too concerned as long as the IR is clear about the version being deposited and supplied to end users and as long as citation details are provided with each version. EPrints 3 provides a version history for each record as well as enabling links to be established between pre- and post-prints where appropriate.

Unnecessary duplication is avoided by the UGL. For instance, if material goes into PubMed, the IR won't take a copy – it will only provide bibliographic metadata with a link for end users to the full text article.

Enlighten has a rapid "take-down" policy, and articles are made available along with a note saying that it has been reproduced in line with publisher's policies – that they have cleared for publication, who the copyright holder is, and where first published etc. Additional useful links of this kind are provided, for example:

"*As part of the drive towards populating Enlighten with full text, staff have been linking bibliographic-only data to full text where it is freely available (PubMed Central for biomedical and life science journals or ArXiv pre-prints in physics, mathematics, computer science and quantitative biology).*"[32]

Development of Repository, Tools and Interfaces

"The Repository itself doesn't need to be visible."

As noted previously, the RAE and the REF Pilot was a key driver for the IR having gained the support of the University for further development in advance of the next research assessment.

GUL staff are seeking to deliver functionality for the IR such that data can be "sliced & diced" for use in different ways to serve the various imperatives and emphases of research evaluation (whether bibliometrics or other measures). Assessors, for instance, may not just want journal data, or data on conference proceedings and the publication of book chapters; they may wish to access the research output of people still at Glasgow, or output from Glasgow during the years 2002-2006.

Staff are looking creatively at how they can extend existing functionality and are committed to IR innovation.

Deposits are currently managed in one of 2 areas: User Work Area and Under Review/Editorial Review (for Bibliographic Services staff) before the record is pushed into the Live Archives. There is a 4th area for retired items which are withdrawn. If a record is flagged as a duplicate, staff assess which is "best" (i.e. which one has more/better information), then merge that information with the live record. Records already made live in the service should always be kept.

UGL staff have also created a staff only field "Enlighten Staff Notes" which Editorial staff can use at the review stage to make notes on such points and in addition to this they are developing a Deposit Issues field which provides a checkbox list of common deposit issues such as "Content Out of Scope", "No Glasgow Author" and "Copyright Checking" which can be searched by the staff for subsequent review This is in place in the EPrints test service, with the range of options currently undergoing refinement.

EPrints 3.1 provides an "Issues Tracking" feature[33] which can identify, for instance records which are 3 years old but still marked as "In Press".

[32] University of Glasgow Library. (2006). *Library Annual Review.* Available online at: http://www.gla.ac.uk/media/media_100619_en.pdf [Last accessed 07 August 2009].

[33] http://wiki.eprints.org/w/New_Features_in_EPrints_3.1#Issues_tracking_system [Accessed 1 September 2009]

For publications, particularly in the sciences which are published online in advance of the print publication Enlighten staff are going to maintain a single record for both the electronic and printed version rather than create separate but linked records. This can be done using the new ISSN (online) field and a new date field for "Advanced Publication Date" which will enable staff to record the online publication date. The record will also be flagged as an "Advance Publication Date" in the Deposit issue field so that it can be retrieved for review. When the publication is published and data including pagination and volume/issue number are available these will then be added to this record.

Other development activities include the use of end user generated metadata (uncontrolled keywords) supplied by researchers and academic staff as a tag cloud instantiated in the author browse view[34]. This is seen as an enhancement to the flat list (e.g. of Glasgow Authors) with the clouds used to represent the frequency of occurrence of keywords and co-authors associated with an author's deposits.

Code has been "dropped in" so that "latest editions" of relevant deposits are linked to social networking sites. They have also enabled Google analytics in the main OPAC, the Encore RDP and the Enlighten IR. It is felt that:

"Here, Google's relevance ranking and faceted classification might be more relevant than OAI-PMH, judging by the number of searches that come in via OAIster and Intute."

Over 75% of the traffic to Enlighten comes from Internet search engines like Google and Bing.

There is a conscious effort to avoid, where possible, the "digital dead end". Enlighten is a "hybrid" service with a mix of full text and metadata only records and options need to be provided for users to enable them to find the full text of the item which they are looking for.

Digital Object Identifiers (DOIs) are added for journal articles which provide a link to the publisher version of the paper, whether it is freely available or not in the repository. This field was re-labelled as "Publishers URL" in Enlighten, in EPrints the default phrase is "Official URL". Staff have also been adding links to records in the library catalogue for books and book chapters now being added to the service. All of this may eventually change how the repository is used and viewed by researchers.

A more flexible option to address this "dead end" issue is the implementation of a link to the library's OpenURL resolver which will enable the user to pass record details to a range of other resources including the library catalogue and journal publishers or to initiate a document delivery request. EPrints includes the code for the SFX and OVID resolvers and Glasgow has modified this to point to its own WebBridge resolver (from III).

Resolver links have already been implemented in DSpace as well as a number of EPrints repositories including the University of Bath's OPUS IR[35] and Northampton's repository, NECTAR.[36] The latter IR links to the OpenURL Router service[37] which will resolve the user to their local service and the items made available through it.

UGL staff are still test driving the library's OpenURL resolvers and exploring how they might use these to add extra value and encourage users to access their online resources; for example they could be embedded in Twitter 'tweets' or blog posts, with Google analytics revealing to staff how users had 'entered' the OPAC or IR.

There is a conscious effort also to get away from the repository being "the end of the line" for those searching for Glasgow's research output; to address the needs of users who may have bounced into the

[34] http://eprints.gla.ac.uk/view/people/Drysdale=3AT=2ED=2E=3A=3A.keywords.html [Accessed 3rd September 2009].
[35] http://opus.bath.ac.uk/ [Accessed 3rd September 2009].
[36] Online at: http://nectar.northampton.ac.uk/ [Accessed 3rd September 2009].
[37] http://openurl.ac.uk/doc/ [Accessed 3rd September 2009].

site from Google searches, and also to consider how the information they harness can serve the needs of future research and funding:

"We are moving away from flat browseable lists to offer end users different views, [such as a] *"group by funder" feature, groupings by keywords or tag clouds, or view the different kinds or sources of funding that an author has received e.g. Scottish Executive etc."*

By doing so they hope to be able to answer queries from researchers/departments/university managers/assessors for research evaluation data, questions such as "how many articles do we have for Faculty X, for month Y?", the aim being to evolve the repository to more effectively report on the level of author outputs and the number of articles downloaded/published by members of a particular department.

In seeking out ideas for service development, Glasgow has also responded to the shift in emphasis to research groups. The Computing Science Department, for example have asked to be able to search by "research group". Research pooling is partly the focus of the ongoing JISC funded ERIS project, and how such data is normalised was part of the challenges of HaIRST.

How might research themes and groups be adequately identified? "Research Group" is currently a free text field but they plan to apply more rigour to how this is mediated. This would involve taking a step back to discern what the university stipulates are its range of previous, current and future research themes – controlled lists could then be drawn up as is the case currently for departments and subjects.

There is an increasing burden on the researcher/department to add sufficient tags when depositing materials but also on the IR provider to apply appropriate fields to support a range of retrieval needs. For instance, the Glasgow Theses service is to extend to map the variety of fields that ETHOS does, exposing doctoral theses. Make them available and searchable by the right fields.

An additional feature that could prove useful to end users would be an Amazon style association service for end users: "if you read this, why not try this?" It could be linked to a spin-off, complementary partition of the IR and improve the synergy between research outputs and the repository.

Part and parcel of the evolution of the repository is its vision to "empower" proxy depositors so that they feel more "enfranchised" and can more clearly explain the benefits of the IR to the staff they administer.

Links to other institutional systems

"The repository is a bit of infrastructure. It is up to the University to decide its functionality."

IR staff are currently involved in a JISC-funded project called "Enrich[38]" which will establish links and interoperability between the university's Research Systems and Enlighten. To quote from the project website:

"*We have recognised that the repository cannot play the range of roles expected by its users and institution if it continues to exist as a separate and disconnected data silo. This project will place the role of the repository as one which is a natural part of the research management cycle rather one which is a separate and disconnected activity. This work will be done in partnership with a range of academic departments and the University's Research and Enterprise Department.*

This integration, in conjunction with the University's Publications Policy[3] *(2008) will enable us to demonstrate the range of benefits offered by the repository and to increase the rate of content (full text and metadata) deposited."*

The research system is interested in interoperating with the IR to be able to retrieve data on the range of publications and funders associated with research pools and projects.

[38] See http://www.jisc.ac.uk/whatwedo/programmes/inf11/enrich.aspx [Accessed 19th August 2009].

As with the shift towards the representation of research themes or pools within the IR to enhance service, they are investigating pulling in data from the Research System to tie up information on research projects with their funders (e.g. "Alexander von Humboldt Foundation, Germany. Project theme: Cardiology research"). The Research System would then find associated publications, link for example to funding documentation and to disclosed data regarding types of award. Such a service would have to be quite granular, as projects can have multiple awards.

Similarly, developers are experimenting with the inclusion of research "Project numbers". Using a unique project number (which will exist within the Institution already) materials could be filtered through the "to be reviewed" file to enable researchers to gather all their publications ready before the end of the project - project work/outputs would then appear close to one another in the review file.

On a technical level, when linking up with the Research System, a file is exported from the RS formatted in a specific way so that there is a flat text file sitting in an EPrints directory and this is the one which is then interrogated. The process can be refreshed on a daily basis, it is fairly efficient but it is not direct interrogation of the RS system or system integration. It is a pragmatic solution which has resulted from the IR team working in relationship with Research and Enterprise over recent years in preparation for the RAE and readiness for the REF.

There is an agreed need for the IR to share data with other systems, as even machine to machine "interoperability" with the LMS will not meet all the university’s needs and requirements: "[The repository needs] *to be plugged into other university systems for funder compliance."*

Linking the LMS and IR

"It’s both a publications database and an open access repository - a pragmatic hybridity."

IRs and catalogues emerged from different backgrounds, so there are a number of difficult issues related to the potential cross-over between the IR, the OPAC and the LMS.

Glasgow is in the fortunate position, however, of having had its IR workflows on cataloguing embedded with Bibliographic Services. IR staff work across both the LMS and IR, bringing with them the standards and practices of trained cataloguers. It is clear that all involved value this arrangement, demonstrating that the IR is operationally integrated if not yet machine to machine interoperable with the LMS/OPAC.

The need to embed this work in the existing operational and organisational framework of the library was a key part of the work of the DAEDALUS project. This enabled the Library to support and to continue the work begun by DAEDALUS and to transition this to a live service. Bibliographic Services received additional funding for posts which included a mix of cataloguing and repository work.

The University Librarian at the time of the IRs inception (Chris Bailey) was always very keen that staff didn't see an artificial distinction between the IR and library services, effectively embedding it into the workflow and activities of the library. However, there are still associated resourcing issues which come with this, with staff being co-located in Bibliographic Services. It remains to be seen if it is a sustainable solution long-term.

It can be the case that IRs become"silo-ised" and not necessarily aligned with institutional strategies but at the University of Glasgow the library is heavily involved in its development.

It is viewed as being a natural extension of the traditional activities of the library into a more granular field, with the capacity to drill down to article level. Libraries have now re-positioned their services.

Approaches to the IR and the LMS are different and functionally must currently remain separate. Standards, however, should be equal in terms of their quality and application. So standards to assist interoperability are viewed as paramount.

"The problems of interoperability had already been encountered in the DAEDALUS project."

In light of their previous experience with DAEDALUS, recent developments, and current experiments, to achieve "interoperability", it is explained that IR staff "*Would export a file from the research system (a flat text file) which would then sit in a directory in EPrints and that is the one which would be interoperable* [with the other systems mentioned] *rather than attempt to directly integrate the two* [or more] *systems."*

They would scope out and bring together relevant materials with the mission of getting the freely available full text version to the end user. Such progress is necessarily a pragmatic environment, a communal process, a continual collaboration:

"[The IR as a kind of]"junction box" where data is being fed from research systems, publications data, data from admin, the author's staff ID from the university's Identity Vault [with the functionality] to re-fashion and re-purpose information for authors and others searching the university's publications, to provide data outputs for the REF and to provide leverage for future funding."

It is important that staff knowledge and expertise be leveraged in the right way, towards best practice. This is something that should be advocated. Building relationships across all levels of the institution and the integration of IR workflows into cataloguing has encouraged reflective practice at the UGL (i.e. "not just slavishly taking publications data from external sources but *doing* things with it", to enable its re-use by the end user).

The IR's core function is to be the publications database and Open Access repository for the university. It is therefore, like the library itself, a hybrid service. IR staff didn't want the service to become "ghetto-ised" either and view it pragmatically as a "junction box" into which data is fed and which can then point end users elsewhere in their discovery of additional resources and data.

Through services such as IRs and RDPs, libraries have an increased role to play in a "post RAE/pre-REF" world and both IR and Bibliographic Services staff at the GUL believe that "*People who previously might not have had much engagement with the library have come to see its real value"*.