

# Implementing HarvestRoad Hive Digital Repository

**Martin Borchert**

*Manager, Access Services, Division of Information Services, Griffith University, Australia*

**Joanna Richardson**

*Digital Repository Administrator, Access Services, Division of Information Services, Griffith University, Australia*

With contributions by

**Geoff Mitchell**

*former Manager, Educational Products and Services, Division of Information Services, Griffith University, Australia*

## Abstract

In 2003 Griffith University entered into a partnership with HarvestRoad to co-develop the Hive digital repository system. Issues relating to the base architecture (structure of collection bureaus, categories and sub-categories), security, permissions, integration with the LMS, granularity of objects to be managed, content reusability, base metadata standard, search facility, workflows, version control, content rendering and format, and timed release and content expiry were worked through. Policies dealing with intellectual property ownership, copyright management, and process ownership were developed. A number of start-up bureaus were established. This paper reports on progress, future applications, and lessons learned at Griffith.

## Introduction

In 2003 Griffith University selected the HarvestRoad Hive digital repository system. Griffith's aim is to use Hive for as many digital repository system applications as the flexibility of the system will allow. Interested readers will find an in-depth discussion of the background to the digital repository project in an earlier paper by Borchert and Richardson (2004). Subsequent presentations by Borchert (2004) and Richardson (2004) have provided project updates.

This paper provides an update on more recent project developments.

Developing and integrating a digital repository system into the University's elearning and digital library environments is proving to be an ambitious, complex and long-term project. Since the project began in earnest in May 2004, significant progress has been made in the areas of repository structure, integration with the Learning Management System, application of metadata standards, use of controlled vocabulary, workflows, object granularity, reusability and archiving, and copyright management and intellectual property issues.

Starting from scratch with little knowledge of either digital repositories in general or the HarvestRoad Hive system in particular, the project team has enjoyed a steep learning curve. The long-term success of the project will depend on the team's ability to develop new skills, and to bring together existing organisational knowledge and skills from areas within the Division of Information Services in order to integrate the digital repository system across the elearning, digital library and various relevant university administrative environments and functions.

Work has begun on a number of 'starter projects' including digitised past exams, course readings, art image collections, ePrints and high economic value teaching and learning objects. The issues listed above are discussed in the context of these starter projects.

Work completed to date lays a foundation for what is a significant long-term project. The size, complexity, and workflow aspects of current elearning and digital library collections require the implementation of a digital repository [object management] system right now. Griffith University, with a philosophy of experimentation and wanting to 'give it a go', fully expects to learn from mistakes, and to increase its expertise in the use of Hive over time. Our elearning and digital library environments have been developed over many years, and so too this is just the beginning of a successful implementation of Hive in its many applications at Griffith.

## **Background to the Hive Digital Repository Project**

Hive is being developed as the solution for as many digital repository purposes as the system will accommodate. Hive is designed to store digital resources such as learning objects, images, readings and other materials. Objects can be grouped into collections or bureaux. Each bureau can be searched individually or by federated searching. A new Digital Repository team was established in 2004 within the department of Flexible Learning and Access Services (Division of Information Services).

## **Background to Learning@Griffith**

Learning@Griffith is a centralised, university-wide online teaching and learning management and delivery system built on the Blackboard LMS, within the portfolio of the department of Flexible Learning and Access Services. Learning@Griffith is currently the most heavily used single application of Blackboard 6.0.11 in the world and is now being migrated to Blackboard 6.2.3.6.

Academics are providing much course content via Web pages within the Blackboard environment and are also loading related digital learning objects directly into the LMS. Academics currently have access to their own content only within Blackboard and are required to manage their own digital learning objects, the result being that they do not currently have an environment in which they can share and re-use learning objects.

## **Digital Repository Structure**

With the aim of multi-tasking the Hive system, Griffith purchased an unlimited license that allows for the implementation of many 'bureaus', i.e. complete clones of the system designed to manage different content sets. Each bureau can be established with customised security permissions, workflows attached to item types, and may consist of many Categories or sub-collections.

The Hive structure needs to be designed taking into consideration the following issues associated with content, its intended use, and required associated levels of control:

- The nature of the content. (e.g. administrative vs. teaching and learning vs digital library collection content)

It is immediately obvious that internal or restricted collections created by Information Services such as training materials, templates, draft policies etc should be stored separately from collections available to Griffith staff and

students and that publicly accessible collections should be stored separately again.

- The structured nature of the submission process (e.g. Information Services staff-centred workflows vs academic or client-centred workflows)

Hive allows for the creation of customised workflows attached to 'item types', i.e. items used for similar business purposes, such that the preparation of past exams for example can be made to follow a different workflow from course readings and a different workflow again from art images. For this reason, item types with similar workflows, drivers and clients, and their associated collections, need to be clumped into appropriate workflows, if the security permissions are to allow relevant staff and clients to access the appropriate parts of the system.

- The intended audience for the content (e.g. University-wide vs. specialist content vs publicly accessible content)

As for the submission audience, the intended end user audience also affects the desire to aggregate or disaggregate content. For example, content suitable for a University-wide audience could sensibly be grouped together while content for more specialised audience might be better managed separately.

Hive allows for cross-category and cross-bureau federated searching of collections; however access to collections along with resource sharing can be more effectively promoted by clumping similar collections together within the same bureau. Once again this has obvious implications for bureau design.

- The persistence of the content (archival/long-term vs. non-permanent/transitory content)

The design of the repository clearly needs to support the distinction between archival resources with an identified long-term value, and transitory resources that either lack economic value and/or persistence or are subject to a short life cycle.

- The integration needs of the presentation layer systems (e.g. administrative systems, teaching & learning systems such as Blackboard, and digital library collections accessible via Web interfaces or the Library Management System)

The needs of associated systems such as Learning@Griffith (using Blackboard) and the Library Management System (GEAC Advance) must be taken into consideration. For example, it is beneficial for collections required to be accessible through Blackboard building blocks to be co-located within the same bureau, thus supporting the deposit and retrieval of objects. Likewise, digital library collections might also best be co-located so content such as ePrints can be either pushed out via collection-specific web interfaces or pulled into the library catalogue.

- The level of control necessary over the content (e.g. digital rights management, copyright and/or licence agreements compliance)

In order to exert control over the management of collections comprised of objects subject to copyright, rights management or license obligations, it is essential to define these collections and control workflows. All course readings must be managed within a single collection, as must purchased or traded learning objects, and readings downloaded to the repository under publisher license.

- The level of security necessary for the content (e.g. public content vs. restricted content)

Hive cascades security permissions throughout a bureau in a top-down fashion, necessitating that only collections with similar security and audience requirements can be clustered as categories within a single bureau. This has obvious implications for bureau design.

With these considerations in mind, the following Hive structure is being implemented at Griffith University.

## Metadata and Controlled Vocabularies

The selection of a core metadata schema—let alone specialised schema—is particularly challenging because of the worldwide lack of consistency among digital repository initiatives. Options include Dublin Core, EdNA, IEEE LOM (Learning Object Metadata), and IMS (based on LOM). The national, DEST-funded ARROW (Australian Research Repositories Online to the World) Project has recently adopted IEEE's LOM (Institute of Electrical and Electronics Engineers' Learning Object Metadata) as its core schema (Harboe-Ree, Treloar & Sabto, 2004). The specificity of the Dublin Core was rejected as insufficient to meet the project's needs. However this isn't without its difficulties. Recent research into LOM adoption suggests that only a limited number of LOM elements are used consistently by various repositories (Friesen, 2004).

Even given the ability to adopt multiple schemas, there is also a growing realisation that the simple adoption of a standards-based schema isn't going to meet all of the metadata needs associated with the various collections identified as a priority within this report. A common challenge shared by universities, for example, is where to enter course details. To-date none of the major metadata schemas have defined a corresponding element.

Therefore the Digital Repository Team soon recognised a need to develop customised schemas or application profiles to support particular Griffith needs. An application profile is an assemblage of metadata elements selected from one or more metadata schemas and combined in a compound schema (CanCore, 2002). Application profiles provide the means to express principles of modularity and extensibility. The purpose of an application profile is to adapt or combine existing schemas into a package that is tailored to the functional requirements of a particular application, while retaining interoperability with the original base schemas. The Team has developed GU-MAP (Griffith University – Metadata Application Profile), which is based on IEEE LOM but also incorporates elements from Dublin Core.

Related to the need for developing applications profiles that are more relevant to local needs is the associated issue of establishing controlled vocabularies that are also locally relevant. A controlled vocabulary is an established list of standardised terminology for use in indexing and retrieval of information (Library and Archives Canada, 2004). An example of a controlled vocabulary is subject headings used to describe library resources. A controlled vocabulary ensures that a field will be described using the same preferred term each time it is indexed and this will make it easier to find all information about a specific topic during the search process. It will also enhance global interoperability for specified collections.

'Resource Type' is a particularly difficult area to create an adequate controlled vocabulary for. There are dozens of value lists in use. These include the lists found in LOM, Smart Learning Design Framework Project (SLDF), Dublin Core Metadata Initiative (DCMI), EdNA (Education Network Australia), GEM (Gateway to Educational Materials, US Dept. of Education), and National Learning Network (UK). There is not as yet a single internationally accepted value list for this field.

With this principle of utilising existing vocabularies, but modifying them to suit specific needs of the collection, we have created a 'General Resource Type' based on a subset of the DCMI 'Type vocabulary', for a broad grouping of item types. We also have a more specific 'Resource Type' that is based on a compilation of SLDF, LOM and other resource type vocabularies. The values in this list create a 'super-set' carefully selected to cover anticipated uses of the digital repository within Griffith University.

## **Initial Projects**

A number of projects have been selected on the basis of their ability to assist the Digital Repository Team to meet key objectives:

- Develop expertise in the Hive software
- Develop base metadata schema and then adapt as required for specialised content
- Develop database architecture to accommodate a range of differing content
- Develop appropriate workflows and business processes
- Identify policy, training and general implementation issues

It is important for the Team to learn to walk before it can run, and as such the projects are being implemented chronologically in order of increasing complexity. By breaking the wide-reaching project into a number of smaller, yet still sizeable projects, the Team is assured of a series of smaller successes. These—when combined—will assure the overall success of the project, including an effective rollout to clients which promotes efficient changes to work practices and positive client acceptance.

## **Previous Examination Papers**

Past exams are the simplest collection to integrate into the digital repository and as such were implemented first. Griffith's collection of 1800 past exams released by academics to students between 1998 and 2003 has been bulk loaded to Hive, thus being a suitable test for the bulk load procedure. No metadata schema was required for this collection, as all required information has been incorporated into the object title field to aid easy identification of objects by academics and students.

Whereas past exams were once accessible via a browsable Web page listing all available documents, they are now fully integrated into Learning@Griffith via a Blackboard building block that 'pushes' the appropriate exams into the 'Resources' section of each course within Learning@Griffith. There is no benefit in offering academics the functionality to manually select the appropriate exams for their courses.

Stage two of the past exams project is planned for 2005. Working with the Examinations and Timetabling (E&T) section of the University, which is responsible for the production and release of exams in conjunction with academics, the project aims to re-design the entire examination management process using Hive workflows. Within Hive, E&T in conjunction with academics will control the examination content creation, re-use, versioning, approval, security, and archival process. Finally, past exams selected for general release by academics will be released by E&T to students via Learning@Griffith using Hive.

## **Course Readings**

This is an area in which all universities have already produced complex workflows and staffing profiles to provide 'ereserve' systems of digital readings supporting both on-campus and online courses. Griffith currently has over 10,000 digitised readings available via the GriffLink library catalogue. However, unlike exams, digitised course readings inherently attract copyright issues. Under the Digital Agenda Act, there are strict guidelines as to how such material is to be communicated electronically. Therefore this project has provided the team with an opportunity to investigate Hive's Copyright Management Module.

The two major issues for Information Services at Griffith are that:

- a) Workflows that attempt to provide for the timely availability of course readings according to the academic calendar are inherently a very labour-intensive process requiring large staffing outlays over the year, especially during peak load periods prior to the commencement of each semester. Griffith simply does not have the staffing resources to support the largely manual processes required for document digitisation and release each semester. Therefore we are designing a more automated system that better leverages our online library collections and existing collection of digitised readings. By developing systems that allow academics to easily select online readings and automatically integrate them into their courses in Learning@Griffith themselves, academics are given greater control, and Information Services staff are given back more time to digitise readings as required. Good system design will ensure that an academic's task of selecting readings using the new system will be no more time consuming than their current task of requesting readings to be digitised.
- b) A very significant investment (over \$3.5 million p.a.) has been made in our digital library collections, and this is not currently being sufficiently leveraged in the course reading process. Currently the vast majority of course readings are requested from materials not available via Griffith's digital library. The process of digitising readings from print sources is a time consuming and costly process. By making the digital library more accessible and convenient in the course reading selection process, Griffith hopes to turn the tide on the digitisation process, by which a much greater proportion of readings will be selected from, and linked to library databases and ejournals. This issue begs pedagogical debate over the selection of the best readings to promote student learning, but there is no doubt the current system is not providing Griffith with the most efficient and timely solution.

To achieve resolution of these issues, a three-stage process is being developed for the selection and availability of course readings by academics using a Blackboard building block (or a product called Sentient DISCOVER currently being investigated) which links courses on Learning@Griffith to our library databases and ejournals using MetaLib/SFX, and also to our database of existing digitised course readings.

The system will require the building block to hold course readings information for each course for the current and next semester, thus allowing academics the opportunity to select readings for the next semester a month before the start of that semester. A month lead time also provides the Digitisation Team sufficient opportunity to digitise print course reading sources where they are really required. The course reading list for the current semester provides the template for the reading list for the same course for the next semester, thus preventing a requirement for re-keying of information, but allowing academics to change readings over time. Simple radio buttons will allow academics to turn readings on and off as required both across and within semesters.

The process will take academics through the following three-step workflow:

- Step 1 Using the 'Select Resources' building block the academic, from one month before the semester begins, will be able to search the Griffith collection of 400+ databases and 30,000+ ejournal titles using MetaLib/SFX to find and select appropriate ejournal articles, ebook chapters, etc. The building block will allow the academic to drag the persistent URL (All SFX URLs are persistent) into the Resources List within Learning@Griffith for their specific course. An unlimited number of readings are available and readings may be easily re-used each semester. Use of these readings is subject to publisher license agreements and so summary information about these agreements will be provided to academics at the point of need.
- Step 2 Academics who have not located appropriate readings may choose to progress to the next stage of the workflow which allows them to search and select readings from Griffith's collection of already digitised course readings. Upon locating a suitable known item using the metadata search provided, academics will be able to determine its availability for their course during the semester or part thereof. Academics will be required to book a reading for their course by stipulating the time period for which the reading is required. Should the reading not be already booked by another academic, then it may be used by the academic for their course. Should the reading be already booked by another academic, then the first academic can be provided with contact details of the other academic to negotiate a suitable compromise by which the reading is used for Course X for weeks 1-6 and Course Y for weeks 7-13, for example.

The competitive process for booking readings is also expected to encourage academics to get their course readings in order well before the start of semester, and thus overcome an age-old problem for libraries.

- Step 3 Should an academic require a course reading that is not available from either step 1 or 2 then they may choose to progress to step 3 of the course readings workflow. Using this process the academic may either copy or enter details of a journal article or book chapter into an online form requesting that the reading be digitised. Before a request is submitted, the details of the reading are automatically passed via the SFX URL Resolver once again to ensure the reading is not available online. The request is also run against the metadata database of existing readings, and those in use in the current semester. The academic is advised whether the reading is/is not copyright compliant at this point and time. This information allows the academic to make the decision whether they wish to progress the request even though the reading may not be made available during the current semester due to a copyright violation involving another reading from the same book or journal issue currently in use.

The building block required to select, book and request course readings will also form the basis of Griffith's copyright management and reporting system. Reports will be available on:

- Use of links to library databases and ejournals made under publisher license agreements, thus informing collection development as well as the improvement of MetaLib/SFX
- Use and availability of digitised readings under the Digital Agenda Act, which data can be made available to the Copyright Agency Limited as required
- Requests for digitised readings, which will assist workflow management

## **Art Images**

Griffith University's Queensland College of Art has an extensive collection of 70,000 art image slides, of which 55,000 have been fully catalogued using the Visual Resources

Association (VRA) metadata schema on the LIDA Filemaker Pro database. A small percentage of art images (2,000) already exist in digital format. However the LIDA database is not networked, thereby limiting potential cross-campus use.

Combing the resources of the Hive project and funds from a teaching and learning grant, the Art Images Project aims to:

- catalogue all images using reduced VRA, which is both affordable in terms of staffing requirements for cataloguing and effective in terms of resource discovery
- bulk load all metadata records to a new collection location within Hive
- digitise all 70,000 slides
- promote the purchase of new art images in digital format
- avail the image collections to all campus and remote locations
- provide for image rendering solutions for thumbnails and full images
- address policy issues concerning preferred file size for images (a modest 5 MB file size has been chosen), preferred suppliers and priorities for continued collection development
- integrate copyright management and CAL reporting systems
- provide a Web-based interface to art collections supporting resource discovery using VRA fields
- provide for practices, and policies that can be easily translated to the application of image collections which support other academic disciplines across the university such as nursing, medicine and architecture

Hive disk space is obviously a key issue for this and other image projects. 300+ GB of high availability disc space will be purchased initially to support the project. A Blackboard building block will allow academics to select art images to be included in resource lists for courses within Learning@Griffith.

## **ePrints**

Many universities and research institutes and bodies have already developed institutional research repositories using application specific software such as eprints.org, Fedora or D-Space software, for example (Sullivan, 2004). Some institutions are reporting difficulty in motivating academics to deposit their research for both the DEST HERDC research publications survey process and then again a second time to the institutional ePrint repository. Some institutions have also reported difficulty using software designed for self-deposit but which Library or Information Services staff are forced to perform the deposit procedures, and software that does not support the bulk-uploading of research documents.

Griffith aims to resolve these integration, client acceptance, and bulk upload problems by using Hive as the basis for our ePrint service. Research publications uploaded to PeopleSoft by academics for the purpose of the Department of Education, Science and Training (DEST) HERDC, i.e. Higher Education Research Data Collection, survey will be migrated in bulk to the ePrint repository on Hive. This will achieve a lower effort threshold for academics and a greater [mandatory] participation rate in the ePrint service.

Policies regarding the mandatory deposit of research publications to the Griffith ePrint service, much like those at QUT, will be developed. Bulk load systems between PeopleSoft and Hive will be developed. A retrospective project could ascertain the suitability for the ePrint service of publications deposited from 2001-2004 into PeopleSoft. This would require that each publication be examined for copyright ownership, and negotiation be undertaken with academics and other copyright owners such as journal and book publishers.

Upon completion of the design and integration of the workflows and policies, approximately 1000 research publications could be made available via the Griffith ePrint service per annum.

### **Teaching and Learning Objects**

Griffith academics as well as staff within the department of Flexible Learning and Access Services have made a considerable investment in the creation of learning objects currently stored in and accessed through Learning@Griffith.

The primary issues faced at Griffith are:

- **Economic value**  
With such a large number of learning objects in the Learning@Griffith system, and only a relatively low percentage of objects being used in any one year (83,000 out of 193,000 were touched in 2004) the differential economic value of different objects becomes apparent. As well as hiding (not linking to) objects not used in the current semester, many academics are deleting objects that may have been created at considerable cost to the University. The high number of objects and fluctuating semester workloads also makes it financially impossible for Information Services to fund the metadata description of all objects, thus requiring the identification of those priority objects to be managed within Hive
- **Sharing and reuse**  
Storing objects within the Blackboard course structure does not allow academics to view, share, and reuse objects created for one course, for the purposes of other courses, thus wasting opportunities for efficiency and achieving greater value for money. Sharing also encourages best practice by example, whereby academics can benchmark against other academics' and FLAS' work
- **Change management**  
Encouraging academics to share and reuse learning objects requires a combination of policy direction, technical feasibility and cultural change (Campbell, Blinco & Mason, 2004). Any system designed to encourage the sharing of objects is expected to confront cultural resistance. Systems that place a significant workload burden upon academics (such as metadata creation) only encourage academics to find the least path of resistance – there is no automated means to prevent academics from continuing to load objects to Blackboard or to prevent them from loading objects to faculty Webservers.

Griffith has elected to manage these issues by initially concentrating efforts on improving the management of learning objects applied initially to objects of high economic value only. Objects created by FLAS will be loaded to Hive, described using Griffith's version of IEEE LOM metadata standard, and stored in categories based upon the Australian Standard Research Classification headings. Academics will be able to search and browse objects via a building block, which will also allow them to link to selected objects from their courses.

Some objects created by academics are easily recognised as having high economic value based upon their file type, and these objects will be copied across to Hive to be available for sharing.

Griffith's short to medium term solution to the management of all learning objects is, at the end of each semester, to bundle the objects within each course in Learning@Griffith into SCORM packages which will then be bulk loaded to Hive for archiving for a period of

three years. Academics will have access to a building block that will allow them to retrieve a copy of the SCORM package for each semester instance of their course, which they can then unpackage to retrieve objects for reuse. This solution thus provides for archiving and reuse but demands little in the way of resources for object description, and is therefore cheap to implement, except for disk space requirements. It also provides for the University's legal obligations to archive course materials during the duration of students' time during a course. The success of this strategy however relies on the assumption that academics will remember in what course and semester/year combination the object they may be looking for resides.

## The Future

The projects described above are expected to fully occupy project staff throughout 2005. In addition to the technical implementation a significant amount of policy work will be required at all levels within the University, and this is expected to be at least as labour intensive as the system side of the project.

After this, resources will become available to investigate and implement other uses of Hive such as managing all learning objects, developing an information commons and eScience data applications, and possibly an Honours and Masters by Coursework digital theses service.

One of the keys to the sustainable adoption of the digital repository will be to ensure that the design of the Hive seamlessly integrates with existing practices or supports the adoption of new practices that improve or enhance information delivery, empower academics to manage their digital resource requirements and ensure that the University's investment in digital resources provides the maximum return.

## References

Borchert, M. (2004). *Critical issues in digital repositories*. [Online]. Available: <http://conferences.alia.org.au/seminars/camqld2004/martin.borchert.html> [22<sup>nd</sup> November 2004]

Borchert, M. & Richardson, J. (2004). *Integrating a digital repository*. [Online]. Available: <https://olt.qut.edu.au/udf/olt2004/> [22<sup>nd</sup> November 2004].

Campbell, L., Blinco, K. & Mason, J. (2004). *Repository management and implementation*. [Online]. Available: <http://www.imsglobal.org/altilab/altilab2004/Altilab04-repositories.pdf> [25<sup>th</sup> November 2004].

CanCore (2002). *CanCore FAQ*. [Online]. Available: <http://www.cancore.ca/faq.html#application%20profile> [26<sup>th</sup> November 2004]

Friesen, Norm. (2004). *International LOM Survey: Report*. [Online] Available: <http://dlist.sir.arizona.edu/archive/00000403/> [25<sup>th</sup> November 2004]

Harboe-Ree, C., Treloar, A., & Sabto, M. (2003). *ARROW: Australian Research Repositories Online to the World*. [Online]. Available: <http://eprint.monash.edu.au/archive/00000046/> [16<sup>th</sup> August 2004]

Library and Archives Canada (2004). *Thesauri and controlled vocabularies*. [Online]. Available: <http://www.collectionscanada.ca/8/4/r4-282-e.html> [26<sup>th</sup> November 2004]

Richardson, J. (2004). *Implementing HarvestRoad's Hive system at Griffith University: practice validates theory*. Available:

<http://conferences.alia.org.au/seminars/camqld2004/joanna.richardson.html>  
November 2004]

[22<sup>nd</sup>

Sullivan, S. (2004). *New models of research publishing*. [Online]. Available:  
<http://ausweb.scu.edu.au/aw04/papers/refereed/sullivan/paper.html> [25th November  
2004].