The CLIF Project: the repository as part of a content lifecycle

Richard Green, Chris Awre and Simon Waddington

Introduction

At the heart of meeting institutional needs for managing digital content is the need to understand the different activities that the content goes through, from planning and creation through to disposal or preservation. Digital content is created using a variety of authoring tools. Once created the content is often stored somewhere different, made accessible in possibly more than one way, altered as required, and then moved for deletion or preservation at an appropriate point. Different systems can be involved at different points: one of these may be a repository. To embed repositories in the content lifecycle, and prevent them becoming yet another content silo within the institution, they thus need to be integrated with other systems that support other parts of this lifecycle. In this way the content can be moved between systems as required, minimising the constraints of any one system.

The JISC-funded CLIF (Content Lifecycle Integration Framework) project, which concluded in March 2011, was a joint venture between Library and Learning Innovation (LLI) at the University of Hull and the Centre for e-Research (CeRch) at King’s College London. It undertook an extensive literature review [1] and worked with creators of digital content at the two host institutions to understand how they would like to deal with the interaction of the authoring, collaboration and delivery of materials using three systems used within Higher Education institutions that are targeted at the management of digital content from different perspectives and for different purposes: the Fedora Commons repository software [2], Microsoft SharePoint [3], and the virtual learning environment, Sakai [4]. Each of these systems addresses a range of lifecycle stages in the functionality provided; yet they were not designed to encompass the whole lifecycle. Armed with this background information, the project team went on to design and produce software that would allow the transfer of digital content between the systems to meet lifecycle requirements: Fedora and SharePoint, on the one hand, Fedora and Sakai on the other. The CLIF software has been designed to try and allow the maximum flexibility in how and when users can transfer material from one system to another, integrating the tools in such a way that they seem to be natural extensions of the basic systems. This open source software is available for others to investigate and work with.

This article draws on several of the pieces of documentation produced by the project.

Background

One might say that the CLIF Project was part of an ongoing journey of discovery at the University of Hull. Ariadne has kindly allowed us to describe two previous repository-related projects at Hull, RepoMMan [5] and REMAP [6]. The first of these, RepoMMan, undertook work to embed the repository in the natural workflow of digital content creators, whilst REMAP looked at how the repository might be made proactive in its own management and development. In developing the scope and position of the repository it became apparent that there was a risk of the repository becoming another institutional silo of digital content, potentially duplicating existing functionality in other systems rather than working in synergy with them. The CLIF Project thus sought to discover how one might best integrate the repository with other content management systems on campus. We were fortunate that colleagues at King's College, London, shared this area of interest and, thus, we undertook the work together. The fact that two campuses were involved gave us access to a wider range of users (academic and administrative) and use cases than either of us could have mustered on our own.

Given that a core output for the project was intended to be software that others could make use of to enable similar lifecycle management we were anxious to adhere to appropriate standards. Acknowledging the technical standards that could be used, a key standard for moving content between systems was how that content was structured. This is particularly the case in the use of Fedora, where, fortunately or unfortunately, the 'F' stands for 'flexible' and readers familiar with the Fedora Commons software will be aware that there are a huge number of ways in which one might construct a digital object around any particular content, some better than others. In parallel with the period of the CLIF project Hull has pursued this issue as one of the founding partners in the Hydra Project [7]. This project is an international collaboration with the University of Virginia and Stanford University to develop a framework for building flexible repository solutions over Fedora, and the work on this has provided guidelines as to how one might sensibly build Fedora digital objects in a way others can share. The term 'Hydra-compliant' is now being applied to digital objects, and referenced in papers and presentations about Fedora repositories. It was only logical that CLIF would take advantage of this and it is Hydra-compliant objects that the CLIF software builds.

Literature review

The first element of the project was a literature review [8]: a piece of cross-disciplinary desk research in liaison with our contributing staff colleagues for their subject and role-related input. This resulted in a document that does not aim to produce or summarise the many different examples of lifecycle in existence, and there are many, but instead addresses the issues that have emerged through developing or examining such lifecycles, particularly focusing on the, limited, literature examining lifecycles across different systems.

In many ways, the literature review highlighted a range of different views, opinions and approaches to dealing with digital content lifecycles. These came from different perspectives and starting points, covering both research and learning teaching, plus others related to different content types, which seemed to emphasise that wherever you were coming from consideration and management of digital content lifecycles is important. It is perhaps surprising that literature on specific system aspects of this management approach was not found, but this may be due to the flux in technology adoption and the rapid pace of change. Nevertheless, the technology involved must be taken into account when implementing a digital content lifecycle management approach as it is core to the day-to-day running of this. Consideration of the appropriateness of different systems is important to inform this, and this can be extended to consideration of appropriate systems at different stages of the lifecycle, the starting point of CLIF.

There were a number of specific points emerging from the literature that were taken into the CLIF project and readers who are interested in this area are encouraged to read our review and, in particular, its concluding section.

One thing the literature review did not attempt was to determine where responsibility lies for overseeing or managing digital content lifecycles. Any implied assumptions that can be read into the literature reviewed, for example about the role of a library, are coincidental. It is acknowledged, though, that as digital content lifecycles become better understood in terms of their design and technical implementation there will be a need to place this in the context of how the lifecycles are managed organisationally and see the lifecycles put into practice.

Case studies

The next part of our work was to gather user requirements [9] from colleagues at our two institutions. There was a recognised difference of emphasis: colleagues at King's College were particularly interested in the requirements for research data whilst at Hull the emphasis was more on text-based materials. The interviewees were from a range of backgrounds, both learning and teaching, research and administrative roles, and were chosen to cover a wide range of possibilities. The interviews sought to discover how people dealt with digital content and what kinds of software were used to manage it. Further, our colleagues were asked to speculate on how they thought a repository might be used to enhance their work around the two identified target systems, Sakai and SharePoint. From all this information two generic use cases, one for experimental data and documentation, the other for essentially textual material, were derived for CLIF to address in terms of providing supporting functionality.

Technical review

Fedora, SharePoint and Sakai all provide a rich and complex set of functionalities; thus it was that the software design and development work was preceded by a technical review [10]. This examined the functionality offered by the three applications and considered how this might be used to support the scenarios identified in the case study work. In particular, the team considered where in the content lifecycle interaction between the systems might usefully take place; after due consideration the team decided to take a somewhat agnostic approach to this question so as to be flexible in addressing the needs in the use cases.

A further major element of the review considered how the software integration might best be carried out. Integrations between systems can be carried out using point-to-point techniques according to specific need. Whilst a loosely coupled approach to point-to-point can enable wider adoption of a solution, such solutions can also be limited by the systems themselves as they change over time. Enterprise Service Buses (ESBs) are an approach to abstract out the ways that systems can communicate with each other, protecting integrations against software changes. The project team was aware that it needed to assess the relative merits of the two approaches in the context of the two partner institutions and also in the context of possible wider adoption of its technical outputs. In the event, whilst the potential of ESBs was clear, the practicalities of the necessary systems integration within the system environments of the partner sites led to the conclusion that implementing an ESB for the purposes of the CLIF project was not realistic.   That decided, a basic, and as far as possible open, point-to-point architecture for the integrations was drawn up.

Software development

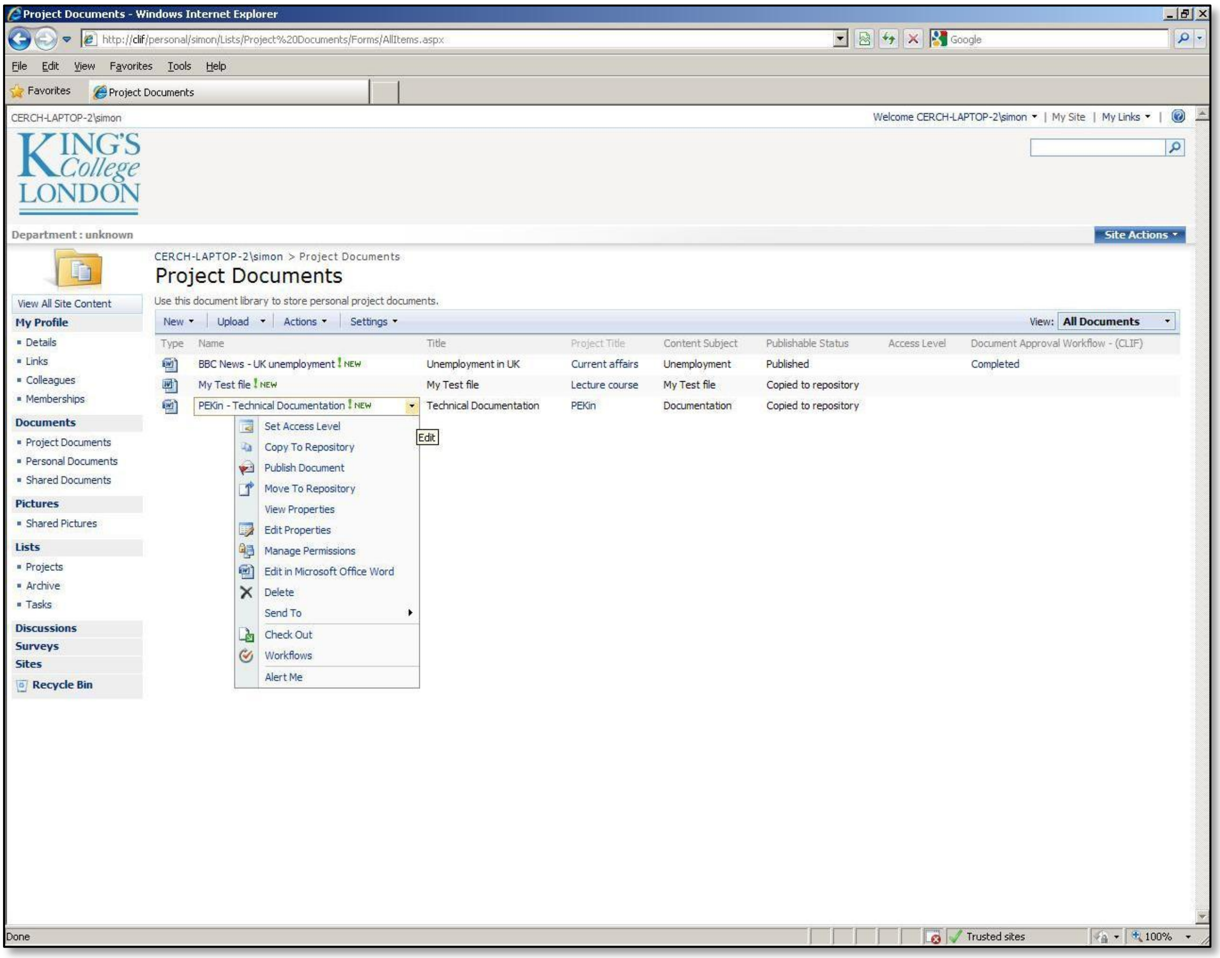
Informed by the technical review, work commenced to develop code that would enable the two integrations: SharePoint with Fedora and Sakai with Fedora. It was decided that the SharePoint work would be based at King's and that the Sakai work should be concentrated in Hull. This is not to say that the two parts of the project proceeded in isolation. There was healthy dialogue between the developers and at times each contributed to the other's code as well as serving as the first port of call for testing. Although rather different in nature the two integrations even share a certain amount of code where that seemed to be appropriate. The detail of the development work, the outcomes and the installation procedures are documented at length in the technical appendix to the CLIF Project's Final Report [11].

SharePoint-Fedora integration

CLIF extends the functionality of SharePoint's MySite. Although MySite was used as the basis for development, the CLIF work could easily be adapted for use for document archiving in other site templates. When a new MySite user is added by an administrator, the CLIF system automatically creates a Fedora account for the user as well as creating a Fedora object as the basis for the user's private repository area under the MySite root object. Within MySite, users can have access to both the existing functionality as well as the additional features provided by CLIF. As part of the document upload process, a certain amount of general metadata is gathered, which can then be appended to the Fedora object that is deposited to the repository.

Users can 'move' a document to the associated repository: this creates a Fedora object in their private archive area of the repository and deletes the SharePoint instance of it. The associated metadata is retained and is re-associated with the object should the file be brought back from archive. This is effectively a short- or medium-term preservation strategy.

Two versions of depositing a copy of a document to the repository are provided (which option(s) are provided to the user is configurable by an administrator). 'Publishing' a document starts a SharePoint workflow which needs to be completed in order for the content to reach the general (public facing) repository (such workflow may require, for instance, approval steps). The list of locations to which a user is able to publish within the general repository can be configured at the top level by an administrator and is presented to the user in a pull-down list on a web form. The second option provided, 'Copy to Repository', takes the object created and places it in a specific place within the repository for further processing by others (this is effective as an accession queue containing materials to be dealt with by repository managers). This repository location typically has restricted access. This second process provides the user with the opportunity to provide significant metadata about the content they are publishing, appropriate to objects being exposed in an institutional repository; this is MODS metadata by default but can be Dublin Core or both. Where possible, default entries in the metadata fields are derived from the user's SharePoint environment.



CERCH-LAPTOP-2\simon > Project Documents

Project Documents

Use this document library to store personal project documents.

Type	Name	Title	Project Title	Content Subject	Publishable Status	Access Level	Document Approval Workflow - (CLIF)
	BBC News - UK unemployment ! NEW	Unemployment in UK	Current affairs	Unemployment	Published		Completed
	My Test file ! NEW	My Test file	Lecture course	My Test file	Copied to repository		
	PEKin - Technical Documentation ! NEW	Technical Documentation	PEKin	Documentation	Copied to repository		

SharePoint screen showing a pop-up menu with options for depositing item to repository

Deposit of multiple documents to the repository, either copy or move, is provided by an additional feature that enables the user to select multiple documents from a document library. This could be used for instance when the user has completed a project and wishes to archive a large number of files.

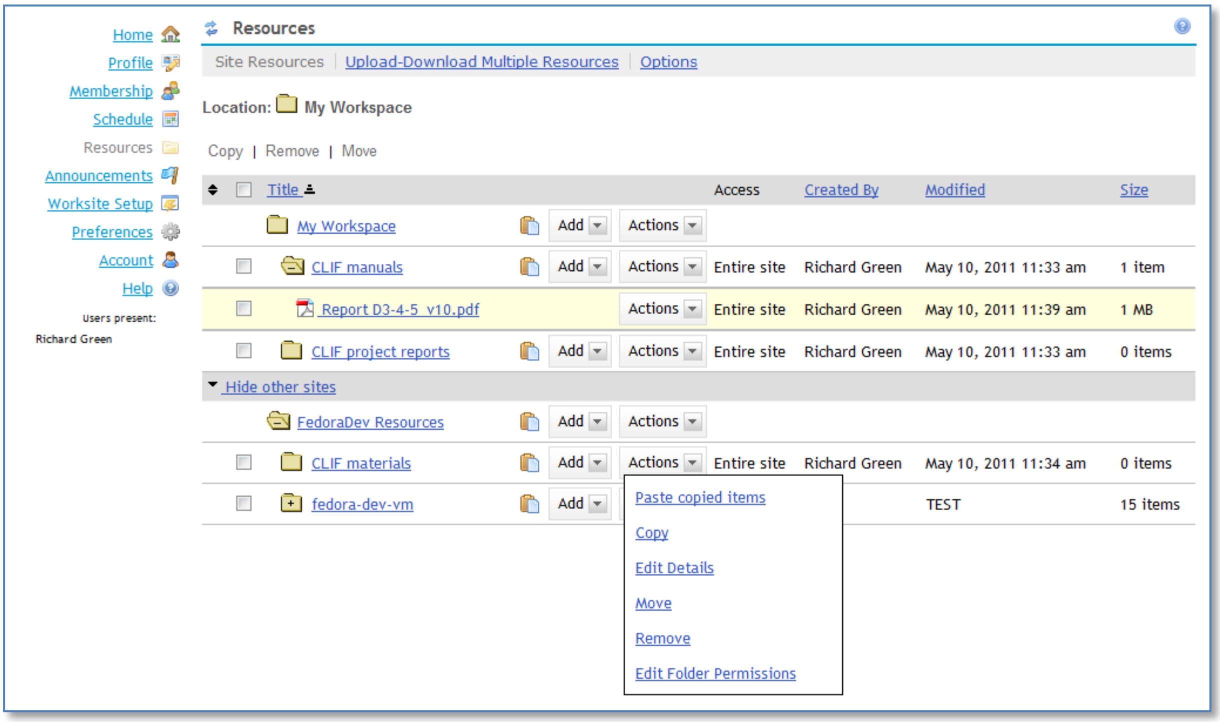
Documents that have been 'moved' from SharePoint to the repository can be retrieved by navigating to the Archive list and selecting the URL of the document. This retrieves the document from the repository to the local file system.

A repository browse functionality is provided that enables the user to browse their private folders in the repository as well as the public repository folders.

A text box on the MySite pages enables the user to enter search queries. Search can be performed across metadata and text of all documents in MySite as well as the metadata of documents that have been moved to the repository (since the metadata is retained in SharePoint). A URL is provided in the search results that retrieves the document to the local file system. Due to time limitations, it was not possible to include free text searching across Fedora, although this could be added using Solr indexing.

Sakai-Fedora integration

The CLIF work also integrates Fedora into the Sakai resources tool. This is an area of Sakai in which users can store digital materials for their own use and, potentially, share them with other users of Sakai. The system allows the creation of a tree structure to aid organisation of the materials held and provides a range of functions to manage them: upload, copy, edit, move, delete, and so on.



Location: My Workspace

Title	Access	Created By	Modified	Size
My Workspace				
CLIF manuals	Entire site	Richard Green	May 10, 2011 11:33 am	1 item
Report D3-4-5_v10.pdf	Entire site	Richard Green	May 10, 2011 11:39 am	1 MB
CLIF project reports	Entire site	Richard Green	May 10, 2011 11:33 am	0 items
Hide other sites				
FedoraDev Resources				
CLIF materials	Entire site	Richard Green	May 10, 2011 11:34 am	0 items
fedora-dev-vm			TEST	15 items

A Sakai resources screen showing the 'paste' stage of a copy into a repository

A configuration setting determines a point in the repository structure below which it will be visible to Sakai users. CLIF then makes this linked Fedora repository area appear as a resources tree with the full range of management functionality allowing movement of content between Sakai folders and repository folders. Operations to transfer content both to and from the repository can be performed on single files or digital objects, folders or indeed trees. By transferring materials into repository folders they are potentially shared outside the Sakai environment and may be in a better location for medium- or long-term preservation.

Unlike the SharePoint integration, the Sakai integration does not have the capability to deal with rich metadata. A certain amount of metadata can be associated with content in Sakai and, where this is available, it is taken with that content into digital objects in the repository. Unfortunately when digital objects are moved in the opposite direction Sakai does not currently provide functionality to insert such metadata into the application environment.

The CLIF code produced within the project assumes that a Sakai user has full read-write capablility within the area of the repository that they can navigate. A somewhat different approach may be needed in a production system and this is discussed further in the section 'Post-CLIF' below.

One might argue that this section on Sakai-Fedora integration is rather short. In a presentation we gave recently we showed a photograph:

© 2008 Richard Green

We liken what we have done to this image of swans: it is all calm and serene on the surface but you can't see what might be going on underneath. We could simply say that we have enabled 'copy and paste' between Sakai and Fedora. Put like that, the functionality seems almost trivial but the comment would maybe overlook our use of standards wherever feasible and the implicit complexity in getting systems that manage content to work in ways they were not originally scoped to do.

To our users the CLIF work has greatly simplified what would otherwise have been a complex process and thus, we hope, it will encourage them to exploit the exchange between the two systems to capitalise on the features offered by each.

Evaluation

CLIF's use cases were identified through interviews with stakeholders. Following the integration work carried out these use cases were re-tested with as many of the original stakeholders as could be contacted. As noted earlier, these were broadly related to teaching, research and administration.

For both SharePoint and Sakai a number of common observations emerged from the evaluation interviews carried out:

- There needs to be a clear understanding and view about where the boundaries are between the different systems being used, to avoid confusion
- There needs to be clarity over why different systems are being used, to overcome concerns about having to work with multiple systems
- There is a need for better preservation and a recognition that integrating the repository could support this, but also a need to be clear about what needs preserving

- There is benefit in being able to access other content stores from within your current working environment in order to see what is available more broadly

The evaluation highlighted that the users were open to combining systems in their working environment, but needed to be re-assured that there was good reason for this and what the different lifecycle stages being enabled through the integration of systems were.

Post-CLIF

Post-project development at Hull will produce a version of the Sakai integration code where users can deposit (write) materials only into a specified part of the repository structure that Sakai exposes, and will be able to browse (read-only) solely those parts of the repository permitted by their level of authorisation. All materials deposited into Hull's institutional repository go through a quality assurance (QA) process and this writable area within the Sakai tree will correspond to the quality assurance 'queue' for learning and teaching materials. Given the limited functionality with respect to metadata being passed into the repository with Sakai content, this QA stage will be a useful opportunity for repository staff to enrich the descriptive metadata ultimately available to end users.

Conclusions

The work undertaken on the CLIF Project leads us to four major conclusions:

- The management of digital content lifecycles has been extensively explored in the literature, from many different perspectives and in many different subject and content domains. The majority of these explorations focus on the processes involved in managing the different steps of the lifecycle, and whilst there is variation there is also a great deal of consensus in the descriptions of digital content lifecycles. This project has not sought to replicate this work or add to the variations in existence, rather to focus on the implementation of the digital content lifecycle across multiple systems. This practical aspect of how a digital content lifecycle can be put into practice is far less explored in the literature. This may be because technologies change and consistency in process is more important that focusing on specific systems; it may be that different domains put their findings into practice using technology designed for that domain, and do not have an identified need to move out of that domain. The literature suggests both. CLIF challenges in particular this latter position by recognising that different systems used to manage digital content within a University do not have to work in isolation, but can be used together.

- The technical integration work carried out has successfully demonstrated that diverse content management systems can be brought together to allow the seamless movement of content between them. Having identified a set of use cases from interviews with local users, we were nevertheless keen to ensure that implementing these use cases did not preclude other uses for the movement of content between the systems, and implemented them in as generic a way as possible. This has resulted in a flexible set of outputs that can be further developed and applied. Our evaluations revealed additional functionality and use cases that could be implemented, and we anticipate further use cases emerging as we implement the project's outputs more widely and more users become familiar with what is feasible.

- The work required to carry out the integration has been extensive and detailed, and it can also be concluded that the lack of the most up-to-date standards in the interfaces for content management presented by both Sakai and SharePoint (e.g., CMIS [12]) does not make the task of getting such systems to work together any easier. It is concluded from this experience that all content management systems should be encouraged to make it as easy to get content out as it is to get content into them in order to facilitate seamless flow and enable the digital content lifecycle across systems.

- An assumption at the start of the project was that we would be agnostic about the direction in which content might flow between the systems once integrated. Evaluation feedback clearly suggests that the repository's archival capability is regarded as one of its strongest assets, and the area that the other systems could not offer comparable functionality on. Hence, the primary flow of content is into the repository. This suggests that the role of the repository within a University will be regarded very much in terms of what it can offer that the other systems cannot, rather than try and compete on all levels. Whilst there is clear benefit in playing to one's strengths there is a challenge to clarify better at an institutional level what functionality is offered by different content management systems, so as to more fully understand how different stages of the digital content lifecycle can be best enabled.

Code

The CLIF code is available on two github sites:

https://github.com/uohull/clif-sakai

https://github.com/uohull/clif-sharepoint

Configuration and installation instructions are to be found in the technical appendix to the CLIF final report. [13]

Acknowledgements

We gratefully acknowledge funding for the CLIF Project from the 'Repositories Enhancement' strand of the Information Environment initiative of the Joint Information Systems Committee (JISC).

We are also grateful to colleagues, academic and administrative, at both the University of Hull and King's College London. Their time in helping us to formulate case studies and, subsequently, to evaluate the project's work against those starting points was much appreciated.

Finally, we acknowledge assistance from the University of the Highlands and the Islands (UHI). Ours was not the first attempt to integrate Sakai and Fedora, colleagues at UHI tried it some years ago as part of the JISC CTREP project. That they were only partially successful was entirely due to

restrictions of the (then) Sakai code. Their work provided a very helpful starting point for the CLIF team and their time in helping us understand it was invaluable.

References:

[1] CLIF literature review: https://edocs.hull.ac.uk/splash.jsp?parentId=hull:1647%26pid=hull:2430

[2] Fedora Commons: http://fedora-commons.org

[3] Microsoft SharePoint: http://sharepoint.microsoft.com

[4] Sakai: http://sakaiproject.org

[5] RepoMMan: http://www.ariadne.ac.uk/issue54/green-awre/

[6] REMAP: http://www.ariadne.ac.uk/issue59/green-awre/

[7] The Hydra Project: http://projecthydra.org

[8] CLIF literature review (*op cit*)

[9] CLIF use case summary: https://edocs.hull.ac.uk/splash.jsp?parentId=hull:1647%26pid=hull:2431

[10] CLIF technical design: https://edocs.hull.ac.uk/splash.jsp?parentId=hull:1647%26pid=hull:2697

[11] CLIF final report: https://edocs.hull.ac.uk/splash.jsp?parentId=hull:1647%26pid=hull:4194

[12] Content Management Interoperability Services (CMIS): http://www.oasis-open.org/committees/cmis/

[13] CLIF final report (*op cit*)

All links tested at 22nd June 2011.

Author Details

Chris Awre
Head of Information Management
Library and Learning Innovation
Brynmor Jones Library
University of Hull
Hull
HU6 7RX

Email: c.awre@hull.ac.uk

Richard Green
Manager
Hydrangea in Hull and Hydra (Hull) Projects
c/o Library and Learning Innovation
Brynmor Jones Library
University of Hull
Hull
HU6 7RX

Email: r.green@hull.ac.uk
Websites: http://hydrangeainhull.wordpress.com/
http://projecthydra.org

Richard Green is an independent consultant working with the Library and Learning Innovation at Hull.

Simon Waddington
Software Development Manager
Centre for e-Research
King's College London
Strand Campus
26-29 Drury Lane
London
WC2B 5RL

Email: simon.waddington@kcl.ac.uk