

Cover Sheet for Proposals*(All sections must be completed)*Name of Capital Programme: **Repositories and Preservation Programme****Bid for Calls** : (Please tick ONE BOX ONLY, as appropriate)**Discovery to Delivery and Interoperability Demonstrators (Strand C)**

Call I – Interoperability Demonstrators

 a) Interoperability demonstrators**Repository Start-Up and Enhancement (Strand D)**

Call II – Repository Start-Up and Enhancement Projects

 a) Repository start-up projects b) Repository enhancement projects c) Rapid innovation projects: enhancing repository content**Digital Preservation Across the Lifecycle (Strand H)**

Call III – Digital Preservation Across the Lifecycle

 a) Digital preservation across the lifecycle

Name of Lead Institution: University of Wolverhampton (UW)

Name of Proposed Project: AIR – Automated Archiving for an Institutional Repository

Name(s) of Project Partner(s): BIOMED Central

Full Contact Details for Primary Contact:**Name:** Prof. Dr. Ruslan Mitkov**Position:** Professor of Computational Linguistics and Language Engineering, and Director of the Research Institute of Information and Language Processing**Email:** R.Mitkov@wlv.ac.uk**Address:** Research Institute of Information and Language Processing, University of Wolverhampton, Stafford St., Wolverhampton. WV1 1SB**Tel No:** 01902 322217**Fax No:** 01902 323543

Length of Project: 18 months

Project Start Date: September 2007 Project End Date: March 2009

Total Funding Requested from JISC: £99,664

Funding Broken Down over Financial Years (Mar – Apr):

Sep07 - Mar08

£44,759

Apr08 – Mar09

£54,905

Total Institutional Contributions: £99,298

Outline Project Description

Manual deposition in institutional repositories –obtaining citation data, encoding them in terms of specific metadata and verifying the copyright status of the publication – is an extremely time- and resource-intensive

process. These costs act as a bottleneck on the fast uptake of large institutional repositories. This challenge has long been recognised and a number of research projects have attempted to develop the technology for unifying disjointed repositories, their efficient management and re-use (e.g., the Bibster project). Very recently, similar industrial products have appeared, such as Symplectic's Publications Management System. Nonetheless, these technologies still fail to address the main problem of acquisition of bibliographic repositories - the discovery of citation data in large text collections, that potentially appear in non-restricted citation formats (such as on home pages of academic staff or in various institutional reports).

The aim of this project is to develop a computer-aided information extraction system that will allow for speedy discovery and extraction of bibliographical data on an institutional website. The project will investigate the degree to which the population of institutional repositories can be automated, in order to maximise the speed of human-supervised compilation of the data, while maintaining its high quality.

To further this aim, the researchers at Research Institute of Information and Language Processing (ILP) of the University of Wolverhampton, in cooperation with the staff of the University's Learning Information Services (LIS) will design a software architecture that helps a user to:

- locate relevant documents on the institutional website,
- extract bibliographical entries from them,
- extract information from each entry and tag it with Dublin Core metadata tags such as Author, Title, and Year,
- export the extracted data into Open Repository or DSpace workflows, and
- facilitate checking of copyright issues using the SHERPA Romeo database.

The system will be integrated with the WIRE ("Wolverhampton Intellectual Repository and E-theses") repository, but it will be designed in a way to facilitate easy adoption of the software by other institutions that use different data encoding standards.

Throughout the project the ILP researchers will work closely with LIS staff who currently run WIRE for the University of Wolverhampton. In addition the project will liaise with Biomed Central, who supply the hosted DSpace-based Open Repository system on which WIRE runs, and the repository community through established links with SHERPA, UKCORR and, if possible, the JISC Repository Support Project run by SHERPA.

The project is expected to bring considerable benefits to the University. Specifically, the project will:

- stimulate significant growth in content in WIRE,
- facilitate the embedding of WIRE within the research process at the University,
- raise the profile of the University research by increasing the likelihood of citation,
- provide opportunities for ILP researchers to gain experience in knowledge transfer,
- free up LIS staff time by introducing mediated deposition process,
- develop the relationship between LIS and RIILP which may lead to further research co-operation on advanced information access technologies.

I have looked at the example FOI form at Appendix A and included an FOI form in the attached bid (Tick Box)	YES ✓	NO
I have read the Circular and associated Terms and Conditions of Grant at Appendix B (Tick Box)	YES ✓	NO

Automated Archiving for Institutional Repositories

C. Appropriateness and Fit to Programme Objectives and Overall Value to the JISC Community

The continued development of OAI-PMH compliant institutional repositories is of vital importance to British HEIs, research and the national economy¹. Many institutions that have implemented repositories have encountered cultural barriers to embedding repositories and challenges surrounding deposit. At the University of Wolverhampton we have successfully established WIRE (Wolverhampton Intellectual Repository and E-theses) open access institutional repository for research output (<http://wlv.openrepository.com/wlv/>). Launched just six months ago this is already populated with over 500² articles and theses (no small achievement for a post' 92 HEI). However, we are acutely aware both from our own experience and through talking to others that we are still well short of achieving comprehensive archiving (or as close as we can get to this while remaining within the constraints of copyright, licencing and embargo restrictions) or a satisfactory level of self archiving by academic staff. Common difficulties facing many repositories include the identification and capture of their institution's research output, time spent providing mediated deposition services or resistance to, and/or lack of engagement with, self archiving and copyright/licence clearing of material for open access.

This overall aim of the proposed project is to address these common issues by:

- Employing state-of-the-art information retrieval technologies to develop an information extraction system to locate and retrieve research outputs and data about research outputs from university web sites to support repository librarians efforts to populate repositories and to support the creation and maintenance of an institutional publication database.
- Adapting existing software to electronically draft Dublin Core metadata tags from research papers that can then be checked by the author and/or repository librarian and exported into to DSpace or e-print workflows.
- Investigating automated checking of copyright/licence issues using the SHERPA Romeo³ database within, or prior to, the repository software workflow to speed deposition and access.

This would facilitate the adoption of a more managed approach to university's intellectual assets and outputs and offers the potential to significantly improve and simplify the management of research repositories. If technically successful and effectively marketed to academic staff these developments would contribute a step change to overcoming cultural barriers and resistance to self archiving and should make a significant contribution to ensuring the population of repositories.

Addressing these issues would support the JISC's aims to develop:

- the effectiveness of scholarly communication and digital resources in support of research, learning and teaching, especially through sustainable content management.
- a common information and communications environment, including addressing intellectual property rights, interoperability standards and appropriate technology

This project brings together a multi-disciplinary team combining the knowledge and creative talents of information retrieval researchers with the vision and practical experience of library staff and managers who are familiar with the challenges of running an institutional repository and engaging academic staff in Open Access. The proposed project team have substantial experience of theoretical and applied research, provision of research support, provision of information services, and the management and administration of an institutional repository. LIS staff in the project team would liaise with the wider repository community to inform others of the developments produced by the project.

In the past few years, in response to the fact that manual creation of tagged bibliographical data is an extremely resource- and time-intensive process, a number of research projects have been carried out

¹ http://www.jisc.ac.uk/news/stories/2007/06/news_repos.aspx

² Not all of these are yet available in full text open access because of copyright/licence/embargo issues.

³ <http://www.sherpa.ac.uk/>

that aim to develop the technology to unify existing disjointed bibliographical repositories (e.g., the Bibster project⁴). Very recently, industrial products enabling efficient management of bibliographical data have appeared. For example, Symplectic's Publications Management System⁵ automatically harvests journal citation data from a broad variety of pre-specified institutions such as CiteSeer, ACM Portal, or BioMed Central. The free-of-charge CiteULike system⁶ aims to do the same by re-using the efforts of a growing community of its individual users. Nonetheless, these systems can only make use of data that has already been encoded in the format with which they have been adapted to deal. As a result, the coverage of such systems is greatly limited. Symplectic's PMS, for example, processes only journal papers of specified publishers, missing out such publications as conference proceedings, books and book chapters. To our knowledge, the only systems that actually extract bibliographical data from the unrestricted Web are CiteSeer⁷ and Google Scholar⁸.

The project will support open standards and the common implementation of these standards to enable future integration and interoperability with both DSpace and Eprints open source software (the two most heavily used systems for research paper repositories) and working to common JISC standards to ensure interoperability with other emerging repository systems and developments.

By testing the partial automation of the population of repositories the project should provide a way forward to reduce the time and effort involved in running and administering repositories and reduce the cost of establishing and maintaining repositories.

The repository developments described above have a wide potential impact. By improving archiving in institutional repositories they will support knowledge transfer. By exploiting university websites they offer the chance to improve integration with other research information systems, but the most significant change will be by easing the time and effort required by academics and/or repository staff to populate/deposit material in the repository.

D. Project Workplan

The activities of the project are organised into seven workpackages (WPs). The chronological relationships between them are shown in the Gantt charts (see Annex 1). Each workpackage is to produce its own deliverables, which will be used as milestones to measure the progress of the project. The following personnel will be employed on the project:

Project Manager (Prof. Dr. Ruslan Mitkov, RM)
ILP Research Associate 1 (Dr. Viktor Pekar, RA1)
ILP Research Associate 2 (RA2)
LIS Team Leader (Frances Hall, FH)
LIS Change Manager (Dr. John Rule, JR)
LIS Community Liaison (John Dowd, JD)
LIS Repository Librarian (RL)

WP1. Project management and quality assurance.

The project will be managed by Prof. Dr. Ruslan Mitkov, who has considerable relevant experience not least his work on the ESRC-funded BiRD ("Building Research Databases") project, which has many overlaps with the current proposal. The aim of the BiRD project was to develop a system to monitor the Web for announcements of resources available for the academic community, extracting information from it, and creating a publicly accessible database on these resources.

In addition, quality assurance on the project will be achieved with the help of a Project Steering Group, that will include John Dowd, John Rule and Frances Hall from Learning Information Services, and Viktor Pekar from ILP. All three have experience of managing library projects and John Rule is

⁴ <http://bibster.semanticweb.org/>

⁵ <http://www.symplectic.co.uk/products/publications.html>

⁶ <http://www.citeulike.org/>

⁷ <http://citeseer.ist.psu.edu/>

⁸ <http://scholar.google.com/>

currently project manager for the WIRE project, which will be the major immediate beneficiary of AIR. Viktor Pekar previously worked as a research associate on the BiRD project. The function of the Steering Group, who will hold bi-monthly meetings, will be to monitor the progress of the project, the quality of its deliverables, and take decisions on alternative strategies. The Steering Group will adopt a quality review as a methodology to assure quality of the workpackages, which will be reported to the Project Manager.

Who: RM, RA1, FH, JR, JD

Duration: 1-18

Deliverables: n/a

WP2. User requirements

The objective of WP2 is to define the requirements that the automated archiving system should satisfy. The system requirements will be based on the needs of its potential users and other stakeholders, and will help determine specific inputs for all the subsequent work packages of the project. In the first phase of the project, colleagues from LIS will be involved in the definition of the major requirements that will guide the development of the AIR system in an application perspective. The WP will seek to define the following groups of user requirements: scope of application (the types of data the system will be operating on), what data/metadata should be extracted, quality requirements (response time, precision of extracted information, degree of integration with other services), the presentation requirements (how the user should interact with the system).

At the end of WP2, a set of key performance indicators will be identified, and for each of them a metric and a threshold/range that identifies the limit of usability of the service will be defined.

Who: RA1, RA2, RL, FH, JR

Duration: 1-3

Deliverables: a report on the user requirements

WP3. Design of the system architecture

WP3 will model and design a web-based automated archiving framework. The framework will consist of three major interacting services – a web crawler, an information extraction component, and the DSpace interfacing component. To ensure flexibility of the system and its easy portability across new application domains, the framework will leverage open data-encoding standards. The AIR infrastructure will be designed in a top-down manner, starting with high-level components (e.g., crawler, DSpace interface), through mid-level ones (e.g., named entity recognition, cascading finite state transducer for information extraction) and going to low-level functionalities (e.g., page parsing, text pre-processing).

Who: RA1, RA2

Duration: 1-3

Deliverables: n/a

WP4. System components

The goal of WP4 is to develop individual high-level components of AIR:

1. Web crawler. To obtain relevant documents from the institutional web-site, the AIR web crawler will aim to discover pages that contain lists of publications for individual members of staff, either on their personal home pages or on lists of publications of a research group, institute, department, or school.
2. Information extraction. The IE component will first extract citation entries from the web pages located by the crawler and then locate strings in them that correspond to DSpace metadata tags, such as AUTHORS, YEAR, and TITLE.
3. Interface with DSpace metadata tagging workflow. The data extracted by the IE component will be passed on to the DSpace interface component, which will ensure that the extracted data are accurately recorded in the DSpace repository. The applicants have discussed this with Biomed

Central, the developers of the Open Repository software, and identified a number of possibilities to achieve efficient integration with DSpace (the underlying software behind Open Repository). The Biomed Central team has agreed to provide help with integrating the system using the DSpace Lightweight Network Interface⁹.

All the system components will be designed in such a way as to enable easy correction of the generated entries by a librarian or author. For example, if the person supervising the output of the system notices, for example, that bibliographical data has been extracted from someone else's web page, then s/he will be able to easily correct the original URL and re-run the system on the new web page. Similarly, the system will allow the supervisor to view the section from the web page from which the entry was extracted, so that they can introduce corrections to the entries.

Who: RA1, RA2, FH, JD, RL

Duration: 4-14

Deliverables: web crawler software, information extraction software

WP5. System Integration

The goal of this workpackage is to integrate the developed components in a working infrastructure. In this phase of the project, issues related to the interoperability between components will be investigated, such as communication efficiency, security policy, and workflow control. There will also be investigation of integration with the SHERPA Romeo database to ease copyright/licence/ embargo checking. The WP will develop several prototypes of the AIR system that will be evaluated in the subsequent evaluation WP.

Who: RA1, RA2

Duration: 10-16

Deliverables: integrated system

WP6. Evaluation

WP6 will be responsible for the evaluation of the AIR system and its components at various stages of the project. It will develop an evaluation dataset, consisting of sample web pages and bibliographical entries manually constructed from them, against which the accuracy of the web crawler and the IE module will be evaluated. The integrated system will also be evaluated in experiments with repository librarians, whereby the time and the amount of edits required for creating database entries will be measured and compared.

Who: RA1, RA2, RL, JR, JD, FH

Duration: 5-18

Deliverables: evaluation datasets

WP7. Dissemination.

The goal of this workpackage is to present the outcomes of the project to the relevant scientific and industrial communities. The major dissemination route will be the project web site, which will contain the software available for download as well as project publications (progress reports, presentations, papers). To ensure visibility of the website to potential users it will be advertised on relevant meetings, conferences, exhibitions, mailing lists, and submitted to major search engines; it will also comply with W3C guidelines in order to maximise its accessibility to wider public.

The project outcomes will additionally be disseminated at various conferences, seminars, exhibitions, and demo sessions at relevant conferences on library science and text mining. Specifically, the following forums will be targeted: JISC programme meetings, SHERPA meetings, UKCORR (UK Council of Research Repositories) meetings and JISC Repository Support project meetings.

Who: RA1, RA2, RL, JR, JD, FH, RM

Duration: 5-18

⁹ <http://web.mit.edu/lcs/www/lni/>

Deliverables: publications, project reports disseminated through the WIRE IR.

Intellectual Property Rights

The AIR participants will adopt the following position with respect to intellectual property (IP), generated by the project:

- All generated IP will be owned by its originators; however, it will be released into the general public under the GPL license, i.e., it will be available free of charge for academic and research purposes.
- For outputs such as project reports, a non-exclusive licence to archive and disseminate the work will be granted to JISC. These outputs will be distributed via the AIR website and WIRE.
- While the developed software will be made freely available to the general public, training and consultancy on its installation, customisation and maintenance will be provided to institutions other than University of Wolverhampton for a fee.
- The project will ensure that any of its outputs do not infringe intellectual property rights of third parties. To that end, it will make use only such software and data released by third parties that itself is being distributed under the GPL license.

Sustainability of project outcomes and the exit strategy

In order to ensure that the end product is relevant to its intended applications, its future users (LIS staff working on the on-going WIRE project) will be actively involved in the project throughout its duration: from the user requirements study, to the development of evaluation materials, to user evaluation within the WIRE environment and quality assurance activities.

To ensure sustainability of the project outcomes, the following measures will be put in place:

- The AIR system will be integrated into the WIRE environment and the ILP staff will ensure that WIRE receives consultations and maintenance services for a period of at least 3 years beyond the end of the project.
- The AIR system will be distributed free of charge to interested parties. However, external institutions will be charged for consultations, training, and services on customising the software, in order to cover the costs of the required resources.
- A website, through which the software will be distributed, will be maintained for at least 3 years upon completion of the project.

Near the end of the project, ILP and LIS will aim to re-employ project staff (subject to satisfactory performance on the project) on other, related projects. This will make it possible to capitalise on the work carried on AIR and to find staff, who will provide system maintenance, training and consultations on its use, after the completion of the project.

Risk Assessment and Mitigation

Risks	Likelihood	Impact	Countermeasures
Failure to employ RAs with sufficient programming and text mining skills within the given time-scale	0.3	0.9	The positions will be re-advertised until suitable candidates are found. Meanwhile, the implementation activities on the project will be carried out by VP, who will devote 100% of his contract time to the project, as well as temporarily employed MSc students at ILP and School of Computing at UW.
Poor scalability (performance standards do not generalise to new data)	0.3	0.7	The system will allow for easy tracing of each step automatically performed by the system, so that appropriate amount of human input can compensate for the errors. This design will guarantee that in the worst cases, 100% of all effort on extracting the information is done by a human, but the risk of inserting erroneous data into the repository is minimised.
Technical problems e.g. an adverse	0.1	1	Web crawler and other software to be thoroughly tested and checked before use on the live

impact on the University website from use of the web crawler.			website. In order not to overload institutional web server, all processing will be done off-line, with the web pages accessed only periodically, during quiet periods between 03.00 and 07.00. Close liaison with UW IT Services colleagues
Copyright on web pages	0.1	1	In order not to infringe the copyright of the owners of web pages, the pages will not be permanently stored by the system locally: it will save only the information extracted from them along with references to original pages and their sections, without storing or re-distributing entire pages.
Poor user friendliness of the system	0.5	0.5	In case the user evaluation indicates that the integrated system does not have a user-friendly interface, the project will attempt, in as much as possible, to allocate more of RAs' time to the development of the front-end of the system (at the expense of its less critical components). More complex activities on customising the system to a specific user's requirements will be done outside the present project.
Tight schedule of the project	0.3	0.7	All desirable functionalities of the system, determined by the user requirements study, will be prioritised, in order to ensure that the most important are implemented within the lifetime of the project.

E. Engagement with the Community

UWs WIRE Project Team have established links with the SHERPA Team at Nottingham University, UKCORR and Biomed Central who supply and develop our DSpace based Open Repository software. The Biomed Central team are interested in supporting this development and have suggested mechanisms such as DSpace Lightweight Network Interface to import data generated by the new software into our Open Repository.

The WIRE project team includes several academics, and through our subject librarians' liaison networks have access to academics across a wide range of disciplines. We would also envisage close liaison with the Repository Support Project led by SHERPA at the University of Nottingham.

Stake holders

- Research active academic staff – the producers of research output and content for institutional repositories. Currently struggling with new challenge of self archiving. This project will support them by partially automating self archiving. The WIRE project team are already planning to engage academic staff in training to boost self archiving. The products of this project will make this task easier to 'sell' to hard pressed academics and by engaging them in its development phase they can help shape it to ensure its value to them.
- LIS staff – managers and operators of the WIRE repository
- Repository community – facing the challenges of improving deposition in IRs but constrained by limited IR staffing budgets and lack of uptake of self archiving by academic staff
- HEFCE – keen to promote the use of IRs to support the competitive edge of UK HE sector research.
- JISC
- Biomed Central - keen to develop and improve the functionality of Open Repository software and to contribute to DSpace community.

F. Budget

Although still in its project phase, there is institutional commitment at University of Wolverhampton to the WIRE institutional repository. Funding for its second year was recently approved by the University's Business Learning and Information Systems Steering Group and there are plans to embed WIRE within the range of services provided by Learning and Information Support Services. The system has also been adopted to electronically archive the institutions doctoral theses.

AIR Project budget			
Non-Staff	October 07 - March 08	April 08 - March 09	TOTAL £
Travel & Expenses	1,067	2,133	3,200
Hardware / Software ¹⁰	4,000	0	4,000
Dissemination ¹¹	0	0	0
Evaluation	0	0	0
Other	333	667	1,000
Total Directly Incurred Staff	5,400	2,800	8,200
Directly Incurred Total	22,753	32,038	54,791
Directly Allocated	October 07 - March 08	April 08 - March 09	TOTAL £
Estates	3,786	6,711	10,496
Other	1,000	2,000	3,000
Directly Allocated Total	21,915	43,007	64,921
Indirect Costs	28,582	50,668	79,249
Total Project Cost	73,249	125,712	198,961
Amount Requested from JISC	44,759	54,905	99,664
Institutional Contributions	40,032	59,266	99,298
Percentage Contributions over the life of the Project	JISC 50%	Partners 50%	Total 100%

A summary of qualitative and quantitative benefits for UW and the repository community

The major qualitative benefits of the project are expected to be the following:

- Significant growth in content in the WIRE institutional repository to better reflect the published research work produced by the UW staff.
- Improved coverage of the repository will increase its popularity with its potential users, and thus will help to further the goals of the WIRE project.

¹⁰ Two computers and accompanying software are requested for implementation activities on the project.

¹¹ Dissemination costs subsumed in travel and expenses line.

- Improve the capture of research activity and thus help embed the WIRE repository within the research process at UW
- Raise the profile of UW research by adding as much as possible to WIRE, which, as an open access repository, will increase the likelihood of citation.
- Enhance opportunities for UW researchers to gain experience in knowledge transfer that will be highly beneficial for their research careers
- Support research informed teaching by improving access to UW research in a form that can be readily integrated within our heavily used VLE.
- Support self-archiving by academic staff by providing partial automation of the deposition process.
- Free up LIS staff time from routine parts of the mediated deposition process, to enable rationalisation or redeployment.
- Develop the relationship between LIS and RILP which may lead to further research co-operation on advanced information access technologies.

In quantitative terms, the project is expected to free up the time of repository librarians, who at the moment manually encode bibliographical data into the repository. According to a conservative estimate (based on the performance of similar information extraction systems), the AIR system will have around 60% accuracy when operating in the fully automatic mode, and a repository librarian will spend half as much time on encoding the data in comparison with doing it completely manually. This will help to free up around 0.25 WTE of the Repository Librarians time. If the system encourages the authors themselves to use the system, the repository will grow much faster, and the reduction of costs on including the data into the archive will be even greater.

Within the wider repository community the project will:

- Produce software that can be used with Open Repository and that can be customised for use with DSpace or Eprints
- Test the concept of partial automation of population and deposition
- Establish the limits of automation
- Establish good practice with regard to automation
- Inform other repositories of the outcomes

G. Previous Experience of the Project Team

Existing RILP staff and roles

Prof. Dr. Ruslan Mitkov leads RILP at UW. He will direct the AIR project team, exploiting his considerable experience of working as the Principal Investigator on a number large-scale research projects. During his Professorship, he has managed numerous high-prestige research projects, including those funded by UK government and charities, the European community, industry, and charitable organizations. Of particular relevance was his leadership of the BiRD project (2003-2006), which was concerned with the closely-related field of information extraction in the domain of computational linguistics research. That project was methodologically similar to the current proposal, requiring the identification of documents in a large collection (the World Wide Web) that are most likely to contain relevant information. Suitable documents were identified by a combination of crawling and text categorization methods, and then processed, with key facts being extracted from them. Successful completion of the BiRD project resulted in the deployment of an online database containing information on relevant conferences held in the field of computational linguistics and the availability of resources such as corpora, software, and datasets together with a GUI to facilitate refined search of the database.

Prof. Mitkov's management activities arise as a result of his success in the research environment. Since his appointment at UW in 1995, he has produced an extensive array of publications, including his monograph on anaphora resolution and the well-received "Oxford Handbook of Computational Linguistics" (2003). Mitkov has supervised PhD students and researchers on a wide variety of research topics including information extraction, named entity recognition, and term extraction.

Dr. Viktor Pekar was appointed as a researcher at UW in 2003 to take on an instrumental role in the ESRC-funded BiRD project. His activities in the period (2003-2006) were undertaken with a view toward the implementation of a system for fully-automatic information extraction from web pages. This

included the development of software for web crawling, text categorization, named entity recognition and terminology processing, as well as software engineering activities. As a result of his experience in the robust identification of key facts in heterogeneous "real-world" web pages, he is well equipped to address the challenges arising throughout the course of the current project, which will similarly require the processing of inconsistently structured documents.

Existing LIS staff and roles

John Dowd, BA (Hons), Hybrid Services Manager, leads Learning and Information Support Services (LISS) within UW's LIS, which supports and develops the LIS infrastructure. This is achieved through managing a network of teams and specialists responsible for systems, including electronic collection development, copyright liaison at University level, cataloguing and classification of resources, the LIS website, and coordination of LIS finance. His focus within this environment is on service development and delivery and on monitoring and assuring effectiveness of the overall service. Recently appointed, John joined the University from OCLC, a non-profit, membership, computer library service and research organisation. Before leaving OCLC he was Regional Account Manager responsible for UK and Southern Africa; in tandem he was Office Manager, and participated in strategy development for the newly formed OCLC PICA. During his 14 years service with the organisation he evolved through a series of posts developing demonstrable skills related to project management – the most recent example being the migration of the South African bibliographic network to a hosted environment.

Dr. **John Rule**, D.Phil. (Oxon.), B.A.(Hons), PGDip ILS, FHEA. Deputy Learning Centre Manager, and Technology Supported Learning Co-ordinator for LIS since 2002, John is a qualified librarian, Fellow of the Higher Education Academy and has prior experience of teaching, research, providing research support and project management in the NHS and university sectors. Formerly project manager for the cross-departmental DRUW (Digital Repositories for University of Wolverhampton Project) and currently project manager for the WIRE (Wolverhampton Intellectual Repository and E-theses) Project. Trained in UW's project management methodology and to Prince 2 foundation level, John has overseen the procurement, launch and ongoing population of the WIRE institutional repository and liaises with other repository managers through SHERPA and UKCORR meetings and discussion lists.

Frances Hall, MA, MCLIP, Hybrid Collections Coordinator in LISS since September 2006. Line manages the Repository Librarian post and has been extensively involved in the WIRE project. She has previously carried out a six month research project into the use of metadata on university websites, encompassing a review of available metadata standards and of the technical and organisational factors in their successful use. From 2001 to 2005, she worked as an Information Specialist in Engineering and Applied Science at Aston University, a liaison role which involved working closely with academic staff on matters relating to information and library provision. In January 2006, she moved into the role of Information Resources Specialist, which was newly created post designed to be focus for development in cataloguing and classification, acquisitions and e-resource management processes at the university.

We are currently in the process of recruiting a replacement for our half time **Repository Librarian** post, and should have filled this by the start of September 2007.

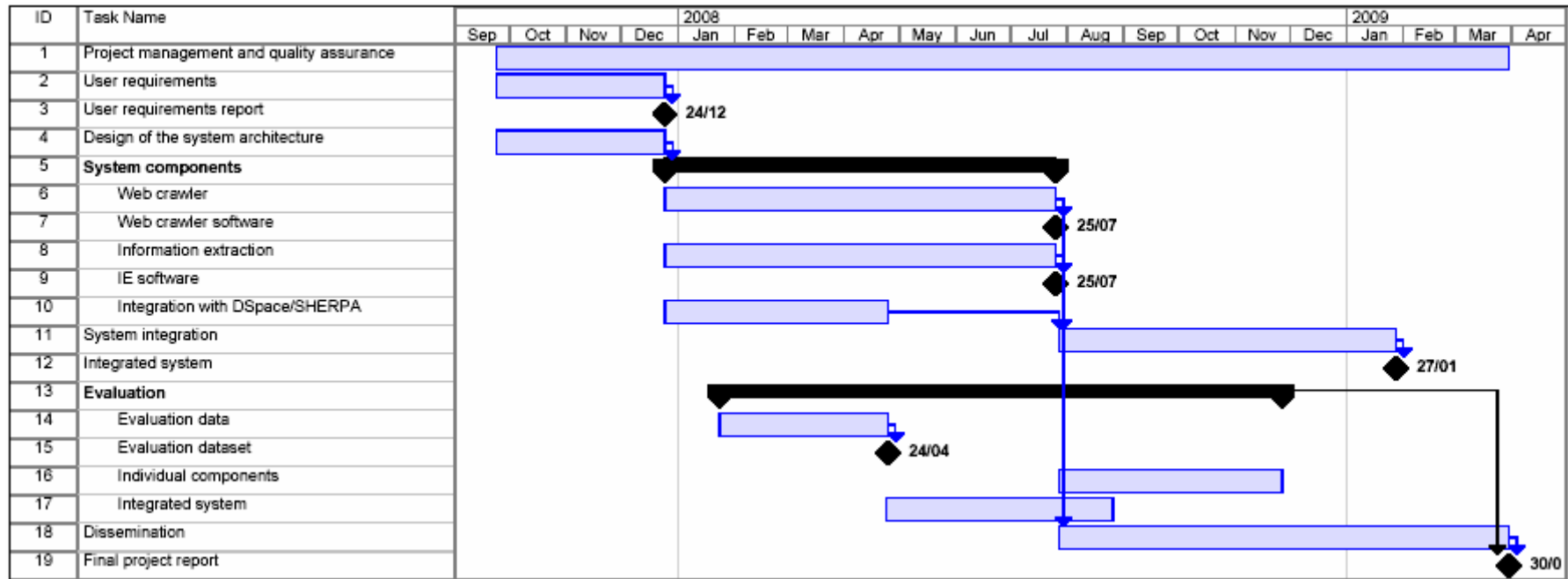
The above LIS staff would help specify user needs for customisation, oversee trialling within UW (through to testing automated population of the existing WIRE repository), contribute to the Project Steering Group and liaise with the repository community through SHERPA¹² or the United Kingdom Council of Research Repositories (UKCORR).

Staff dedicated to the project

A dedicated researcher will be recruited (on the UW Lecturers' Payscale point 12 to ensure appropriate skills) to produce, integrate and test the software to be used for automation in conjunction with Dr. Pekar, under the oversight of Prof. Mitkov. Employed with JISC funds and based in RIILP under RIILP line management, they will be recruited by the end of November 2007.

¹² University of Nottingham based repository support body, home of the JISC funded Repository Support Project which this project should link into.

Appendix A. Gantt chart.



Appendix B

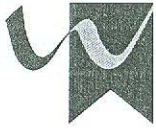
FOI Withheld Information Form

We would like JISC to consider withholding the following sections or paragraphs from disclosure, should the contents of this proposal be requested under the Freedom of Information Act, or if we are successful in our bid for funding and our project proposal is made available on JISC's website.

We acknowledge that the FOI Withheld Information Form is of indicative value only and that JISC may nevertheless be obliged to disclose this information in accordance with the requirements of the Act. We acknowledge that the final decision on disclosure rests with JISC.

Section / Paragraph No.	Relevant exemption from disclosure under FOI	Justification
None	No exemption requested	Not applicable

Please see <http://www.ico.gov.uk> for further information on the Freedom of Information Act and the exemptions to disclosure it contains.



UNIVERSITY OF
WOLVERHAMPTON

Professor Sally Glen PhD MA
Pro Vice-Chancellor (Academic)

Executive Suite
University of Wolverhampton
Wulfruna Street
Wolverhampton
WV1 1SB
United Kingdom

20 June 2007

Telephone Codes
UK: 01902 Abroad: +44 1902

Direct Line: 322399
Switchboard: 321000

Fax: 322632

E-mail: s.glen@wlv.ac.uk

JISC
Northavon House
Coldharbour Lane
Bristol, BS16 1QD

Dear Sir/Madam,

RE: Funding application for a project titled "AIR – the Automated Archiving Institutional Repository"

I am writing in support of the application, prepared jointly by the Research Institute of Information and Language Processing and the Learning Information Services Department of the University of Wolverhampton, to the JISC for funding of a research project titled "Automated Archiving for an Institutional Repository".

I would like to confirm my strong support for the proposed research. The project promises to bring considerable benefits to the University, both in qualitative and quantitative terms. We expect it will increase the coverage and impact of WIRE, our institutional repository, while improving the efficiency of Learning Information Services support for WIRE, by overcoming reliance on time consuming searching and manual entry of bibliographical data into the repository. From this point of view, the collaboration with the researchers from the Research Institute of Information and Language Processing in developing an automated archiving system appears a very attractive possibility to overcome these difficulties. This research also has the potential to have a much wider impact across the repository community.

I confirm that the University will also invest in the development of the project.

Yours faithfully,

Sally Glen