

# JISC DEVELOPMENT PROGRAMMES

## *PROJECT PLAN*

### Project

<b>Project Acronym</b>	ROAD	<b>Project ID</b>	
<b>Project Title</b>	Robot-generated Open Access Data		
<b>Start Date</b>	1 <sup>st</sup> April 2007 (subject to recruitment)	<b>End Date</b>	31 <sup>st</sup> March 2009
<b>Lead Institution</b>	University of Wales Aberystwyth		
<b>Project Director</b>	Dr. Mike Hopkins		
<b>Project Manager &amp; contact details</b>	Mr. Stuart Lewis Information Services Llandinam Building Penglais Campus Aberystwyth Ceredigion SY23 3DB  01970 622860  Stuart.lewis@aber.ac.uk		
<b>Partner Institutions</b>			
<b>Project Web URL</b>	<a href="http://www.inf.aber.ac.uk/projects/road/">http://www.inf.aber.ac.uk/projects/road/</a>		
<b>Programme Name (and number)</b>	<i>Repositories and Preservation (03/06)</i>		
<b>Programme Manager</b>	Mr Phil Vaughan		

### Document

<b>Document Title</b>	<i>Project Plan</i>		
<b>Reporting Period</b>	N/A		
<b>Author(s) &amp; project role</b>	Stuart Lewis – Project Manager		
<b>Date</b>	30/11/2006	<b>Filename</b>	ROAD – Project Plan – v0a.doc
<b>URL</b>	N/A		
<b>Access</b>	<input checked="" type="checkbox"/> Project and JISC internal		<input type="checkbox"/> General dissemination

### Document History

Version	Date	Comments

---

# Project Plan

---

## Overview of Project

### 1. Background

The aim of the ROAD project is:

- **To investigate the use of current open-source digital repository software to enable the automatic curation of robot-generated experimental data and metadata.**

The intention is to demonstrate the feasibility of using current open-source digital repository software for management of data acquired directly from automatic integrated laboratory equipment, specifically the Robot Scientist<sup>1</sup> created at UWA. It is expected that the project will also contribute to research on the sharing of experimental data as well as providing a case study for similar scientific installations in other institutions and for other scientific domains.

The vision for data repositories in *Digital Repositories Roadmap: looking forward*<sup>2</sup> has an information environment in which raw research data is made available on an open access basis. This vision includes the idea of direct linking between laboratory equipment and a departmental or institutional repository. The link with the Robot Scientist constitutes a sophisticated demonstration of this vision.

Professor Ross King of the Department of Computer Science at UWA has built a “Robot Scientist” that is capable of automatically carrying out cycles of scientific experimentation<sup>3</sup>. That is, the Robot Scientist generates its own hypotheses to explain observations, devises experiments to test these hypotheses and carries out the experiments. It is capable of initiating over 1000 experiments and making over 200,000 observations per day. It generates over 1 Gigabyte of data and experimental metadata a day. The Robot Scientist is designed to investigate gene function in yeast, but the approach is applicable to other fields of scientific investigation and this high level of laboratory automation is likely to become increasingly important in the future.

The Robot Scientist uses laboratory robotics to physically implement a closed-loop scientific discovery system. The Robot Scientist represents the beginning of a new breed of automated laboratory. In addition to devising hypotheses, each constituent part of the robot is connected to a single data-gathering computer. This allows much richer metadata to be gathered and linked to the resulting data. This extra layer of metadata includes data such as hypotheses, the goals of the experiment and laboratory environmental data. A final layer of data includes a video stream of the experiment taking place. Whilst this high level of integrated data-gathering is currently unusual, it is likely to become much more common in the future. The final benefit of storing the complete rich metadata is that it aids reproducibility of experiments, which, in the bioinformatics world is highly important.

The automated nature of the robot bypasses the major problem of data gathering and storage in repositories, that of integrating it with the workflow of the scientist and the associated sociological problems.

---

1 <http://www.aber.ac.uk/compsci/research/bio/robotsci>

2 [http://www.jisc.ac.uk/uploaded\\_documents/rep-roadmap-v15.doc](http://www.jisc.ac.uk/uploaded_documents/rep-roadmap-v15.doc)

3 R. D. King et al, “Functional genomic hypothesis generation and experimentation by a robot scientist” <http://dx.doi.org/10.1038/nature02236>

## 2. Aims and Objectives

The broad aims of the ROAD project are to investigate the viability of using common open source repository platforms to curate large amounts of automatically generated data in a way which is efficient and allows easy access to the data once it has been stored.

The ROAD project will examine the following key areas:-

- **Evaluation of the ability of current open-source digital repository software (for example GNU Eprints, DSpace and Fedora) to automatically ingest and store large quantities of experimental data and metadata.**
- **Development of tools where applicable to assist with the automated ingest of large amounts of scientific data into an open access repository.**
- **Issues regarding placing data into an open access repository as soon as it is created (such as IPR, persistent identifiers and date stamping)**
- **The curation of such data and metadata to manage its preservation and its accessibility to diverse communities of interested parties.**

## 3. Overall Approach

The approach that will be taken is to examine each issue in turn, splitting the issues up into work packages. The issues build upon each other, starting with requirements gathering, move on to product evaluation and implementation, and finish with dissemination.

### 1. Investigation

- Investigate the requirements of capturing large quantities of experimental data
- Investigate the requirements of capturing metadata and how it may be used to allow the exploration of the final repository
- Research possible repository structure designs to hold the data in a useful configuration
- Investigate the in-built functionality in current open source repository platforms with respect to ingest mechanisms and scalability
- Investigate the overall suitability of each of the repository platforms studied with respect to their ability to be used for the project

### 2. Implementation

- Based on the repository platform suitability report, choose a repository platform to build the pilot repository on
- Implement an automated submission process for the chosen repository platform
- Design and implement a useful repository interface allowing easy access to the data, possibly with suitable visualisations of the data
- Examine possible approaches to assisting with the downloading of large number of items, and implement a suitable method, taking into account interoperability

### 3. Evaluation

- Evaluate the final repository
- Disseminate results to the community

The work plans are described fully in section 15.

The project aims to create a pilot repository which will ingest data and metadata created by the Robot Scientist. The pilot repository will serve the short-term needs of making the data available, but will not be a production service. The aim of the project is to gain experience in the area of submitting large amounts of automatically generated data into repositories. To make the pilot service a production service would require further funding to be sought from relevant bodies. As such, servers with enough

computational power and storage capacity will be purchased to run a pilot service, but these would not have sufficient storage capacity to run a long-term full-scale production repository.

It may be that none of the current open-source digital repository software platforms (e.g. Fedora / DSpace / GNU EPrints) are powerful enough to hold such large amounts of data over a long time period. If this is the case, work packages can be changed to investigate the limitations of the current software to identify the problems. This information can then be fed back into the products to try to improve them, or could be used to justify the need for an open source digital repository software package to be created with the specific aim of being able to fulfil the requirements of large data sets.

## 4. Project Outputs

The following is a list of deliverables that this project will create:

### Reports

- A report detailing the ability of different repository platforms to ingest and store the requisite data in a suitable fashion
- Quarterly project reports giving updates on the progress of the project.
- Final report detailing the investigations undertaken, the software written, and the evaluation performed.

### Software (specific to chosen repository platform)

- A suitable data repository to hold the data from the Robot Scientist
- Software to perform the ingest of large amounts of data into the chosen repository platform
- Suitable interfaces to allow people and machines to easily download the data help within the repository

## 5. Project Outcomes

The major outcome of the project will be an open access repository holding the data generated by the Robot Scientist. This pilot repository will give public access to the vast amounts of data created by the Robot Scientist.

A further important outcome from the project will include a detailed report detailing the suitability of different open source repository platforms to hold large data sets such as generated by the Robot Scientist. This report will be useful to both potential creators of data repositories, and to the repository platform creators.

It is hoped that the data repository will be an exemplar data repository advocating the benefits of making data available openly in machine readable formats. It is also hoped that the project will contribute to community discussion, knowledge and understanding in the area of data repositories.

## 6. Stakeholder Analysis

Stakeholder	Interest / stake	Importance
The Wolfson Bioinformatics group, department of Computer Science, UWA	<p>As partners in the project, creators of the Robot Scientist, and owners of the data, the research group are key stakeholders in the work.</p> <p>We shall work closely with the group to ensure our systems are mutually compatible, and will work together. We will reply on their advice and expertise in the areas of the Robot Scientist and its data, and data ontology representation.</p>	High
Repository Platform creators	<p>One of the main thrusts of the work is to critically evaluate the suitability of current open source repository platforms with respect to providing the functionality required to provide a repository interface to the Robot Scientist.</p> <p>We will work with the repository creators to ensure our reports are accurate, and will freely offer our findings back to the creators to allow them to address any issues we discover.</p>	Medium
Other data owners	It will be important to disseminate our findings through relevant email lists and groups to ensure work we undertake is given good publicity. This will help to ensure any lessons we learn can be fed into similar projects.	Low

### Relationships with other JISC projects

Due to staff overlaps, the ROAD project can work closely with the Repository Support Project (RSP) to share best practise in the area where appropriate.

## 7. Risk Analysis

Risk	Probability (1-5)	Severity (1-5)	Score (P x S)	Action to Prevent/Manage Risk
Staff recruitment	5	2	10	Recruiting a suitable project officer will be essential to the success of the project. To ease recruitment problems, candidates can be drawn from the pool of resources trained by UWA's Computer Science department, and its Computational Biology Group and Wolfson Bioinformatics Unit.
Staff retention	2	4	8	Were key staff to leave, this could impact the project. This risk cannot be mitigated, but by following strict documentation procedures, replacement staff could pick up work more easily.
Technical knowledge	1	4	4	Rick can be prevented by ensuring that a technically competent 'Project Officer' is employed.
Technical advancement	3	3	9	Open source repository packages are rapidly changing pieces of software. So where possible, any software written will sit 'around' these projects rather than being embedded within them. Where sensible, embedded code within such software will be offered to its respective code-base for consideration of incorporation into the wider software project.
Systems and data change	4	2	8	The Robot Scientist is still under final development. This means that some details such as the rate of data output has yet to be finalised. The project will have to work around any changes, but will help the project produce flexible outputs.

## 8. Standards

Where possible the project deliverables will use open standards such as OAI-PMH (Open Archive Initiative Protocol for Metadata Harvesting) formats. These will be researched and evaluated during the course of the project. Where appropriate during the project, designs will be formalised using descriptive notations such as UML.

If appropriate, techniques aligned with the JIE will be utilised such as Web Services and RSS feeds.

The project is likely to use emerging techniques such as EXPO<sup>4</sup> expressed through OWL<sup>5</sup>.

<sup>4</sup> <http://expo.sourceforge.net/>

<sup>5</sup> <http://www.w3.org/TR/owl-ref/>

## 9. Technical Development

It is envisaged that all technical developments undertaken throughout the project will be written using Java, unless it would be more sensible to use an alternative language (e.g. Perl for GNU eprints). As the software written will be distributed to other organisations, in line with open-source best practice the software will be packaged appropriately (e.g. using other open source software such as ant). Standard versioning techniques will be used to assist with code management.

Once the requirements have been defined, an iterative development lifecycle will be used as much of the technical background of the systems that have to integrate, are at present unknown. Early iterations will concentrate on simple importing objects, and this will be built upon until a finished system is completed. By adopting this method, some development risks will be reduced as small parts of the project will be completed at a time.

## 10. Intellectual Property Rights

All of the repository platforms to be investigated are 'Open Source' pieces of software where use is openly offered on certain conditions which will be adhered to by this project (retaining ownership comments in code etc). Because the projects are open source, other bodies wishing to use these component parts of the finished project can do so freely. As such, where deliverables constitute code, these shall be made available under a Berkeley Software Distribution (BSD) licence. It is not anticipated that there will be commercial exploitation of the project deliverables.

If any code written were to be included into platforms where the BSD is unsuitable, the licence would have to be changed to a similar and suitable licence.

### **BSD licence to be used:**

Copyright (c) 2005-2006, University of Wales Aberystwyth  
All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

Neither the name of the University of Wales Aberystwyth nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

## ***Project Resources***

### **11. Project Partners**

The main staff involved in the project are:

**Project Director: Dr. Mike Hopkins**

Information Services  
Hugh Owen Library  
Penglais Campus  
Aberystwyth  
Ceredigion  
SY23 3DZ

01970 622391

Mike.Hopkins@aber.ac.uk

**Project Manager: Mr Stuart Lewis**

Information Services  
Llandinam Building  
Penglais Campus  
Aberystwyth  
Ceredigion  
SY23 3DB

01970 622860

Stuart.Lewis@aber.ac.uk

Staff collaborating from the department of Computer Science include:

**Professor Ross King**

rdk@aber.ac.uk

**Dr. Amanda Clare**

afc@aber.ac.uk

### **12. Project Management**

The project will be overseen by a project management team consisting of staff members from Information Services and Computer Science. Day-to-day running of the project will be managed by Stuart Lewis, the project manager.

A project officer will be employed to undertake the majority of the project work, including research, documentation, technical developments and report writing. The project officer will report directly to the project manager. The project officer will work full time and the project manager will work 1 day per week on the project. There are no known training needs of project staff.

### **13. Programme Support**

We are not aware of any specific areas where support from the programme manager will be required for the project, but if support is later required, we shall request it.

Once the project has finished, support may be requested to find more funding if there are new issues that arise that would benefit from further research.

### **14. Budget**

See appendix A.

## Detailed Project Planning

### 15. Workpackages

See appendix B.

### 16. Evaluation Plan

Timing	Factor to Evaluate	Questions to Address	Method(s)	Measure of Success
End of each quarter	Achievement of aims and objects detailed in project plan for that quarter.	Have milestones been met? If not, why not?	Compare outputs against project plan.	All milestones are met on time.
		Is the project management working effectively?	Ask all members of the management committee.	All members of the management committee are happy with the project management.
		Are stakeholders aware of, and in agreement with decisions made?	Ask all members of the management committee.	All members of the management committee are in agreement with decisions made, and no major decision is made without the approval of the management committee.

End of project	Success of project	Does the system work well?	Execute user and system tests to prove (or otherwise) the success of the system.	All tests completed successfully.
		Have the lessons learnt from the project been widely disseminated?	See what coverage the project has achieved in the form of papers, press releases, dissemination materials, listserv emails and conference papers.	
		Have any findings from the repository suitability report been acted upon by the repository platform creators?	Contact repository platform creators to see if any changes were required, and if they have been acted upon.	Any problems found whilst testing the suitability of different open source repository platforms will have been investigated and potentially fixed by their corresponding communities.
		Are the deliverables of production quality?	Run a pilot or system.	Pilot system running smoothly and reliably.

## 17. Quality Plan

Explain the quality assurance procedures you will put in place to ensure that project deliverables meet quality expectations and acceptance criteria. Complete the table below for each of the major deliverables providing as much detail as possible. Repeat the table as many times as necessary to accommodate all deliverables.

<b>Repository platform suitability report</b>				
<b>Output</b>				
<b>Timing</b>	<b>Quality criteria</b>	<b>QA method(s)</b>	<b>Evidence of compliance</b>	<b>Quality responsibilities</b>
End of year 1	Accurate and correct	Work with repository platform creators to ensure accurateness and correctness of findings	Emails and discussion documents from the partners and the repository creators	Project Manager

<b>Automated ingest system</b>				
<b>Output</b>				
<b>Timing</b>	<b>Quality criteria</b>	<b>QA method(s)</b>	<b>Evidence of compliance</b>	<b>Quality responsibilities</b>
Q5	Complete and error free	Work with Robot Scientist researchers to ensure all data is being ingested completely and correctly	Test reports	Project Manager
	Adheres to common standards	Verify methods against standards as used by bodies such as the JIE	Detail technologies used	Project Manager
	Fitness for purpose	Ensure the ingest system can cope with the large amounts of data	Test reports	Project Manager

<b>Implemented repository design</b>				
<b>Output</b>				
<b>Timing</b>	<b>Quality criteria</b>	<b>QA method(s)</b>	<b>Evidence of compliance</b>	<b>Quality responsibilities</b>
Q6	Complete and error free	Work with Robot Scientist researchers to ensure all data is being ingested completely and correctly	Test reports	Project Manager
	Adheres to common standards	Verify methods against standards as used by bodies such as the JIE and the chosen repository platform creators	Detail technologies used	Project Manager
	Accessibility legislation	Ensure WAI compliance to web-based outputs	Test reports	Project Manager

## 18. Dissemination Plan

<b>Timing</b>	<b>Dissemination Activity</b>	<b>Audience</b>	<b>Purpose</b>	<b>Key Message</b>
<b>Start of project</b>	Press releases and emails	Local institution and wider repository community.	To promote awareness of the project.	The project has started.
<b>End of year 1</b>	Formal project contact (probably via email lists)	Repository platform creators	To ensure repository suitability findings are accurate, and to feedback findings.	The system works.
		Other large-scale data creators	To ensure proposed system designed	The system works, can bring benefits to the institution, but requires policy changes.
<b>End of project</b>	Information dissemination	Other large-scale data creators	Share findings of the project.	As appropriate.
		DSpace, Eprints and Fedora email lists and user conferences	Share findings of the project.	As appropriate.
		Conferences	Share findings of the project.	As appropriate.
		Journals	Share findings of the project.	As appropriate.
<b>Ongoing</b>	Project website	Anyone	To allow people to keep up to date with project aims and progress.	Key information about the project.
	Programme meetings	Other repository projects	To meet and interact with similar projects.	How can we share information?

## 19. Exit and Sustainability Plans

<b>Project Outputs</b>	<b>Action for Take-up &amp; Embedding</b>	<b>Action for Exit</b>
Open source Repository suitability report	Encourage repository platforms creators to engage with the findings and act upon them if they consider them to be an issue.	Encourage active debate on the findings so that the repository platform creators can
Pilot open access data repository	Solicit views from potential users of the repository, and examine access statistics to see if it is proving a well-used site.	Pass the completed pilot system to the Wolfson Bioinformatics team with appropriate documentation to allow them to continue running the system.
Project web site	Keep website alive for 3 years. Update as appropriate if resources allow.	Keep website alive for 3 years.

<b>Project Outputs</b>	<b>Why Sustainable</b>	<b>Scenarios for Taking Forward</b>	<b>Issues to Address</b>
Pilot open access data repository	The pilot system will be holding real data which may be of use to bioinformaticians.	Pass the completed pilot system to the Wolfson Bioinformatics team with appropriate documentation to allow them to continue running the system.	Further funding to support a production service, passing of system and documentation to Wolfson bioinformatics group.

## Appendix B. Workpackages

<b>WORKPACKAGES</b>	<b>Mon th</b>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1: Data requirements capture	1-3	X	X	X																					
2: Metadata requirements capture	4-6				X	X	X																		
3: Repository design	7-9							X	X	X															
4: Selection of repository software	10-12										X	X	X												
5: Automated submission process	13-15													X	X	X									
6: Interface design	16-18																X	X	X						
7: Data download	19-21																			X	X	X			
8: Dissemination	13-24													X	X	X	X	X	X	X	X	X	X	X	X

Project start date: 01-04-2007

Project completion date: 31-03-2009

Duration: 24 months

Workpackage and activity	Earliest start date	Latest completion date	Outputs (clearly indicate deliverables & reports in bold)	Milestone	Responsibility
YEAR 1					
<p><b>WORKPACKAGE 1: Data requirements capture</b></p> <p><b>Objective:</b> Investigate the requirements for capture of experimental data, especially in view of the likely volume of such data and of making it available to different interested communities.</p>					
1. Data capture requirements	1/4/2007	30/6/2007	Data capture requirements report		PO
<p><b>WORKPACKAGE 2: Metadata requirements capture</b></p> <p><b>Objective:</b> Investigate the capture of metadata and how this can best be used to allow the exploration of the repository and its data in ways that support different communities.</p>					
2. Metadata capture requirements	1/7/2007	30/9/2007	Metadata capture requirements report		PO

<b>WORKPACKAGE 3: Repository design</b>					
<u>Objective:</u> Following earlier investigations establish an effective design for the Robot Scientist data repository.					
3. Repository design specification	1/10/2007	31/12/2007	Repository design specification.		PO
<b>WORKPACKAGE 4: Selection of repository software</b>					
<u>Objective:</u> Evaluate current open-source digital repository software both in terms of how well it can support the proposed design of the repository and also perform load testing of the software to see how it copes with the import of the required quantities of data and metadata. The ability of the software to operate as a service as part of a wider Service Oriented systems approach, as used within the e-Framework, will be evaluated.					
4. Repository platform suitability	1/1/2008	31/3/2007	Repository platform suitability report	MS1	PO / PM
5. End of year report	1/3/2008	31/3/2008	End of year report.		PO / PM / PD

<b>YEAR 2</b>					
<b>WORKPACKAGE 5: Automated submission process</b>  <u>Objective:</u> Design and implement the processes and any required tools that link the repository to the Robot Scientist to support the automatic ingest and curation of the experimental data and metadata.					
6. Automated submission process	1/4/2008	30/6/2008	Implemented and working automated ingest system for the chosen repository platform	MS2	PO
<b>WORKPACKAGE 6: Interface design</b>  <u>Objective:</u> Design and implement interface components for external users of the repository for effective extraction of the data held therein.					
7. Design and implement repository interface	1/7/2008	30/9/2008	Implemented repository design	MS3	PO

<b>WORKPACKAGE 7: Data download</b>					
<u>Objective:</u> Explore approaches to sharing the data and metadata across the interested communities. This might also include exploring relations with domain based repositories or archives for long term storage of the large quantities of experimental data, as suggested in the vision for data repositories in the <i>Digital Repositories Roadmap: looking forward</i> .					
8. Data download investigation	1/10/2008	31/12/2008	Download investigation report		PO
<b>WORKPACKAGE 8: Dissemination</b>					
<u>Objective:</u> Disseminate findings of the project to the community. This will involve writing papers and attending relevant academic conferences in the fields of repositories, data creators, and data users.					
9. Disseminate project findings through appropriate channels.	1/4/2008	31/3/2009	As appropriate		PO / PM / PD
10. Submit JISC final and completion reports		31/3/2009	JISC final and completion reports	MS4	PO / PM / PD

Members of Project Team: *PD - Project director / PM - Project manager / PO - Project officer*