

Project Acronym: Deposit Plait  
 Version: 0a  
 Contact: Stuart Lewis (stuart.lewis@aber.ac.uk)  
 Date: 24<sup>th</sup> July 2008

Project Information			
<b>Project Acronym</b>	Deposit Plait		
<b>Project Title</b>	The Deposit Plait		
<b>Start Date</b>	1 <sup>st</sup> May 2008	<b>End Date</b>	30 <sup>th</sup> April 2009
<b>Lead Institution</b>	Aberystwyth University		
<b>Project Director</b>	Dr. Christine Urquhart		
<b>Project Manager &amp; contact details</b>	Stuart Lewis  Stuart.lewis@aber.ac.uk 01970 622860	Information Services Llandinam Building Aberystwyth University Penglais Campus Aberystwyth Ceredigion SY23 3DB	
<b>Partner Institutions</b>	None		
<b>Project Web URL</b>	<a href="http://www.inf.aber.ac.uk/projects/deposit-plait/">http://www.inf.aber.ac.uk/projects/deposit-plait/</a>		
<b>Programme Name (and number)</b>	Digital Repositories		
<b>Programme Manager</b>	Dr. Neil Jacobs		

Document Name			
<b>Document Title</b>	Project Plan		
<b>Reporting Period</b>	N/A		
<b>Author(s) &amp; project role</b>	Stuart Lewis (Project Manager)		
<b>Date</b>		<b>Filename</b>	Deposit Plait – Project Plan.doc
<b>URL</b>	N/A		
<b>Access</b>	✓ Project and JISC internal		

Document History		
Version	Date	Comments
0a	24 <sup>th</sup> July 2008	First draft



## JISC Project Plan

### *Overview of Project*

#### **1. Background**

One of the main issues affecting institutional repositories today is the lack of content, and one of the common reasons expressed by academic staff is the resentment of re-keying data.

The concept of a 'deposit engine' as outlined in the call could help to address this problem by making use of interoperable systems that can assist by gathering metadata for a given paper, aiding the academic in easing the overhead of depositing, and thus increase the corpus of works in institutional repositories.

There are many systems and techniques that could be employed by such a deposit engine. This project will investigate two of these in detail:

- What information can be easily extracted from a version of the deposited document once converted to an 'open format' which can easily be read by software.
- What interfaces are currently provided to publisher (commercial, learned society, open access), funder catalogues and bibliographic software that can be queried to provide metadata.

Combining the three strands of metadata derived from within the document, metadata about the document from external databases, and input in the form of verification or extra metadata from the depositor should create a plait strong enough to perform a deposit with sufficient metadata about the document. The investigations can provide information about possible interfaces a deposit engine could use to make the action of depositing easier.

#### **2. Aims and Objectives**

The five elements of the project aim to:

- 1. Investigate the metadata requirements of repositories and service providers of deposited items**

The repository project will first document the metadata needs, both minimal and ideal, of a deposit, from the perspective of the repository itself (taking into account standards such as SWAP) and from the perspective of service providers such as Intute Repository Search who harvest and re-use the data.

- 2. Investigate what metadata can be easily extracted from the deposited document, and if useful, whether a web service can be built to provide such metadata from a file**

The project will investigate existing tools that examine the document to automatically extract metadata, select candidate metadata for selection by the depositor, or ask depositors to highlight the metadata.

Interfaces into documents (via software products or APIs) will be investigated to see what is easily possible with existing tools, or what might be feasible with further work. Software and APIs to be investigated include:

- Xena<sup>1</sup> '(Xml Electronic Normalising for Archives') is a piece of software developed by the National Archives of Australia to aid long term preservation by converting files to open formats.
- OpenOffice<sup>2</sup> is a well known desktop office application, which has the ability to read and write files in an OpenDocument format, and a machine interface to allow the conversion of a document.
- Metadata Aggregator<sup>3</sup> is a tool which can read and extract metadata from documents in an OpenDocument file.
- OpenXML tools<sup>4</sup>.
- Other software and systems as recommended by the OpenDocument Fellowship<sup>5</sup>.

The project team will investigate, and if possible build, a web service that can perform metadata extraction from OpenDocument or OpenXML files. In addition this element will demonstrate the interoperability of candidate systems identified during the project.

### **3. Investigate what metadata can be obtained from online metadata sources and personal bibliographic management software**

The project will investigate current and possible interfaces into publisher, bibliographic software and funder catalogues of metadata which could be used to provide some of the metadata required by repositories. Any obvious licensing issues will be included in the investigation. Systems and interfaces to be investigated include:

- Web of Knowledge – ISI Proceedings
- ArticleFirst from OCLC
- ZETOC
- Thomson's Dialog and Datastar
- SWETS online / BIDS / OpenSIGLE / NTIS
- EndNote, RefWorks, and open source bibliographic software

Project staff will seek to work closely with systems providers and will work within the terms and conditions of the typical institutional licences we have to these systems.

### **4. Investigate the possibility of building a 'Meta-API' using the open document format web service and bibliographic data sources**

Once candidate bibliographic information provider systems have been identified, and an OpenDocument format web service created, the possibility of a Meta-API can be investigated. Such a Meta-API, perhaps provided as a web service could in theory be used to query such systems using either sparse information provided by an author, or from a file in an OpenDocument format.

---

<sup>1</sup> <http://xena.sourceforge.net/>

<sup>2</sup> <http://www.openoffice.org/>

<sup>3</sup> [http://books.evc-cit.info/odf\\_utils/aggregator.html](http://books.evc-cit.info/odf_utils/aggregator.html)

<sup>4</sup> <http://openxmldeveloper.org/>

<sup>5</sup> <http://opendocumentfellowship.com/applications>

## 5. Investigate what metadata needs to be verified or provided by the depositor

Finally the results of each of the investigations will be brought together to see what, if any, metadata might be missing which would need to be provided by the depositor, or from other interfaces with the deposit engine.

The main objective of the project is to understand to what extent metadata can be extracted from documents and bibliographic services in order to assist with the population of institutional repositories.

## 3. Overall Approach

### Methodology

The initial element of the project will document the metadata needs of repositories and service providers (e.g. Intute Repository Search, OAIster, Google Scholar) from a repository.

Following that, three strands of a potential plait of data providers will be investigated to see which can provide each of the metadata needs identified. The investigations will focus on what systems and software could provide this information, what interfaces there are into these systems, and what interfaces would be useful. Individual reports will be compiled for each strand and the overall findings will be written up in a final report.

### Important issues

The important issues that the project will be focussing on are:

- Metadata requirements
- Automated creation of metadata
- Quality control of metadata
- Interoperability

### Scope and boundaries

The project will restrict itself to the deposit of Scholarly Works in the form of research papers. Books, data sets, or presentations are examples of works not covered by the project.

### Critical success factors

Ability to gather metadata for content from external sources

Licence restrictions on gathering metadata from potentially useful but commercial resources.

## 4. Project Outputs

Each of the main work packages will result in a published report:

- Report 1: A report into the metadata requirements for repositories and service providers, aimed at facilitating the evaluation of the other work packages.
  - This will not only provide an evaluation mechanism for the later investigations, but will be a useful document in its own right for repository managers to know what metadata to collect.
- Report 2: A report detailing the findings of the investigation into metadata that is available from information stores.
  - This will provide a useful overview of what systems may be able to easily provide metadata to depositors about the items they are depositing.
- Report 3: A report detailing the findings of the investigation into metadata that can be derived from a document.
  - This will inform possible future work looking into metadata extraction and the structure of information within documents to support this task.

- Final report: A report pulling together the contents of the first three reports, and subsequent investigations into the possibility of a meta-API.
  - This report may inform future funding calls in this area.

Depending upon the result of the investigations into a meta-API, a prototype service may be created.

## 5. Project Outcomes

The outcomes of the project will be:

- A model for use in the repository community describing a minimal level of metadata required for an item in a repository, in order for the metadata to be sufficient for searching, retrieval and harvesting (re-use).
- An understanding of the possibilities of extracting metadata from documents, and from bibliographic services.

## 6. Stakeholder Analysis

Stakeholder	Interest / stake	Importance
Repository managers	Knowing the metadata they need to collect for items in their repository.	High
Repository software developers and technical staff	Knowing the metadata requirements for an out-of-the-box configuration.  Knowing what interfaces they may be able to build into their software to enable metadata extraction from submitted files.	High
Repository research community and JISC	Future possibilities for research in this area. A knowledge of which avenues look like they could provide promising sources of metadata.	Medium
External providers of metadata (funding bodies, commercial systems etc.)	How to build useful interfaces and licences to metadata.	Medium
Academics	How they can write documents in a way that enabled the extraction of key metadata.	High

## 7. Risk Analysis

As with any project of this nature, there are risks associated. These have been summarised below:

Risk	Probability (1-5)	Severity (1-5)	Score (P x S)	Action to Prevent/Manage Risk
Staff recruitment and retention	1	3	3	Staff recruitment should not be an issue as the project will use internal pre-existing staff.
Staff retention	1	3	3	Due to the short nature of the project, were retention should not pose an issue, or other departmental staff could be used.
Technical challenges	3	3	9	If some of the technical issues do not provide useful outcomes, this will still provide useful knowledge in the area.
Changing environment	3	3	9	The repository environment is a constantly changing landscape and staff will use their professional knowledge to stay abreast of these changes throughout the project.

## 8. Standards

Metadata will be accessed in whatever format it can be, with a aim of getting it transformed into Dublin Core as this is the most widely used metadata standard used in UK institutional repositories.

Document (file) standards to be investigated are OpenXML and Open Document Format (ODF). The project may also investigate other common standards such as PDF or DOC, and their conversion to an open xml format.

## 9. Technical Development

Technical development will follow an agile methodology with quick prototyping in order to quickly generate usable software. It is not envisaged that this project will create any production quality code, rather only working prototypes.

## 10. Intellectual Property Rights

All reports created by the project will be licensed under a suitable Creative Commons licence, and will be deposited in a local Open Access repository, and the JISC Information Environment repository.

All software produced by the project will be licensed under a suitable open source licence, and will be deposited in a repository (source code repository or institutional repository).

There may be issues with metadata licenced via subscription or commercial bibliographic sources which will be dealt with as they arise.

Project Acronym: Deposit Plait  
Version: 0a  
Contact: Stuart Lewis (stuart.lewis@aber.ac.uk)  
Date: 24<sup>th</sup> July 2008

## ***Project Resources***

### **11. Project Partners**

#### **Main contact:**

**Name:** Dr. Christine Urquhart

**Position:** Director of Research, Department of Information Studies, Aberystwyth University

**Email:** cju@aber.ac.uk

**Address:** Department of Information Studies, Llanbadarn Fawr, Aberystwyth, Ceredigion, SY23 3AS

**Tel:** 01970 622188

**Fax:** 01970 622190

Other partners are Information Services at Aberystwyth University, and CASIS the commercial arm of the Department of Computer Science at Aberystwyth University.

### **12. Project Management**

The project will make use of the expertise of three departments at Aberystwyth University: Information Services (repositories), Department of Information Studies (library systems), and the Centre for Intelligent Systems and Advanced Software (intelligent software). The staff undertaking the work are:

#### **Christine Urquhart (Project Director)**

Department of Information Studies / cju@aber.ac.uk / 01970 622162

#### **Stuart Lewis (Project Manager)**

Information Services / sdl@aber.ac.uk / 01970 622860

#### **Neil Taylor**

CASIS / nst@aber.ac.uk / 01970 621528

#### **Jackie Knowles**

Information Services / jak@aber.ac.uk / 01970 628490

#### **Talat Chaudhri**

Information Services / tac@aber.ac.uk / 01970 622396

#### **Hannah Payne**

Information Services / hep@aber.ac.uk / 01970 628490

#### **Rhian Thomas**

Department of Information Studies / ret@aber.ac.uk / 01970 622974

#### **Suzana Barreto**

CASIS / szb@aber.ac.uk / 01970 621528

The staff members undertaking the project will form a project board to direct the project. In addition to this, the project will work closely with the CRIG and WO-CRIG projects to ensure the work is complimentary to any work undertaken by these projects.

There are currently no perceived training needs for the staff members.

### **13. Programme Support**

It is not perceived that there will be much support required from the programme. Some advice may be required relating to the publishing of the reports.

Page 7 of 11

Document title: Deposit Plait Project Plan

Last updated: July 2008

## 14. Budget

See Appendix A. There are no changes to the project budget from the budget submitted in the project proposal.

## *Detailed Project Planning*

## 15. Workpackages

The project will be split into three phases, first defining the requirements for a deposit, followed by investigations of possible methods of metadata population, followed by a report writing phase. Each element of the project will be a workpackage.

See Appendix B for further detail.

- **Workpackage one: Repository and service metadata needs:** An investigation undertaken by repository research staff from Information Services into the metadata requirements of repositories and search services.
  - **Deliverables:** A report into the metadata requirements for repositories and service providers, aimed at facilitating the evaluation of the other workpackages.
  - **Effort:** 2 weeks.
- **Workpackage two: Metadata from information stores:** An investigation undertaken by staff from the Department of Information Studies into the availability, licensing issues and interfaces to metadata information stores owned by publishers and funders. Staff from CASIS will then create light-weight demonstrators of the most potential systems.
  - **Deliverables:** A report detailing the findings of the investigation into metadata that is available from information stores.
  - **Effort:** 4 weeks + 4 weeks interoperability testing by CASIS.
- **Workpackage three: Metadata from OpenDocument and OpenXML files:** An investigation undertaken by CASIS staff into the possibility of easily extracting metadata from OpenDocument and OpenXML files, or by tools allowing easy selection of metadata by the depositor. If possible, a web service will be created to provide this functionality, or a definition of a suitable service or API defined.
  - **Deliverables:** A report detailing the findings of the investigation into metadata that can be derived from a document.
  - **Effort:** 6 weeks.
- **Workpackage four: Meta-API: Investigating a Meta-API:** An investigation undertaken by CASIS in conjunction with the other partners into the possibility of creating a Meta-API to query bibliographic systems and to make use of an OpenDocument file format metadata extractor.
  - **Deliverables:** A report and if possible a prototype of a Meta-API.
  - **Effort:** 6 weeks.
- **Workpackage five: Plaiting the strands:** An investigation undertaken by repository research staff from Information Services comparing the required metadata defined by workpackage one with that which can or could be made available by the methods investigated in workpackages two and three. Where relevant the project will work with CRIG and disseminate findings via CRIG.
  - **Deliverables:** A report comparing the findings of workpackage one with the findings of workpackages two and three.
  - **Effort:** 2 weeks.

- **Workpackage six: Project Management:** Undertake the standard management and reporting functions required by JISC.
  - **Deliverables:** Reports, plans and attendance at meetings as required by JISC.

## 16. Evaluation Plan

Timing	Factor to Evaluate	Questions to Address	Method(s)	Measure of Success
End of project	The potential of collecting metadata from external and local bibliographic sources	Is it likely, taking into account licensing and cost issues, that metadata can be easily captured from external or local bibliographic systems in order to ease the workload required when depositing a journal article into a repository?	Examine report written during workpackage two.	Whether an answer has been given to the question. (N.B. A negative answer does not in this case mean a failure, just that no potential is shown).
End of project	The potential of extracting metadata programmatically from xml-based file formats.	Is it likely, perhaps depending upon the template used, that metadata can be easily extracted from new xml-based file formats in order to ease the workload required when depositing a journal article in a repository?	Examine report written in workpackage three.	Whether an answer has been given to the question. (N.B. A negative answer does not in this case mean a failure, just that no potential is shown).
End of project	The potential of creating a meta-API to assist in the deposit of journal articles into a repository.	Is it likely that a meta-API could be created to extract metadata from new xml-based file formats and to capture metadata from external or local bibliographic systems in order to ease the workload required when depositing a journal article into a repository?	Examine report written in workpackage four.	Whether an answer has been given to the question. (N.B. A negative answer does not in this case mean a failure, just that no potential is shown).

## 17. Quality Plan

Output Timing	Quality criteria	QA method(s)	Evidence of compliance	Quality responsibilities	Quality tools (if applicable)
End of each workpackage	Quality, accurate report written.	Hold meeting to sign-off report	Signed-off report	Project manager / project director	
End of project	Good quality code	Code review meeting	Report from code review meeting	Project manager / project lead at CASIS	

## 18. Dissemination Plan

Timing	Dissemination Activity	Audience	Purpose	Key Message
As appropriate	Attendance at CRIG events	CRIG members	To raise awareness of the project and its findings within the repository interoperability community.	The project exists, and is carrying out the work as defined in the project plan.
As appropriate	Attendance at relevant events	Event attendees	To raise awareness of the project and its findings.	The project exists, and is carrying out the work as defined in the project plan.
End of project	Dissemination of project reports	All stakeholders	To disseminate the findings of the project.	The findings of the project.

## 19. Exit and Sustainability Plans

Project Outputs	Action for Take-up & Embedding	Action for Exit
Project reports	Dissemination to the repository community	Deposit in repository
Prototype software	Dissemination to the repository community	Deposit in repository

Project Outputs	Why Sustainable	Scenarios for Taking Forward	Issues to Address
Prototype software	If code is useful, further projects may be able to continue their development into production service.	If the code is able to perform the jobs it may be able to, then future projects of services may be able to use the code.	Hosting and licensing of code.

Project Acronym: Deposit Plait  
Version: 0a  
Contact: Stuart Lewis (stuart.lewis@aber.ac.uk)  
Date: 24<sup>th</sup> July 2008

## Appendix A. Project Budget

Below is a copy of the project budget:

<b>Amount Requested from JISC</b>	<b>£49579</b>
-----------------------------------	---------------