

Repositories and Preservation Proposal Cover Sheet

Cover Sheet for Proposals (All sections must be completed)		JISC Capital Programme
Name of Capital Programme: Repositories and Preservation Programme		
Bid for Call Area : (Please tick ONE BOX ONLY, as appropriate)		
Tools and Innovation (Strand B)		
<input checked="" type="checkbox"/>	Call Area I – Tools and Innovation Projects	Please specify area of proposed project eg ' <i>metadata generation and validation</i> ' <i>metadata generation and validation</i>
Discovery to Delivery (Strand C)		
<input type="checkbox"/>	Call Area II – Discovery to Delivery Projects	<input type="checkbox"/> a) Version identification framework <input type="checkbox"/> b) Persistent identifier interoperability demonstrator <input type="checkbox"/> c) Federated access management and repositories <input type="checkbox"/> d) Semantic interoperability demonstrator
Repository Start-Up and Enhancement (Strand D)		
<input type="checkbox"/>	Call Area III – Repository Start-Up and Enhancement Projects	<input type="checkbox"/> a) Repository start-up projects <input type="checkbox"/> b) Repository enhancement projects
Digital Preservation and Records Management (Strand H)		
<input type="checkbox"/>	Call Area IV – Digital Preservation and Records Management Projects	<input type="checkbox"/> a) Digital preservation across the lifecycle <input type="checkbox"/> b) Models and implementation of preservation services <input type="checkbox"/> c) Preservation tools development
Shared Infrastructure Services (Strand I)		
<input type="checkbox"/>	Call Area V – Shared Infrastructure Services Projects	<input type="checkbox"/> a) Pilot implementation of licence registry <input type="checkbox"/> b) Pilot national name and factual authority service <input type="checkbox"/> c) Scoping an architecture to support digital policy management <input type="checkbox"/> d) Scoping a terminology registry
Name of Lead Institution:		AHDS (Arts & Humanities Data Service)
Name of Proposed Project:		MetaTools - Investigating Metadata Generation Tools
Name(s) of Project Partner(s):		AHDS (Arts & Humanities Data Service)
Full Contact Details for Primary Contact:		
Name: Dr. Malcolm Polfreman Position: Information Officer Email: malcolm.polfreman@ahds.ac.uk Address: Arts & Humanities Data Service (AHDS), King's College London, 26 - 29 Drury Lane, LONDON, WC2B 5RL Tel: 0207 848 1985 Fax: 0207 848 1989		

Length of Project: 18 months		
Project Start Date: 15 March 2006		Project End Date: 15 September 2007
Total Funding Requested from JISC: £96679.66		
Funding Broken Down over Financial Years (April – March):		
Apr06 – Mar07	Apr07 – Mar08	Apr08 – Mar09
£3,837.05	£60430.25	£32412.38
Total Institutional Contributions:		
Percentage Contributions over the Life of the Project:	JISC 71.6 %	PARTNERS 28.4 %
Outline Project Description		
<p>Repositories and portals are struggling to provide resource discovery metadata for the rapidly growing number of new digital resources. Without it, resources remain hidden and unused and much of the original investment is wasted. Increasingly, repositories entertain the hope that automated metadata generation will provide a solution. However, there is no single tool or suite of tools to which portal and repository managers can go to meet most of their metadata generation requirements. The available tools generally handle a narrow range of digital formats, generate a restricted element set and, in the case of extraction algorithms, are mostly effective within narrow subject domains or for documents of a predictable layout or genre.</p> <p>There is no registry or trusted body of documentation that rates the quality of metadata generation tools or identifies the most effective tool(s) for any given task. Benchmarks and reliable evaluation studies are conspicuously lacking.</p> <p>Moreover, the ad hoc nature of interfaces and the wide variations in the format of APIs, when they exist at all, mean that it is not possible to call these tools automatically in a flexible manner, as differently formatted calls need to be explicitly programmed for each interface. A single metadata record will therefore usually require the merging of output from several tools each of which must be invoked separately.</p> <p>The project aims to:</p> <ol style="list-style-type: none"> 1. Develop a methodology for evaluating metadata generation tools 2. Compare the quality of currently available metadata generation tools 3. Develop, test and disseminate prototype web services that integrate the best metadata generation tools and functionality. <p>Our approach will be to transform the best tools into services with well-defined service interfaces, based on XML-based standards such as SOAP and WSDL. We will develop an ontology that allows the metadata generation web services to be given machine-interpretable semantic annotations and descriptions, explicitly representing knowledge about the services in a flexible and extensible way. These annotations will allow the dynamic discovery of appropriate services by software agents, which will be able to invoke these services, combine them into workflows, and monitor the results, with no (or minimal) user interaction.</p> <p>The effectiveness of the prototype metadata generation Web services will be tested on a broad range of digital resources held by AHDS and in co-operation with other repositories and portals.</p>		
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	YES	NO

Appendix B (Tick Box)	✓	
------------------------------	---	--

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.

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

MetaTools - Investigating Metadata Generation Tools

Submitted under: JISC Circular 04/06: Repositories and Preservation Programme Strand B: Tools and Innovation, Call Area 1 - Tools and Innovations Projects.

Submitted by: Arts and Humanities Data Service, King's College London

1. Introduction

1.1 Background

Resource discovery metadata (as it is known) is a crucial component of the lifecycle of digital resources. Without appropriate metadata, resources remain hidden and unused and much of the original investment is wasted. Standardised metadata is crucial to interoperability, since metadata is a powerful tool that enables the discovery and selection of relevant digital resources quickly and easily. Poor quality or non-existent metadata on the other hand is equally effective at rendering resources unusable, since without it a resource is essentially invisible within a repository or archive and thus remains undiscovered and inaccessible. However, with the ever increasing amount of digital resources being made available, finding the time and resources necessary for ensuring metadata of appropriate quality is created is becoming a more and more difficult task. This is true for both formal digital repositories and community or domain specific portals which seek to select resources specifically useful for their user base.

Increasingly, repositories are entertaining the hope that automated metadata generation will provide a solution. Indeed, without automation, it may be impractical to describe resources at item-level or any finer level of granularity than for the collection as a whole. Automated metadata generation is still in its infancy but several approaches have emerged, including metatag harvesting, content extraction, automatic indexing or classification, text and data mining, social tagging, and the generation of metadata from associated contextual information or related resources. Some technical metadata captured by tools developed by the preservation community can also contribute to resource discovery: e.g. JHove, Droid, and the NLNZ Metadata Extraction tool.

Most of the resource discovery metadata found within the JISC IE is manually created either by authors, depositors and/or repository staff. A major obstacle to portals and repositories incorporating metadata generation tools is the absence of common, standardised interfaces. Effective metadata generation functionality is spread thinly across a dozen or so tools each of which offer only a partial solution and which must be called up separately. Such tools have generally been developed for specific institutions or in response to particular commercial opportunities and, consequently, handle a narrow range of source formats or generate a restricted element set. Metadata extraction algorithms are generally effective within narrow subject domains or for documents of a predictable layout or genre. Current tools are not easily integrated into efficient portal or repository workflows. The generation of a single, complete metadata record is likely to entail the merging of output from several tools plus some subsequent manual enhancement – each stage of which, in the absence of a plug-in or Web services architecture, involves the use of a separate online interface. There is no single tool or suite of tools to which portal and repository managers can go to meet most of their metadata generation requirements. Nor is there a registry or trusted body of documentation available to repository managers that rates the quality of metadata generation tools or identifies the most effective tool(s) for any given task – let alone a way for the most appropriate tool to be triggered and executed automatically.

1.2 Length of Project

This is a project of eighteen months' duration, starting on 15 March 2007 and ending on 15 September 2008.

2. Project Description

2.1 Aims, objectives and methodology

The project will take as its starting point the US *AMeGA report*¹, which identifies the generic functionality required of metadata generation tools, JISC's unpublished *Metadata Generation for Resource Discovery* study, which has identified key tools, gaps and areas for future research and development, and the JORUM *Automated metadata*² report.

The project aims to:

- **Develop a methodology for evaluating metadata generation tools.**

Infomine's libiViaMetadata suite of tools has gone some way towards developing an evaluation, as well as an assignment, module for the elements that it generates but is very much the exception among readily available tools. The study will develop appropriate benchmarks for comparing metadata generation tools, such as the source format(s) that they support, their method(s) of extraction and range of elements generated, the output format(s), encodings and bindings that are incorporated, their standard vocabularies and terminology control, configurability, ease of use and licensing constraints. The study expects to develop benchmarks on an element-by-element approach. This is necessary because, although key metrics are likely to be accuracy, recall, and precision³, some metadata elements are typically assigned a single metadata value, in which case accuracy may be a good performance measure, but others typically have multiple values, and precision and recall may be more informative. Some fields are even more complex. For example, accuracy may not be a useful way to measure the assignment of text to the Description element.

- **Compare the quality of currently available metadata generation tools**

In the absence of accepted benchmarks, testing has been sporadic and largely unsatisfactory. The project aims to improve upon the limited testing of metadata generation software that has hitherto been carried out. Much of it has been auto-evaluation rather than human evaluation, which is more thorough as well as being more expensive. This project is likely to use both human and machine-based methods of evaluation⁴.

The project is particularly important because tools have rarely been tested specifically in relation to JISC resources and studies have almost always had a narrow focus and small sample size. For instance, Greenberg's comparison of DC-dot and Klarity, like Irvin's research, covered only NIEHS environmental health web pages and involved a tool, Klarity, that no longer exists. The study will use a variety of AHDS resources as a test bed. The AHDS has wide subject coverage across the arts and humanities and its resources are of varied origin because they are received as deposits from the whole community. The AHDS archives a broad range of text resources (e.g. MS Word, PDF, plain text, html, xml), still digital images, moving images, sound, and datasets. We also intend to extend the testing phase to include other cooperating repositories.

Tools have not been tested within real-life workflows. Little is known about how long it takes to manually upgrade/edit the partial output from metadata generation tools compared with a person cataloguing the resource manually from scratch. This is a difficult area but, where it is appropriate,

¹ Automatic Metadata Generation Application (AMeGa) Project final report
http://www.loc.gov/catdir/bibcontrol/lc_amega_final_report.pdf

² Baird, K. & JORUM Team. (2006), *Automated metadata: A review of existing and potential metadata automation within JORUM and an overview of other automation systems*. JORUM, p.23. Available at http://www.jorum.ac.uk/docs/pdf/automated_metadata_report.pdf

³ Accuracy measures the proportion of the time that autogenerated values for a given element exactly match those assigned by an expert cataloguer (after simple normalizations have been done). Recall is the proportion of relevant documents retrieved out of the total number of expected relevant documents in the entire collection. Precision is the ratio of relevant documents retrieved to the number of documents retrieved.

⁴ For other weaknesses of testing, see Paynter, G. *Developing Practical Automatic Metadata Assignment and Evaluation Tools for Internet Resources*. Date unknown, The INFOMINE Project, University of California, p. 292, <http://ivia.ucr.edu/projects/publications/Paynter-2005-JCDL-Metadata-Assignment.pdf>

the study aims to test metadata generation tools on new deposits as they are ingested into the AHDS repository.

The project will also redress the problem of testing generally having been conducted by the developers of the tools involved – even when conducted professionally, as when Infomine's PhraseRate team compared the Keyphrase output generated by the Infomine libiViaMetadata tool for 101 websites with that of Kea, DC-dot, and Turney's Extractor. The current project will be demonstrably impartial.

Properly independent and widespread testing of metadata generation tools is a prerequisite for the development of a web services solution to the problem of metadata generation. It is expected that the selection of tools for this aspect of the study will be informed by the JISC study, *Metadata Generation for Resource Discovery*.

- **Develop, test and disseminate prototype web services that integrate metadata generation tools.**

The metadata generation tools considered for this project have a variety of ad hoc interfaces. The available APIs, when they exist at all, vary widely in format. In most cases no API is provided; rather, the interface is via a web-based form or a user-invoked stand-alone client. The variety of APIs means that it is not possible to call these tools automatically in a flexible manner, as differently formatted calls need to be explicitly programmed for each interface. In the case of screen-based interfaces, it is either necessary to copy information manually, or to use "screen scraping" techniques, which are not very reliable.

As observed in the *Metadata Generation for Resource Discovery* study, metadata generation services are highly specialised, as regards both the types of input on which they are most effective, and the types of output produced, and this situation is likely to continue. A typical use case for such tools would be to automatically generate resource discovery metadata for digital content in circumstances where it is not feasible to do so manually. However, the disparity or non-existence of programmatic interfaces militates against this automatic tool invocation. The manual use of the tools will fail to scale as the environment of available formats & tools expands, particularly with the advent of complex multimedia formats that are frequently encountered in the arts and humanities. What is required is a mechanism for dynamically discovering services for generating resource discovery metadata appropriate for particular digital content, and automatically invoking them as part of a workflow, e.g. on ingest into a repository.

Our approach will be to transform these tools into services with well-defined service interfaces, based on XML-based standards such as SOAP and WSDL. By using such standards, web services provide a mechanism for enabling interoperability between distributed applications. However, these standards typically enforce only weak or implicit typing of data, and do not allow the semantics of web services to be represented, which restricts the potential for dynamic discovery and invocation of such services. One approach would be to require web service definitions to be more strongly validated against a set of schemas, but this would be rather inflexible, and infeasible in an environment containing many service providers.

Instead, our approach will be to use an ontology that allows the metadata generation services to be given machine-interpretable semantic annotations and descriptions, explicitly representing knowledge about the services in a flexible and extensible way. Given suitably extended registry functionality, these annotations will allow the dynamic discovery of appropriate services by software agents, which will be able to invoke these services and integrate them into workflows. We will pay close attention to previous initiatives in semantic web service description and semantic registries, e.g. PANIC⁵, myGrid/Feta and Grimoires⁶. In particular, we will examine existing web service description ontologies such as OWL-S⁷, WSMO⁸ and the myGrid ontology⁹,

⁵ <http://metadata.net/panic/>

⁶ <http://www.grimoires.org/>

http://www.mygrid.org.uk/index.php?module=pagemaster&PAGE_user_op=view_page&PAGE_id=57&MMN_position=64:51:63

⁷ <http://www.daml.org/services/owl-s/>

as we intend if possible to use such an ontology as a foundation, extending it for semantics specific to metadata generation.

Part of the project would involve investigating, developing and testing common interfaces for various metadata generation services. This relates to Paragraph G40, bullet 10, of the Tools and Innovations strand.

2.2 Interoperation with other activities

The AHDS is submitting a concurrent proposal under Strand H, Call Area IV(c): Preservation Enhancement Tools of the current call, which is investigating the use of semantically annotated web services for the automated generation of descriptive metadata. The proposals are independent, in the sense that the funding of one does not pre-suppose the funding of the other, although with their service-oriented approaches they would combine provide a powerful framework for generation of metadata that supports both preservation and discovery of digital resources.

Although this proposal is being submitted under Programme Strand B: Tools and Innovation, Call Area 1 - Tools and Innovations Projects, it has links with other strands of the programme. The second phase investigates the use of semantically annotated and semantically aware web services for the dynamic creation of service workflows, and as such is directly related to the e-Infrastructure Programme, specifically Call VI: Semantically Coordinating Resources and Services Across Registries (see Paragraph E102 of the call).

3. Work Packages

3.1 Work Package 1: Project Management

Months 1-19

This work package will act to manage and coordinate activities, to prepare and report as required, and to assess risks and opportunities as the project progresses.

Tasks:

- Develop a detailed work plan with timescales, deliverables, and milestones
- Monitor progress, identify corrective actions where necessary, and ensure compliance with the schedule
- Prepare periodic management reports

Deliverables:

- Detailed work plan
- Progress and risk assessment reports
- Website and dissemination activities

3.2 Work Package 2: Scoping of appropriate tools and collections

Months 2-7

Using the 'Metadata Generation for Resource Discovery' report as a point of departure, this work package will identify and prioritise the most promising metadata generation tools for inclusion in the testing phase of the project.

Tasks

- Identify software tools that may be used to extract or generate resource discovery metadata and analyse their system requirements.
- Review the capabilities of the tools on the basis of their documentation and any third party testing and their suitability for provision as web services.

⁸ <http://www.w3.org/Submission/WSMO/>

⁹ <http://www.mygrid.org.uk/>

- Identify AHDS collections that have attributes suitable for testing.

Deliverables

- Software tools specification
- Specification for data analysis and indication of relevant AHDS collections

3.3 Work Package 3: Development of comparative framework and analysis

Months 4-10

This work package is designed to serve as a practical exercise in comparing the capabilities of the software tools identified by the project. It will serve to identify the circumstances in which the tools may be applied and differences in the output.

- Review the research literature of metadata generation to identify appropriate metrics for measuring the effectiveness of metadata generation on an element-by-element, attribute-by-attribute basis.
- Assess the evaluation modules of the few metadata generation tools that contain them (e.g. Infomine's libiViaMetadata)
- Develop a practical framework for the comparative testing of software tools, detailing methods of benchmarking and comparison of performance, and taking into account issues such as amount of information produced by the tools, accuracy of the information and quality, etc.
- Execute the software tools identified in work package 2 against the subset of AHDS resources as part of a controlled testing programme.
- Analyse results and identify the most effective tools for particular circumstances e.g. for particular types of content, digital format, output format, environment and other factors.

Deliverables

- Framework for the comparison of resource discovery metadata generation tools
- Analysis of software tools using framework and recommendation of useful applications.

3.4 Work Package 4: Repository modelling

Months 4 – 10 Analysis of repository operation in the AHDS

Months 10 – 19 Development of user guide for use by other repositories.

This work package is designed to examine how the generation of metadata via web services may be integrated into the operation of a digital repository. It will examine how integration of web services is likely to affect existing workflows and develop a generalised model for repository ingest processes.

Tasks

- Analyse existing repository operations to generalise workflows.
- Identify circumstances under which web services may be integrated and how this can be achieved.
- Develop use cases and scenarios.
- Develop existing documentation into a user guide that details scenarios for use of the web services in other

Deliverables

- User guide for use by repositories wishing to implement web services for metadata generation.

3.5 Work Package 5: Development and integration of metadata generation services

Months 10 – 15 Development of web services

Months 16 – 19 Revision to web services

This work package is intended to analyse the capabilities of the preferred metadata generation tools from a developer viewpoint. Technologies for developing them as web services, such as wrapping the tools using SoapLab, will be investigated and implemented. The work package will also identify likely methods of integrating the services into repository workflows.

Tasks

- Investigate web service technologies.
- Analysis of APIs provided by the software tools or services that may be used to pragmatically call the application and perform appropriate tasks.
- Produce prototype web services that integrate metadata generation tools.
- Integrate web services into repository workflows.

Deliverables

- Documentation on APIs provided by the software tools and proposal for integration into web services
- Prototype web services & associated documentation

3.6 Work Package 6: Testing and dissemination

Months 16 – 19

This work package tests the web services developed by the project. They will be tested in relation to AHDS resources identified in work package 2 and the scenarios identified in work package 4. A subset of JISC repositories and portals will be invited to participate in this testing stage. A set of training materials, including a step-by-step guide to installation and use, will be created as part of this work package.

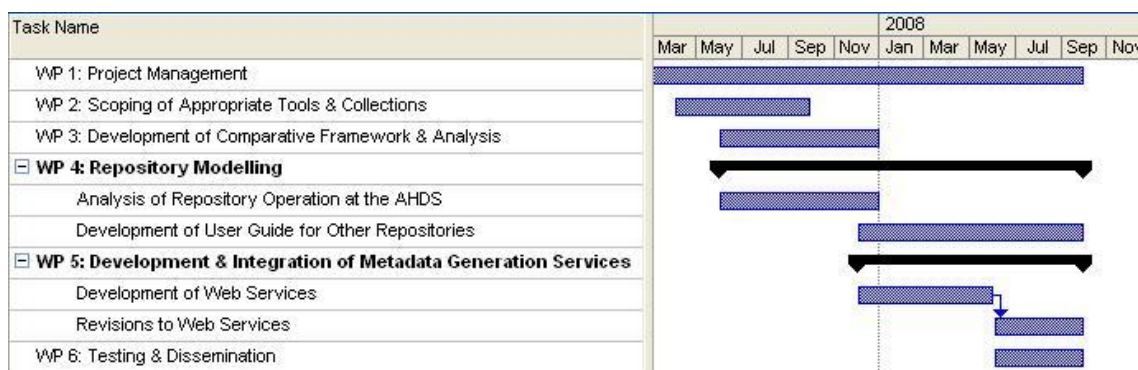
Tasks

- Test the metadata generation web services on AHDS collections identified in work package 2 and undertake comparison with previously reported test results
- Apply developed web services to use case scenarios.

Deliverable

- Step-by-step guide to installation and use of metadata generation web service prototype(s).
- Comparative study of metadata generation test results.

4. Outline Project Plan



5. Risks

Risk	Probability (1-5)	Severity (1-5)	Score (P x S)	Action to Prevent/Manage Risk
Staffing Problems (difficulty of recruiting and retaining)	2 ¹⁰	4	8	Spread expertise throughout the project; document the project

¹⁰ The nominated personnel are already employed at the AHDS.

staff with appropriate skills and experience)				sufficiently so that other staff can take over if/when project partners are unable to continue with their work; advertise vacancies in timely fashion.
Technical hardware and software issues	2	4	8	Complete evaluation of hardware and software to be deployed. Ensure adherence to standards and best practice. Provide training as necessary.
Key stakeholders (e.g. test bed repositories) do not buy in to/support the project	1	5	5	Ensure regular information flow to all stakeholders; seek feedback on direction and progress.

6. IPR

IPR in all reports and other documents produced by the project will be retained by the authors and host institutions but made freely available on a non-exclusive licence as required by JISC.

All software created during the project will be made available to the community on an open-source basis on the GPL licence. We will respect the licence model of all third party software used during the project, most of which is made available under open source licences.

7. Sustainability, dissemination and take-up

- 1) All software developed during the project will be made available on an open-source basis, in accordance with the *Policy on Open Source Software for JISC Projects and Services*¹¹. It will be made open to enhancements & patches contributed by the user community.
- 2) Appropriate metadata generation tools will be incorporated into the AHDS repository ingest workflow.
- 3) To encourage take-up, we will create an installation package (and associated user guide) that will simplify the installation and configuration process as much as possible, and we will undertake publicity activities to promote it. Widespread use of the software will encourage the creation of a self-sustaining user and developer community, ensuring longer-term sustainability.
- 4) In particular, we will liaise with UKOLN to promote the availability and use of the tools.
- 5) Papers, presentations and posters at conferences and workshops.

8. Key Personnel & Partner details

The Arts and Humanities Data Service (AHDS) will conduct all parts of the project. The Arts and Humanities Data Service is a national service funded by the Joint Information Systems Committee (JISC) and the Arts and Humanities Research Council (AHRC), to collect, manage, catalogue, preserve and promote the use of digital resources in research, teaching and learning in the arts and humanities. The AHDS provides advice and guidance in the creation of digital resources to quality standards that ensure their suitability for use in research and teaching and their long-term viability. The AHDS identifies and accessions a wide range of digital resources of many different types, and evaluates, validates, adds metadata, and incorporates the collections into its resource discovery, delivery and preservation systems.

Project Manager, 0.4 FTE, 18 months, based at the AHDS Executive

Dr. Malcolm Polfreman, AHDS Information Officer, is responsible for resource discovery and metadata strategy across the AHDS and for co-ordinating metadata-related activity within its five Centres. He is co-author of the recent JISC Metadata Generation for Resource Discovery study, which is the point of departure for the project proposed here. He has written QA Focus briefing papers

¹¹ http://www.jisc.ac.uk/about_opensourcepolicy.html

on metadata-related issues¹². In addition to developing significant in-house expertise in relation to metadata issues and processes, the AHDS has been a regular contributor to national and international metadata initiatives (e.g. the JISC Information Service Registry) and has written on metadata within recent JISC-funded reports, such as the Digital Image Archiving Study and the Feasibility and Requirements Study for Preservation of E-Prints (James et al, 2003)¹³.

Research Officer, 0.2 FTE, 18 months, based at the AHDS Executive

Gareth Knight, (AHDS) is Digital Preservation officer for the Arts & Humanities Data Service. He is responsible for the investigation and implementation of AHDS preservation strategy, performing actions necessary to ensure digital collections stored by the AHDS are managed correctly and are fit for purpose. In this role, he provides technical advice to depositors and AHRC applicants wishing to manage their digital resources. Gareth is currently preservation officer for the JISC funded SHERPA-DP project, investigating a distributed OAIS-compliant model for the preservation of digital objects stored by institutional repositories participating in the SHERPA project.

Technical Officer, 1 FTE, 18 months, based at the AHDS Executive

The Technical Officer will be responsible for all software design, development and testing, including the production of associated reports and other documentation.

To be appointed.

Consultant

The project will also receive informal input from Dr Mark Hedges, Technical Manager of the AHDS. He has extensive experience of technical and project management, gained from 17 years work in the software industry. For the last 2 years, he has been manager of AHDS technical services, in particular managing a number of projects.

¹² <http://www.ukoln.ac.uk/ga-focus/documents/briefings/briefing-64/briefing-64-A4.doc>

¹³ http://www.jisc.ac.uk/uploaded_documents/e-prints_report_final.pdf