

JISC DEVELOPMENT PROGRAMMES
Project Document Cover Sheet
BIANNUAL PROGRESS REPORT

Project

Project Acronym	HaIRST	Project ID	
Project Title	Harvesting Institutional Resources in Scotland Testbed		
Start Date	08/2002	End Date	07/2005
Lead Institution	Centre for Digital Library Research (CDLR), Strathclyde University		
Project Director	Gordon Dunsire		
Project Manager & contact details	Fabio Simeoni (fabio.simeoni@cis.strath.ac.uk)		
Partner Institutions	St.Andrews University, Napier University, Glasgow College Group (GCG), John Wheatley College		
Project Web URL	http://hairst.cdlr.strath.ac.uk		
Programme Name (and number)	FAIR		
Programme Manager	Chris Awre		

Document

Document Title	Biannual Report		
Reporting Period	09/2003-02/2004		
Author(s) & project role	Fabio Simeoni (Project Manager), Gordon Dunsire (Project Director), Jane Barton (Project Team)		
Date	01/01/2003	Filename	HaIRST-FAIR-BR0204.pdf
URL	http://hairst.cdlr.strath.ac.uk/document/ HaIRST-FAIR-BR0204.pdf		
Access	<input type="checkbox"/> Project and JISC internal		<input checked="" type="checkbox"/> General dissemination

JISC Progress Report Template for FAIR Projects

February 2004

Overview of Project

Grant Statement

We confirm that project development is being conducted under the terms agreed with JISC in the letter of grant and the JISC Terms and Conditions attached to it. No changes to the original award are to be reported.

2. Aims and Objectives

The project retains the aims, objectives, and methodology described in the original proposal and further elaborated in the current version of the Project Plan and in the previous project reports.

The objectives previously targeted for this reporting period were:

1. *the roll out of the a pilot e-print repository service at the University of Strathclyde and an analysis of the eprints.org to report to JISC;*

Strathprints, a pilot implementation of an Institutional Repository (IR) service for the University of Strathclyde is now available at <http://eprints.cdlr.strath.ac.uk:8080> as an experimental, non-operational service under constant development. An informal analysis of the first experiences gathered when customising the code base of the eprints.org system (version 2.1) and the principles adhered to has also been offered to JISC in October 2003. For further details on Strathprints, see Section 3.

2. *The disclosure of Napier's resources via the DLESE OAI Data Provider;*

The publication in October 2003 of the beta version of the 'Specification for an OAI Static Repository and an OAI Static Repository Gateway'¹, of the Open Archive Initiative, has soon identified a preferred solution to facilitate the participation of some project partners – including Napier University, the Glasgow College Group (GCG) and John Wheatley College – into the OAI-based federation promoted by the project. In particular, the experimental service of the OAI Static Repository Gateway at the Research Library of the Los Alamos National Laboratory² has been chosen for the purpose and successfully tested. The DLESE OAI Data Provider remains nonetheless within the scope of project investigation as an option available to partners to upgrade to more functional and heavyweight OAI participation models. For further details on the OAI Static Repository approach, see Section 3.

3. *The installation of an a OAI-compliant harvester;*

After a first successful installation of the DLESE OAI Service Provider, the ARC Cross Archive Search Service³ has become the preferred harvesting solution for the current phase of the project. This is mostly due to database-driven, out-of-the-box discovery service that ARC makes available. All the unqualified Dublin Core metadata made locally available at partners is currently searchable via the ARC installation.

¹ <http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm> (beta version).

² <http://libtest.lanl.gov/registry.htm> .

³ <http://arc.cs.odu.edu> .

4. *Further dissemination activities (a column for Library Review on OAI developments has already been arranged);*

During November and December 2003, the objectives and methodology of HalRST have been disseminated through a number of talks presented at internal research groups at Strathclyde University. Some of the IR-oriented research perspectives developed within the project have also been presented at the last eFair cluster meeting in November 2003. A journal column comparing the harvesting and distributed searching models from a technical perspective has been accepted in December 2003 for publication in a future issue of Library Review and is currently being extended for submission to a peer-reviewed forum (yet to be decided). Finally, a presentation of some of the metadata-related hypotheses advanced within HalRST is scheduled for a CETIS Metadata SIG meeting early in March 2004.

3. Overall Approach

While the project approach remains largely unaltered, the terms of the internal relationship between the project and the University of Strathclyde have been further refined to more adequately reflect both available resources and project findings. On the basis of current experiences, in particular, it has become apparent that an operational IR service for the University of Strathclyde is largely beyond the limited resources currently available to the project. Perhaps more significantly, we believe that such service would be in fact premature and may potentially compromise the success of future IR initiatives.

As repeatedly noted in the literature – and then further corroborated by the experiences gathered at St. Andrews University and, outside the project, at other, more narrowly scoped FAIR projects – institutional commitment and strategy are both key factors for the success of an Institutional Repository. To date, the University of Strathclyde has not yet deliberated on a clear strategy, or in fact commitment, to support the launch and maintenance of an operational IR service. We feel that, without authoritative plans for functionality, scope, access policies, etc., the scope for generalised advocacy and data collection plans can only achieve limited success.

Accordingly, the Strathprints pilot offers a minimal interpretation of an IR service focused around the core functionalities of resource discovery and resource deposit. Similarly, the content policy of the pilot focuses on the research, management, and administration output of the CDLR and thus ensures that – by offering access to a compact, varied, and yet functionally consistent collection – the pilot may be immediately useful to an institutional sub-community of users. We expect that, in turn, a restricted community of users will offer high degrees of control for the purpose of pilot testing and evaluation. The overall goal is to provide the Strathclyde partner with a focused body of knowledge on the design of IR-oriented digital services which may be eventually considered for the definition of a future full-blown IR initiative. At the same time, we expect that, for its diversity, this approach will allow us to report to JISC on issues that are not normally encountered by other FAIR projects.

Given the narrower scope of the pilot, the technical work on the eprints.org code base has acquired further significance. System soundness, flexibility, and most significantly extensibility have now become increasingly important properties to offer to Strathclyde's future use, customisation, and extension of the pilot implementation. The current system is now considerably autonomous with respect to the original installation, for it relies on completely reengineered modules for – among other components – (i) the configuration layer (which is now far more flexible and fully XML-based), (ii) the Web control layer (which is based on a unifying notion of data views and offers templating and screen-flow management), and (iii) the data model layer at the foundations of the system. For the targeted functionality (e.g. search and deposit), the current system offers significant more modularity than the original system, which results in: (a) much smaller and manageable code (savings up to 80%), and (b) better support for both programming and non-programming skills that may be made available to a full-blown IR initiative.

4. Project Outputs

In the previous report, we documented delays to the project schedule due to the uneven rate of development across inter-sector and intra-sector partners. We largely motivated the delays with the genuine difficulty of appointing local metadata officers with the required combination of IT and managerial skills. These obstacles have been now overcome by accommodating a mixture of

technical and organisational solutions within the envisioned proxy-based model of collaboration between partners and the leading institution.

In early December 2003, Stephen Winch from SLIC has been seconded to the project as the joint metadata officer for the GCG group and John Wheatley College. During the month of December, the officer has been inducted on the goals of the project and the details of the interface between the FE partners and the leading institution. In January 2004, the first sample of metadata records from the FE partners has reached the leading institution where it has been rapidly made available to the HalRST installation of ARC – as well as any other harvester – through what, to the best of our knowledge, is one of the earliest applications of the Static Repository framework recently specified by the Open Archive Initiative and briefly introduced below. A similar approach has also been used to expose the metadata records previously received from Napier University, so that all project partners are currently acting – whether directly or by mediation – as data provider in the OAI community defined within the scope of the project.

OAI Static Repositories (SRs) are intended to further ease the participation of data providers to OAI-based distributed digital library services. In the typical OAI-based scenario, a data provider exposes its metadata through a software layer normally layered on top of some pre-existing content management system – e.g. the eprints.org system – which the provider already employs to satisfy local services, such as a local resource discovery or resource deposit. In other words, OAI servers are commonly found as distribution-oriented features of existing systems

This scenario, however, does not adequately serve the need of data providers that want to expose metadata without having – and often wanting – an existing system of local services already in place. In this case, providers need to manually install an OAI server capable of deriving metadata records from whatever back-end the provider has or makes available for the purpose (e.g. an Access database, plain files in the file system, etc.). However, re-usability of off-the-shelf OAI servers is difficult, for different back-ends require different OAI servers, or different customisation of a single OAI server. In both cases, OAI participation is predicated on development costs that the data provider may not wish to afford.

A Static Repository is meant to simplify this scenario by: 1) standardising the back-end to a plain XML file, and 2) allowing a remote OAI server, called an OAI gateway, to use the XML file and act as a mediator for the data provider. Specifically, the provider needs only to populate and maintain an XML file of metadata records, after having registered it with the gateway. Harvesters may then talk to providers via the gateway, which uses the registered XML file to respond. The task of the provider becomes then that of populating the XML file with records in the right format, as they become available. Given the initial focus on unqualified Dublin Core, this may require translations from the native format of the metadata in the XML encoding of Dublin Core dictated by the OAI specs. As our testing confirmed, these processes are relatively inexpensive compared with the development required to set up or customise a full-fledged OAI server.

Overall, the static repository approach suits the case of metadata collections which do not change too frequently, are not too big (up to 5000 records, approximately), and cannot be managed with high implementation and maintenance costs. We have found that, in this phase of the project, the collections and resource models available at Napier University and our FE partners identify exactly this type of scenario. Both partners have provided us with MARC records which either originate from pre-existing collections (e.g. Napier's records) or have been recently created with remote cataloguing facilities provided by OCLC (e.g. Connexion for the records contributed by the FE partners). We have then mapped the MARC records to the OAI encoding of unqualified Dublin Core (using the freely available MarcEdit) and then uploaded them into two dedicated Static Repositories via per-partner Perl scripts specifically developed for the task. Subsequently, the Static Repositories have been registered with the experimental gateway service offered by the Research Library of the Los Alamos National Laboratory to be finally harvested by our ARC installation. The solution developed for the first sample of records has been designed to be reused for incremental repository updates and the process has been streamlined in collaboration with partners.

Meanwhile, St. Andrews University has further populated the local eprint repository service previously launched. At this stage, progress is slow and the experience gathered is not dissimilar from that collected and disseminated by similarly focused FAIR projects, namely the reluctance of faculty and

other interested stakeholders to convert an initial interest in the aims of an Institutional Repository into a more substantial engagement. Accordingly, content gathering cannot yet rely on self-archiving and thus retains traditional high costs and lack of homogeneity. In turn, this prevents the formation of a user base and thus the tangible indicators of usefulness which are required to stimulate the allocation of further resources. Indeed, this rather common scenario has strongly contributed to the desire of differentiating the approach of the Strathprints repository along the lines discussed in Section 2.

5. Project Outcomes

See Section 2, Section 3, and Section 4.

6. Stakeholder Analysis

See Section 3 and Section 4.

7. Risk Analysis

See Section 3 and Section 4.

8. Standards

See Section 3 and Section 4.

9. Technical Development

See Section 3 and Section 4.

10. Intellectual Property Rights

The project Steering Group has agreed that the model IPR agreement produced by JISC is too detailed for the purposes of the project, and contains much that is irrelevant. We are currently working on a simpler, more focused IPR agreement for circulation and ratification by the project partners.

Project Resources

11. Project Partners

No changes to the institutional project partners or subcontractors are to be reported.

12. Project Management

During this reporting period, the following changes to project staff are to be reported:

1. as mentioned in Section 4, Stephen Winch from SLIC has assumed the role of HaIRST joint metadata officer for the Glasgow College Group and John Wheatley College;
2. Paul Cunnea has been replaced by Sara Brown and Lynn Corrigan as the HaIRST contacts for Napier;
3. the Steering Group has experienced a genuine difficulty to meet and it has agreed to conduct most of its business by email, including: circulation of regular progress reports for comments and queries, notification of draft reports and other project materials made available on the project website, and the possibility of requesting centrally coordinated face-to-face meetings.

13. Programme Support

Project management is constantly in contact with the Programme Manager and other projects within the project's cluster via the FAIR and eFAIR mailing lists, eFAIR biannual meetings, the annual Joint Programme Meetings, and other events (e.g. the last OAI3 Workshop in Geneva). Particularly close is the cultural and technical exchange with the Daedalus project, also based in Glasgow.

14. Budget

Period reported on: 1/8/03-31/1/04

Total JISC Grant: £175,000

Co-funding: £20,000 (donated by CDLR)

	Forecast budget for this reporting period (from project plan)	Budget for this reporting period (including any underspend or overspend)	Spend for this reporting period	Balance for this reporting period
Staff (<i>list all staff with FTEs and salary scale range</i>)				
Grade 2A Researcher FT	£15000	-£896.66	£15885.75	-£16782.41
Part time systems installation support and training	£5000	£15066	0	£15066
Travel & Subsistence	£1000	£751.09	£302	£449.09
Equipment (<i>items over £10k</i>)				
Dissemination activities				
Evaluation activities				
Other				
<i>List headings as in project plan budget</i>				
Total	£21000	14920.43	16187.75	-£1267.32

Funds allocated for part time systems installation support and training paid for the services of a FT Grade 2A Researcher with relevant training over the initial four months of the project. The overall overspend of £1267.32 can be attributed to slightly higher than expected staff costs over this initial phase.

Detailed Project Planning

15. Workpackages

As per project plan, this reporting period has concentrated primarily on the deposit (workpackage DSI) harvesting (workpackage H) and discovery (workpackage DS1) of DC-encoded metadata records, and on the definition of a pilot organisational and technical framework for supporting those activities. In the process, progress has naturally been made in the other ongoing workpackages, 'Involvement, Advisory Service, & Collection Development' (IACD), 'Cultures, Policies, Strategies & Structures' (CPSS), 'Exit Strategy', and 'Dissemination'. Key activities, methodology, and objectives are discussed in Section 2, Section 3, and Section 4.

Objectives for the next reporting period concentrate on the second, post-DC phase of metadata definition (workpackage MD II), and the refinement of the related processes of resource deposit and discovery (workpackages DP II and DSII). We have begun investigating a range of potential extensions to the thin layer of current metadata agreements. Clearly, DP II and DSII add further technical constraints to the range of options identified in MD II as suitable to partners. Similarly, the future need of further disclosing the harvested records both via hierarchical harvesting and distributed searching – thereby turning the pilot harvester into an aggregator (workpackage 'Disclosure' – is to be taken into account early in the definition of richer metadata agreements and more costly deposit and disclosure processes. For all these reasons, we are and will be seeking a unified strategy to the issues raised by each of the previous workpackages.

No changes to the current version of the project plan are to be reported.

16. Evaluation

The project team has met with Peter Brophy of CERLIM, who will be carrying out an evaluation on behalf of HAIRST. It has been agreed that we do not have sufficient resources to carry out a meaningful evaluation of the project as a whole, and that we should focus instead on an area of the project which will benefit from a more detailed formative evaluation. Several possible areas have been identified; these will be discussed with the management group, after which detailed plans will be drawn up for an evaluation exercise to be carried out later this year.

17. Quality Assurance

As reported in Section 3, the implementation of Strathprints adheres to a number of software engineering principles, which have been presented separately to JISC in October 2003.

Sample content is now being added to the system for testing and demonstration purposes. Metadata for sample content is standards compliant but will not be subjected to QA procedures at this stage. Work on quality assurance for metadata creation within distributed systems is being carried in parallel with the HAIRST project; QA procedures based on the findings of this work will be put in place at an appropriate point in the development of the HAIRST system.

18. Dissemination

- F.Simeoni, *The case for Metadata Harvesting* (preprint), Library Review, (to be published)
- For a brief discussion of where HaIRST fits in the JISC Information Environment model see Dunsire, G. & Macgregor, G. *Clumps and collection description in the information environment in the UK with particular reference to Scotland*. Program: electronic library and information systems, v.37 no.4, 2003.;
- HAIRST is given as an example in the key prompt "Participation in national or local initiatives, organisations and professional activities which encourage resource sharing", of the quality indicator "Staff have an appropriate awareness and involvement in resource sharing initiatives" of Element 1: Learning resource organisation, in [Resources & services supporting](#)

[learning: a service development quality toolkit](#). The involvement of Glasgow Colleges Group in the HAIRST project was given as evidence of prior experience when bidding for the project which produced the toolkit.”;

As already mentioned in Section 2, a presentation of some of the metadata-related hypotheses advanced within HaIRST is scheduled for a CETIS Metadata SIG meeting early in March 2004.

19. Exit/Sustainability

The project has recently submitted a response to the Fair sustainability questionnaire. The lead site has a policy of preserving project websites, appropriate deliverables and supporting documentation for two years after the end of a project, after which material will be archived offline. Appropriate deliverables such as reports will continue to be available online for as long as they are relevant. The project plan seeks to create a sustainable and scalable service, and pilot and testing components are created as the project progresses with these aims in mind.