

SHERPA DP

Final report of the SHERPA DP project

Gareth Knight, Sheila Anderson

Contact: Gareth Knight (gareth.knight@ahds.ac.uk)

29th March 2007

Acknowledgements

The SHERPA DP project was funded by JISC and CURL under the programme, 'Supporting Digital Preservation and Asset Management in Institutions' strand. Further information on the strand is available at: http://www.jisc.ac.uk/index.cfm?name=programme_404.

SHERPA DP was led by the Arts & Humanities Data Service at King's College London and the University of Nottingham, and the project partners were the University of Glasgow, University of Edinburgh, the White Rose Consortium, and the London Leap consortium. Further information may be found on the project web site (<http://www.sherpadp.org.uk>).

We are grateful to both JISC and CURL for their support for this project. The CURL contribution provided the necessary funding for the SHERPA partner repositories who took part in the project.

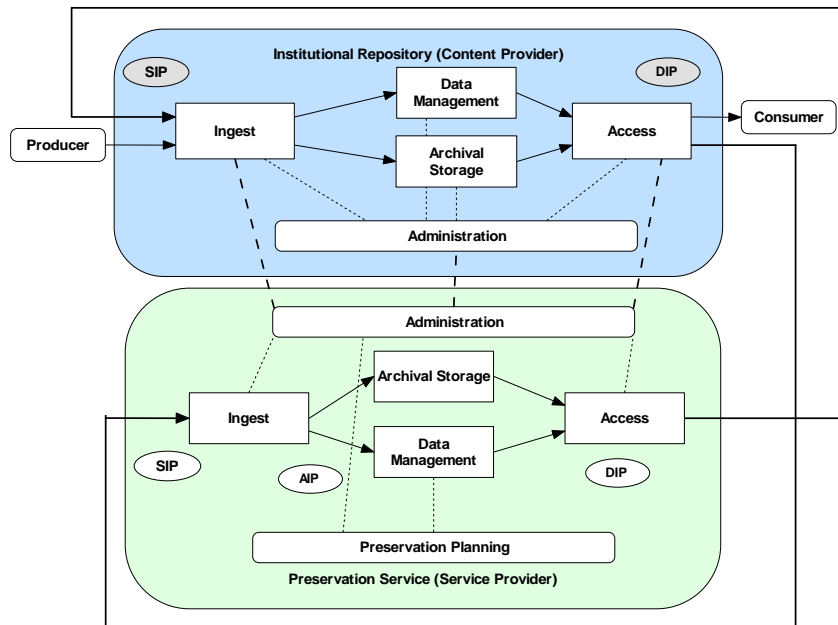
Table of Contents

Acknowledgements	1
Table of Contents	1
Executive Summary	2
Background	3
Aims and Objectives.....	4
Methodology.....	5
Implementation.....	5
Outputs and Results.....	7
Outcomes	7
Conclusions.....	9
Implications.....	9
Recommendations	9
References	11

Executive Summary

The SHERPA DP project (2005 – 2007) investigated the preservation of digital resources stored by institutional repositories participating in the SHERPA project. An emphasis was placed on the preservation of e-prints – research papers stored in an electronic format, with some support for other types of content, such as electronic theses and dissertations.

The project began with an investigation of the method that institutional repositories, as Content Providers, may interact with Service Providers. The resulting model, framed around the OAIS, established a Co-operating archive relationship, in which data and metadata is transferred into a preservation repository subsequent to it being made available.



The model identified two types of institution – Content Providers and Service Providers that perform different tasks in the workflow. The institutional repositories participating in the project serve as content providers, taking responsibility for accepting data and making it available to their user community. The service provider, in turn, takes responsibility for the long-term management of the digital objects, essentially serving as a centralised “dark archive”. The Arts & Humanities Data Service produced a demonstrator of a Preservation Service, to investigate the operation of the preservation service and accepted responsibility for the preservation of the digital objects for a three-year period (two years of project funding, plus one year).

The most notable development of the Preservation Service demonstrator was the creation of a reusable service framework that allows the integration of a disparate collection of software tools and standards. The project adopted Fedora as the basis for the preservation repository and built a technical infrastructure necessary to harvest metadata, transfer data, and perform relevant preservation activities. Appropriate software tools and standards were selected, including JHOVE and DROID as software tools to validate data objects; METS as a packaging standard; and PREMIS as a basis on which to create preservation metadata. These were selected as the best available tools for the purpose of the project. If better tools are developed in the future, these applications may be removed and the replacements integrated into the service environment with little disruption.

A number of requirements were identified that were essential for establishing a disaggregated service for preservation, most notably some method of interoperating with partner institutions and

the establishment of appropriate preservation policies. Institutional repositories operate different types of repository software, store different types of digital content or metadata that must be considered. It would be extremely limiting for a Preservation Service to support only a single type of repository or content type. In its role as a Preservation Service, the AHDS developed a repository-independent framework to support the EPrints and DSpace-based repositories, using OAI-PMH as common method of connecting to partner institutions and extracting digital objects.

The SHERPA DP project has developed a practical and cost effective alternative to performing preservation activities in the institutional repository. In the coming year the demonstrator will be further developed to provide a commercial preservation service provider for institutional repositories.

Background

The JISC-funded Feasibility and Requirements Study for Preservation of E-Prints (James et al, 2003) argued that there is a unique window of opportunity to address the preservation requirements of repositories at the beginning of their adoption rather than leaving it until the lack of preservation management becomes an issue and content is no longer accessible. A key recommendation of the report was the establishment of a repository infrastructure based upon the OAIS reference model and a practical investigation that includes implementation of preservation practices (p.56, James et al, 2003). It noted that, in the absence of staff and services with practical digital preservation skills and expertise, a sensible approach may be to disaggregate functions and activities in the OAIS Reference Model and seek collaborative arrangements in which different repositories and specialist services take responsibility for different functions.

The collaborative model proposed for the SHERPA DP project took advantage of the skills provided by institutional repositories in the SHERPA consortium and the preservation expertise of the Arts and Humanities Data Service (AHDS). By extending this collaboration into a full preservation service the project removes from each individual institutional repository the burden of adding a preservation layer to their repository, and the need for them to seek to employ scarce preservation management skills and expertise. It was recognised that institutional repositories often have priorities that must be achieved in the short term – the embedding of their digital repository in the wider institutional infrastructure, and the collection and distribution of research data to the research community. At their current stage of development, many institutional repositories lacked sufficient time and resources to make long-term decisions on the management of the data in their repository. In planning and working on the project, we wished to assist IRs through the modelling and development of a third-party preservation service that would consider these requirements and provide appropriate guidance. The preservation service would take a uniform approach to preservation across institutions that would compliment and enhance the existing operation of the institutional repository.

The AHDS worked with representatives from Edinburgh University, Glasgow University, and Nottingham University, as well as the consortiums of the White Rose Research Online (Leeds, Sheffield, and York) and London Leap to develop a collaborative model for preservation and test it in a practical environment. Through subsequent work, we developed a cross-repository infrastructure that enabled Fedora to interact with EPrints and DSpace-based repositories.

The methodology used by the SHERPA DP project has similarity with other projects that were funded under the JISC 4/04 funding strand. The PRESERV project, managed at the University of Southampton, also investigated the development of preservation services in the institutional repository sphere. The project takes a different, but complimentary approach, considering the type of information that may be stored by an institutional repository itself that may prove useful. The Repository Bridge project also has some similarities through an investigation of the interaction between DSpace/EPrints and Fedora as a method of preserving electronic theses.

Aims and Objectives

The aims and objectives, as listed in the project proposal are as follows:

Demonstrate a collaborative model using the OAIS reference model that brings together local repositories with national services

1. Use the OAIS reference model to develop a persistent preservation environment for the SHERPA consortium, assigning rights and responsibilities and establishing protocols and work flow processes that will ensure the long-term preservation of the repository content.
2. Explore the use of METS as the framework for packaging and transferring metadata held within the institutional repositories, including the preservation metadata created by the preservation service.
3. Establish a coordinated set of protocols and software to be implemented as a working preservation service for a group of institutional repositories.
4. Explore the use of open source software and tools to add functionality to and extend the storage layer of repository software applications.
5. Draw together the experience gained into a Digital Preservation User Guide that will complement the 'The Preservation Management of Digital Material Handbook' created by Maggie Jones and Neil Beagrie, and act as a practical user guide to implementing this type of preservation environment

Methodology

The methodology outlined in the proposal and subsequently developed in the project was based on the premise that it was not cost effective for all institutional repositories to perform preservation activities, and to develop the range of expertise necessary to do so. Instead, the proposal was to develop a shared service that would bring together institutional repositories with a national service that had significant preservation expertise. Although there was recognition that specific actions necessary for preservation (e.g. choice of appropriate file format) must be performed at the point of submission, the approach taken by the project was to limit the amount of preservation actions that the institutional repositories must perform. It proposed the development of a disaggregated service model, in which preservation services are provided by a third-party. The repository operated by the Service Provider serves as the hub for preservation services.

The focus of work in the early stages of the project was the creation of a detailed diagram, framed around the OAIS Reference Model that demonstrated the established the potential interaction between the Content and Service Provider in a distributed environment. It was influenced significantly by the operation of institutional repositories themselves – the workflow they have introduced, the repository software in use, the type of content that is stored, and the method it is made available. Several models were developed, eventually resulting in the adoption of a co-operating archive approach, where the Service Provider harvested the OAI output of the institutional repository, transferring digital objects into a preservation repository subsequent to them being made available. Subsequent effort was focussed on the development of an appropriate technical infrastructure to support the process and appropriate policies and guidance that would support the process. These coalesced into the creation of a Preservation handbook for institutional repositories. In addition, a business model was developed that outlined the costs associated with the operation of a Preservation Service.

The next step beyond this project is to develop a full business case to launch a preservation service that may operate beyond the confines of project funding. In addition, the shared services model has been created for application to institutional repositories that store e-prints and other textual data and, as a result, may have limited use for other types of repository. Subsequent work is necessary to consider a wider range of repositories and content types, in order to develop a more robust model. We hope to investigate both the full business model and the expansion of content types in future projects.

Implementation

To identify the method in which the institutional repository and Preservation Service would interact, we created a high-level that indicated the workflow of both institutions. The method of interaction changed throughout the project. In the initial model, the Preservation Service connected directly to the storage facility of the institutional repository and would extract data directly after submission by a depositor. However, this would require significant modification of the repository software and many partners were hesitant to support a communication that bypassed their existing security arrangement. A subsequent model was developed in which the Preservation Service extracted data from the institutional repository subsequent to it being released for wider use. The project proposal identified EPrints and DSpace-based repositories in the SHERPA project as the target for preservation, which established boundaries on the method in which content may be obtained. The project sought to find a consistent interoperable method of connecting to participating repositories. Several methods of transferring metadata and data between partner institutions were identified, including RSS feed and OAI output. However, they were not fully supported by the partner institutions and we eventually settled on the use of OAI-PMH. The use of OAI as a method to transfer data between the Content and Service Provider had the advantage of being widely supported by many different repositories. However, it meant that the data made available for preservation was the same as that output for dissemination – in most cases, PDF documents. Further work was necessary to extend the metadata output a full metadata record, including information that would not previously be made available. The DSpace community have produced a METS patch that allows the export of detailed metadata that is MODS conformant. However, an equivalent patch did not exist, until recently, for EPrints. The OAI output was subsequently modified to produce a 'dp' output, alongside the existing oai_dc output. The

modification allowed the Service Provider to harvest a record of all information stored by the institutional repository. The extended metadata set may be used for preservation and backup, the complete record may contain information, such as passwords, that is essential for preservation. Due to the method in which the OAI output is created in the EPrints software, this required manual identification of each database field and its inclusion in the OAI output for each repository.

In the disaggregated model envisaged for the project, the Service Provider is responsible for performing preservation activities. It was immediately clear that automated services would need to be developed to perform the various actions associated with preservation. In the absence of an appropriate 'out of the box' system, a framework appropriate to the project was developed. The framework, composed of various interoperable web services, is designed to perform preservation tasks utilising appropriate software applications developed for use in the wider research community. These services include: object validation (i.e. creation of checksums); format validation; format characterisation; creation of preservation metadata; as well as the management of disparate metadata types and multiple manifestations of digital objects.

A key component of the preservation workflow was the creation of an archival package suitable for preservation. The archival package contains all data and metadata created by the institutional repository and service provider. After some review of the various options available (MPEG21, METS), the project adopted the latter as a packaging method. The METS (Metadata Encoding and Transmission Standard) allows the storage of disparate metadata types, including resource discovery metadata, administrative information and preservation metadata. Similar to other projects, the PREMIS Data Dictionary was adopted as the basis for development of a preservation metadata scheme. This is supported by various format-specific metadata types, such as MIX for images and TextMD for text documents.

The project was not restricted by the software tools that it used and an initial investigation was performed to ensure those selected were fit for purpose. For the project, the software developer chose to adopt JHOVE and DROID as tools for format validation. Although both tools were developed to perform a similar task, albeit in different ways, it was considered that the use of both would be beneficial. Neither tool is able to provide information that is completely accurate – JHOVE is able to extract a wider range of information, but recognises just a small number of file formats, while DROID has been designed to perform the opposite function, of producing a small amount of information regarding a greater number of file types. It has also been identified that neither tool can produce results with complete accuracy – DROID makes tentative suggestions, if it is unable to identify the file format type or version; JHOVE also has well documented flaws that occasionally causes it to misidentify a file format if it contains unexpected characteristics. During the ingest workflow for the preservation service DROID performs an initial assessment of the digital objects, indicating the most likely format. The value produced by DROID, if it is one of the data formats supported, is passed to JHOVE as an attribute. The outputs produced by both tools are also compared, to ascertain that any notable differences are identified. In these circumstances, the error is reported to a member of staff who is responsible for reviewing the erroneous results.

The standards, data dictionaries and software tools serve as the building blocks for a Preservation Service. Through combination of the respective capabilities of each one, we began to construct a workable Preservation Service that fit the requirements of institutional repositories.

Outputs and Results

The project was able to develop an effective model for the provision of preservation services to institutional repositories.

A disaggregated Service Model built on the OAIS

We developed a high-level disaggregated service model that may be used to frame discussion of a Service Provider infrastructure, based on the OAIS. The extension of the OAIS Co-operating archives model defines two types of institution, Content Providers and Service Providers, supported by additional services, such as format registries and federated search facilities that comprise an Open Archival Information System.

A reusable framework for the provision of preservation service

We have developed and tested a framework of interconnected services that may be used to automate data extraction from a remote Content Provider and perform actions necessary to manage and preserve it in a controlled environment. The SHERPA DP implementation is architected to be modular and extensible. This facilitates changes or upgrades to be made to the repository software and associated Java-based services with minimal impact to the overall system. The Fedora repository, DROID, JHOVE and other applications may be replaced with other software with only small changes to the software code. The SHERPA DP preservation services automate the production of preservation metadata, validation of data objects, and other activities necessary to manage the data in the long-term.

Application of preservation metadata standards

We defined a refined set of metadata elements considered essential for preservation of e-prints provided by institutional repositories. The PREMIS Data Dictionary served as a basis on which to build an appropriate metadata scheme. The project identified refinements necessary to support the operation of the preservation service. The broad requirements of PREMIS were further supplemented by format-specific metadata (e.g. MIX for image metadata), as appropriate.

Business model for preservation in a distributed service model

The project has developed a business and cost model that may be used to cost preservation action. The model uses the lifecycle cost model developed by the LIFE Project as a basis, indicating how it may be applied to Service Providers offering a service to institutional repositories, or indeed anyone, who wish to assess the cost of preservation on a year-by-year basis. Costs are calculated on a three-tier basis – set-up costs; service costs; and exit costs that occur at different time period during the Service contract. Service costs serve as the focus of investigation, with the cost of storage, preservation and limited technical support being provided as examples of the method in which different levels of service will incur varying costs.

An investigation of metadata that institutional repositories currently store

An investigation of metadata that is currently stored by institutional repositories participating in the SHERPA DP project. The objective was to establish if it were possible to store disparate metadata provided by many institutional repositories in a consistent manner and identify if any metadata that may be used to support preservation activities was being produced. The investigation identified that basic metadata produced by IRs was, to some degree, consistent. However, repositories had developed specific-schemes that provided granular metadata for different types of research paper (e.g. journal articles).

Outcomes

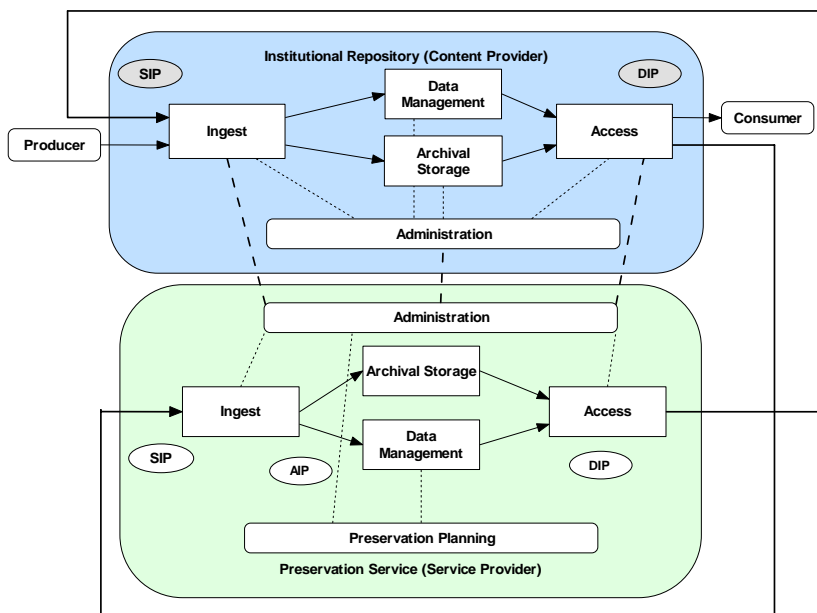
The amount and diversity of content stored by institutional repositories will inevitably increase over the next few years. The funding of projects, such as SHERPA DP and PRESERV demonstrate an increasing recognition that institutional repositories are beginning to consider the long-term management of digital objects for which they are responsible. The establishment of Service Providers may be considered one of many methods that preservation may be established and an alternative to the traditional understanding that it must be performed in-house. In the SHERPA DP project, we have

sought to gain a greater understanding of the type of services that may be provided, the infrastructure that must be developed, and a business model that may fund such services.

The most useful outcome of the project was the development of a Preservation Service demonstrator. By working with operational institutional repositories – members of the SHERPA consortium, the project gained a practical understanding of the services that a Preservation Service may perform and the method in which it may interact with partner institutions that serve as Content Providers. Most notably, the project identified different types of service, beginning at simple remote data storage and extending to full preservation, through migration and the creation of preservation metadata. Complimentary services may also be considered, such as metadata enhancement. We have sought to gain a greater understanding of the type of services that must be developed to support preservation function. A number of technologies and standards exist that support preservation functions, but work was necessary to combine the capabilities of these applications. Most notably, the use of JHOVE and DROID, in conjunction with the PREMIS data dictionary and METS (Metadata Encoding and Transmission Standard) provides a practical method of creating and storing metadata appropriate for preservation.

The work produced some surprising results, demonstrating the need for interoperability between different repository software and the need to consider different types of user. The project identified that OAI-PMH is sufficient to introduce a basic method of transferring data between repositories. However, the basic oai_dc is insufficient for extracting the required amount of information – institutional repositories may store essential information, such as passwords to remove printing protection, in administrative metadata that is vital for preserving the digital object. To supplement the oai_dc output, the OAI output module was modified to output a second set of metadata. The ‘dp’ output is tailored to the metadata stored by each institutional repository, making available a full record of each e-print. It is the ‘dp’ output that is subsequently harvested by the Preservation Service and used to transfer the associated data.

The disaggregated Service model for institutional repositories is likely to prove useful, allowing repositories to conceptualise and understand the interaction between an Institutional Repository and a Preservation Service Provider.



The development of a business model to support Preservation Services is also likely to benefit the wider community. The model builds on the investigation performed in the LIFE Project to provide a method of calculating year-by-year cost, associated with offering a preservation service for e-prints held in Eprint and DSpace repository software. The model uses a three-tier structure, establishing costs for establishing the partnership, including any software development costs (setup costs); Service costs that establish the year-by-year expense of service provision; and Exit costs. The

business model may prove useful for any institution that is seeking to establish a Preservation Service.

Conclusions

There is no single out-of-the-box solution to preservation. Long-term management requires the repository to consider the most cost effective method of accessing and decoding content stored in digital objects, and to decide the appropriate strategy. It is also evident that the location in which the preservation activities are performed is a matter for strategic decision making – preservation actions may be performed by a third-party Service Provider or by appropriate staff members inside the institutional repository. The primary requirement is that appropriate expertise and services exist to support these activities. The SHERPA DP project has produced a practical implementation of a Preservation Service maintained by a third-party institution. Through the development of the demonstrator, the project has identified a method of integrating preservation activities into the institutional repository, without disruption to the existing workflow or placing an increased workload on repository staff.

It may be concluded that preservation management is a feasible activity, through the combination of the PREMIS Data Dictionary, an appropriate metadata packaging standard (METS, or indeed MPEG-21), and appropriate software tools (JHOVE, DROID). Furthermore, preservation management may be achieved in a semi-automated manner, thereby supporting the development of sustainable Preservation Services.

Implications

The development of an operational Service Provider for institutional repositories is a practical solution to preservation. The research community may support a number of preservation services that operate in and are tailored to the requirements of different research communities and different organisational models. For example, a preservation service may be established to cater for the needs of different repositories in an institution, or different repositories that share a common research theme. However, the project, by definition, has considered preservation services only in the context of Institutional repositories. Further questions that may be addressed are: Can the disaggregated service model be applied to other types of repository? What are the implications for repositories that interact in different ways and use different technologies? The AHDS has bid for and received funding for a follow-on project, SHERPA DP2 that will address these questions. This is likely to result in the refinement and enhancement of the high-level and business model, through consideration of a wider range of repositories. The AHDS has also received funding for the 'MetaTools' Project that aims to develop a methodology for evaluating metadata generation tools; compare the quality of currently available metadata generation tools; develop, test and disseminate prototype web services that integrate the best metadata generation tools and functionality. The results of this project will feed into and improve the SHERPA DP shared services preservation model.

Recommendations

The need for preservation planning by institutional repositories remains important. Although Preservation Services may provide the infrastructure for preservation action to take place, the capabilities of a Service Provider to preserve content is limited by the data types provided by the partner institution. It is recommended that IRs consider the development of a preservation policy that encourages submission of digital research in the file format in which it was created. A distinction should be made between source format (the format that the author used to create their research data); the archival format intended for preservation; and a format intended for use by distribution to the wider community.

Further investigation into alternative options for the provision of preservation services is necessary, in particular to investigate models that are financially sustainable. The disaggregated Service model for preservation developed by SHERPA DP and PRESERV is one approach to the solution, but there are others. These include performing in-house preservation (supported by the various modules available for EPrints v3 and DSpace); and research into the LOCKSS project.

The project has confirmed the importance of interoperability between different types of repository. Although there is some degree of compatibility through oai_dc, the metadata that is supported in repository software is tailored to their own requirements and does not consider the possibility that an institutional repository will wish to migrate to other repository software. Further work is necessary to improve repository interoperability, particularly between Fedora, DSpace and EPrints. This may be achieved through a comparison of the various capabilities of each repository software may be performed to identify functionality that is currently lacking in a repository. E.g. repository software typically support oai-pmh, but may not support RSS. Work may be commissioned to develop software patches to support the required functionality. Alternatively, the esoteric features of each repository software may be standardised. The EPrints repository software, for example defines a large number of database fields in the base installation that are not defined in other metadata schemes. Work should be undertaken to document these terms and develop an appropriate metadata scheme.

References

Beagrie, N. & Jones, M. (2001). Preservation Management of Digital Materials: A Handbook. The British Library.

James, H, Ruusalepp, R. Anderson, S & Pinfield, S. (2003). Feasibility and Requirements Study on Preservation of E-Prints. http://www.jisc.ac.uk/uploaded_documents/e-prints_report_final.pdf

University of Nottingham (2006). SHERPA Consortium web site. <http://www.sherpa.ac.uk/>

The Library of Congress (2007). Metadata Encoding and Transmission Standard. <http://www.loc.gov/standards/mets/>

Consultative Committee for Space Data Systems (2002). Reference Model for an Open Archival Information System (OAIS). <http://public.ccsds.org/publications/archive/650x0b1.pdf>

MIT Libraries & Hewlett Packard (2007). DSpace Federation. <http://www.dspace.org/>

University of Southampton (2007). EPrints for Digital Repositories. <http://www.eprints.org/>

JSTOR (2006). JHOVE - JSTOR/Harvard Object Validation Environment. <http://hul.harvard.edu/jhove/>

Stanford University et al (2007). Lots of Copies Keep Stuff Safe
<http://www.lockss.org/lockss/Home>