

LOCKSS Quick Start Reference Guide

January 2012

Version 1.2

LOCKSS Quick Start Reference Guide	1
What is LOCKSS and what does it do?.....	1
Responsibilities.....	1
Once You Join	3
UK LOCKSS Alliance Communication	3
Installing LOCKSS.....	3
Reference Guide for Librarians	4
LOCKSS User Interface.....	4
1. Content Administration.....	4
<i>How do I add content for collection?.....</i>	<i>5</i>
<i>What content is available through LOCKSS?.....</i>	<i>5</i>
<i>How do I monitor the progress of content retrieval?.....</i>	<i>6</i>
<i>How do I remove content from a LOCKSS box?.....</i>	<i>5</i>
2. Access to Content	8
3. System Administration.....	8

What is LOCKSS and what does it do?

LOCKSS is a system for archiving a local copy of scholarly material such as e-journals so that the content can be accessed in the long term, even if institutional subscriptions change or the publisher ceases to provide access to the content. To assure the long-term stability of the content, a LOCKSS box communicates with other LOCKSS boxes using an automated 'audit and repair' mechanism so that the content preserved on each box does not suffer from bit-rot and degrade over time.

The LOCKSS software is designed to run on a box in your IT infrastructure while requiring the minimum of maintenance.

This document is aimed at new participating institutions and describes some of the planning and investment necessary to get up and running as part of the UK LOCKSS Alliance. Setting up and maintaining a LOCKSS box is an essential part of being a member, and requires a small outlay for human resources and hardware. Each participating institution is responsible for the provision of hardware, installation and maintenance of LOCKSS.

Responsibilities

Running a LOCKSS box requires staff input on a variety of levels and we outline rough responsibilities below.

- ***IT Support and Linux System Administration***
 - A member of staff is responsible for the installation and maintenance of the LOCKSS system, which runs on Linux, typically CentOS.
 - The LOCKSS installation CD results in a customized installation of CentOS, installing and configuring the LOCKSS software and setting basic system security levels.



- The LOCKSS software requires direct access to the Internet on a restricted set of ports (that is, avoiding any institutional proxy). The system administrator may need to negotiate the opening of these with the institutional Network Administration team.
- Maintenance is intended to be lightweight but we strongly recommend that **responsibility be explicitly assigned** to a member of staff, as institutional problems tend to arise when no one takes ownership of the LOCKSS installation. Most major maintenance activities arise in the event that either the preserved content fills the capacity of the hard disk, or in the event of a hardware failure. The software is updated regularly, and the upgrades can be performed automatically or manually.
- **Collection Specialist**
 - A separate member(s) of staff is responsible for configuring content for preservation in a LOCKSS box in accordance with local collection development policies. Periodically it will be necessary to configure the box to schedule the collection of newly released content, and to review the collections to ensure content is being collected correctly.
 - Content should be scheduled for collection through the LOCKSS user interface. Periodically, the collection specialist should confirm that content scheduled for collection is indeed being preserved as expected, diagnosing the reason why it hasn't been collected and contacting the EDINA helpdesk for further advice if needed.
 - Collection staff should register on the UK LOCKSS Alliance mailing list to monitor announcements regarding new content. When new content is made available for preservation that meets the institution's collection criteria, staff should schedule new content for collection where appropriate.
 - In order to remain abreast of recent developments and to highlight institutional requirements we recommend the collection specialist attends face-to-face UK LOCKSS Alliance user meetings and participates in online surveys.
- **Institutional Support and Administration**
 - We also recommend that someone, possibly at management level, takes overall responsibility for the allocation of collection specialist and IT support time. This same person will be responsible for ensuring that budget is available for membership renewal and hardware expansion or replacement.
 - We recommend that institutions coordinate a review of LOCKSS on at least an annual basis. We have found that where LOCKSS is regularly discussed within the institution, better use is made of the resource and institutional support continues.
 - In light of the above, we recommend institutions consider how to embed support for LOCKSS into institutional policy and job descriptions. This will help ensure the long-term participation in preservation initiatives, rather than relying on the good will of administrators and librarians (which may be lost during staff turnover).
 - Note that membership options are available from <http://www.jisc-collections.ac.uk/Catalogue/Overview/Index/879>. JISC Collections will send out renewal notices in the month of April (approximately).

Once You Join

UK LOCKSS Alliance Communication

- Relevant staff should be added to the UK LOCKSS Alliance mailing list. A list archive is available to subscribed members on the JISCMail site. Please email the list of staff names and their email addresses to edina@ed.ac.uk.
- A twitter account for the UK LOCKSS Alliance and related EDINA activities is maintained at http://twitter.com/EDINA_ejournals. If you are posting tweets about the UK LOCKSS Alliance, please use the hashtag #lockss.
- The public facing UK LOCKSS Alliance website lives at <http://www.lockssalliance.ac.uk/>. This page contains general information about the UK LOCKSS Alliance, including briefing updates (<http://www.lockssalliance.ac.uk/documents-and-publications/>) and outputs from steering committee meetings (<http://www.lockssalliance.ac.uk/steering-committee/>).
- We suggest new institutions get involved in the Steering Committee to shape the development of the LOCKSS approach. To find out more or to get involved please email edina@ed.ac.uk.

Installing LOCKSS

- Once you've joined your system administrator will need to bring online a LOCKSS box. For the latest instructions on how to do so, please visit: http://www.lockss.org/lockss/Installing_LOCKSS.
- Please note that before your LOCKSS box will operate correctly, the LOCKSS team must register its static IP address. Please email the IP address of your LOCKSS box to edina@ed.ac.uk.
- Guidance on hardware can be found in the installation instructions above. The use of RAID is recommended. Hard disk capacity is perhaps the most important part of a LOCKSS box. The LOCKSS Linux installation will attempt to partition your disks using software RAID, which improves the availability of your LOCKSS box by storing your data redundantly across multiple hard disks. If a single hard disk in the software RAID fails, the LOCKSS box will continue to provide content and the possibility of data loss is minimal. Without RAID, a hard disk failure will mean data loss and an extended outage period for your LOCKSS box until it can be rebuilt from scratch. A combination of RAID along with the preservation capabilities of LOCKSS ensures maximum availability (minimum impact of total disk loss) and maximum reliability (minimum probability of data loss).
- You can also build a LOCKSS box using virtual machine technology, such as a VMWare product, as long as you can allocate the resources to match the minimum requirements described in the installation guidance.
- When the installation is complete, the collection specialist will be able to add content to the LOCKSS box.
- If you need support either when selecting hardware or during installation, please email edina@ed.ac.uk. If preferable, we can phone you to provide one-on-one support.

Reference Guide for Librarians

- This section will describe the key operations a collection specialist is required to undertake to maintain a LOCKSS box.
- Staff should note that LOCKSS is intentionally lightweight and meant to be a simple activity scheduled within weekly routines. You may feel that you do not have much interaction with the LOCKSS box: as long as the content configured for collection is being collected, that's fine.
- An overview of LOCKSS can be found at: <http://www.lockss.org/lockss/How It Works>
- Documentation on the LOCKSS interface is available at <http://lockss.org/lockss/Cache Help>
- If you need help or support regarding the operation of LOCKSS, email edina@ed.ac.uk

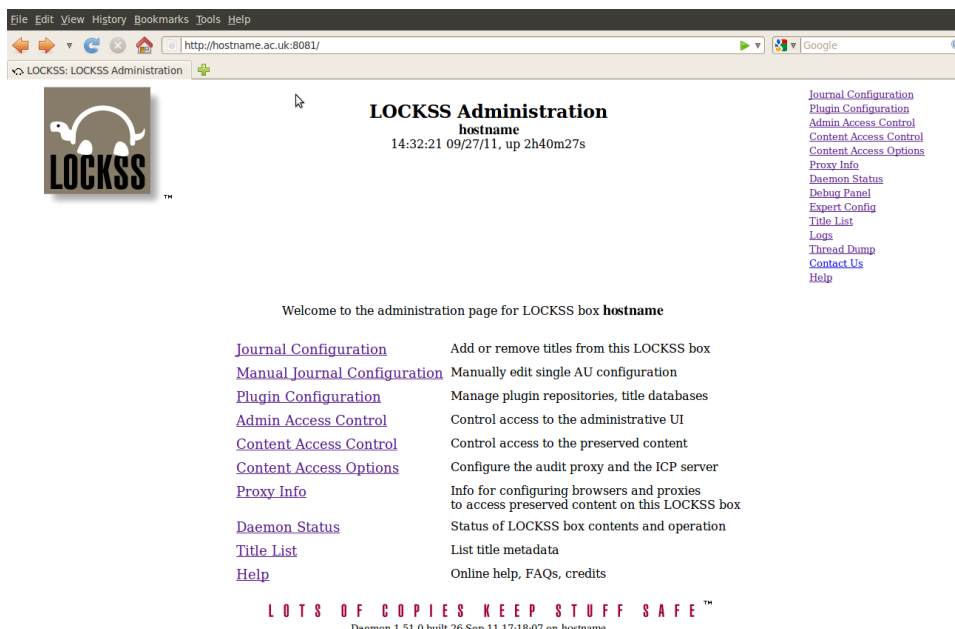
LOCKSS User Interface

The LOCKSS interface allows interaction with the LOCKSS box. We have broadly categorised three types of interaction a user may wish to undertake.

1. **Content Administration**
2. **Access to Content**
3. **System Administration**

1. Content Administration

Point your browser to the interface available at <http://hostname.ac.uk:8081/>, replacing the URL with that of your LOCKSS box, and enter your username (lockss) and password. This is the page from which all LOCKSS cache administration functions can be performed.



LOCKSS Administration
hostname
14:32:21 09/27/11, up 2h40m27s

Welcome to the administration page for LOCKSS box **hostname**

Journal Configuration	Add or remove titles from this LOCKSS box
Manual Journal Configuration	Manually edit single AU configuration
Plugin Configuration	Manage plugin repositories, title databases
Admin Access Control	Control access to the administrative UI
Content Access Control	Control access to the preserved content
Content Access Options	Configure the audit proxy and the ICP server
Proxy Info	Info for configuring browsers and proxies to access preserved content on this LOCKSS box
Daemon Status	Status of LOCKSS box contents and operation
Title List	List title metadata
Help	Online help, FAQs, credits

LOTS OF COPIES KEEP STUFF SAFE™
Daemon 1.51.0 built 26-Sep-11 17:18:07 on hostname

How do I add content for collection?

- To add content, follow the simple process outlined below:
 - (1) Choose **Journal Configuration** and then select **Add Titles**.
 - (2) Select relevant publishers from the list of collections and press **Select Titles** to proceed; you will have the opportunity on the following page to select specific titles from those publishers.
 - (3) Each entry on this list represents an Archival Unit (AU), which usually corresponds to a single volume of a single journal. Select the individual volumes you wish to collect in your LOCKSS machine (shift-click to select title ranges), and press **Add Selected AUs**.
- The LOCKSS daemon will shortly begin to collect the content and store it within your LOCKSS box. Note that scheduling content for collection does not guarantee collection: you must have an active subscription directly with the publisher and so it is necessary to periodically monitor the progress of content retrieval.
- As the quantity of material available through LOCKSS increases, we suggest that librarians add content according to their collection priorities by selecting content from publishers of interest only. Note that only the specific volumes you configure for collection are preserved, and you will need to continue to add new volumes as they are released.
- Watch for email announcements about newly available content and content that is due to become obsolete; this should inform how you configure your LOCKSS box.
- Administrators should keep journal configurations up to date in line with institutional subscriptions. Experience suggests this requires approximately two hours of a librarian's time per month.
- Further guidance on adding content is available at:
http://lockss.org/lockss/Adding_Content

What content is available through LOCKSS?

- The document at http://lockss.org/lockss/Publishers_and_Titles lists the set of content committed by publishers, listing both the titles made available for preservation and titles where the publisher has made a commitment but some technical work is still to be completed.
- In order to see the content currently available for preservation in your LOCKSS box, login to the interface and select the **Title List** feature available in the navigation menu. This provides two options: available and configured titles. Selecting 'available' titles generates a KBART-compliant report for the library listing all the titles and volumes currently available for preservation in LOCKSS. Selecting 'configured' titles generates a KBART-compliant report of the specific titles and volumes a library has configured for preservation in LOCKSS.
- The set of volumes available for preservation does not always extend to complete title runs. The reason for this can vary: either the publisher has not given permission for complete preservation or the LOCKSS team has not yet undertaken the semi-automated quality assurance process needed prior to release where every AU is tested to confirm completeness.



- You will not be able to collect and preserve content to which you are not permitted access (for example, where you don't have a subscription). Publishers control access to a permission statement hosted on the publisher's server, which LOCKSS must access in order to know which content to collect. IP authentication is performed before access is granted to this permission statement. Without access, the system is unable to access the permission statement and in turn is unable to complete a crawl.
- It will not cause significant problems if you attempt to add all available content, but you will receive many '*No permission from publisher*' crawl results. Establishing the lack of permission will inevitably take time which could be better used in collecting the content to which you do have access. Once you can identify the titles failing to collect, you may find it helpful to clean up your interface by removing them.

How do I monitor the progress of content retrieval?

- Select **Daemon Status**, and then choose **Crawl Status** from the drop-down. Content being retrieved for the first time will have a status of *New Content* in the Crawl Type column. If content is being re-crawled as a result of the LOCKSS polling mechanism, the Crawl Type status will read *Repair*. LOCKSS allows two concurrent crawls to take place at any one time.
- The Status column will indicate collection progress for an Archival Unit, with positive progress being either 'Active' or 'Successful'. A status of 'Pending' means the content is scheduled for collection but this has not yet taken place.
- Common statuses to watch out for that may indicate problems are:
 - **No permission from publisher:** LOCKSS first attempts to access a pre-defined permission statement, called a manifest page, before proceeding with a content crawl. This status means the LOCKSS machine cannot access the permission statement. This may mean you do not have a subscription to the journal and have been redirected to a login page. In some cases, the publisher's platform may have changed, requiring an update to the plugin.
 - **Can't Fetch Permission Page:** This most likely means you do not have appropriate rights to archive the content, and the remote server has responded to a LOCKSS crawl with a '403 Forbidden' error.
 - **Interrupted by Crawl Window:** Some publishers only allow LOCKSS crawls outside peak traffic hours. Content collection will resume at a later stage outside the peak times.
- To list the content that has been configured for collection on your LOCKSS box, use the **Title List** feature. After navigating to this screen, select the 'Configured' radio button, choose your desired output option, and press 'List Titles'.
- You can monitor the status of individual Archival Units in more detail by interpreting the status information in the **Daemon Status -> Archival Units** page. At first glance, this information can be overwhelming. The Archival Units page provides a detailed view into what is happening, but you generally only need to watch out for a few specific scenarios:
 - The Disk Usage column is a useful reference to indicate that the content has been collected (how large this is depends on the size of the AU, but generally several MB at minimum). If the Disk Usage is blank or 0, it is likely that the AU has not been collected.



- Under the Status column, the agreement percentage is used to compare content across boxes: a low percentage value most likely means that dynamic content is not being filtered correctly, but in less likely circumstances may also mean that you are not entitled to collect the journal. "Waiting for Crawl" indicates collection of the AU has not yet started, and will likely have a Disk Usage of 0 accompanying it.
- If the content has not yet been collected, the 'Last Crawl Result' column will be blank. Otherwise it will report one of the statuses reported on the Crawl Status page, as described above. 'Successful' is a healthy status that indicates a crawl has completed successfully.
- You will likely encounter some common combinations:
 - Content has been collected successfully if the following is encountered:
 - Content size = (big number)
 - Status = 100% agreement
 - 'Last Crawl Result' = Successful
 - It is likely that the content has been collected successfully but there is a problem with the filtering of dynamic content if the following are encountered:
 - Content Size = (big number)
 - Status = mixed % agreement
 - 'Last Crawl Result' = Successful
 - Content has not yet been crawled if the following are encountered:
 - Content Size = 0
 - Status = 'Waiting for Crawl'
 - Last Crawl Result = (blank)
- If you click into the status page for an individual AU, all known information about the AU will be presented, including a list of the files archived.
- It is worth paying particular attention to the 'Has Substance' flag, which reports whether an AU has collected content of substance (ie. full-text HTML or PDF articles).
- Finally, it is worth noting that many of the warnings and errors encountered are transient and/or are displayed to help the LOCKSS team monitor development. The most significant errors to watch out for are crawl errors that report '*No permission from publisher*' or '*Can't fetch permission page*' as these errors prevent collection of content, or 'Has Substance' = No, as this indicates that content is not being collected correctly.
- In a future iteration, we will document how to test access to preserved content to confirm the content that has been preserved in your box.
- If you encounter a status and think it may represent an error, please contact the EDINA helpdesk at edina@ed.ac.uk.



How do I remove content from a LOCKSS box?

- Select **Journal Configuration -> Remove Titles**. Titles are categorised into collections, so select the relevant collections and click **Select Titles**.
- Now, select the AUs you wish to remove and select **Remove Selected AUs**.
- This process *only removes configuration information* from the LOCKSS system and does not delete the content from the hard disk. If you wish to delete content to create some free space on your hard drive please contact LOCKSS support. This two-step process is designed to prevent an administrator from accidentally removing content via the user interface (as after deletion it might not be possible to restore).

2. Access to Content

This section will describe how to integrate a LOCKSS box with link resolver systems. The work to make this possible is nearing completion and we are currently negotiating with the major link resolver vendors to finalise the integration.

[Coming Soon]

3. System Administration

System administration tasks will allow a user to control who has access to the user interface and the preserved content, to enable various optional or beta (test) features, to configure plugins and proxies, and to view logs and detailed status information.

[Coming soon]