TECHNICAL SPECIFICATIONS

The MetaArchive Cooperative

2015-01-20

This document provides an overview of the current recommendations and requirements for administering a preservation cache for the MetaArchive Cooperative, including staffing and hardware. This document was last revised on January 20, 2015.

1. Skills Recommendations

Our Member institutions have identified three key roles that they assign to their local staff members in order to effectively run their caches and prepare their content for ingest into the MetaArchive Preservation Network: cache administration, plugin development, and data wrangling. It is possible that a single technical staff member may be responsible for one or more of these roles. For example, the staff member assigned to plugin development and data wrangling may be the same person (if not they should, if possible, work closely together). The MetaArchive staff regularly provides training for these roles. The anticipated time commitments, skill-sets, and common tasks, as based on current Member experiences, are documented below.

1.1 Cache Administration

Time required	Between 2-10 hrs/mo (average 5 hrs/mo)
Required skills	Basic level administration of UNIX-based platforms; ability to run and maintain servers, proxy servers, firewalls, and RAID configurations
Helpful skills	Knowledge or experience in digital libraries or library IT
Common tasks	Installing a MetaArchive-LOCKSS cache; assisting with content ingest; performing updates for the cache; monitoring the cache; documenting procedures

1.2 Plugin Development for Content Ingest

Time required	Between 2-25 hrs on 1st plugin (average 15 hrs) Between 1-6 hrs on additional plugins (average 3 hrs)
Required skills	Familiarity with XML; familiarity with file structuring on widely used platforms (Windows/Unix/Linux); understanding of regular expressions; solid understanding of web technologies (e.g., browsers and plugins)
Helpful skills	Familiarity with metadata standards; programming experience; VMWare
Common tasks	Writing/testing plugins

1.3 Data Wrangling

Time required	Between 15-40 hrs per collection (depends on existing repository solution)
Required skills	Familiarity with file structuring on widely used platforms (Windows/Unix/Linux); basic understanding of web technologies (e.g., web servers)

Helpful skills	Experience with re-formatting digital content and media; experience with archival appraisal and selection methods; familiarity with metadata standards and cataloging; programming experience; database management
Common tasks	Creating manifest pages; re-naming and re-sizing files; preparing web servers to deliver content; creating collection level metadata

2. Operational Requirements

2.1. Preparing the Technical Environment

- Member system administrators (or designated technical staff members) should have ready access and authorizations to access their MetaArchive-LOCKSS caches to the fullest extent possible.
- Member system administrators (or designated technical staff members) should have the ability to effectively coordinate with staff members that are responsible for configuring institutional firewalls to allow MetaArchive-LOCKSS caches to participate in the MetaArchive Preservation Network.

2.2. Necessary Cost Expenditures

- Designated Members must purchase hardware that meets the specifications below to operate a MetaArchive-LOCKSS cache.
- Member institutions must be prepared to adequately staff the necessary roles (see Skills Recommendations above) to implement and maintain a MetaArchive-LOCKSS cache throughout the period of their membership.

3. Support and Equipment Life Cycles

3.1. Member Obligations

- Members agree to purchase and maintain the necessary technical hardware (as described below) required to operate a MetaArchive-LOCKSS cache throughout their membership period.

- Members also agree to update their technical hardware on a three-year cycle using the current MetaArchive-LOCKSS cache specifications. This ensures that all of the MetaArchive Preservation Network's equipment is replaced in a manner consistent with industry best practices. This rolling cycle also enables the Cooperative to avoid network-wide uniformity of technical hardware.

- Unless otherwise negotiated with the Central Staff and Steering Committee, Members also agree to repurpose their technical hardware for the MetaArchive test network when it has reached its three-year end-of-life cycle so long as it is still functioning. Caches re-purposed for the MetaArchive test network will not be subjected to any recovery actions in the event of disk failures. If a member has retired more than one cache, only its most recently retired cache should be part of the test network; older caches may be repurposed as needed by the member for other functions.

3.2. Replacement Option

- In the case of catastrophic circumstances, Members have the ability to request technical and financial assistance with the restoration of a preservation site's caches, software, and collections by the MetaArchive Cooperative. These requests will be reviewed and, at the discretion of the Steering Committee, either approved or denied.

4. Technical Specifications

<table>
<tr><th>Required</th><th>Recommended</th><th>Notes</th></tr>
<tr><td>Machine architecture capable of running an operating system with a Sun or Sun-compatible JVM</td><td>Intel Core i7 processor and Intel Core i7 Extreme Edition compatible</td><td>Quad Core Processors based on this architecture are the current standard for new LOCKSS caches.</td></tr>
<tr><td></td><td>Rack-mountable server chassis</td><td>Not a hard requirement, but standard for most PLNs.</td></tr>
<tr><td>At least 8 GB of RAM</td><td>8 or more GB of RAM</td><td>At a lower level, excessive swapping may occur.</td></tr>
<tr><td>Standalone Server Storage Space
• 32 TB Raw
• ~22 TB Usable
SAN Storage w/ hardware RAID
• ~36 TB max usable
• Can start at a negotiated minimum and scale up to max usable over 3-year term as required and requested</td><td>32-48 TB Raw
32 TB of raw space should be considered the minimum for a new production cache.
An optimal standalone server configuration would involve a 32 TB base purchase using 2U enclosures with 12 bays (i.e., eight 4 TB disks). This would facilitate expansion to 48 TB as needed.</td><td>Use of RAID is required. Software RAID 5 is an option for 32 TB configurations but may require more system administrator time/attention when it comes to disk replacement/syncs. RAID 6 is preferred.
<table>
<tr><th>Raw</th><th>Software RAID 5</th><th>RAID 6</th></tr>
<tr><td>32 TB</td><td>~22 TB usable</td><td>~22 TB usable</td></tr>
<tr><td>48 TB</td><td>~33 TB usable</td><td>~29-36 TB usable*</td></tr>
</table>
* Upper limit on usable space (36 TB) is likely available if 12 (4 TB) disks are purchased up-front. Assumes fewest number of configured file systems.</td></tr>
<tr><td>RPM-based Linux Distribution</td><td>CentOS 6.x</td><td>The highly recommended distributions are 1) CentOS; 2) Red Hat Enterprise Server.
RHEL 6.x is also possible so long as the member purchases their own RHEL license</td></tr>
<tr><td>Java Virtual Machine</td><td>OpenJDK 1.7</td><td>LOCKSS software requires a Java Virtual Machine</td></tr>
<tr><td>LOCKSS software</td><td></td><td>The Cooperative provides an RPM repository with the current version that is in use. All caches run the same version of the LOCKSS daemon and we synchronize any upgrades.</td></tr>
<tr><td>LOCKSS caches should be physically secure, accessible only to appropriate staff members, and climate controlled</td><td></td><td>Temperature of 40-80˚F and humidity of 10-80%.</td></tr>
<tr><td>A firewall should be used to block access to all unused ports</td><td></td><td>Any services not required for functionality and maintenance of a cache should be protected by a firewall. Appropriate ports to leave open include ports used by: 1) the Meta-Archive Preservation Network, and 2) the administrative servers to communicate with the node itself. For Unix machines, ssh should be considered an appropriate method of remote access. Telnet and VNC are not considered secure methods.</td></tr>
<tr><td>User accounts should be kept to a minimum</td><td></td><td>Only the system administrators who need to maintain the server should have user accounts. These accounts should have strong passwords.</td></tr>
<tr><td>There should be no direct remote administrative access</td><td></td><td>In the case of SSH, this constitutes disabling root logins.</td></tr>
<tr><td>Security patches should be applied promptly</td><td>Always follow local site security update policies.</td><td>For users of RPM-based Linux distributions, this can be achieved by periodically running yum or up2date for software updates. In general this is handled by the creation of a cron job when the server goes through the Kickstart procedure.</td></tr>
</table>