Disaster Recovery

1. Disaster Recovery Steps by Disaster Type

2. How to Recover from an Amazon Service Outage (Chris)
   - 2.1 Many Items Are Now Permanently in Place to Deal with DR
   - 2.2 There are NEW IP Addresses
   - 2.3 When a Disaster Recovery is Taking Place

3. How to Recover from an Amazon Security Breach (Chris)
   - 3.1 Analysis
   - 3.2 Recovery

4. How to Recover from Overdue OS and Software Updates (Chris)
   - 4.1 Profile the system in question
   - 4.2 Reboot
   - 4.3 Take a snapshot
   - 4.4 Update rpm, yum and related repositories
   - 4.5 Update everything else

5. How to Recover from Cache Disk Failure(s) (Clay)

6. How to Recover from Cloud Backup Failures (Chris)

7. How to check for failures
   - 7.1 Log Files
   - 7.2 Backup locations
   - 7.3 Automation and scheduling
   - 7.4 Scripts

8. How to Recover from Overdue Database and Server Maintenance (Monika)

9. How to Recover from User Permissions Failures (Chris)
   - 9.1 EC2 Instances
   - 9.2 User/Group Errors

10. How to Recover from SSL Malfunctions (Monika)

11. How to Recover from Material Breach (Matt)

12. How to Recover from Member Web Server Corruption (Matt)

13. How to Recover from a Major LOCKSS Software Bug (Matt)

14. How to Recover from Cache Hardware Failure (Clay)

15. How to Recover from Member Hosting System Outages (Matt)

16. How to Recover from Member Listserv Hosting Outage (Matt)

Disaster Recovery Steps by Disaster Type

The following list includes instructions for recovery process how-tos that MetaArchive central staff (and members as appropriate) can follow when facing certain disruption impacts. This set of resources are supplemental to the Educopia/MetaArchive Contingency Plan.

How to Recover from an Amazon Service Outage (Chris)

The MetaArchive systems administrator should create a new set of admin, public, and data server instances on Amazon West. This can be done by restoring AMIs and db/files from backups stored in S3. Thorough steps and instructions for the MetaArchive system administrator to consult can be found here:

Many Items Are Now Permanently in Place to Deal with DR

- We will be ready with:
  - Title Database for access to caches
    - It was discovered that the caches get to the correct server after about 2-6 hours. No action is needed on the caches by their local admins.
    - Our partners have loaded the DR IPs to their security in advance for BOTH CONTENT and contacting the caches.
    - The caches have the WC IP addresses in their Admin Allow List, this includes the test caches.
  - S3 Data buckets are used daily for system backups. There is also a bucket just for AMI packages. These buckets were also created and left ready on the west coast.
  - An IP security group is needed. One was created for the west coast.
  - The Amazon keys and certs were copied to ~root so they will exist on the WC instances.

There are NEW IP Addresses

- Most important are configurations related to a new IP. The IP can not be relocated.
- We have reserved IP address(es) so they are known ahead of time. It costs a little bit per month but is needed, otherwise we could not know the IP address until the WC server boots.

When a Disaster Recovery is Taking Place

- When the time comes, actual changes need to be made to the following:
  1. Network Solutions DNS entries - Have user/pass convenient
  2. Disable the watchdog timers on the caches. (This can be scripted.)
3. Start the WC instances
4. Nagios settings have to change [/etc/nagios/hosts.cfg ]
5. MA staff contact the caches via a proxy server on the cloud servers. MA staff will also have to wait until the IP changes propagate to their own DNS servers to use the proxy. If needed, the MA staff Ubuntu servers can be updated and set to use their own local /etc/hosts file for DNS, and thus start using the proxy service right away.
6. Restore from backups on the WC.
   - The DR servers, and for that matter any newly started AMI, will need a few items
   1. A dir for backups to be created in /mnt /mnt/bkups
   2. A dir for AMI bundles to be created in /mnt /mnt/bundles
   3. The Amazon security cert files are copied to /mnt of the DR servers.
   - Other items were set up on the East Coast servers, so they would start when brought up from an AMI.

How to Recover from an Amazon Security Breach (Chris)

The MetaArchive system administrator should immediately power down or destroy the impacted EC2 instance(s), patch or update any software, and restore the instance from backup AMIs and db/files. Thorough steps and instructions for the MetaArchive system administrator to consult can be found here:

- Shutdown the EC2 instance
- Assess which resources have been lost and notify the admin group
- Restore services based on business impact

Analysis

- Identify the EBS volume associated with the compromised instance
  - CLI: ec2-describe-instance
  - Or via AWS console by selecting the block device
- Create a new snapshot, make note of the snapshot ID
- Create a new EBS volume from the snapshot ID
- Mount the EBS volume as /dev/sdX on dev.metaarchive.org for analysis

Recovery

- Launch a new Amazon linux based AMI to replace the lost instance
- Make sure the instance is fully updated and patched
- Create user accounts and make firewall allocations as needed
- Only restore from known good backups, and what could be salvaged from the compromised machine.

How to Recover from Overdue OS and Software Updates (Chris)

It is important that the MetaArchive system administrator evaluate and, where deemed critical, apply updates within the first two weeks of notification via software messages and alerts. Thorough steps and instructions for the MetaArchive system administrator to consult can be found here:

The cloud systems are either based on the Amazon linux AMI (data) or CentOS 5 (www,admin). Both linux releases utilize rpm (RPM Package Manager) and yum (Yellowdog Updater Modified) to manage system packages.

Profile the system in question

- What services are running?
- Check dependencies. Perl applications may require a specific library or a group of libraries to run. RPM perl packages are notorious for squashing libs installed via CPAN which will break your application (RT).

  ps -aux (List of running processes)
  /sbin/chkconfig --list (Which services are enabled)

Reboot

- After creating the initial system profile; reboot.
- Double check your list of running process against what you viewed earlier.

Take a snapshot

- All of the EC2 instances are backed by an EBS volume
- AWS web console https://aws.amazon.com/
- CLI EC2 API tools: ec2-create-snapshot

Update rpm, yum and related repositories

- Make sure you packing sub-system is up to date and working properly
yum update yum rpm

- Look for non-standard yum repositories defined in /etc/yum.repos.d These "extra" repositories may be out of date and should either be updated via a new RPM or disabled to avoid problems.

**Update everything else**

- If possible, try breaking down services in to groups. Such as system (kernels, libs), web services (PHP, Apache), databases (MySQL), notification services (MTA), RPM management (yum, rpm).
- Make a note of version numbers. It is best to know from whence you started and to what version you are about to upgrade to.
- reboot and reload services frequently if you find a large number of updates

**How to Recover from Cache Disk Failure(s) (Clay)**

Disk failures on member hosted caches can occur for a number of reasons ranging from longevity issues to poor manufacturing. Though the full range of disk-related issues cannot be fully predicted, a number of issues have already been experienced and dealt with. In addition, a number of monitoring measures have been put into place to detect the more common issues that arise. These experiences and monitoring measures have provided the MetaArchive central staff and members with a reliable process for detecting, diagnosing, and responding to most disk failures.

Due to the potential impacts that disk failures represent to member collections and the overall stability of the network, it is vital that central staff and members act quickly, yet conscientiously and thoroughly, when a disruption impact arises.

1. Central staff detects problem through Nagios and contacts Member systems administrator for further action, opening an RT support ticket as necessary; or
   a. Member systems administrator detects problem locally and contacts central staff via an RT support system ticket
2. Member systems administrator and central staff partner to diagnose and analyze problem.
   a. Member systems administrator may pull in hardware vendor support (per support contract between hardware vendor and member).
3. Disks are fixed (replaced) as needed.
4. Central staff and member systems administrator consult on disk failure impact.
   a. In most cases software RAID5 can be used to restore data to healthy disks
   b. In some cases, Member may need to reinstall entire system from scratch with help from central staff.
   c. In cases of partial or full data loss; either
      i. Member systems administrator will grant temporary SSH access to central staff, central staff will repair/rebuild AU configuration using last good bi-monthly AU backup as captured on Cloud Server, and data will be re-crawled from source; or
      ii. Member systems administrator and central staff will work together to copy data from peer caches using standard methods such as scp or rsync
      iii. Central staff will validate AU inventory, crawl status, and/or data copies and take additional actions as necessary until central staff determines that server is back at healthy function and AU inventory is correct.
5. Central Staff will make final notes in Support ticket and close ticket.

**How to Recover from Cloud Backup Failures (Chris)**

**How to check for failures**

**Log Files**

- Log files are maintained with each incremental and full backup.
- Nagios checks for the existence of current log files by looking in to the corresponding S3 bucket. $YEAR/$MONTH/$DAY-*incremental|full.log
  - Nagios script running on data: check_backup.sh
  - The system administrator should periodically monitor logs for failures.

**Backup locations**

All backups are stored in one of the following S3 buckets

- admin.metaarchive.org - Incremental/Full backups for admin
- dev.metaarchive.org - Incremental/Full backups for dev
- data.metaarchive.org - Incremental/Full backups for data
- www.metaarchive.org - Incremental/Full backups for www
- status.metaarchive.org - Backup files relevant to each LOCKSS cache

**Automation and scheduling**

- All backup scripts are launched via the root crontab
- Current (as of 12/2012) scheduling is as follows
  - Incremental backups: Monday-Saturday 0100
  - Full backups: Sunday 0100

**Scripts**

- Nagios monitoring
How to Recover from Overdue Database and Server Maintenance (Monika)

There are a number of network components and software applications that have database and underlying server functioning dependencies in the MetaArchive network. For example, the ability to query information about the health and status of the collections in the network is contingent upon things like the Cache Manager and Conspectus databases--information needs to be complete and current in those tables. Our websites and wikis also have underlying databases that need to be maintained. Likewise, the availability of all of the resources are contingent upon proper server upkeep.

The MetaArchive central staff currently rely on a series of cron jobs to ensure that standard database maintenance scripts are being applied to all of the databases in the network. If those cron jobs fail to run they will be reported in the nightly mails and will receive attention immediately within the 1 week update period. As a secondary notification, many of the software systems themselves are setup to deliver system messages when updates or maintenance jobs are needing to be performed.

The process of responding to lack of maintenance problems includes:

1. Check for email status
2. Verify with any software system messages
3. Identify failed cron job
4. Investigate any needed software updates (update as needed)
5. Apply software updates
6. Restart cron job
7. Monitor for proper functioning (watch logs and for emails)
8. Modify and re-run as needed
9. Continue monitoring for proper functionality

For server maintenance, much of what is needed falls under the recovery processes documented in Appendix C: How to Recover from Lack of OS or Software Updates (see above).

How to Recover from User Permissions Failures (Chris)

The MetaArchive has an Authorizations and Authentication Policy and Access Policy that requires central staff and members to have only those permissions granted to them by their role in the Cooperative and according to their responsibilities. Ensuring that no party performs unauthorized or damaging operations, for whatever reason malicious or otherwise, on MetaArchive systems is of critical importance.

Instances of permissions failures can have repercussions ranging from problematic yet recoverable to the outright catastrophic. In all such cases regardless of severity, immediate attention is required. The proper response is one of regaining control of all systems if necessary, modifying user permissions, and recovering any disruptions.

EC2 Instances

Access to all EC2 instances are controlled via SSH key pair combinations. Private keys can not be replaced nor can they be recreated via the AWS console. If you are completely locked out of an instance:

1. Shutdown the instance
2. Make note of how the root volume is mounted
3. Detach the volume
4. Attach the volume to a new instance that has been launched with a new key pair or attach to an existing instance
5. Create a new key pair (ssh-keygen), add the public key to user X (~/.ssh/authorized_keys), save the private key
6. Detach volume, re-attach to original instance, and start

User/Group Errors

1. Manipulate group membership via usermod
2. Manipulate groups via groupmod
3. Change directory permissions via chmod
4. Use sudoers wisely, visudo

How to Recover from SSL Malfunctions (Monika)

In addition to restricting communications between the caches to only the IP addresses of other caches via the use of firewalls, the communications between those caches are also encrypted through the use of SSL. This means that each cache carries a public/private keystore that must be updated with the keys generated for each cache. If caches do not update their keystores, for example when a new cache is added to the network, then the new cache will not be allowed to communicate (vote/poll/repair) with those caches. This is the primary cause of malfunctions with SSL and it is avoidable if all caches keep their keystores up-to-date.
As of December 2012 the MetaArchive central staff and the LOCKSS team at Stanford are in communication about SSL error exceptions that are being detected in both the MetaArchive and in the CLOCKSS network. Discussion and analysis are on-going and will resume in January 2013. Steps for detecting and recovering from such malfunctions will be thoroughly documented here upon resolution.

How to Recover from Material Breach (Matt)

The MetaArchive and its use of LOCKSS provides no intelligent mechanism for detecting bit corruption on members’ collections as they reside on their local web servers. To date, this responsibility falls to the member and their routine processes for validating the bit integrity of their collections. When LOCKSS re-crawls member collections, if a file has changed at the bit level (for good, bad, or otherwise), that file will be archived and replace the previously collected version. If a member is concerned that the integrity of their files could be compromised over time, there are a number of audit approaches that can be implemented locally. In 2013, the MetaArchive Content & Preservation Committees will discuss best practices that can be recommended.

In the meantime, if a file or set of files are found to be corrupted on a member's web server, the steps to recovering include:

1. Central staff should consult their local backup storage for healthy restoration copy(ies)
2. Consult MetaArchive central staff for an archived copy(ies) as needed
3. Replace corrupted copy(ies) on local web server
4. Request re-crawl from MetaArchive

How to Recover from a Major LOCKSS Software Bug (Matt)

The MetaArchive is dependent upon the LOCKSS team at Stanford for keeping the software updated and bug free. To date, only minor issues have arisen and been forwarded to LOCKSS for addressing in future releases. The MetaArchive makes use of supported operating systems. All updates to the LOCKSS software are thoroughly tested at Stanford. In addition, all updates to the LOCKSS software are tested in the MetaArchive test network prior to release in the production network. Incidences of major bugs impacting the MetaArchive should be rare.

A major bug would be classified as any software vulnerability that compromises the stability and security of managing the preserved content, the host machine, or the network at large. If such vulnerabilities arise, the steps for recovering include:

1. Central staff should work with any impacted member(s) to diagnose the bug and the scope of its impact
2. If necessary, the impacted machine(s) should be quarantined by taking it/them off-line
3. The machine(s) should then have its/their public/private keys removed from /cacheadmin/lockss/keys directory on the admin server
4. The machine(s) should also have its/their IP address removed from the list of allowable IPs
5. Central staff should submit a bug fix request with relevant machine/system details included to LOCKSS and mark it urgent
6. While waiting for the bug fix, data and log files on the impacted machine(s) should be analyzed for signs of corruption or tampering
7. Upon release of the bug fix, MetaArchive central staff will update the test network to validate success
8. Upon success, the update will then be released to the production network
9. The impacted machine(s) should be brought online (w/communication to the admin server only) to receive the update
10. The impacted machine(s) should be monitored for persistent problems
11. Once approved the machine(s) should have its/their public/private keys regenerated and its IP address added once again to the allow list
12. Any impacted data will be repaired by peers or re-crawled from source web servers

How to Recover from Cache Hardware Failure (Clay)

Hardware failures on member hosted caches can occur for a number of reasons ranging from longevity issues to poor manufacturing. Though the full range of hardware-related issues cannot be fully predicted, a number of issues have already been experienced and dealt with. In addition, a number of monitoring measures have been put into place to detect the more common points of failure. These experiences and monitoring measures have provided the MetaArchive central staff and members with a reliable process for detecting, diagnosing, and responding to most failures.

The process for updating keystores can be found here: http://wiki.metaarchive.org/metawiki/index.php/CachesMaintain#Updating_the_Public_Keystore
Due to the potential impacts that hardware failures represent to member collections and the overall stability of the network, it is vital that central staff and members act quickly, yet conscientiously and thoroughly, when a disruption impact arises.

1. Central staff detects problem through Nagios and contacts Member systems administrator for further action, opening an RT support ticket as necessary; or
   a. Member systems administrator detects problem locally and contacts central staff via an RT support system ticket
2. Central staff and member systems administrator consult on hardware failure impact. 
   a. In some cases, Member may need to reinstall entire system from scratch with help from central staff.
   b. In cases of partial or full data loss, member systems administrator will grant temporary SSH access to central staff, and central staff will repair/rebuild AU configuration using last good bi-monthly AU backup as captured on Cloud Server.
3. If an outage persists beyond 1 week, the central staff should monitor the overall copy protection and voting/polling successes on other peer caches.
4. Central staff and member systems administrator consult on hardware failure impact.
5. Central staff will validate AU inventory and crawl status and take additional actions as necessary until central staff determines that server is back at healthy function and AU inventory is correct.
6. Central Staff will make final notes in Support ticket and close ticket.

How to Recover from Member Hosting System Outages (Matt)

MetaArchive members host LOCKSS caches with preserved content in both our production and test networks and make their content servers available for crawling by these caches. Occasionally, unforeseen system outages do occur and if prolonged can begin to impact the health of both their own as well as other members' collections. If the outage is due to a MetaArchive or LOCKSS related software/network problem, the central staff and members may need to work together to bring systems back online in a timely fashion. Some examples of disruption impacts above that would fall under this category include:

- How to Recover from a Major LOCKSS Software Bug
- How to Recover from Cache Hardware Failure
- How to Recover from Cache Disk Failure(s)
- How to Recover from Overdue OS and Software Updates

If a system outage is due to local disruptions that are non-MetaArchive or LOCKSS related, the MetaArchive central staff will likely not be permitted to get directly involved. General system outages due to intermittent power issues or maintenance will likely be temporary and should not require drastic stop-gaps. Nevertheless, safeguarding the health of the network and collections must remain a priority. The steps for recovery include:

1. Central staff and members should communicate beforehand or very soon thereafter when a member system outage is experienced--nature /causes/severity should be discussed
2. If the member knows how long the outage may persist this should be communicated to the central staff
3. Central staff should make note of the nature/causes/severity in Nagios
4. The member should commit to alerting the central staff when the outage has been resolved or if more down-time will be required
5. When systems are returned to normal operation the cache and web server(s) at the member location should be monitored for proper connectivity and communication via Nagios

How to Recover from Member Listserv Hosting Outage (Matt)

The MetaArchive listservs (Main-L, Admin-L, Dev-L, Pres-Comm, Content-Comm) are all currently hosted on a mail server at one of our member locations (University of North Texas). There are a number of things that could lead to a listserv outage, including failure to apply L-Soft software patches/updates, general system outages at the member location, and a temporary disruption in service as the current member hosting site transitions out of the Cooperative and a new hosting partner (member or service provider) is identified.

Each of the above-mentioned disruption impacts will likely require their own specific intervention measures. A general and simplified process follows:

1. Central staff detects problem through Nagios and contacts Member systems administrator for further action, opening an RT support ticket as necessary; or
   a. Member systems administrator detects problem locally and contacts central staff via an RT support system ticket
2. Central staff and member systems administrator partner to diagnose and analyze problem.
   a. Member systems administrator may pull in hardware vendor support (per support contract between hardware vendor and member).
3. Hardware is fixed (replaced) as needed.
4. Central staff and member systems administrator consult on hardware failure impact.
   a. In some cases, Member may need to reinstall entire system from scratch with help from central staff.
   b. In cases of partial or full data loss, member systems administrator will grant temporary SSH access to central staff, and central staff will repair/rebuild AU configuration using last good bi-monthly AU backup as captured on Cloud Server.
5. Central staff will validate AU inventory and crawl status and take additional actions as necessary until central staff determines that server is back at healthy function and AU inventory is correct.
6. Central Staff will make final notes in Support ticket and close ticket.

How to Recover from Member Hosting System Outages (Matt)

MetaArchive members host LOCKSS caches with preserved content in both our production and test networks and make their content servers available for crawling by these caches. Occasionally, unforeseen system outages do occur and if prolonged can begin to impact the health of both their own as well as other members' collections. If the outage is due to a MetaArchive or LOCKSS related software/network problem, the central staff and members may need to work together to bring systems back online in a timely fashion. Some examples of disruption impacts above that would fall under this category include:

- How to Recover from a Major LOCKSS Software Bug
- How to Recover from Cache Hardware Failure
- How to Recover from Cache Disk Failure(s)
- How to Recover from Overdue OS and Software Updates

If a system outage is due to local disruptions that are non-MetaArchive or LOCKSS related, the MetaArchive central staff will likely not be permitted to get directly involved. General system outages due to intermittent power issues or maintenance will likely be temporary and should not require drastic stop-gaps. Nevertheless, safeguarding the health of the network and collections must remain a priority. The steps for recovery include:

1. Central staff and members should communicate beforehand or very soon thereafter when a member system outage is experienced--nature /causes/severity should be discussed
2. If the member knows how long the outage may persist this should be communicated to the central staff
3. Central staff should make note of the nature/causes/severity in Nagios
4. The member should commit to alerting the central staff when the outage has been resolved or if more down-time will be required
5. When systems are returned to normal operation the cache and web server(s) at the member location should be monitored for proper connectivity and communication via Nagios

How to Recover from Member Listserv Hosting Outage (Matt)

The MetaArchive listservs (Main-L, Admin-L, Dev-L, Pres-Comm, Content-Comm) are all currently hosted on a mail server at one of our member locations (University of North Texas). There are a number of things that could lead to a listserv outage, including failure to apply L-Soft software patches/updates, general system outages at the member location, and a temporary disruption in service as the current member hosting site transitions out of the Cooperative and a new hosting partner (member or service provider) is identified.

Each of the above-mentioned disruption impacts will likely require their own specific intervention measures. A general and simplified process follows:

1. Central staff detects problem through Nagios and contacts Member systems administrator for further action, opening an RT support ticket as necessary; or
   a. Member systems administrator detects problem locally and contacts central staff via an RT support system ticket
2. Central staff and member systems administrator partner to diagnose and analyze problem.
   a. Member systems administrator may pull in hardware vendor support (per support contract between hardware vendor and member).
3. Hardware is fixed (replaced) as needed.
4. Central staff and member systems administrator consult on hardware failure impact.
   a. In some cases, Member may need to reinstall entire system from scratch with help from central staff.
   b. In cases of partial or full data loss, member systems administrator will grant temporary SSH access to central staff, and central staff will repair/rebuild AU configuration using last good bi-monthly AU backup as captured on Cloud Server.
5. Central staff will validate AU inventory and crawl status and take additional actions as necessary until central staff determines that server is back at healthy function and AU inventory is correct.
6. Central Staff will make final notes in Support ticket and close ticket.