Harvest Ingest Guidelines

Overview:

CLOCKSS can use web harvesters to programmatically discover and collect content from your website. In order to support this method of preservation, the following general conditions must be met:

1. Programmatically-discoverable permission statements for each base host used for substantive content.
2. Programatically discoverable finite sets of content. Preservation is based on collecting a unit of content with components that cease to grow at some point. This could be a volume or year of articles for one journal, or a single book with all its chapters, or all the datasets deposited over a certain period of time.
3. A website that enables a preservation crawler to discover and collect all the items in the defined set of content at static URIs. This could be through a browsable website that uses a logical consistent link structure and in which every item to be preserved can be reached through at least one static text link even if site behavior is enhanced with dynamic elements. Alternatively, this could be through the use of an access API, such as OAI-PMH, that allows the crawler to generate a list of access points for all the items to be preserved.
4. Programmatically-discoverable metadata for each preserved item that can be used to identify the preserved contents.

Details:

Permission statements

There must be a permission statement for each domain serving substantive content. It is not required for CDN support for auxiliary files.

For open access content, the permission statement could be a legal Creative Commons license including the ‘rel=“license”’ attribute.

Alternatively, and for subscription content, the permission statement needs to be:

- CLOCKSS system has permission to ingest, preserve, and serve this Archival Unit

This does not need to be visible and could be a comment in the html of a web page.

When all the content at the site is open access, it is sufficient to have one permission statement somewhere under the domain.

If all content is available for preservation, it is sufficient to have one permission statement somewhere under the domain instead of at the starting point for each set of content.

Defined content sets and starting points for collection

For harvest, the crawler needs to be able to collect content and have a point in time at which that collection is complete with a repeatable, defined set of components. The crawler needs to be able to determine a start URI or URIs for this content based on information unique to the content. This might be a URI defined by the domain, the journal identifier, and a particular volume or year, such as:

- https://www.publisher.com/journals/xyz/12
- https://www.publisher.journalid.com/content?year=2017
- https://www.publisher.com/ebooks/isbn/97811111111
- https://www.publisher.com/ebooks?search&year=2017&order=newest_first

In some cases, such as when the website does not normally provide content in defined units, the start page might be an artificial manifest page provided for the purposes of preservation, such as:

- https://www.publisher.com/xyz/lockss_manifest?volume=27

Or, if the website supports an API, the crawler can generate a request that returns and XML response that can be parsed for a list of all the article URIs associated with a given set of content, such as:

- https://www.publisher.com/api/search?type=data_sets&publication_start=2019-03-01&publication_end=2019-04-01

Site behavior

- Make your content discoverable by a crawler by building your site with a logical link structure. Every page should be reachable from at least one static text link or have consistent identifiable tags in the html from which links can be generated. You can use a text browser, such as Lynx, to examine your site. If dynamic web features or server side processing keep you from accessing your entire site in a text browser, then preservation is likely to be impaired.
- Use unchanging canonical links for content even if the content is served through single-use redirection. Preservation requires that the same URI will serve equivalent content in the future.
- Use HTTP codes as expected: 5xx codes to indicate temporary errors that should be retried soon, 4xx codes to indicate permanent errors that should not be retried for some time, and 3xx redirects for each URI to its new location in the event that you need to move content.
- Use if-modified-since headers so new content is identified and preserved.
- Content from CDNs should have a stable URI.

Required metadata

- For scholarly publication content, such as articles or books, you need to provide basic bibliographic metadata in a non-proprietary machine-readable format, such as the HTML metadata tags, RIS, or XML files.
- The metadata for scholarly content should include: DOI, publication, publication date, ISSN or ISBN and as appropriate: publication title, item title, series title, volume, issue and page number.
- For items other scholarly published content, the metadata should provide enough information to uniquely identify the item - DOI if applicable or proprietary identifier, provider and so forth.

IP Addresses

Enable IP address access to all content you wish preserved:

- 171.66.236.0/24 (171.66.236.0 through 171.66.236.255) located at Stanford University
- 128.42.174.11 located at Rice University
- 128.42.174.12 located at Rice University
- 156.56.241.164 located at Indiana University
- 156.56.241.166 located at Indiana University

Note that you may need to separate these IP addresses from the Rice, Indiana, and Stanford subscription records and set up a new CLOCKSS subscription. These IP addresses are used for all development and testing as well as preservation for the CLOCKSS Archive.