Quality Assurance
Quality Assurance Steps

- Check the crawl report
- Enable Proxy Mode
- Crawl a sub-page as a seed
- Use Wayback QA
Checking the Crawl Report
Checking the Crawl Report

What to look for in the Crawl Overview:

- Did the crawl finish because of a limit?
- Is the data in line with your expectations and budget?
Checking the Crawl Report

What to look for in the Seed Report:

- Is the **Seed Status** something other than crawled?
- Is the **data** in line with your expectations?
Checking the Crawl Report

What to look for in the Host Report:

- Are there documents in the **Blocked** column?
- Are there documents in the **Queued** column?
- Are there documents in the **Out of Scope** column?
From the Host Report

- Capture content blocked by robots.txt by running a **patch crawl**
- **Edit rules** for future crawls
Checking the Crawl Report

From the Host Report

- Capture content blocked by robots.txt by running a patch crawl
Viewing sites in “Proxy mode”
Viewing sites in “Proxy mode”

- Replay archived documents without interference from the “live web”
- Use the Firefox add-on or a manual configuration
- Only available for sites beginning with http, not https
- Not currently available for most social media sites
Viewing sites in “Proxy mode”

- Replay archived documents without interference from the “live web”
- Use the Firefox add-on or a manual configuration
- Only available for sites beginning with http, not https
- Not currently available for most social media sites
Wayback QA
## Using Wayback QA

<table>
<thead>
<tr>
<th>Wayback QA</th>
<th>Patch Crawls</th>
</tr>
</thead>
<tbody>
<tr>
<td>● Discovers missing documents from already captured pages</td>
<td>● Only crawl the individual documents in the Wayback QA list</td>
</tr>
<tr>
<td>● Only looks for missing documents on the page you’re viewing</td>
<td>● Does not follow links off of documents, or capture embedded content in them</td>
</tr>
<tr>
<td></td>
<td>● Don’t adhere to collection or seed scoping rules</td>
</tr>
<tr>
<td></td>
<td>● Data available 24 hours after crawl finishes</td>
</tr>
</tbody>
</table>
Using Wayback QA
Crawling a page as a seed
Seeds are captured by Umbra:

- Runs alongside Heritrix on only the seed page of each archived site
- Helps to capture embedded and dynamic content like carousels, video, and infinite scrolling
Seeds are captured by Umbra:

- Runs alongside Heritrix on only the seed page of each archived site
- Helps to capture embedded and dynamic content like carousels, video, and infinite scrolling
Crawling with Brozzler
Crawling with Brozzler

Brozzler opens all pages in-browser and transcribes interactions to WARCs
Quality Assurance resources

In the Archive-It Help Center:

- Quality Assurance steps
- Reading your crawl report
- Modify scope and run patch crawls from your report
- Access your archives in “Proxy mode”
- How to perform quality assurance (QA) and use Wayback QA