Books in Browsers

Enable Links in Archive.org Books

Problem Statement
To date, the Internet Archive has preserved millions of print books by making them digitally accessible. While now digitally accessible, these books still lack modern enhancements that could extend their utility beyond their physical counterparts in the digital age of the web. For example, many of these books contain static URLs that have not yet been archived and are unable to be clicked within the browser.

A spot-check shows there are:

- 265,000 books that reference “http://www”
- 1.4M books reference “http”

Goals

1. **Preservation.** Books are a high quality source of urls. Let’s make sure these urls are archived in the Wayback machine
2. **Accessibility.** Paper books are flat; So are our images in Book Reader. Let’s augment them with the ability to click/follow semantic content, like urls and references

Preserving URLs in Wayback
Starting with open access material, work with staff and use the `obgp` role account s3 keys to:

- Identify 10,000 books likely to contain URLs
    - Consider pulling down the existing 250k public book genome files from items and pulling out some of those URLs into a list to see how many of them currently resolve, can be easily fixed, and see if we can submit those to wayback.
- **Maintain a global list of [item_id, filename, page, url]** so we can count the total number of items we’ve processed, the total number of urls discovered, etc.
    - so this might look like: a text log of item_id,filename,page_num (where page_num = -1 when the document is finished) – this lets us stop/start/resume our script gracefully and also count the number of completed items and/or subitems.
- **For each book…**
    - **Fetch the OCR Full Text** (as djvu.xml) for these books and identify, extract, and repair URLs ∞ The Open Book Genome Project (OBGP) Sandbox
    - **Test & Clean URLs** Check whether the format of the url is valid (e.g. regex) and possibly fix common url errors, e.g. extra spaces, make an http request (e.g. a HEAD) to check whether the url is valid / living
    - For each URL that resolves, ensure it is saved using the wayback machine Save Page Now (SPN2) API
    - Make sure the repaired, archived URLs are saved in the book’s archive.org item as urls.json

Clickable URLs in Books

In 2020, Giacomo (GSoC 2020) & drini@archive.org implemented a hidden text layer for the Internet Archive book reader, enabling text selection and highlighting for public domain books (e.g. https://archive.org/details/autobiograp00fran?ref=ol&view=theater). We also have a rudimentary book reader **prototype** that detects urls within our OCR and allows clickable urls within bookreader #743**.**

Background: Right now, the Read Aloud feature of our BookReader enables a full page of text to be retrieved from the server. We also (via Read Aloud and Search-Inside features) have the ability to highlight regions of BookReader pages.

The proposed solution (version 1) is to extend the BookReader with a plugin which, on page-load, to:

1. Pull a page of region-labeled, OCR'd + text using the https://api.archivelab.org/books/{identifier}/pages/{page}/ocr?mode=words API
2. Hit a new **entities** endpoint which identifies urls (and later other semantic entities) which returns a list of:
   - type: e.g. `url`
   - location: (x, y, w, h)
   - value: e.g. https://archive.org
3. Highlight the corresponding region on the book containing the link and make the region clickable to a Save Page Now version of the link
   - I.e., once clicked, capture the webpage if we don't already have it, or, in either case, bring the patron to a viewable version of this url

Support

John Gilmore expressed interest in supporting this project