

Image Files

The Internet Archive BookReader needs JPEG image files for rendering the displaying the book in a web browser. The best option for generating these JPEG's is to create JPEG derivatives of each page and store them as individual datastreams under the typical datastream ID naming methodology, i.e. JPEG-1, JPEG-2, JPEG-{n pages}.

Generating JPEG's

The JPEG's that were created for testing and development we generated using ImageMagick installed on a workstation. The derived TIFF's were used to generate the JPEG image files. During the JPEG generation process images were downscaled to 1100 pixels wide at a quality setting of 80%. This offered acceptable quality in most cases. The following command was used:

```
convert -strip -interlace Plane -quality 80% -thumbnail 1100 {input} {output}
```

During quality control reviews if a page requires less compression or downscaling a new JPEG page can be generated with higher quality or resize settings. The new JPEG can replace the existing JPEG datastream.

For search term highlighting to continue to work all JPEG datastreams that are generated from the source TIFF files need to retain their original proportions.

Searching Inside the Yearbook

The Internet Archive BookReader has the ability to search inside of an item and highlight the text when a hit is found. That functionality has been interfaced with the resources source OCR file to provide the BookReader with page number and coordinate information.

OCR Source File

An OCR source file is needed to determine when and where search hits have been found. Currently this information is stored in a DjVu file. The derived XML-1 datastream does not contain the coordinates of the words that were discovered by the OCR engine. It is the recommendation of this specification that the OCR source file be extracted from the DjVu file and stored as a separate datastream. This will enable simpler interaction with the OCR source file and its word/coordinate pairs. Possibly storing the OCR source file using the datastream identifier OCR-1 would be suitable.

OCR Correction Process

For this implementation corrected OCR XML can be added and updated once it is exported from the editing tool being used by the digitization specialist. Since no DjVu datastream is necessary or needed for these resources the only datastreams that would need updating would be the OCR-1 and XML-1 datastreams.

Text Searching

Text searching inside of a resource is available when an OCR datastream also exists in the resource. If an OCR datastream is not present, searching is not available. Currently searching works by treating the submitted term as a

phrase. Matches are found by compiling paragraphs of OCR text together and looking for a case-insensitive match. When a match is found it is returned as part of a JSON that the BookReader uses to render hits along the resources navigation bar with text highlighting and pin markers with contextual text bubbles. The user is automatically taken to the first hit in the resource the searched. See Figure 1.

Ingest Workflow

On the test/development system two yearbooks have been ingested for testing and development. These yearbooks were ingested using the WMS on the test system. All presentation JPEGs were created offline and upload, not generated by the system, from the derived archival master TIFF files. After ingest the OCR-1 datastreams for both resources were added using dlr/EDIT. During testing no XML-1, search index ready, datastreams were prepared.

Double-page Spreads

When scanning some yearbooks double-page spreads have been scanned as a single page. These double-page scan will need to be either rescanned or cropped into single-pages and the corresponding OCR text layer will need to be updated.

If this is not done then when viewing the yearbook in 2 page view you might occasionally present the user with three pages, a double-page spread and a single page, or even worse four pages, two double-page scans.

mods:typeOfResource & URL Implementation

It is the recommendation of this specification that all resources in RUcore with mods:typeOfResource equal to “text” that have more than one JPEG datastream direct the users to the BookReader functionality. The JPEG datastream links will be replaced with one link labeled “Read Online” followed by the number of pages the resource is comprised of.

Following along with the syntax used for accessing resources and their files the following syntax is recommended for accessing the full and embedded versions of the BookReader.

Full BookReader - /rutgers-lib/1234/jpeg/read/

Embedded BookReader - /rutgers-lib/1234/jpeg/read/embed/

Information Icon and Title Link

Along the top of the BookReader the title of the item being displayed is present. Clicking on that link will take the user back to the resources landing page. Selecting the information icon link will provide the user a modal window with the title and abstract. If a PDF of the resource is available a link to download that PDF will be provided as well. See Figure 2.

Embedding Documents

The BookReader offers the ability to embed playback of a resource in a webpage. This is accomplished by using an iframe. If a user would like to embed a resource into a webpage the URL and instructions for embedding are available in from the BookReader interface by clicking on the share icon. See Figure 3 & 4.

Statistics

When JPEG's are accessed during viewing each JPEG viewed will not be recorded as a download. Instead a statistic for viewing the item using the BookReader will be recorded; this is similar to the recording of a full record view. Displaying that usage metric would be manageable from the public statistics display. The reserved statistic entry will be "readbook" and this will need to be added to the Statistics API configuration file.

Gesture Support

The Internet Archive BookReader natively supports iOS devices for page swiping and zooming. The delivered solution does not support Android devices however. The BookReader has been modified to support Android devices by changing the JavaScript library used for binding gestures to interactions with the BookReader. The JavaScript library is the same one that has been used with the image slideshow feature.

Sample Yearbooks

Two sample yearbooks have been ingested on the test system and can be viewed using the following URL's.

1892 Yearbook <http://rucore-devel.libraries.rutgers.edu/rutgers-lib/201614/jpeg/read/>

1953 Yearbook <http://rucore-devel.libraries.rutgers.edu/rutgers-lib/201609/jpeg/read/>

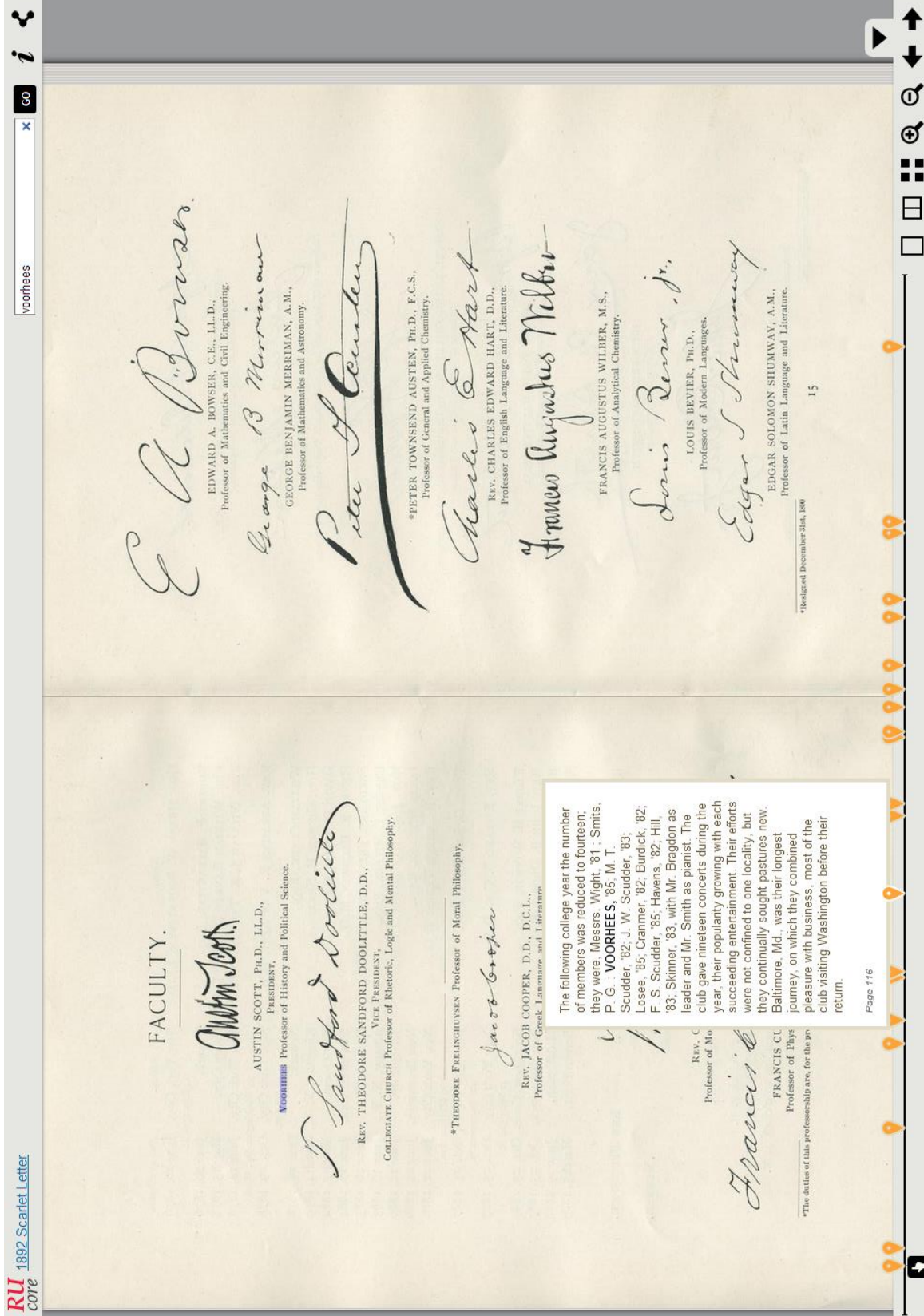


Figure 1: Search example

The screenshot displays the Internet Archive BookReader interface. At the top left, there is a search bar with the text "Search inside" and a "GO" button. To the right of the search bar is an information icon (a lowercase 'i' in a circle) which is highlighted with a yellow circle. Further right are navigation icons: a left arrow, a right arrow, a magnifying glass, a plus sign, a square, and a double square. The main content area shows a page from a 1892 Tiffany & Co. catalog. The page features the company name "TIFFANY & CO." and "UNION SQUARE, NEW YORK." at the top. Below this, there are sections for "WATCHES," "CLASS CUPS," and "TIFFANY & CO.'S 'BLUE BOOK' OR CATALOGUE FOR 1891 SENT UPON REQUEST." A pop-up window is overlaid on the page, titled "About this book" with a close button (X) in the top left corner. The pop-up contains the text "1892 Scarlet Letter" and a link: "For more information on this resource visit the following link <http://rucore-devel.libraries.rutgers.edu/rutgers-lib/201614/>". The background page also includes a circular logo for "J. COLEMAN'S 1107 BROADWAY" and the word "Hats.".

Figure2: Information Icon

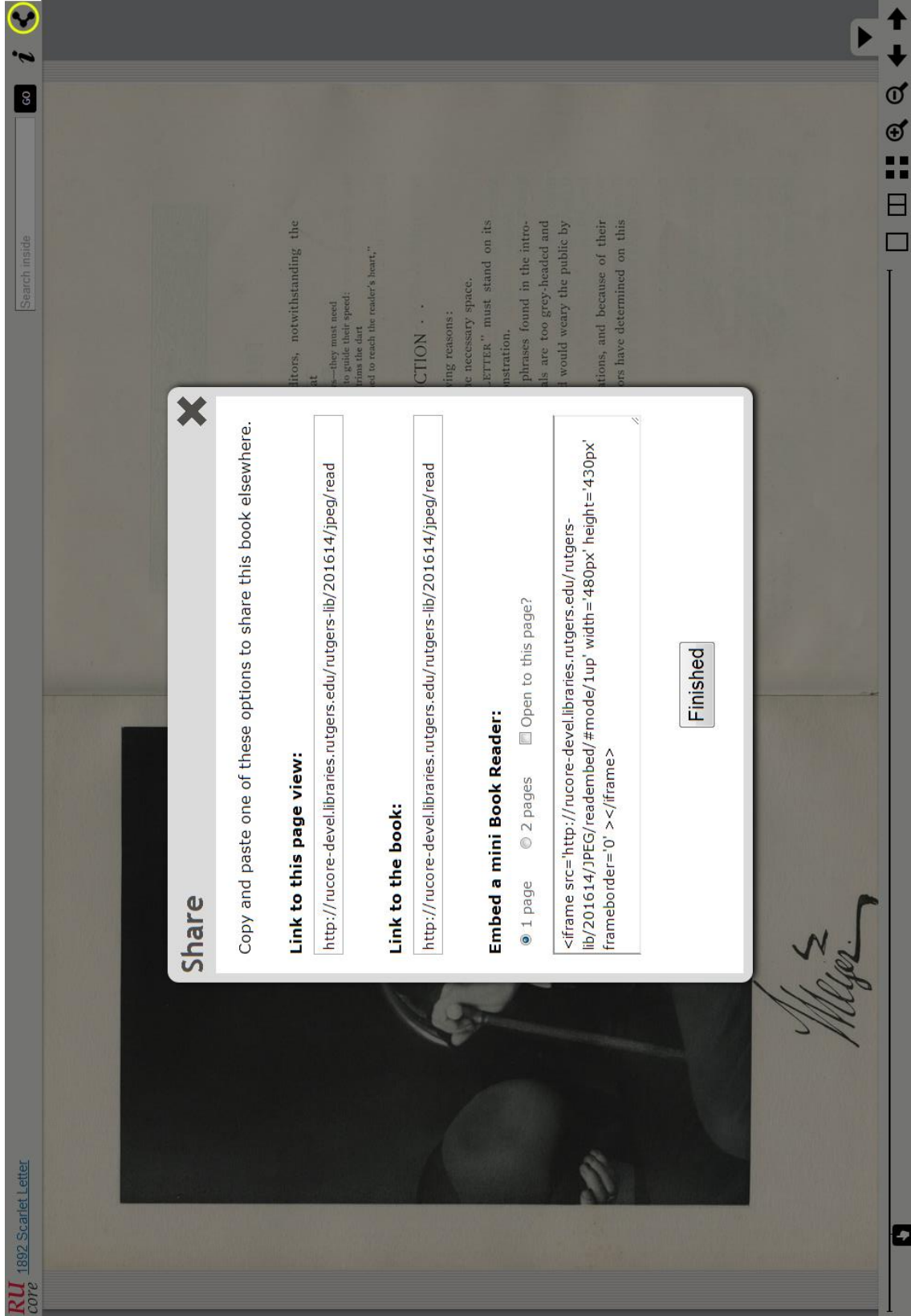


Figure 3: Share Icon

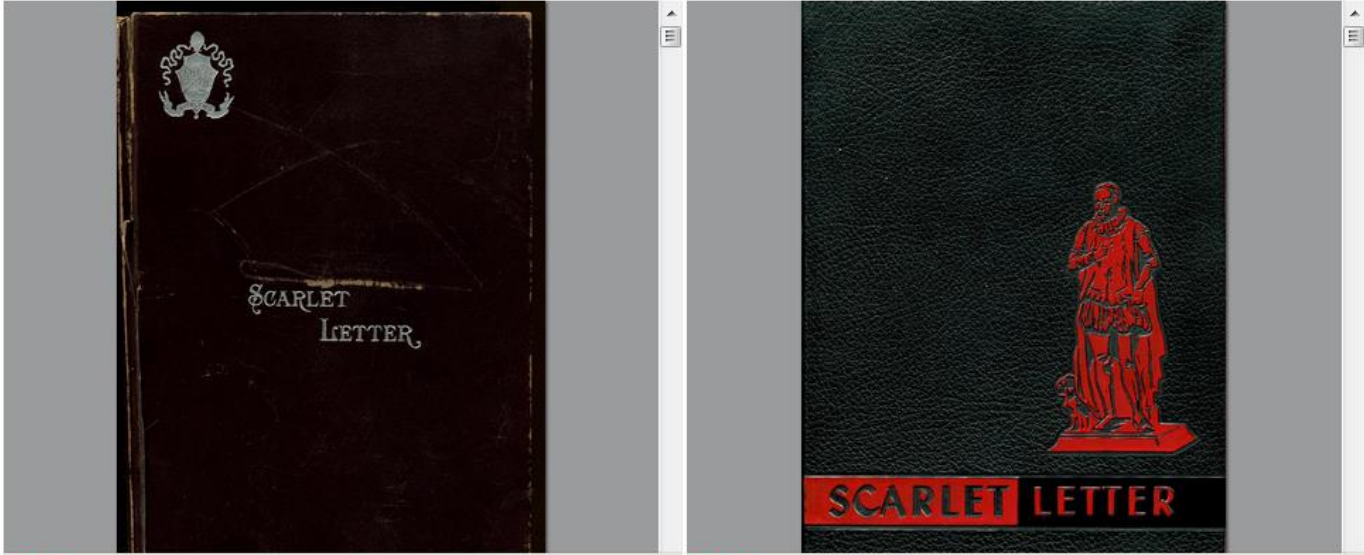
RUcore
Rutgers University Community Repository


Help • Contact Us • Site Search


Search Faculty Collections Collaborations

Home Search Portals Rutgers Yearbooks Prototype


Rutgers Yearbooks Prototype



RUcore [1892 Scarlet Letter](#) 

RUcore [1953 Scarlet Letter](#) 

About Us How does RUcore work? Policies Services Collections	Infrastructure Trusted Repository Preservation Understanding Metadata Technical Glossary	Open Source OpenWMS OpenETD OpenMIC OpenWAAND	Developers Reference Materials Web Services/APIs Schemas Harvesting
---	---	--	--

 [Statistical Profile](#) • [Version 7.2](#)

Rutgers University Libraries • [Privacy Policy](#) • [Copyright ©2013](#)

Figure 4: Embedding Example