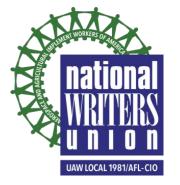
What is the Internet Archive doing with our books?

National Writers Union (NWU.org)

Edward Hasbrouck, Co-Chair, Book Division

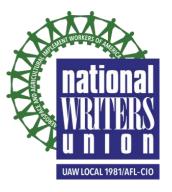
webinar first presented April 27, 2020; updated for May 5, 2020, presentation



Is there a national emergency and a need for distance learning material? Yes!

"The COVID-19 pandemic has created a huge increase in demand for... materials for distance learning.... If authors were fairly compensated, most of us would be eager to participate in projects to digitize our backlists and distribute them to students for distance learning. But authors need to be involved in these projects, and we need to be fairly paid."

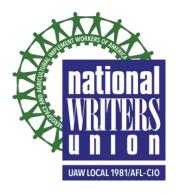
("We Need Federal Funding for Distance Learning, During the Pandemic — and After", NWU blog, April 22, 2020)



The Internet Archive

(Brewster Kahle, Founder) San Francisco, CA





• "Based on some of the misinformation by the lobbyists on the other side, they've basically inflamed a bunch of authors.... The lobbyists on the other side... are paid a lot of money by some people to go and spread misinformation."

(Brewster Kahle, Internet Archive, April 20, 2020)

- The NWU has no paid lobbyists. We are unpaid volunteers lobbying on behalf of ourselves and our fellow writers.
- Average annual compensation of the five highest-paid employees of the Internet Archive (2017 IRS Form 990): \$192,000
- Median annual writing-related income of full-time book authors: (2018 Authors' Guild survey): \$20,300

What is the Internet Archive doing with our books?

"National Emergency Library"?

"Controlled Digital Lending"?

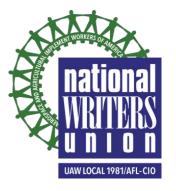
These are neither <u>accurate</u> nor <u>complete</u> descriptions of the Internet Archive's book scanning and distribution programs.

Rather than talking about what they <u>say</u>, let's start by talking about what they <u>do</u>.



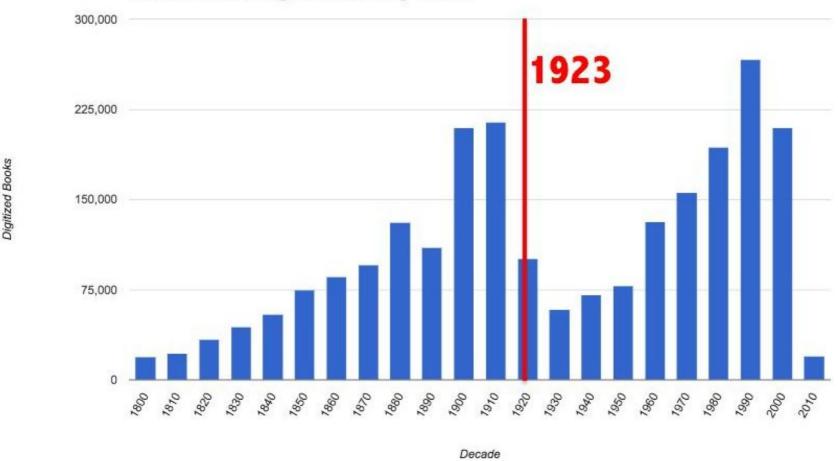
The Internet Archive starts with printed books. Where do the books come from?

- 1. Library discards (Trent University, Ontario)
- 2. Closed libraries (Marygrove College, Detroit)
- 3. Donations of used books (through libraries, etc.)
- 4. "Sponsored" book purchases ("Scan On Demand")
- 5. Library collections of printed books, later returned to library shelves (Phillips Academy Andover, MA)



Some books are out of copyright, but the largest percentage by decade were published in the 1990s:

Internet Archive Digitized Books by Decade



What types of books are included?

- Old books
- New books
- Books that are in copyright
- Books that are available for sale in printed editions
- Books that are available for licensing as e-books
- Books that are available for licensing as audiobooks
- Books that are out of print but that include works that are available in other editions or formats
- Fiction, non-fiction, prose, poetry, journalism, etc.
- Art and photography books
- Books in many languages, published in many countries



What happens to the books?

- 1. Packed in containers & shipped to scanning centers
- 2. Scanned (scanning centers in San Francisco; Hong Kong and Shenzhen, China; Cebu, Philippines; etc.)
- 3. Shipping containers filled with scanned books returned and warehoused (Richmond, CA; etc.)
- 4. Some scanned books returned to library shelves (Phillips Andover Academy, etc.)



How some key words are used:

- When we refer to a "book", we might mean a physical object, or we might mean the collection of words and images it contains. And we may speak colloquially of "borrowing" an e-book, even though "e-lending" is actually very different from lending or borrowing of physical books.
- So when the Internet Archive talks about "books", it's important to ask ourselves whether they are talking about the physical objects, or digital copies.
- And when they talk about "lending" or "borrowing", it's important to ask whether they are talking about actual lending or about "e-lending".

- 1. Books (physical objects) \neq e-books (digital files)
- 2. Books ≠ works included in books (text, photos, etc.)
- 3. Lending of books ≠ "e-lending" of e-books

The Internet Archive does not lend books.

The Internet Archive makes and distributes digital copies of works that have been included in books.

Copies sent by Web server to readers, listeners, & viewers:



E-book downloads

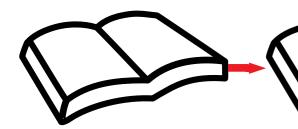
Audiobooks (scan \rightarrow OCR \rightarrow text-to-speech)

Page images on OpenLibrary.org

Page images on Archive.org (unique URL for each page)

APIs for automated download of page images

Web Server (page images, e-books, audiobooks)



Physical

Storage

(books)

Copy 1 (paper)

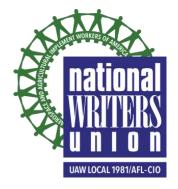






How does the Internet Archive distribute images of pages of the books it scans?

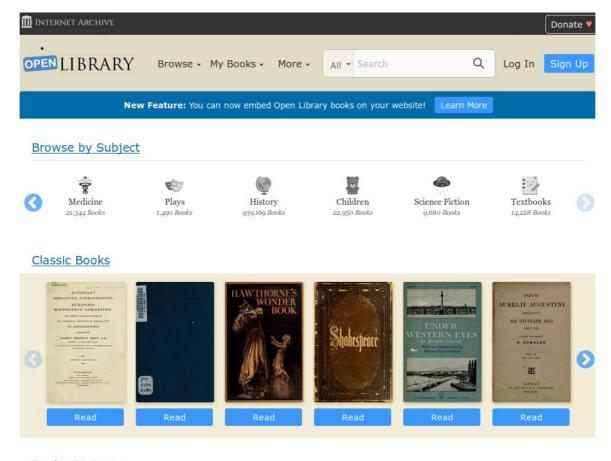
- 1. Downloads via OpenLibrary.org of e-books assembled from page images
- 2. Audiobooks generated from page images
- 3. Viewing of page images on OpenLibrary.org
- 4. Viewing of page images on Archive.org
- 5. APIs for automated downloads of page images



1. Downloads via OpenLibrary.org of e-books assembled from page images:

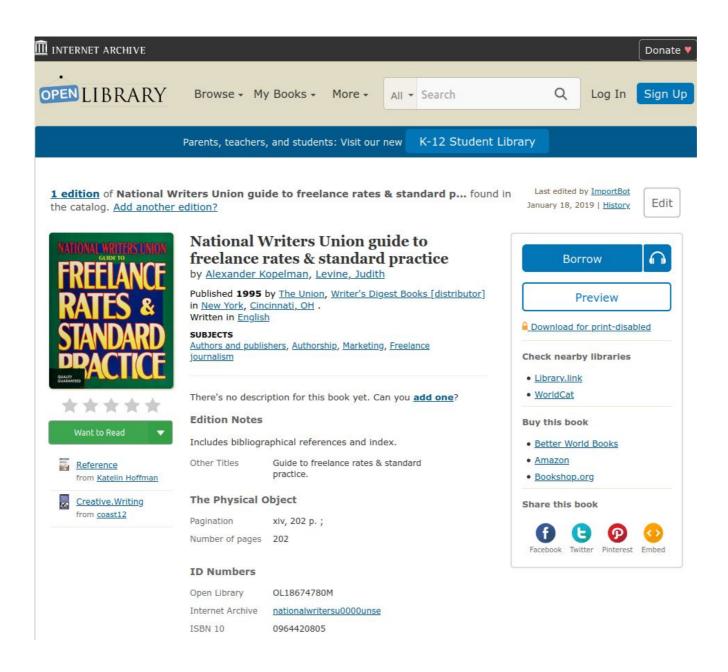
"Borrowers" Copy 3 **Physical** Web Server (digital) Storage Copy 4 (digital) Copy 1 Copy 2 Copy 5 (paper) (digital) (digital) ...and another copy for each "borrower"

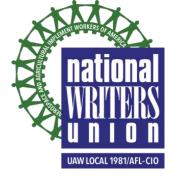
OpenLibrary.org



Books We Love







It is incumbent on

us to become our own business

managers

election, and when some

we dare not risk alienating

nally invested in the busi-

td the best way to do that

n the publisher's best inter-

needy people who live to

for the most part, honest

writing at the lowest pos-

or is thus inherently prob-

e cases

Q)



The Business of Writing

Writing is a difficult business. A

exposure, rejection, isolation,

cal hobgoblins. But, as difficult

are talking about pay rates and

The major reason

economic plight

relationship with

that hire us.

for writers'

remains our

publishers

Even many h

time paying I

alone. In a sur

ducted in 199

of responding

than \$10,000

85 percent de

erately establis

much in the I

WRITERS AND PUBLISHERS

Samuel Johnson once said that "no man but a blockhead ever wrote except for money." His sentiment is in direct conflict with the widespread image of the writer as a gaunt artist pounding away at the keyboard in blind dedication to the craft, oblivious to cold, hunger, and all other adversity. This metaphor of writing as a self-sacrificing search for meaning and self-expression is deeply rooted in our culture and in the hearts of many writers.

The metaphor is certainly flattering, since it communicates the discipline and commitment required of a writer. But it is also deceptive and dangerous,

because it pemetuates the view of writing as the fulfillment

Select Download Format

Your loan will last two weeks from the initial borrow date.

writers will agree that the tour **Encrypted Adobe PDF** from their craft. High Quality Page Images We are not talking about gr

Why are writers so poorly paid? The causes are complex and varied, but

This relationship is defined by two interconnected factors: the absence of

minimum standards for compensation and working conditions, and a pro-

found power imbalance between writers and publishers. (Note: For the sake

of convenience, publisher is used to mean "anyone who hires writers" and

freelancer to mean "any independently employed writer."]

the major underlying reason for freelance writers' economic plight remains our relationship with the publishers and other businesses and organizations Encrypted Adobe ePub Smaller File, May Contain Errors

Install Adobe Digital Editions

In order to access your downloaded book you will need Adobecompliant software on your device. The Internet Archive will administer this loan, but Adobe may also collect some information.

gainst those of a corporate ertain degree, they are at odds in some crucial respects. And big corporations, especially, with their large financial and human resources, are invariably in a better position to pro-

The problem is exacerbated by the fact that many writers feel uncomfortable with the business of writing-setting and negotiating fees, review-

ing contracts, selling rights, and so on. By shying away from these issues, we often leave the decisions about how much we get paid, and on what conditions, to the publishers. Not surprisingly, the compensation and conditions we are given are frequently not in our favor.

WRITERS AND PUBLISHERS







This is what the Internet Archive calls "Controlled Digital Lending" (CDL)

Is this like lending of printed books, or authorized digital "e-lending" of e-books? No:

- 1. Lending of printed books does not require copying.
- 2. "E-lending" is a type of <u>licensed copying</u>. The copying required for "e-lending" is specifically <u>authorized by the license</u>.
- 3. CDL is a name for a type of <u>unauthorized copying</u>.



Position Statement on Controlled Digital Lending

Last Updated September 2018

Introduction

This Position Statement on Controlled Digital Lending by Libraries ("Statement") offers a good faith interpretation of U.S. copyright law for American libraries considering how to perform traditional lending functions using digital technology while preserving an appropriate balance between the public benefit of such lending and the protected interests of private rights holders. This Statement only applies to in-copyright works, as public domain works may be distributed without restriction. This Statement is not intended to describe the upper limits of the fair use or other rights of libraries, bind the signatories to any legal position, or constitute legal advice. Because the following analysis is general, any library considering implementing controlled digital lending should consult a competent attorney to develop an appropriate program responsive to the specific needs of the institution and community.

Controlled Digital Lending ("CDL")

One of the most fundamental and socially beneficial functions of libraries is providing broad access to information by lending books and other materials to their

Co-Authored By:

Lila Bailey

Policy Counsel Internet Archive

Kyle K. Courtney

Copyright Advisor Harvard University Library

David Hansen

Associate University Librarian for Research, Collections & **Scholarly Communications Duke University Libraries**

Mary Minow

Fellow 2017-18

Berkman Klein Center for Internet and Society, Harvard University

Iason Schultz



The White Paper in support of "Controlled Digital Lending" is fundamentally flawed:

- 1. There are many discrepancies between what it describes and what the Internet Archive is actually doing.
- 2. The authors of the White Paper claim to have a good-faith belief that authors' incomes aren't being affected. But they don't appear to have any idea where our incomes actually come from, or to have made a good-faith attempt to find out which would require that they ask authors.



Controlled Digital Lending (CDL):

An appeal to readers and librarians from the victims of CDL

February 2019

































































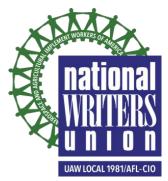














You are here: Home / Book Division / Controlled Digital Lending (CDL) / FAQ on Controlled Digital Lending (CDL)

FAQ on Controlled Digital Lending (CDL)

(download as PDF)



















































Book Division

NWU-BOOK mailing list

Controlled Digital Lending

- FAQ on Controlled Digital Lending (CDL)
- · Appeal from the victims of Controlled Digital Lending (CDL)

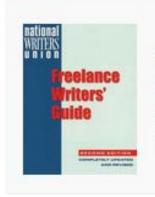


An appeal to readers and librarians from the victims of Controlled Digital Lending (CDL):

"CDL is not a victimless crime. As the victims of CDL, we want librarians, archivists, and readers to understand how they are harming the authors of the books they love by participating in CDL projects, even if they have the best of intentions....

"We appeal for a dialogue among writers, authors, publishers, and librarians on how to enable and create the digital libraries we all want, in ways that fully respect authors' rights."

2. Audiobooks generated from page images (streamed from OpenLibrary.org)



National Writers' Union freelance writers' guide

by National Writers Union (U.S.)

1 edition - first published in 2000



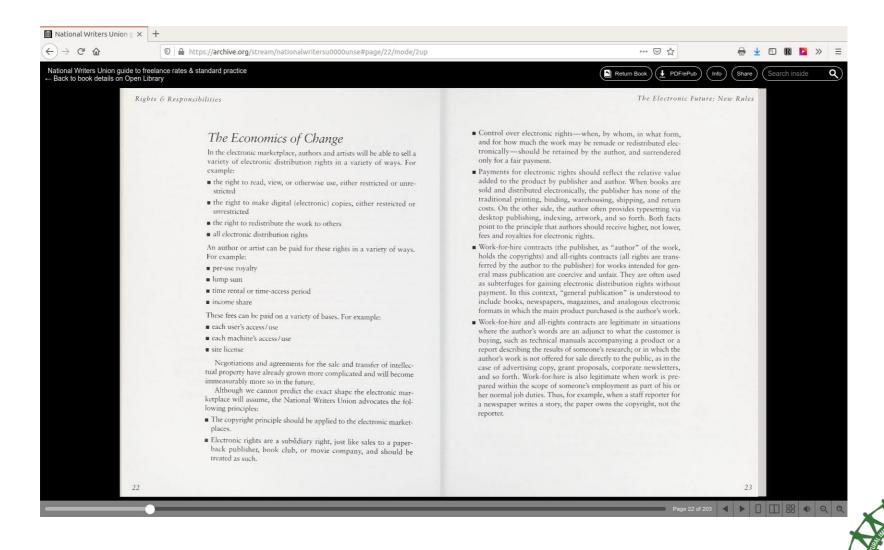




The Internet Archive has given itself (i.e. has taken) a series of unpaid upgrades:

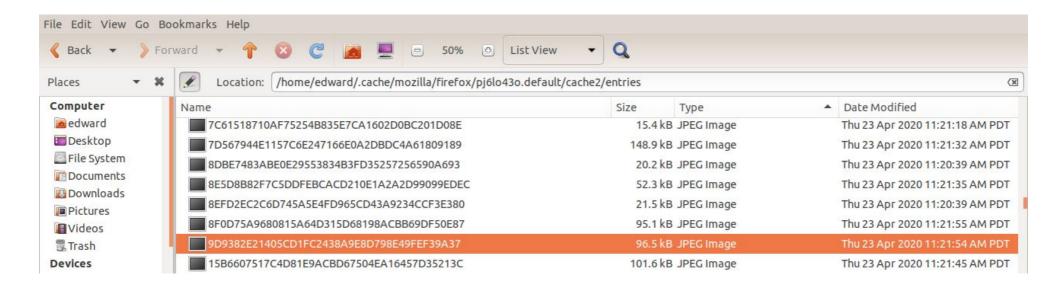
- 1. Printed book
- 2. Printed book + e-book (scanning)
- 3. Printed book + e-book + audiobook (OCR, text-to-speech)
- 4. Printed book + e-book + audiobook + "single-user" life-ofcopyright e-book and audiobook e-lending license ("Controlled Digital Lending")
- 5. Printed book + e-book + audiobook + unlimited-users e-book and audiobook e-lending license ("National Emergency Library")

3. Viewing of page images on OpenLibrary.org



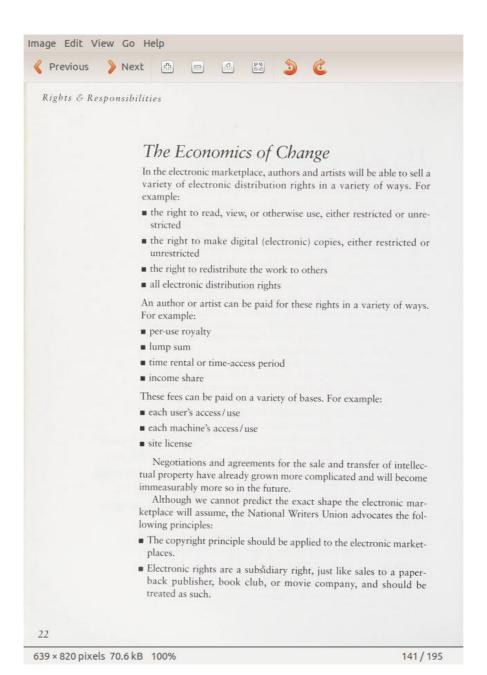
UAW LOCAL 1981/AFL-CIO

Page images (unencrypted JPG files) remain in the browser cache of each "borrower" after the e-book is "returned":



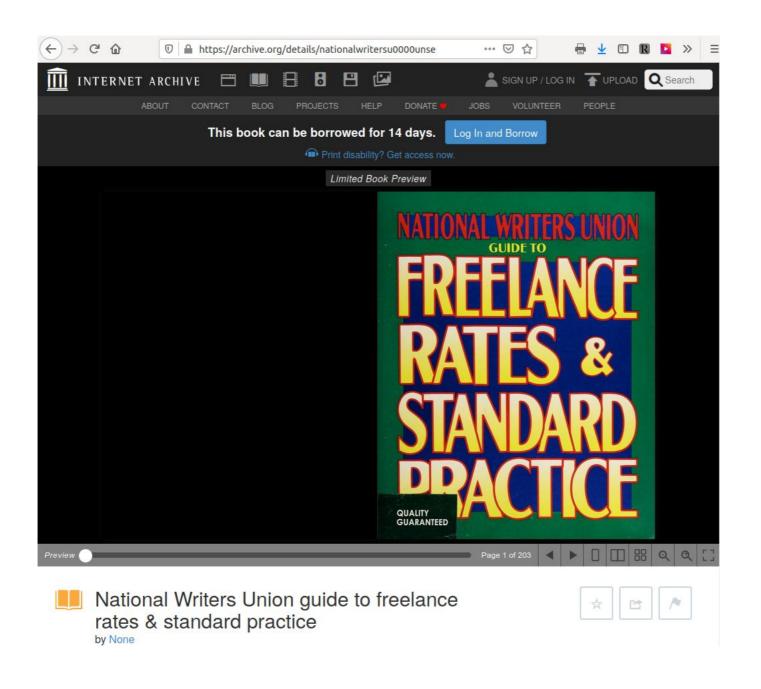


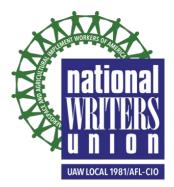
Each image file is retained by each reader after the e-book is "returned":



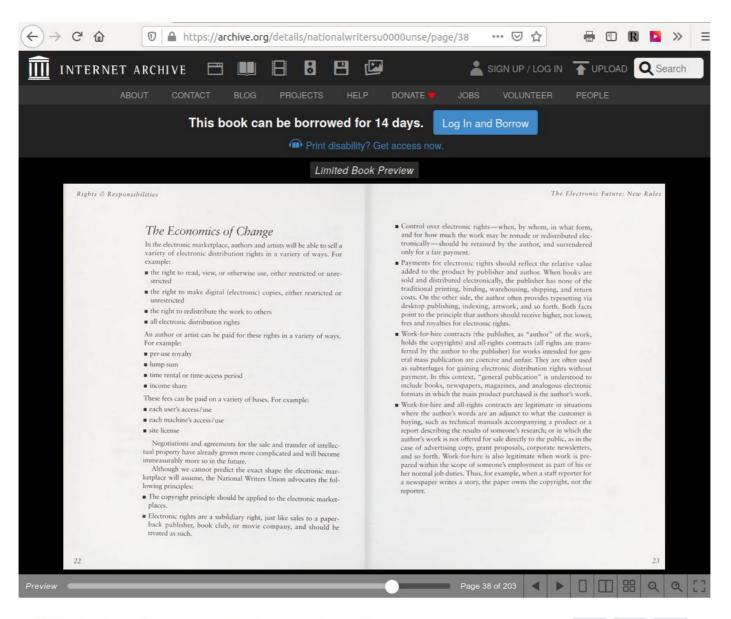


4. Viewing of page images on Archive.org





"One Web page for every page of every book"









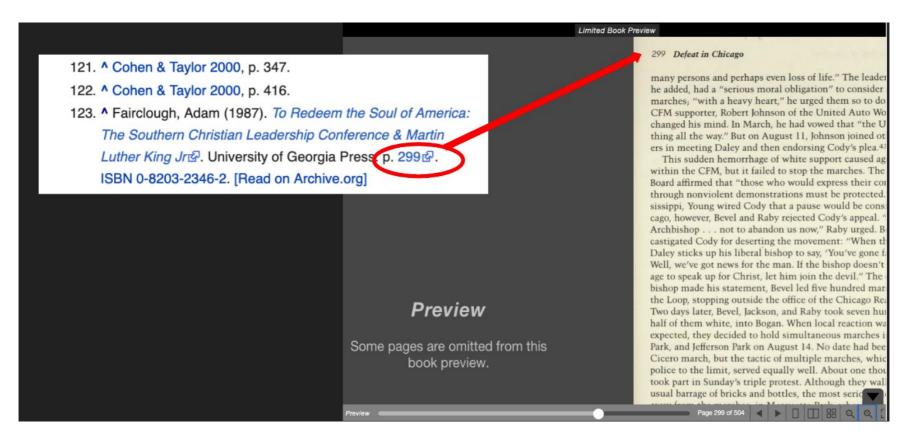




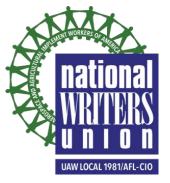
"One Web page for every page of every book" on Archive.org:

- No OpenLibrary.org or Archive.org account needed.
- No log-in or "borrowing" needed.
- Unencrypted JPG files of full-page images.
- No DRM or special software needed.
- No limit on number of simultaneous readers.
- Unique static URL on Archive.org for each page.
- Every page is included in "limited preview" (any page or pages can be viewed if you clear browser cookies).
- Includes all books on OpenLibrary.org, not just those in "National Emergency Library".

"One Web page for every page of every book" is a feature, not a bug.



(Example from Internet Archive blog of how URLs for individual pages of books are intended to be used.)



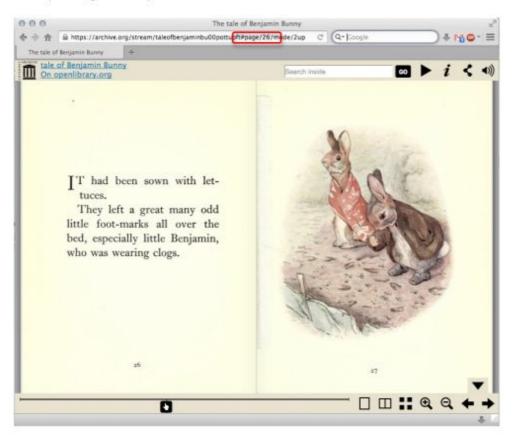


Citing Open Library on Wikipedia

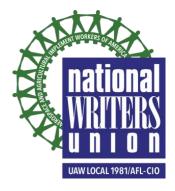
WP:BOOKLINKS

Books that are readable on Open Library can be cited directly by page. We have over one million free ebooks.

http://en.wikipedia.org/wiki/Wikipedia:BOOKLINKS



https://archive.org/stream/taleofbenjaminbu00pottuoft#page/26/mode/2up

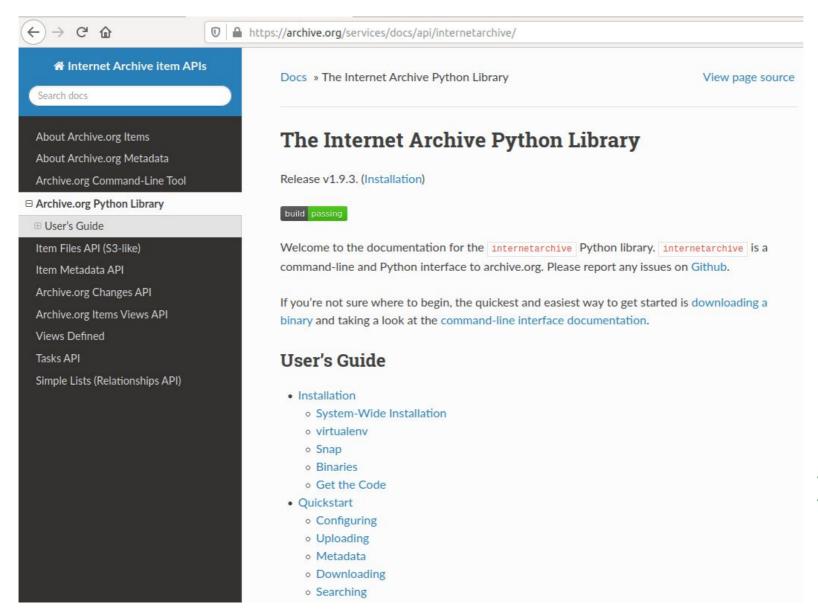


A single page of a book, or a few pages, can contain an entire "work", such as:

- Photographs
- Illustrations
- Poetry
- Flash fiction
- News articles
- Columns and essays
- Etc.

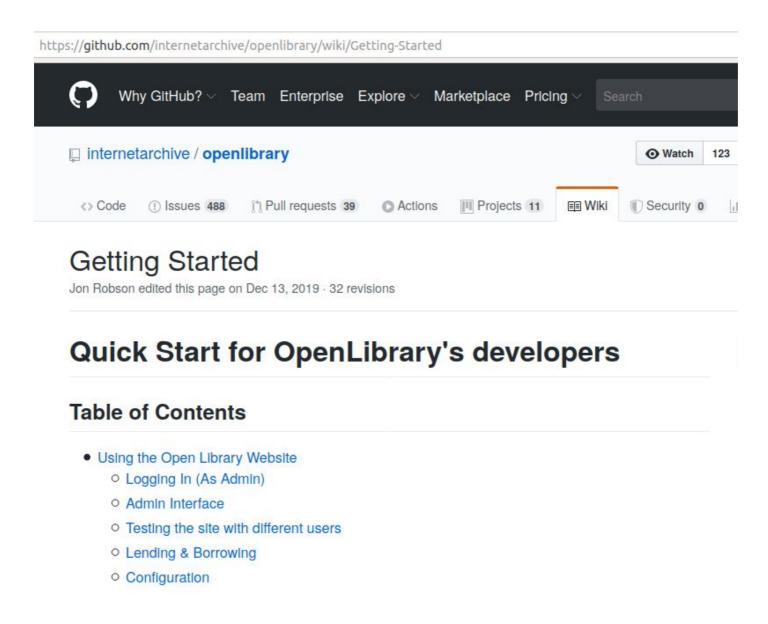


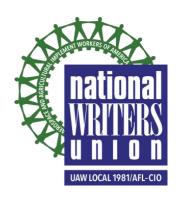
5. Application Programming Interfaces (APIs) for automated downloads of page images



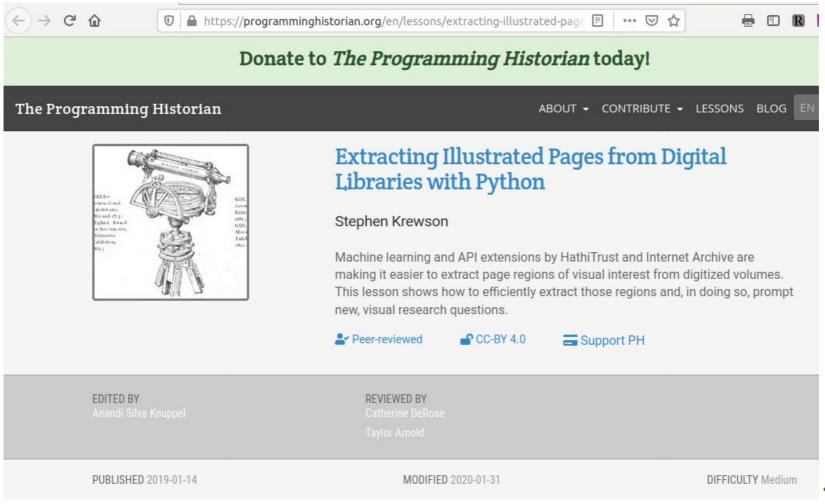


Programmers can use these APIs to automate downloads of images of book pages.





What can programmers do with these APIs?





This program identifies and downloads images of pages with illustrations.



Internet Archive

API Access

We connect to the Python API library using an Archive.org account email and password rather than API tokens. This is discussed in the Quickstart Guide. If you do not have an account, register for your "Virtual Library Card."

In the first cell of the internetarchive.ipynb notebook, enter your credentials as directed. Run the cell to authenticate to the API.

Create Volume List

The IA Python library allows you to submit query strings and receive a list of matching key-value pairs where the word "identifier" is the key and the actual identifier is the value. The syntax for a query is explained on the Advanced Search page for IA. You can specify parameters by using a keyword like "date" or "mediatype" followed by a colon and the value you want to assign that parameter. For instance, I only want results that are *texts* (as opposed to video, etc.). Make sure the parameters and options you are trying to use are supported by IA's search functionality. Otherwise you may get missing or weird results and not know why.

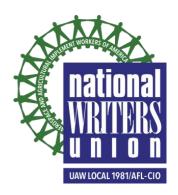
In the notebook, I generate a list of IA ids with the following code:

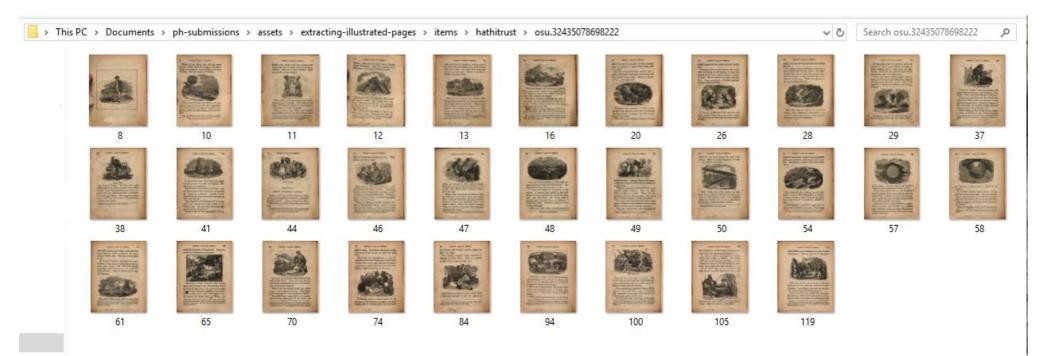
```
# sample search (should yield two results)
query = "peter parley date:[1825 TO 1830] mediatype:texts"
vol ids = [result['identifier'] for result in ia.search items(query)]
```

Visual Feature: Picture Blocks

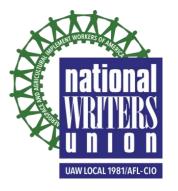
Internet Archive does not release any page-level features. Instead, it makes a number of raw files from the digitization process available to users. The most important of these for our purposes is the Abbyy XML file. Abbyy is a Russian company whose FineReader software dominates the OCR market.

All recent versions of FineReader produce an XML document that associates different "blocks" with each page in the scanned document. The most common type of block is **Text** but there are **Picture** blocks as well. Here is an example block taken from an IA Abbyy XML file. The top-left ("t" and "l") and bottom-right ("b" and "r") corners are enough to identify the rectangular block region.





(Example of images of pages with illustrations, automatically identified and downloaded. This example uses books scanned for Hathi Trust, but the article includes code to do the same thing with Archive.org.)





Prize-winning photos like those included in this book scanned by the Internet Archive may be the crown jewels of a photographer's portfolio, included in many publications and generating licensing revenue year after year.



Some patterns in the ways the Internet Archive distributes images of pages of the books it scans:

- None of these five distribution modes involve "lending". They don't send out the printed book, or the master digital copy. They make a new digital copy for each reader.
- Only the first <u>one</u> of these five distribution modes (e-books) fits the Internet Archive's descriptions of "Controlled Digital Lending" and the "National Emergency Library".
- Only the first <u>two</u> of these five distribution modes (e-books and audiobooks) are similar (except that libraries pay, and the Internet Archive doesn't pay), to licensed "e-lending".

Is any of this legal? Are authors' incomes being harmed?

These two questions depend on many of the same facts, because one of the factors that determines whether unauthorized copying is "fair use" is the effect of the copying on the market(s) for, and value of, the work.

This is a <u>factual</u> question that must be answered before even the best lawyer can apply the <u>legal</u> tests for fair use.

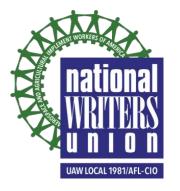
Many of the ways authors monetize rights to our backlists are invisible to lawyers, librarians, and publishers. The only way to get the facts needed to asses a fair use claim is to <u>ask authors</u> about our income sources.

Is this "fair use" under U.S. law?

"In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include —

(4) the effect of the use upon the potential market for or value of the copyrighted work."

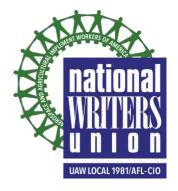
(U.S. Copyright Act, 17 U.S. Code § 107)



Can this be permitted under international law?

"It shall be a matter for legislation ... to permit the reproduction of such works in certain special cases, provided that such reproduction does not conflict with a <u>normal exploitation of the work</u> and does not unreasonably prejudice the legitimate interests of the author."

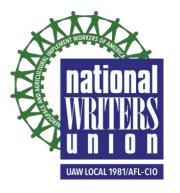
(Berne Convention on Copyright, Article 9)



What are the markets for works included in books?

The only market considered by the supporters of "Controlled Digital Lending" is the sale of books:

"The final fair use factor looks at the market effect.... For CDL ["Controlled Digital Lending"], the arguable negative impact is the loss of sales [of books] due to lending as a substitution."



Markets for works included in books:

- Sales of physical books
- Licensing of e-books and audiobooks
- Licensing by syndication and stock photo agencies, etc.
- Subscriptions to Kindle Unlimited ("Netflix for e-books")
- Downloadable digital "offprints" and excerpts
- Web sites (advertising, subscriptions, syndication fees, etc.)
- Smartphone/tablet apps (sales, in-app advertising, etc.)
- E-mail newsletters (subscriptions, advertising, etc.)
- Licensing for use in film, TV, electronic games, etc.
- Etc.



Authorized Web pages and sites through which written and graphic works that have been included in printed books are available include:

- Authors' own websites
- Other websites to which content has been licensed
- Websites for distributors of e-books, audiobooks, and other paid downloads
- Websites of stock photo licensing agencies, article syndication services, and other reproduction rights organizations (licensing agencies).

How much damage is being done?

- For works available in authorized e-books or audiobooks, the damage is at least equal to the number of times the unauthorized e-books or audiobooks are "borrowed" times the average cost per reader of an "e-lending" license.
- For works available through authorized Web sites, clicks are money. Page views can be monetized in many ways. Clickstream diversion deprives legitimate sites of revenues even if the pirate site is operated by a nonprofit entity and distributes copies for free. The damage to authors' incomes from distribution of Web pages with unauthorized images scanned from books is at least equal to the number of such page views times the average value of a page view.

How does this relate to the pandemic?

Online sales are unlikely to make up for the reduction in sales of printed books through closed bookstores.

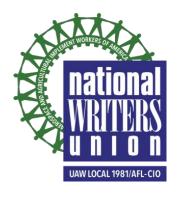
More people are reading online, however, and are searching online for reading material. Links from library catalogs to OpenLibrary.org rather than to authorized "e-lending", and links to images of pages of books on Archive.org rather than to legitimate Web pages where these works can be found, are <u>directly</u> diverting some of the readers and Web traffic that would otherwise give authors our best chance to make up, at least in part, for the reduction in our incomes from book sales.

What do we want?

- 1. Dialogue between authors (writers, photographers, illustrators, etc.) and the Internet Archive and its partners and defenders.
- 2. Engagement with the specific issues we have raised.
- 3. Dialogue with librarians about how to create digital libraries and help readers find existing digital versions of our works, while respecting authors' rights.
- 4. Participation in planning of digital libraries and in library acquisition strategies, as authors hold many of the rights and much of the metadata that libraries want.
- 5. Federal funding for fair pay for digital library use of our work, through licensing and/or a public lending right (PLR).
- 6. To speak for ourselves. Neither publishers nor well-meaning "allies" are adequate proxies for authors' own voices.

More about library book-scanning:

- NWU.org/book-division/cdl
- OpenLibrary.org
- OpenLibraries.online
- LibraryLeadersForum.org
- ControlledDigitalLending.org
- Archive.org



National Writers Union (UAW Local 1981, AFL-CIO)

nwu@nwu.org

https://www.nwu.org

Edward Hasbrouck
Co-Chair, NWU Book Division
+1-415-824-0214 (San Francisco)
edward@hasbrouck.org

