

# PDF to Web: benefits

## 1 Navigation

Section sidebar, stable URLs

## 2 Search

"fair use" across all 24 chapters

## 3 Performance

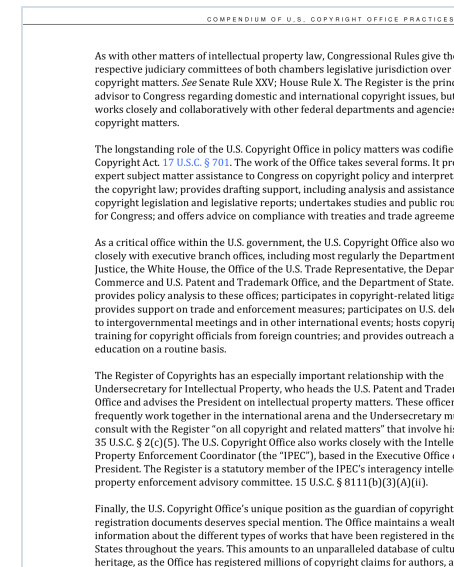
287 KB page vs. 23.6 MB PDF

## 4 Accessibility

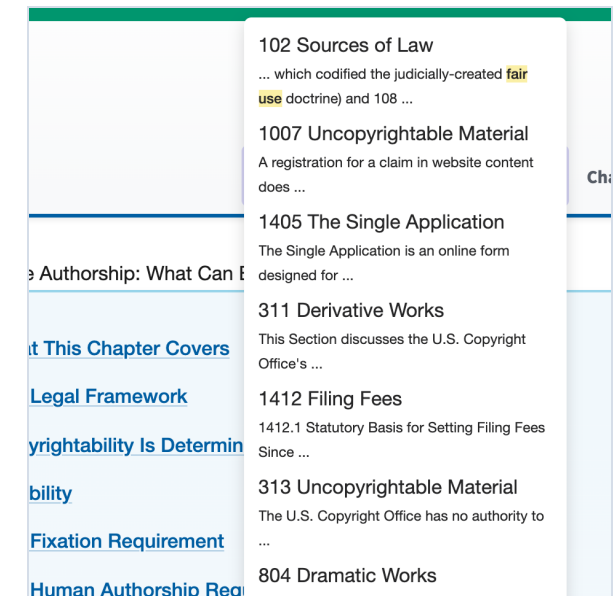
Headings, reflow, 12 languages

## Freedom for PDFs

### PDF



### WEB



# How: Regex to LLMs

## One-shot conversion

### 1 Extract the text layer

Modern PDFs embed text. The structure still has to be parsed into XHTML.

### 2 LLM writes structured XHTML

LLMs are built on transformers, which model structure across a whole chapter.

### 3 Deterministic quality checks

Programs diff the PDF text against the HTML to confirm fidelity.

2012–2024

**Regex + NLP**

`r'^SEC\. (\d+)\.'`

Rules per edge case: BillMap, House Comparative Prints



2026

**LLM + QA**

`<section id="sec-101-1">`

One pass per chapter, then QA

# Examples

## Same text, two formats

### Copyright Compendium 1,396-page PDF

### Commonwealth Drafting Manual 189-page PDF

### Statutes at Large Scanned volumes 1–125

#### PDF

The longstanding role of the U.S. Copyright Office in policy matters was codified in the Copyright Act. [17 U.S.C. § 701](#). The work of the Office takes several forms. It provides expert subject matter assistance to Congress on copyright policy and interpretation of the copyright law; provides drafting support, including analysis and assistance for copyright legislation and legislative reports; undertakes studies and public roundtables for Congress; and offers advice on compliance with treaties and trade agreements.

As a critical office within the U.S. government, the U.S. Copyright Office also works closely with executive branch offices, including most regularly the Department of Justice, the White House, the Office of the U.S. Trade Representative, the Department of Commerce and U.S. Patent and Trademark Office, and the Department of State. It provides policy analysis to these offices; participates in copyright-related litigation;

#### 6.5 The aims of plain English in legislation

- To limit the ideas expressed in a legislative sentence
- Not to allow sentences to get too long so as to interfere with communication of the central message of the provision
- Where necessary, to break up the format of sentences into paragraphs (see Chapter 7)
- To follow the standard usage and styles of modern English

110 STAT. 4196

PUBLIC LAW 104-333—NOV. 12, 1996

(B) Section 503 of the Act of December 19, 1980 (Public Law 96-550; 94 Stat. 3228; 16 U.S.C. 410ii-2).

(C) Subsections (b) and (c) of section 4 of the Act of October 15, 1982 (Public Law 97-335; 96 Stat. 1628; 16 U.S.C. 341 note).

(D) Section 7 of Public Law 89-671 (96 Stat. 1457; 16 U.S.C. 284f).

(E) Section 3(c) of the National Trails System Act (Public Law 90-543; 82 Stat. 919; 16 U.S.C. 1242(c)).

(F) Section 4(b) of the Act of October 24, 1984 (Public Law 98-473; 98 Stat. 1709; 16 U.S.C. 1261).

#### WEB

The longstanding role of the U.S. Copyright Office in policy matters was codified in the [Copyright Act](#). [17 U.S.C. § 701](#). The work of the Office takes several forms. It provides expert subject matter assistance to Congress on copyright policy and interpretation of the copyright law; provides drafting support, including analysis and assistance for copyright legislation and legislative reports; undertakes studies and public roundtables for Congress; and offers advice on compliance with treaties and trade agreements.

#### 6.5 The aims of plain English in legislation

- To limit the ideas expressed in a legislative sentence
- Not to allow sentences to get too long so as to interfere with communication of the central message of the provision
- Where necessary, to break up the format of sentences into paragraphs (see Chapter 7)
- To follow the standard usage and styles of modern English
- To use words in common use rather than elaborate or obscure ones (see the list of preferred words in the appendix)

[Public Law 104-333 — To provide for the administration of certain Presidio properties at minimal cost to the Federal taxpayer, and for other purposes.](#)

Begins inside Sec. 814 — [Read section 814 in full](#)

110 STAT. 4196

(B) Section 503 of the Act of December 19, 1980 (Public Law 96-550; 94 Stat. 3228; 16 U.S.C. 410ii-2).

(C) Subsections (b) and (c) of section 4 of the Act of October 15, 1982 (Public Law 97-335; 96 Stat. 1628; 16 U.S.C. 341 note).

(D) Section 7 of Public Law 89-671 (96 Stat. 1457; 16 U.S.C. 284f).

[compendium.adhocteam.us](https://compendium.adhocteam.us)

[aih.github.io/  
commonwealth-leg-drafting-  
manual](https://aih.github.io/commonwealth-leg-drafting-manual)

[statutes.linkedlegislation.org](https://statutes.linkedlegislation.org)