Scanning Techniques to Create Accessible PDF Documents

Many instructors provide their classes with materials in the form of scanned copies of journal or newspaper articles or book chapters, usually as PDF files. While this practice provides information that may be much more recent than textbook chapters, it does bring to the forefront the issue of how to scan these items to create accessible PDF files.

Accessible PDFs have been rendered into searchable text through the Optical Character Recognition (OCR) process available through Adobe Acrobat and some scanner software. Once the scanned image of text is recognized, it can be read aloud by screen reader technology, such as VoiceOver, NVDA, JAWS, etc. These screen reader applications are essential tools for people with various disabilities, including visual, learning, cognitive, or mobility disabilities.

Guidelines for Good Scans:

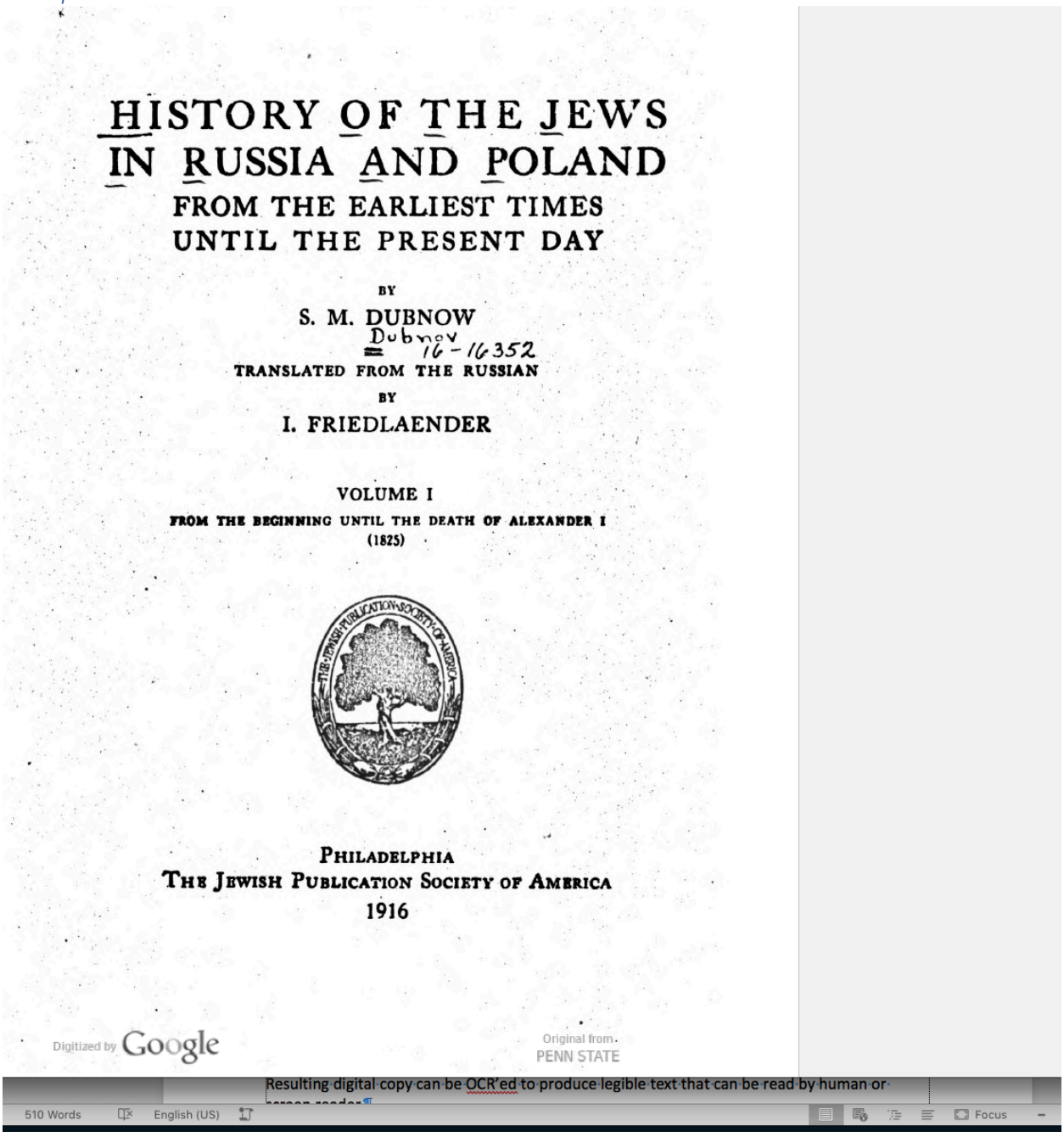
- Use "clean," high quality source material with good contrast and without blurriness. Determine if an accessible, digital version of the material already exists. Using an accessible, electronic version of the text is preferable to scanning, which introduces a certain level of replication error.
- Choose source material that is an original printing of the text, not a scan that must be scanned another time. Scanning a scan creates another level of replication error, decreasing the crispness of the dots that create text and images and increasing the likelihood and intensity of printer or paper artifacts. To make text that can be read by screen readers, programs use Optical Character Recognition (OCR) to render a scanned image into text characters that can be recognized as words by the screen reader. Printer or paper artifacts can confuse the OCR process, either stopping it if too many errors are present or creating misreads.
- Review the source material to verify that the text is complete, and clean, and not cut off at gutter or edge. Scanned documents will not improve on the quality of the original. Skewed or crooked source material, highlighting, underlining, marginalia, or stains complicate the OCR process, leading to errors or inability to produce text that can be read by screen readers.
- The OCR process can also be hampered by highly textured or aged paper, which creates artifacts in the scanned copy.
- When scanning multiple pages from a bound source, press the book firmly against the scanner bed to prevent visible gutters and scan such that only one page of source material is copied per digital page. This technique can reduce or eliminate problematic artifacts introduced from the gutter formed at the spine of bound documents and the page's curvature that can interfere with optical character recognition. Avoid shadows from subsequent or previous pages of bound source material.

- Some scanners allow users to adjust the scanning resolution. The optimal scanning resolution is usually 300-400 DPI.
- Scan in color (24-bit) ONLY if color is essential for the purpose of the document.
- If possible, save the scan as “Searchable PDF.” Name the file clearly and understandably, preferably with more than one identifier in the title, such as the course identifier, semester, author, etc.

Analyzing Scans

Problematic Scans

Example 1:



HISTORY OF THE JEWS
IN RUSSIA AND POLAND
FROM THE EARLIEST TIMES
UNTIL THE PRESENT DAY

BY

S. M. DUBNOW

Dubnov
16-16352

TRANSLATED FROM THE RUSSIAN

BY

I. FRIEDLAENDER

VOLUME I

FROM THE BEGINNING UNTIL THE DEATH OF ALEXANDER I
(1825)

PHILADELPHIA
THE JEWISH PUBLICATION SOCIETY OF AMERICA
1916

Friedlaender, I. (1916). *History of the Jews in Russia and Poland: From the Earliest Times Until the Present Day*. Philadelphia, PA: The Jewish Publication Society of America.

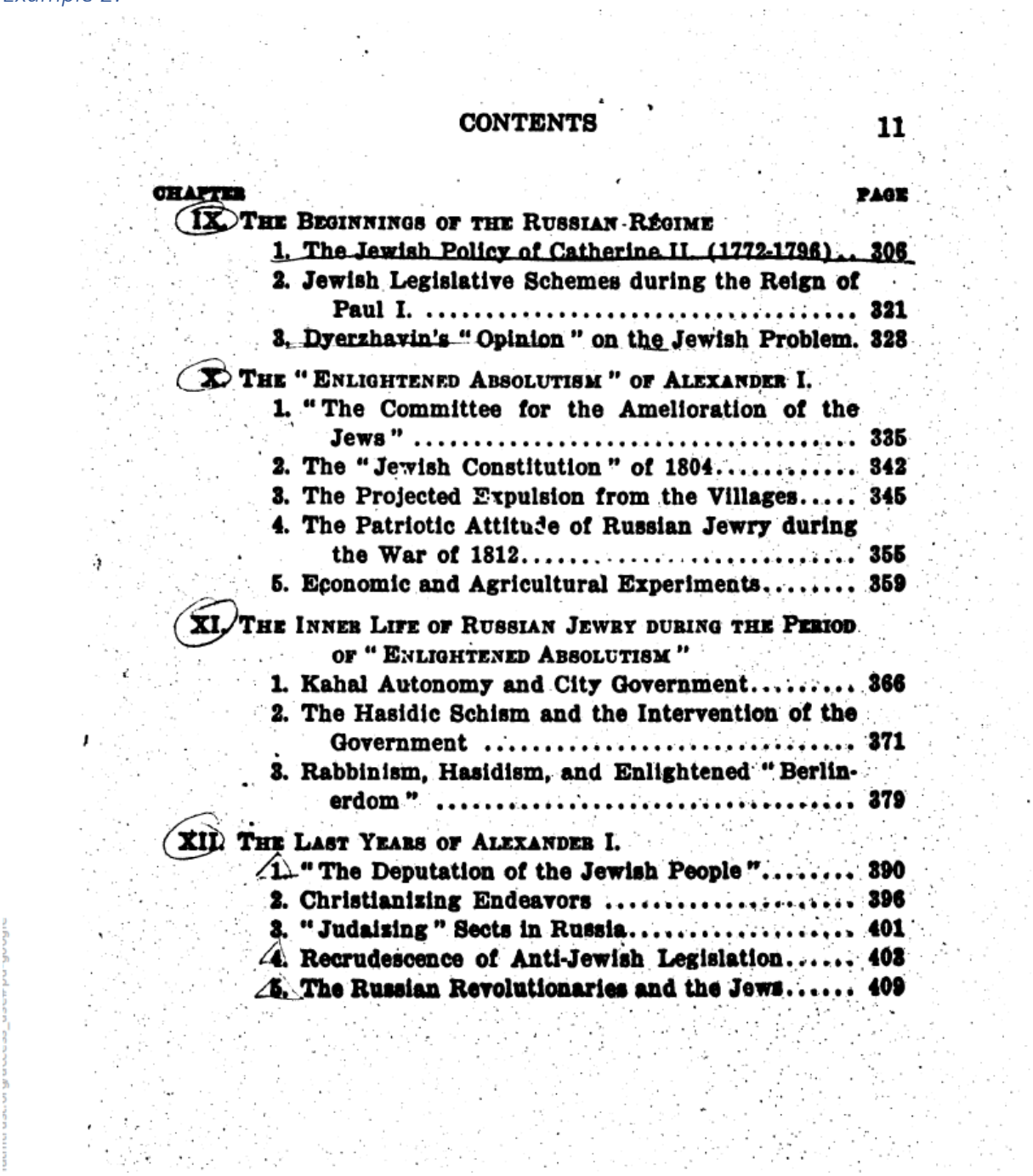
Scan Analysis: The scan above is from an old book published in 1916. The paper age or quality has caused the speckling artifacts in the scan. The page also has handwritten notes. The modern notes indicating that the page was digitized by Google and Originated from Penn State

may cause problems for the OCR because of the font color used, particularly if the scanner reads “Google” as colorized and the technician has not indicated that the scan should be processed in black and white.

Verdict: When this scan undergoes OCR, these artifacts will generate misreads, areas in which the OCR process is unable to correctly render the image into accessible text.

Correction: To create an accessible version of this text, the page should be rescanned from a cleaner copy of the book or retyped.

Example 2:



CONTENTS 11

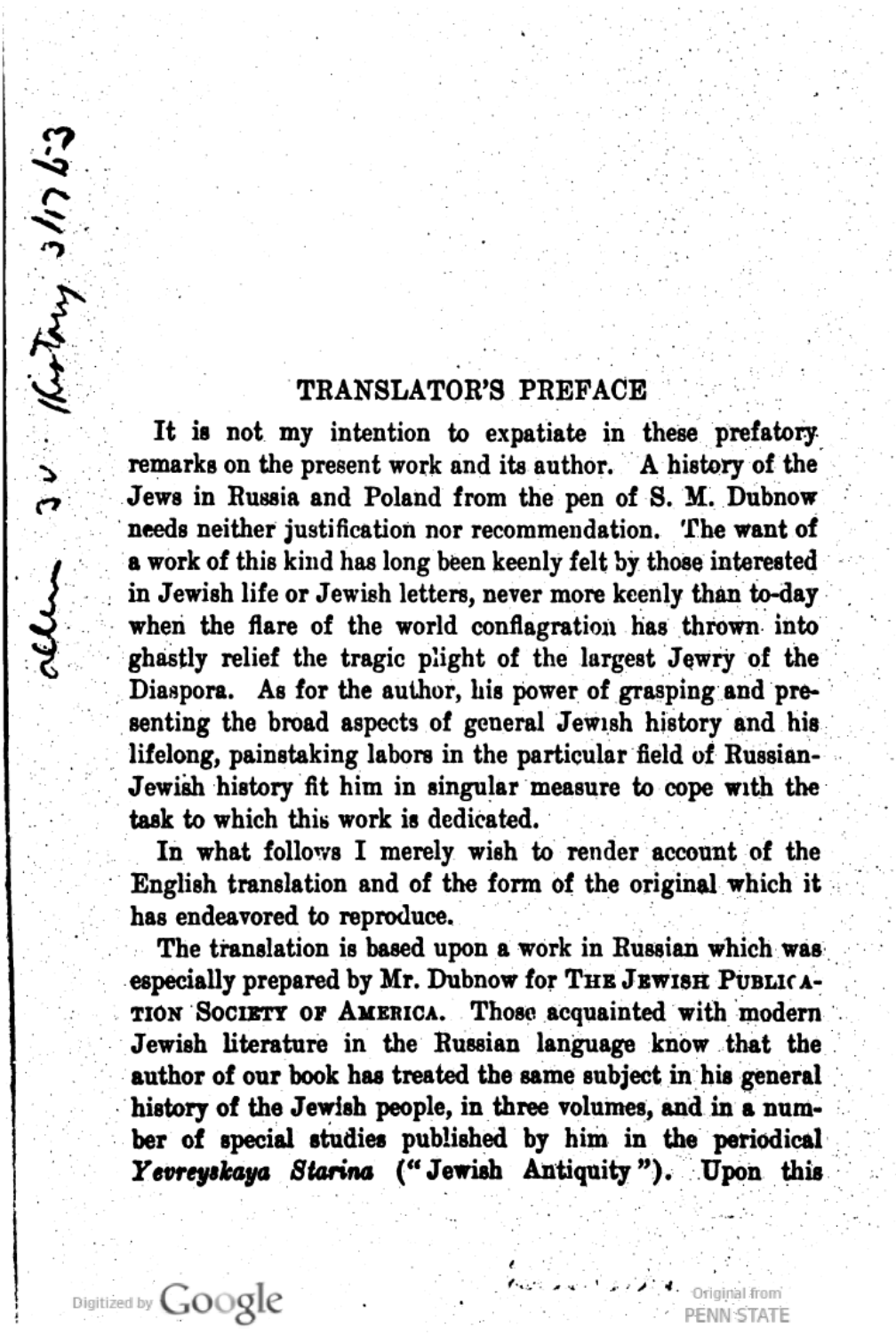
Friedlaender, I. (1916). *History of the Jews in Russia and Poland: From the Earliest Times Until the Present Day*. Philadelphia, PA: The Jewish Publication Society of America.

Scan Analysis: This scanned page also exhibits speckling caused by the age and texture of the source material, and it also has highlighting and underlining. An edge of the book's gutter is also evident in the vertical line at the lower left. Notice the blurriness of the text as well.

Verdict: The artifacts, highlighting, and blurriness of the text will cause OCR misreads.

Correction: To generate an accessible document, this text should be rescanned from a cleaner copy of the source material or retyped if one is not available.

Example 3:



TRANSLATOR'S PREFACE

It is not my intention to expatiate in these prefatory remarks on the present work and its author. A history of the Jews in Russia and Poland from the pen of S. M. Dubnow needs neither justification nor recommendation. The want of a work of this kind has long been keenly felt by those interested in Jewish life or Jewish letters, never more keenly than to-day when the flare of the world conflagration has thrown into ghastly relief the tragic plight of the largest Jewry of the Diaspora. As for the author, his power of grasping and presenting the broad aspects of general Jewish history and his lifelong, painstaking labors in the particular field of Russian-Jewish history fit him in singular measure to cope with the task to which this work is dedicated.

In what follows I merely wish to render account of the English translation and of the form of the original which it has endeavored to reproduce.

The translation is based upon a work in Russian which was especially prepared by Mr. Dubnow for THE JEWISH PUBLICATION SOCIETY OF AMERICA. Those acquainted with modern Jewish literature in the Russian language know that the author of our book has treated the same subject in his general history of the Jewish people, in three volumes, and in a number of special studies published by him in the periodical *Yevreyskaya Starina* ("Jewish Antiquity"). Upon this

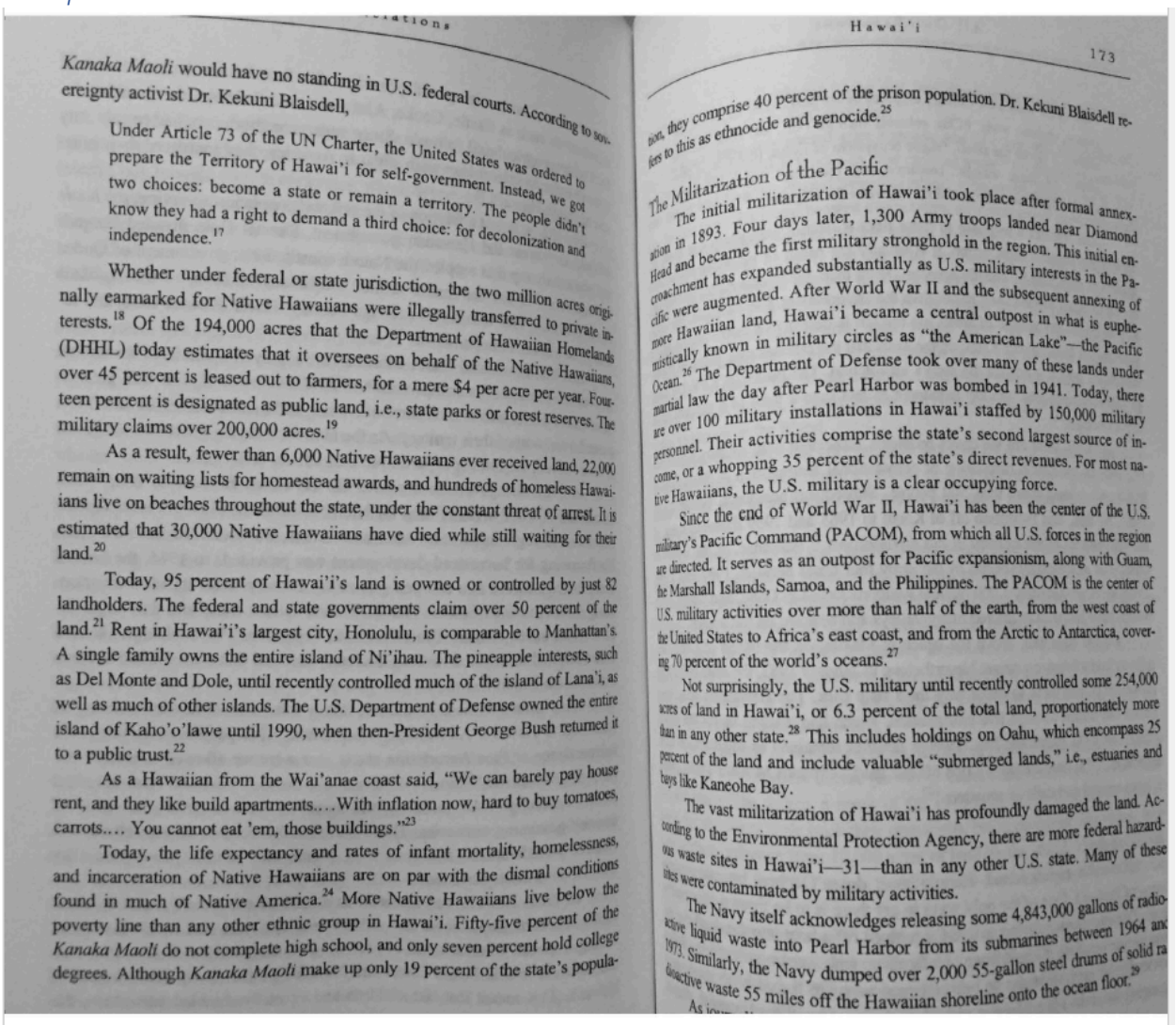
Friedlaender, I. (1916). *History of the Jews in Russia and Poland: From the Earliest Times Until the Present Day*. Philadelphia, PA: The Jewish Publication Society of America.

Scan Analysis: While this page exhibits many of the same flaws as the previous example scans, it also has vertical text in a very light font color and vertical, blurry handwriting. Near the "Original from Penn State" notation is an area of strong black dots that the OCR will attempt to resolve into readable text.

Verdict: The OCR process on this page will result in misreads in multiple, disjointed areas that will need considerable manual rework.

Correction: Correction: To generate an accessible document, this text should be rescanned from a cleaner copy of the source material or retyped if one is not available.

Example 5:



Kanaka Maoli would have no standing in U.S. federal courts. According to sovereignty activist Dr. Kekuni Blaisdell,

> Under Article 73 of the UN Charter, the United States was ordered to prepare the Territory of Hawai'i for self-government. Instead, we got two choices: become a state or remain a territory. The people didn't know they had a right to demand a third choice: for decolonization and independence.[17]

Whether under federal or state jurisdiction, the two million acres originally earmarked for Native Hawaiians were illegally transferred to private interests.[18] Of the 194,000 acres that the Department of Hawaiian Homelands (DHHL) today estimates that it oversees on behalf of the Native Hawaiians, over 45 percent is leased out to farmers, for a mere $4 per acre per year. Fourteen percent is designated as public land, i.e., state parks or forest reserves. The military claims over 200,000 acres.[19]

As a result, fewer than 6,000 Native Hawaiians ever received land, 22,000 remain on waiting lists for homestead awards, and hundreds of homeless Hawaiians live on beaches throughout the state, under the constant threat of arrest. It is estimated that 30,000 Native Hawaiians have died while still waiting for their land.[20]

Today, 95 percent of Hawai'i's land is owned or controlled by just 82 landholders. The federal and state governments claim over 50 percent of the land.[21] Rent in Hawai'i's largest city, Honolulu, is comparable to Manhattan's. A single family owns the entire island of Ni'ihau. The pineapple interests, such as Del Monte and Dole, until recently controlled much of the island of Lana'i, as well as much of other islands. The U.S. Department of Defense owned the entire island of Kaho'o'lawe until 1990, when then-President George Bush returned it to a public trust.[22]

As a Hawaiian from the Wai'anae coast said, "We can barely pay house rent, and they like build apartments....With inflation now, hard to buy tomatoes, carrots.... You cannot eat 'em, those buildings."[23]

Today, the life expectancy and rates of infant mortality, homelessness, and incarceration of Native Hawaiians are on par with the dismal conditions found in much of Native America.[24] More Native Hawaiians live below the poverty line than any other ethnic group in Hawai'i. Fifty-five percent of the *Kanaka Maoli* do not complete high school, and only seven percent hold college degrees. Although *Kanaka Maoli* make up only 19 percent of the state's popula-

tion, they comprise 40 percent of the prison population. Dr. Kekuni Blaisdell refers to this as ethnocide and genocide.[25]

The Militarization of the Pacific

The initial militarization of Hawai'i took place after formal annexation in 1893. Four days later, 1,300 Army troops landed near Diamond Head and became the first military stronghold in the region. This initial encroachment has expanded substantially as U.S. military interests in the Pacific were augmented. After World War II and the subsequent annexing of more Hawaiian land, Hawai'i became a central outpost in what is euphemistically known in military circles as "the American Lake"—the Pacific Ocean.[26] The Department of Defense took over many of these lands under martial law the day after Pearl Harbor was bombed in 1941. Today, there are over 100 military installations in Hawai'i staffed by 150,000 military personnel. Their activities comprise the state's second largest source of income, or a whopping 35 percent of the state's direct revenues. For most native Hawaiians, the U.S. military is a clear occupying force.

Since the end of World War II, Hawai'i has been the center of the U.S. military's Pacific Command (PACOM), from which all U.S. forces in the region are directed. It serves as an outpost for Pacific expansionism, along with Guam, the Marshall Islands, Samoa, and the Philippines. The PACOM is the center of U.S. military activities over more than half of the earth, from the west coast of the United States to Africa's east coast, and from the Arctic to Antarctica, covering 70 percent of the world's oceans.[27]

Not surprisingly, the U.S. military until recently controlled some 254,000 acres of land in Hawai'i, or 6.3 percent of the total land, proportionately more than in any other state.[28] This includes holdings on Oahu, which encompass 25 percent of the land and include valuable "submerged lands," i.e., estuaries and bays like Kaneohe Bay.

The vast militarization of Hawai'i has profoundly damaged the land. According to the Environmental Protection Agency, there are more federal hazardous waste sites in Hawai'i—31—than in any other U.S. state. Many of these sites were contaminated by military activities.

The Navy itself acknowledges releasing some 4,843,000 gallons of radioactive liquid waste into Pearl Harbor from its submarines between 1964 and 1973. Similarly, the Navy dumped over 2,000 55-gallon steel drums of solid radioactive waste 55 miles off the Hawaiian shoreline onto the ocean floor.[29]

LaDuke, Winona. (2017). *All Our Relations: Native Struggles for Land and Life.* Chicago, IL: Haymarket Books.

Scan Analysis: This scan was generated by using a scanning application from a mobile device. Two pages of the source material is visible on the single scan. The curvature of the book is very evident, skewing the text. Some of the text has been cut off on the lower right side, resulting in

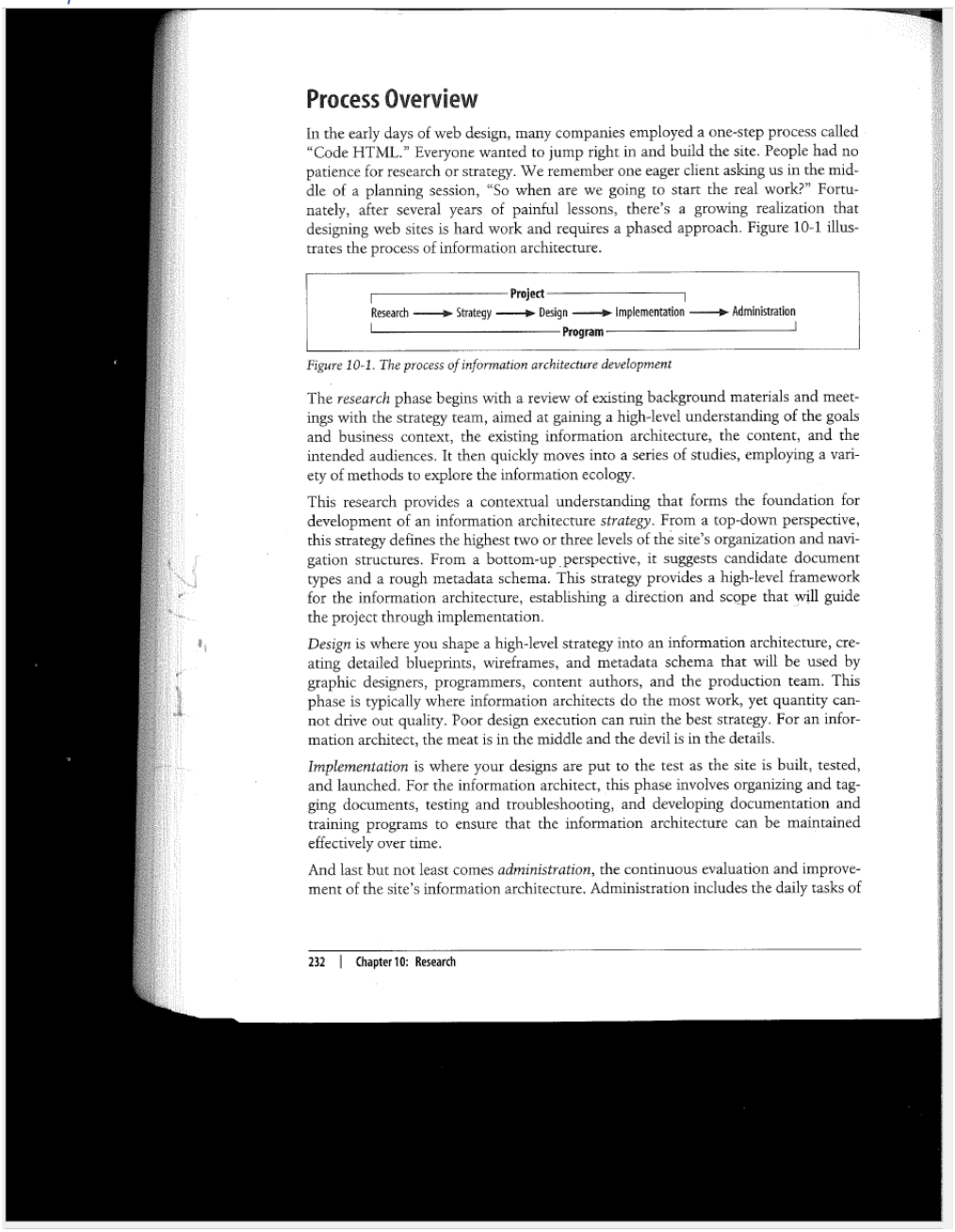
partial words. The page is pigmented, contributing to the darkness of the scan. Shadows of text from the previous and subsequent pages show through.

Verdict: The OCR process on this page will result in misreads. Incomplete words will require manual editing.

Correction: Rescan the source material on a scanner, pressing down firmly to reduce the curvature of the book. Scan to create a single page of source material per page of scanned image. Increase the scanner's brightness setting to decrease the pigmentation of the page and shadows of previous and subsequent page text.

Usable Scans

Example 1:



Process Overview

In the early days of web design, many companies employed a one-step process called "Code HTML." Everyone wanted to jump right in and build the site. People had no patience for research or strategy. We remember one eager client asking us in the middle of a planning session, "So when are we going to start the real work?" Fortunately, after several years of painful lessons, there's a growing realization that designing web sites is hard work and requires a phased approach. Figure 10-1 illustrates the process of information architecture.

Figure 10-1. The process of information architecture development

The *research* phase begins with a review of existing background materials and meetings with the strategy team, aimed at gaining a high-level understanding of the goals and business context, the existing information architecture, the content, and the intended audiences. It then quickly moves into a series of studies, employing a variety of methods to explore the information ecology.

This research provides a contextual understanding that forms the foundation for development of an information architecture *strategy*. From a top-down perspective, this strategy defines the highest two or three levels of the site's organization and navigation structures. From a bottom-up perspective, it suggests candidate document types and a rough metadata schema. This strategy provides a high-level framework for the information architecture, establishing a direction and scope that will guide the project through implementation.

Design is where you shape a high-level strategy into an information architecture, creating detailed blueprints, wireframes, and metadata schema that will be used by graphic designers, programmers, content authors, and the production team. This phase is typically where information architects do the most work, yet quantity cannot drive out quality. Poor design execution can ruin the best strategy. For an information architect, the meat is in the middle and the devil is in the details.

Implementation is where your designs are put to the test as the site is built, tested, and launched. For the information architect, this phase involves organizing and tagging documents, testing and troubleshooting, and developing documentation and training programs to ensure that the information architecture can be maintained effectively over time.

And last but not least comes *administration*, the continuous evaluation and improvement of the site's information architecture. Administration includes the daily tasks of

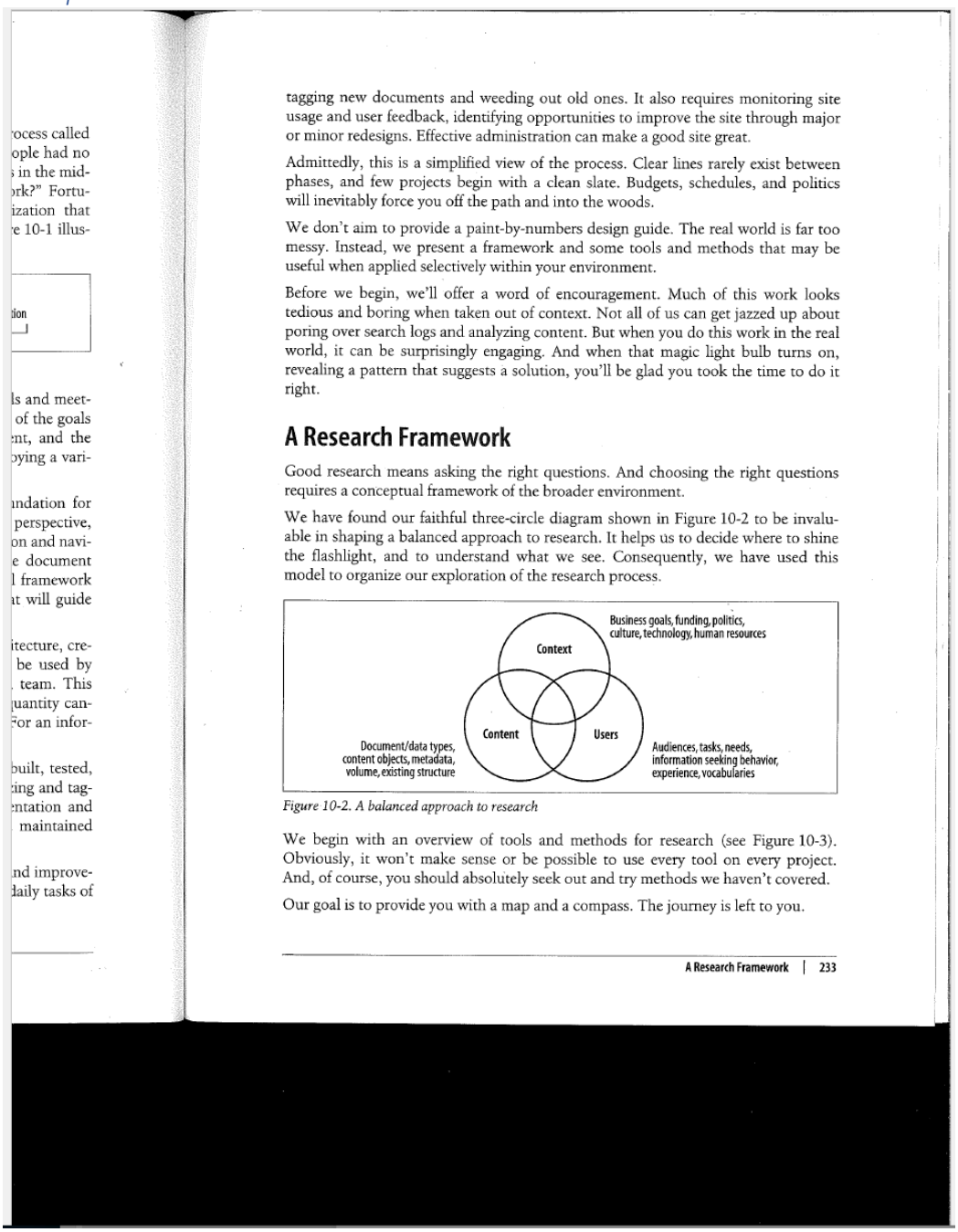
Morville, Peter, et. al., (2006). *Information Architecture for the World Wide Web: Designing Large-Scale Web Sites, 3rd Edition*. Sebastopol, CA: O'Reilly Media.

Scan Analysis: The scan has black banding from the difference between the book's size and the scanner's scanning area. Shadows of previous pages and some artifacts are visible along the left side. A shadow of the gutter between the left and right pages of the book can be seen along the right margin. The text itself is crisp and clean.

Verdict: Although the page is not perfectly scanned, it can be edited by Adobe Acrobat before the OCR process to remove the problematic aspects.

Correction: Adobe Acrobat's Crop Tool can extract the text from the banding and shadowing, resulting in a page that can be resolved through OCR into accessible text.

Example 2:



tagging new documents and weeding out old ones. It also requires monitoring site usage and user feedback, identifying opportunities to improve the site through major or minor redesigns. Effective administration can make a good site great.

Admittedly, this is a simplified view of the process. Clear lines rarely exist between phases, and few projects begin with a clean slate. Budgets, schedules, and politics will inevitably force you off the path and into the woods.

We don't aim to provide a paint-by-numbers design guide. The real world is far too messy. Instead, we present a framework and some tools and methods that may be useful when applied selectively within your environment.

Before we begin, we'll offer a word of encouragement. Much of this work looks tedious and boring when taken out of context. Not all of us can get jazzed up about poring over search logs and analyzing content. But when you do this work in the real world, it can be surprisingly engaging. And when that magic light bulb turns on, revealing a pattern that suggests a solution, you'll be glad you took the time to do it right.

A Research Framework

Good research means asking the right questions. And choosing the right questions requires a conceptual framework of the broader environment.

We have found our faithful three-circle diagram shown in Figure 10-2 to be invaluable in shaping a balanced approach to research. It helps us to decide where to shine the flashlight, and to understand what we see. Consequently, we have used this model to organize our exploration of the research process.

Figure 10-2. A balanced approach to research

We begin with an overview of tools and methods for research (see Figure 10-3). Obviously, it won't make sense or be possible to use every tool on every project. And, of course, you should absolutely seek out and try methods we haven't covered.

Our goal is to provide you with a map and a compass. The journey is left to you.

Morville, Peter, et. al., (2006). *Information Architecture for the World Wide Web: Designing Large-Scale Web Sites, 3rd Edition*. Sebastopol, CA: O'Reilly Media.

Scan Analysis: A partial page of the source material is visible along the page's left. Some shadows of subsequent pages can be seen on the right margin. A black band lies along the bottom of the page from a size mismatch between the book's size and the scanning area. The text, however is clean and free from artifacts.

Verdict: This scan can be corrected using Adobe Acrobat's tools.

Correction: Adobe Acrobat's Crop Tool can extract the text from the banding and shadowing, resulting in a page that can be resolved through OCR into accessible text.