

# Web Scraping & Provenance Logging Procedures for the Criminal Justice Administrative Records System

---

## Introduction

The Criminal Justice Administrative Records System (CJARS) works to ensure that all of its data were acquired through transparent and legitimate means. CJARS complies with both the requirements established by the University of Michigan Institutional Review Board and the U.S. Computer Fraud and Abuse Act (CFAA). This document outlines the workflow for developing web scrapers in accordance with the CJARS standards.

## Scraping Target Criteria

The two factors that determine whether an agency's website is "scrapeable" are its structure and security measures in place. Any agency that provides public access to their offender or criminal court data via (i) Application Programming Interfaces (APIs),<sup>1</sup> (ii) direct downloads,<sup>2</sup> or (iii) search forms is considered for web scraping. Agencies who provide subscription services or bulk purchase agreements are reviewed by a project manager or the principal investigator for acquisition appraisal. Agency websites using CAPTCHA or other security measures to prevent access by robots or non-human entities are not targeted; this would be a direct violation of the website's terms of use. Instead, data from such agencies should be acquired through a Freedom of Information Act (FOIA) request or a data use agreement with the agency.

---

<sup>1</sup>One agency that provides API service to access court data is the Wisconsin Circuit Court Access system. For such API-enabled targets, save the full JSON response as a \*.json file.

<sup>2</sup>Some agencies allow users to directly download their database. For an example, see Nebraska Department of Correctional Services ([https://dcs-inmatesearch.ne.gov/Corrections/COR\\_download.htm](https://dcs-inmatesearch.ne.gov/Corrections/COR_download.htm)).

## Scraping Practice Guidelines

All project web crawlers must conform to specific guidelines to ensure scraping efforts do not jeopardize other data collection efforts or CJARS overall. To ensure compliance with the Access Provision laid out in the CFAA, scraping efforts must consider both a website's terms of service (ToS) as well as the privacy policy of using personal information. Although ToS typically contains brief disclaimers, some of the more scraper-friendly agencies may also share additional guidelines specifically pertaining to scraping, such as when to run a scraper (Figure 1). Violating ToS can expose the project and the university to legal liabilities and reputational concerns. Therefore, reviewing ToS carefully is critical to ensure CJARS avoids improper data collection behavior.

In addition, the robots exclusion protocol (robots.txt) provides instructions about the listed directories of a website to robots, an automated program for harvesting web data by emulating human behavior. Figure 2 provides an example of such a file. From these text files, the only field that will be singled out in this document is the `Crawl-delay` field for rate limiting crawl requests. As an example, if an agency's robots.txt has `Crawl-delay: 10`, then for each interaction the crawler makes with the website that causes the page to refresh or reload, the crawler should wait at least 10 seconds before proceeding with the next task. For websites that do not explicitly specify a rate limiter, the default strategy should be to use `from selenium.webdriver.support.ui import WebDriverWait` function to wait up to 30 seconds for an HTML element to load.

To avoid denial of service, IP blocking, and thereby jeopardizing data use negotiations with agency, CJARS limits the number of threads (or workers) when running a crawler in parallelization or distributed queues. A crawler can be assigned a maximum of 4 threads out of consideration for the CJARS' scraper server's resource constraints as well as the target's capacity constraints. If 4 concurrent threads is causing target servers to slow and potentially crash, the number of concurrent threads should be reduced as needed. Under no circumstance should a scraper be repeatedly launched against a target if the script consistently leads to target server crash.

**Figure 1:** Example of ToS with Scraping Guidelines

The screenshot shows a search interface with the following fields: Court Registry Account, Party SPN, Public Image Number, Case Status (dropdown), and Defendant Status (dropdown). Below the fields are 'Search' and 'Reset' buttons. A notice below the buttons states: 'Several County Criminal Courts at Law and District courts have reported intermittently slow system performance for Harris County District Clerk web site users. These slowdowns are causing longer than usual wait times for external users such as public users, law firms, and commercial vendors. We ask commercial customers who run data scraping scripts to help ease the system slowdown time by moving their large data pulls to between the hours of 6:00 P.M. Central Standard Time and 6:00 A.M. Central Standard Time.' A note below this states: 'Note: The Public Access to the Harris County District Clerk Court Electronic Records, its Help Desk, its Call Center and/or the Harris County District Clerk reserves the right to suspend/reduce service or restrict access to any account causing an unacceptable level of congestion or disrupting operations for the following: -- Harris County District Clerk Court Electronic Records -- the Court system The District Clerk supports -- its Call Center -- its Help Desk -- Another Public user'.

Texas Harris County District Clerk website provides guidelines for the best time to run web scrapers.

## Documentation & Provenance

CJARS is required to provide documentation of legal provenance for all data brought to the U.S. Census Bureau. To satisfy these requirements, the crawler should archive the following files within the relevant “./documentation/provenance” directories:

**Figure 2:** Example of robots.txt

```
User-agent: *
Disallow: /criminalBackgroundSearch/
User-agent: *
Allow: /
User-agent: *
#Crawl-delay: 30
```

- Robots exclusion protocol (robots.txt)
  - If there is no robots.txt, save the contents of the page as “robots\_404.txt”
- Screenshot of source webpage (PDF or PNG)
- Perma.cc link of source (HTML page)
  - If the main page does not contain ToS information, a separate Perma.cc link should be created for the page with ToS details

Perma.cc provides a permanent third-party hosted snapshot of the ToS when scraping is launched. This important documentation serves as proof that our collection efforts were not improper even if ToS are modified at a future date by a target site.

When a website also provides a data dictionary, this information should be saved in the documentation folder for future reference. See Figures 3 and 4 below for examples of relevant folder layout.

**Figure 3:** Example of “documentation” folder contents

Name	Date modified	Type	Size
provenance	6/19/2019 12:42 PM	File folder	
MDOC - Definitions_Glossary.pdf	5/7/2019 5:08 PM	Adobe Acrobat D...	160 KB
MDOC - Documentation.pdf	5/7/2019 5:08 PM	Adobe Acrobat D...	198 KB

The “MDOC - DOCUMENTATION.pdf” provides instructional information on the website search engine as well as its ToS while “MDOC - Definitions\_Glossary.pdf” is the data dictionary that explains the definitions of the variables.

**Figure 4:** Example of “provenance” folder contents

Name	Date modified	Type	Size
permacc_9HEN-FVYL_2019061909191023...	6/19/2019 9:19 AM	Internet Shortcut	1 KB
robots.txt	6/19/2019 9:19 AM	Text Document	3 KB
screenshot-2019-06-19_09-19-16.png	6/19/2019 9:19 AM	PNG File	588 KB

The three files that should always be included in the provenance.