Ossi Mätäsaho

The Regulation of Web Scraping

A brief Literature Review on Legal Frameworks and Access Control Mechanisms

School of Technology and Innovations
Bachelor's thesis in Technology
Data Architecture

Vaasa 2025

Statement of AI Usage

In accordance with academic integrity standards and institutional guidelines regarding artificial intelligence use in academic work, I hereby declare the following utilisation of AI tools in the preparation of this bachelor's thesis:

Artificial Intelligence Assistant Used

- Provider: Anthropic
- AI Model: Claude 3.7 Sonnet

Areas of Application:

- AI was used as a sounding board when interpreting EU directives and CJEU rulings.
- Questions critical for filtering the source material were asked from the AI used.
- Some of the source material was given to AI and tasked to summarise the content.
- AI translated one Russian research article into English.

Responsibility Statement

I acknowledge full responsibility for:

- The accuracy and validity of all content presented in this thesis.
- The originality of analysis.
- The scholarly contribution and academic integrity of this work in its entirety.
- All conclusions, interpretations, and recommendations.

This declaration has been prepared in compliance with the University of Vaasa guidelines for the use of artificial intelligence in teaching and learning.

Legal Disclaimer

This thesis examines web scraping regulation for academic purposes only. The legal analysis does not constitute legal advice and reflects the author's understanding as of May 2025. Interpretations of legislation, case law, and technical implementations should not guide decisions about specific web scraping activities. Consult qualified legal counsel before engaging in web scraping. Neither the author nor the University of Vaasa accepts liability for actions based on this thesis.

UNIVERSITY OF VAASA
School of …
Author: Ossi Mätäsaho
Title of the Thesis: The Regulation of Web Scraping : A brief Literature Review on Legal Frameworks and Access Control Mechanisms
Degree: Bachelor of Science in Technology
Programme: Data Architecture
Supervisor: Maarit Välisuo
Year: 2025 **Sivumäärä:** 58

ABSTRACT:
Web scraping on vakiintunut olennaiseksi keinoksi digitaaliseen tiedonkeruuseen, mutta sen oikeudellinen asema on edelleen epäselvä. Tässä tutkielmassa tarkastellaan web scrapingin sääntelyä kahdesta vuorovaikutteisesta näkökulmasta kirjallisuuskatsauksen keinoin. Aineistona on käytetty tieteellistä kirjallisuutta, oikeudellisia lähteitä sekä teknisiä raportteja, joiden pohjalta on muodostettu kokonaiskuva nykyisestä sääntely-ympäristöstä ja sen haasteista.

Tarkastelu keskittyy tutkimuskysymyksiin: millaisia oikeudellisia haasteita web scrapingiin liittyy, sekä mitä menetelmiä yleisesti käytetään verkkosivustojen automatisoidun käytön rajoittamiseen. Tutkimuksen tavoitteena on lisätä ymmärrystä web scrapingin sääntelyn nykytilasta, sekä oikeudellisten ja teknisten ratkaisujen keskinäisestä suhteesta.

Laillinen viitekehys rakentuu erityisesti Euroopan unionin tietokantojen suojaa koskevasta sääntelystä, kuten tietokantadirektiiveistä ja teksti- ja datanlouhintaa koskevista poikkeuksista direktiivissä 2019/790. Lisäksi huomio kiinnitetään eri oikeudenkäyttöalueiden hajanaisiin tulkintoihin ja siihen, kuinka nämä vaikuttavat web scrapingin laillisuuden arviointiin. Laillisen viitekehyksen lisäksi esitellään yleisiksi havaitut pääsynhallintamekanismit, jotka on jaettu teknisiin ja hallinnollisiin menetelmiin.

Työssä on havaittu, että web scrapingin oikeudellinen asema on useimmiten tulkinnanvarainen, ja eri mekanismien oikeudellinen sitovuus vaihtelee myös tapauskohtaisesti. Tekniset pääsynhallintamekanismit eivät aina estä tehokkaasti kehittyneitä automatisoituja järjestelmiä ja hallinnollisten menetelmien, kuten käyttöehtojen oikeudellinen painoarvo riippuu niiden teknisestä toteutuksesta. Näin ollen web scrapingin sääntely on nykytilassaan hyvinkin epäselvä, jolloin tulkinnat sen laillisuudesta voivat vaihdella paljonkin.

Tutkielma auttaa ymmärtämään automatisoitua tiedonkeruuta koskevaa monimutkaista sääntely-ympäristöä ja havainnollistaa kuinka lainsäädäntö, ja pääsynhallintamekanismit ovat olennaisesti sidoksissa toisiinsa. Jatkotutkimusehdotuksina esitetään: syksyllä 2025 sovellettavaksi tulevan EU Datasäädöksen vaikutuksia web scrapingiin, AI-kehityksen vaikutusta teknisten rajoitteiden toimivuuteen, vankkojen eettisten viitekehysten muodostamista web scrapingin harjoittamiseen.

KEYWORDS: web scraping, access control methods, law, database protection, text and data mining

Contents

Figures

Abbreviations

REP - Robots Exclusion Protocol

ToS - Terms of Service

CAPTCHA - Completely Automated Public Turing test to tell Computers and Humans Apart

IP - Internet Protocol

HTML - HyperText Markup Language

XML - eXtensible Markup Language

HTTP - HyperText Transfer Protocol

CJEU - Court of Justice of the European Union

1 Introduction

The digital marketplace has fundamentally transformed consumer decision-making processes, with price comparison services emerging as critical tools for informed purchasing decisions. Stankevich (2017, p.8) observes that consumers increasingly seek solutions that integrate multiple functionalities. In response, digital marketplace is seeing an increase in services that have adapted to this change in customer requirements. Specifically, price comparison services have evolved from simple price aggregators into comprehensive platforms that encompass various stages of traditional consumer decision-making process (model illustrated in Figure 1). The three middle stages of this process – Information Search, Alternatives Evaluation, and Purchase Decision – are also particularly crucial for successful implementation of a digital marketplace solution. For example, market leaders such as Skyscanner and Booking.com both command significant market share (Curry, 2025) by providing efficient solutions to customers through integrating the three middle stages of the customer decision-making process.

Figure 1 Traditional model of the consumer buying process (Stankevich, 2017, p.10).

However, the implementation of services like Skyscanner and Booking.com is a process characterized by complexity as it includes, in addition to the technical challenges, also legal and ethical perspectives that must be addressed. Data sourcing for these services can be done either with APIs (application programming interfaces) or through web scraping, usually both. Sourcing data by web scraping poses legislative and ethical challenges as it is not a transaction between service provider and a customer, which is the case with APIs.

Digital marketplace is just an example of the various fields where web scraping can be utilized. As Luscombe et al. (2022, p. 1024) note, some of the fields where web scraping is now crucial include the likes of:

- Criminology
- Communication Science
- Economics
- Organization Studies
- Policy Studies
- Political Science
- Psychology
- Sociology

The growing importance of web scraping extends beyond the listed and rather academic fields, finding critical applications in journalism, market research, and public policy analysis. Its ability to extract vast amounts of digital data has revolutionized information gathering across both scholarly and professional domains. In contrast, to the growing importance of web scraping the legal landscape of it has not been addressed thoroughly. The motivation of this research is to synthesize information about some of the legal considerations regarding web scraping and relevant access control mechanisms.

1.1 Research Objectives and Scope

This thesis aims to explore how web scraping is regulated, focusing on how legislative frameworks address this practice and what mechanisms are used to limit or permit it. The central objective is to examine how current legislation interprets and governs automated data collection, particularly highlighting the legal interpretations of explored access control mechanisms. The legislative perspective specifically considers directives of the European Parliament and the Council, as they establish a harmonised legal foundation across EU member states. Additionally, select U.S. legal papers are included to

provide insights into the enforceability of website terms of use. The research objective will be accomplished through answering the following questions:

- What are the primary legal challenges associated with web scraping under current legislative frameworks and judicial interpretations, and how do these challenges affect the permissibility of automated data collection?
- What are the general techniques used to control automated access to website data? And what legal implications constitute from circumventing them?

To provide a clearer understanding of the research context and findings, this thesis includes an overview of the fundamentals of web scraping. The focus is limited to defining web scraping, outlining its general implementation process, explaining its relevance in digital marketplace environments and overview of other application areas. A comprehensive technical analysis of various implementation technologies is beyond the scope of this thesis. However, selected examples of commonly used tools will be presented to illustrate different aspects of the foundational concepts.

1.2 Research Methodology

This thesis will be conducted as a narrative literature review that applies the theoretical process presented by Juntunen & Lehenkari (2021, p. 336) in practice. The research methodology used in this thesis is illustrated in Figure 2.

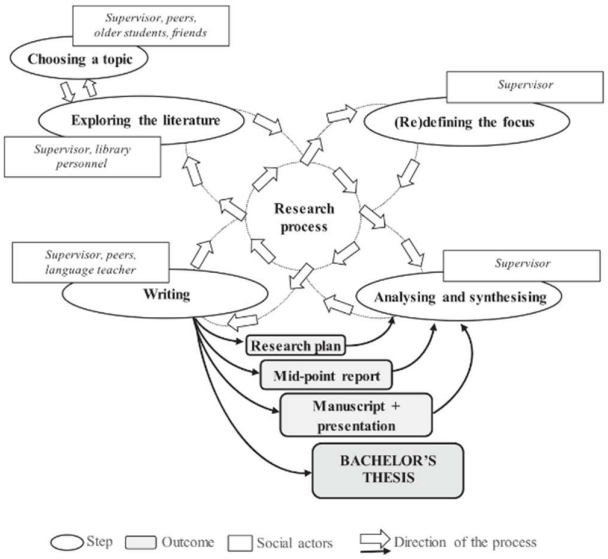


Figure 2 The narrative literature review process by Juntunen & Lehenkari (2021).

The primary literature search was conducted using three databases:

- Tritonia-Finna: A multidisciplinary academic database offering access to the collections and information services of the Tritonia Academic Library, primarily supporting higher education and research institutions in Finland.
- Google Scholar: A freely accessible search engine that indexes scholarly articles across a wide range of disciplines and sources, providing broad international coverage to complement specialized databases.
- ResearchGate: An academic networking platform where researchers share publications, collaborate, and often provide open access to their work, including preprints and full-text articles.

In addition to the primary literature search, snowball sampling methodology was employed to ensure comprehensive coverage of the relevant literature within the research

domain. Snowball sampling involves the systematic identification of additional relevant sources through examination of the reference lists contained within previously identified literature, thereby expanding the collection of relevant texts through bibliographic networks. *Zotero* was used to manage all the source material for this thesis, including its browser extension for efficient integration to the database.

1.3 Research Structure

This thesis is structured to six main chapters, each of which builds progressively towards answering the research questions defined before. The structure listed below ensures a comprehensive approach, starting from foundational understanding of web scraping and then advancing into the relevant legislation and related literature. Brief explanation of the structure and contents of each chapter:

- Chapter 1: Introduction
    - This chapter presents the background, motivation, and relevance of the topic. It also outlines the research objectives, scope, and methodology used to conduct this study.
- Chapter 2: Fundamentals of Web Scraping
    - This chapter outlines the core concepts, definitions, and general principles of web scraping, providing a foundation for understanding the technology and its typical use cases across various domains.
- Chapter 3: Legislative Framework regarding Web Scraping
    - This chapter explores the European Union's laws on web scraping, analysing directives like Directive 96/9 on database protection and Directive 2019/790 on text and data mining, with case law to illustrate legal implications.
- Chapter 4: Access Control Mechanisms
    - This chapter discusses various methods used to control automated access to websites, including administrative measures like REP and ToS, and technical measures such as CAPTCHA, IP blocking, and rate limiting.

- Chapter 5: Literature Review
    - Synthesises academic perspectives on legal challenges and access control mechanisms.
- Chapter 6: Conclusions
    - Summarises key findings through structured tables, discusses relationships between legal frameworks and technical implementations, and identifies future research directions.

Each chapter is designed to build on the previous one, ensuring a coherent and logical progression of ideas from introduction to conclusion. The structure supports the thesis aim by balancing technical explanation with legal analysis and academic synthesis.

2 Fundamentals of Web Scraping

Web Scraping is a sophisticated way of gathering and structuring data systematically from the web. According to Zhao (2017, p. 1), it is the process of gathering data from the World Wide Web using HTTP protocols or browsers, then transforming the data into a form of a dataset that is easily analysed. The process uses software agents also known as web robots to simulate the act of browsing and pulling information from websites (Glez-Peña et al., 2014, p. 789).

In general the web scraping process can be divided into two sequential phases (Zhao, 2017, p. 1):

1. HTTP Request Phase: The initial communication with the website occurs through either:
    a. A URL (Uniform Resource Locator) containing a GET query, or
    b. An HTTP message containing a POST query.

    During this phase, the requested resource is retrieved from the website and transmitted back to the web scraping application.
2. Data Extraction Phase: Once content is retrieved, the scraper proceeds to:
    a. Parse the HTML/XML structure
    b. Identify and extract relevant data points
    c. Reformat and organise the extracted data into a structured format (e.g., CSV, JSON, database records).

Modern web scraping implementations achieve these processes through two well-defined software modules:

1. Request and Interaction Module: Responsible for composing the HTTP requests and controlling the web interactions (Glez-Peña et al., 2014, p. 789-790), such as:
    a. *Urllib2* which defines set of functions to dealing with HTTP request like authentication, redirections, cookies and session management (Zhao, 2017, p. 1).

   b. *Selenium* which is a web browser automation framework that enables programmatic control of browsers to interact with dynamic websites using JavaScript (Zhao, 2017, p. 1).
2. Parsing and Extraction Module: Responsible for parsing and extracting the data from the content fetched (Glez-Peña et al., 2014, p. 790), such as:
   a. *Beautiful Soup* which is designed to scrape HTML and other XML documents (Zhao, 2017, p. 2).
   b. *Pyquery* which provides *jQuery*-like functionalities for parsing XML documents (Zhao, 2017, p.2).

The significance of web scraping is highlighted by its effectiveness in handling the rapidly growing amount of data available across the internet in various formats. As Khder (2021, p. 165) concludes it is an essential tool for companies in various fields in their attempts to maintain market position in today's digital market.

Singrodia et al. (2019, p. 5) have identified distinctive application areas for web scraping:

- Data mining: Extracting patterns and insights from large datasets collected across multiple websites.
- Research: Gathering data for academic studies, market analysis, and trend identification.
- Marketing: Monitoring competitor pricing, product offerings, and promotional strategies
- Competitive Intelligence: Tracking industry developments and competitor activities to inform business decisions
- Personal Tools: Creating customized alerts, dashboards, and aggregators for individual use
- Data Integration: Combining information from multiple sources into unified datasets

In addition, one of major application areas is consumer targeted digital marketplaces such as already mentioned Skyscanner and Booking.com.

3 Legislative Framework regarding Web Scraping

Web scraping exists in a complex legal framework that intersects multiple areas of law and its interpretations. The legal complexity surrounding web scraping stems from its technological nature, which often outpaces existing legal frameworks. While the scraping of publicly available content in Europe is regulated variably depending on the country (Fontana, 2025, p. 203), the direction of national regulations is coordinated at the European Union level with various directives. The decisions made by the Court of Justice of the European Union in various web scraping related cases also shape the legislative framework. This chapter will examine web scraping related directives and some examples of case law to foster a basic judicial knowledge of the matter.

3.1 The Legal Protection of Databases

This chapter will examine the contents of Directive 96/9 of the European Parliament and of the Council of 11 March 1996 on the Legal Protection of Databases, which provides two distinctive layers of protection to databases. Article 1(2) defines databases in the context of Directive 96/9 as systematically or methodologically arranged collections of independent works, data or other materials individually accessible by electronic or other means.

The first layer of protection is copyright protection for original databases. Article 3(1) (Directive 96/9) grants copyright to “databases which, by reason of the selection or arrangement of their contents, constitute the author’s own intellectual creation”.

Sui generis right acts as the second layer of protection as it provides the creator of the database right to “prevent extraction and/or re-utilization of the contents of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database” (Directive 96/9 art 7 para 1). This protection is only provided if it is shown that there has been a substantial investment in creating the database. Together copyright protection of database structure and *sui generis* right create a protection system

where either the intellectual originality or substantial investment put into creating the database or both may be protected by law.

To understand the effects of *sui generis* right to web scraping, it is imperative to examine the definition of extraction and re-utilization. Article 7(2) of Directive 96/9 defines them in context as follows:

> a) 'extraction' shall mean the permanent or temporary transfer of all or a substantial part of the contents of a database to another medium by any means or in any form;
> (b) 're-utilization' shall mean any form of making available to the public all or a substantial part of the contents of a database by the distribution of copies, by renting, by on-line or other forms of transmission

If the database has been put to public availability, no matter the manner, the creator does not have the right to prevent a lawful user from "extracting and/or re-utilizing insubstantial parts of its contents, evaluated qualitatively and/or quantitatively, for any purposes whatsoever" (Directive 96/9 art 8 para 2). In the cases where a lawful user has authorization to only a part of the database, this paragraph applies only to that part. However, Article 7(5) still protects the database from unreasonable exploitation as it defines repeated and systematic extraction of insubstantial parts of the database as non-permitted actions (Directive 96/9).

Although providing strong protection to databases Directive 96/9 also explicitly recognizes that the protection does not extend to their contents. Article 3(2) separates these two entities – a database and its contents – in a way that the protection of a database does not extend to its contents, but the contents still preserve all existing rights

(Directive 96/9). This creates a system where different legal protections can coexist at different levels:

- The database structure may have copyright protection.
- Individual content may have its own separate protection.

In addition to the separation of entities it is also clarified in recital 45 of Directive 96/9 that the *sui generis* right does not in any way extend copyright protection to data or mere facts.

3.2 Text and Data Mining Exceptions in the Digital Single Market

Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market of the Union represents clear advancement on clarity of the legislation related to web scraping, as it establishes legal framework for text and data mining activities which are inherently connected to web scraping. The directive includes important exceptions to copyright and database protection rights.

To lay the foundation for the examination of this directive it is important to cover the key definitions made in the directive. Article 1(1) defines *research organisations* as non-profit or public interest entities primarily conduction scientific research or education, where research results cannot be preferentially accessed by any controlling commercial entity (Directive 2019/790). It is also crucial to note Article 1(2) as it gives clear definition for text and data mining (TDM):

> 'text and data mining' means any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations;

The directive makes a distinction between general TDM and TDM conducted for scientific research. Research organisations enjoy broader exceptions as they have a mandatory exception to the following rights (Directive 2019/790 art 3 para 1):

- From Directive 96/9 (Database Directive):
    - Article 5(a): The reproduction right for databases protected by copyright
    - Article 7(1): The sui generis extraction right for databases
- From Directive 2001/29 (Information Society Directive):
    - Article 2: The reproduction right for copyright works
- From Directive 2019/790 (Digital Single Market Directive):
    - Article 15(1): The new press publishers' rights for online uses of press publications

Article four is applicable generally to any entity and provides the following exceptions in paragraph one “for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining”:

- From Directive 96/9 (Database Directive):
    - Article 5(a): The reproduction right for copyright-protected databases
    - Article 7(1): The sui generis right preventing extraction from databases
- From Directive 2001/29 (Information Society Directive):
    - Article 2: The reproduction right for copyright works
- From Directive 2009/24 (Software Directive):
    - Article 4(1)(a): The permanent or temporary reproduction of computer programs
    - Article 4(1)(b): The translation, adaptation, arrangement or other alteration of computer programs
- From Directive 2019/790 (Digital Single Market Directive):
    - Article 15(1): The press publishers' rights for online uses of press publications

While providing very similar exceptions to article three, article four does not provide as strong of a protection, as the exception is only applicable on condition that the rightholders have not reserved their rights in appropriate manner (Directive 2019/790 art 4 para 3).

3.3 Nature of websites – databases or not

Whether or not websites qualify as databases has significant implications for web scraping. The definition of a database from Directive 96/9 (art 1 para 2) establishes three key criteria:

- A collection of independent works, data or other materials
- Arranged in a systematic or methodical way
- Individually accessible elements

Recital seventeen of the same directive clarifies that databases should be understood to include literary, artistic, musical or other collections of works or collections of other material such as texts, sound, images, numbers, facts, and data.

Although not addressing this question directly the Court of Justice of the European Union has had related cases, which have given clear implications about the judicial interpretations of European Union law.

3.3.1 Case Fixtures Marketing Ltd v Organismos prognostikon agonon podosfairou

In the Case C-444/02 (2004) Fixtures Marketing Ltd v Organismos prognostikon agonon podosfairou AE (OPAP) it was stated that the definition of database in Article 1(2) of the Directive 96/9 refers to *any* "collection of works, data or other materials, separable from one another without the value of their contents being affected, including a method or system of some sort for the retrieval of each of its constituent materials" (Case C-444/02 para 32). The specifics of this case were that:

- Fixtures Marketing Ltd represented English and Scottish football leagues who claimed intellectual property rights in their fixture lists.
- OPAP used information from these fixture lists for organizing betting games.
- Fixtures Marketing sued OPAP, claiming violation of the *sui generis* right.

While this case did not directly address whether websites are databases or not, as it only interpreted Directive 96/9 from a general perspective and specifically regarding football

fixture lists, it gives perspective on how the directives are interpreted by judicial entities. The preliminary ruling of the CJEU, in addition to elaborating the definition of a database, in this case was that:

- A fixture list for a football league constitutes a database within the meaning of Article 1(2) of Directive 96/9
- The expression "investment in … the obtaining … of the contents" of a database in Article 7(1) of Directive 96/9 must be understood to refer to the resources used to cover the resources invested in seeking out *existing* data not investment in *creating* the data.

Therefore, although a fixture list for a football league constitutes as a database, they are not protected by *sui generis* right as the court ruled that: "In the context of drawing up a fixture list for the purpose of organising football league fixtures, therefore, it does not cover the resources used to establish the dates, times and the team pairings for the various matches in the league." The court's ruling emphasises the implications of its elaboration on the definition of a database, which seems to imply a very broad applicability of the term database.

3.3.2 Case Ryanair Ltd v. PR Aviation BV

The ruling and implications of Case C-444/02 were further affirmed in Case C-30/14 (2015) Ryanair Ltd. v PR Aviation, where the CJEU issued a preliminary ruling on whether contractual restrictions can be imposed on databases not protected by copyright or sui generis rights (Case C-30/14 para 28).

Ryanair had brought an action against PR Aviation for extracting flight data from its website for use in a price comparison service, in breach of Ryanair's terms of service that prohibited such commercial "screen scraping." After mixed rulings in Dutch courts, the Supreme Court referred the question to the CJEU. The Court held that Directive 96/9 does not apply to databases lacking copyright or sui generis protection. Consequently, Articles 6(1), 8, and 15 do not prevent database creators from imposing contractual

restrictions on third-party use. This led to the paradoxical outcome that unprotected databases may offer their creators more contractual freedom than protected ones, allowing Ryanair to enforce its terms despite lacking legislative protection for its database.

Referring to paragraph 33 of Case C-30/14, the Court emphasized that Article 1(2) of Directive 96/9 gives a broad definition of "database," unconstrained by technical or formal considerations. It also cited the judgment in Case C-444/02 in support of this interpretation. In essence Case C-30/14 further established the interpretation that the definition of database is in fact very wide, but only to determine what Directive 96/9 might potentially cover – not what it actually protects.

4 Access Control Mechanisms

To mitigate automated access and data sourcing, websites implement various mechanisms that can be categorized as either administrative or technical. Administrative access control refers to non-technical mechanisms that establish boundaries through policy-based approaches rather than technological barriers. These mechanisms rely on voluntary compliance and may carry legal implications despite lacking technical enforcement. In contrast, technical access controls implement actual barriers that must be overcome for scraping to occur.

This chapter examines both types, beginning with administrative measures such as the Robots Exclusion Protocol (REP) and Terms of Service (ToS), followed by technical implementations with CAPTCHA systems and IP-based controls as examples.

4.1 Robots Exclusion Protocol

Robots.txt is a fundamental implementation of REP, which represents a standardized methodology for website administrators to communicate their preferences regarding automated client access to their web resources (Koster et al., 2022, p. 1). This protocol, initially conceptualized in 1994, has evolved to a de facto standard for advisory regulation of web scraping.

The architecture of robots.txt comprises two primary structures:

1. User-Agent Specification: The protocol employs User-Agent directives to identify specific crawlers or crawler groups. As detailed in RFC 9309 (Koster et al., 2022, p. 4-5), these identifiers must conform to precise formatting requirements:
    a. Case-insensitive matching for crawler identification.
    b. Permissible characters limited to letters (a-z, A-Z), underscores (_), and hyphens (-).
    c. Universal wildcard (*) functionality for comprehensive crawler designation.

2. Access Control Directives The protocol implements two primary directive types:
    a. Allow: Explicitly permits access to specified Uniform Resource Identifier (URI) paths.
    b. Disallow: Restricts crawler access to designated URI paths.

These directives operate within hierarchical precedence framework, where rule specificity determines application priority (Koster et al., 2022, p. 6).

The following example demonstrates the standard syntax for access control implementation, illustrating the hierarchical relationship between general and specific rules (Koster et al., 2022, p. 10):

```
User-Agent: *
Allow: /example/page/
Disallow: /example/page/disallowed.gif
```

In this example the initial User-Agent specification indicates universal applicability of the following directives to all automated clients. The access control directives establish hierarchical relationship, where `Allow: /example/page/` gives baseline permission for URI paths beginning with the specific prefix. `Disallow: /example/page/disallowed.gif` implements a specific restriction for a single resource.

While being the de facto standard it is not in any way a security measure for websites as complying with it is totally voluntary, and as Koster (1994) states in his initial definition it is not enforced at all nor can any guarantees be given that any current or future robot respects it.

4.2 Terms of Service

Website Terms of Use constitute contractual documents that establish the conditions under which website resources may be accessed and used. While these terms, like robots.txt file, operate on a voluntary basis, they carry out distinctive legal implications as

the website owner can accuse the user of breach of contract. According to Dreyer and Stockton (2013) a viable way to prevent scraping is to include prohibitions against it in the website's terms of use, but the implementation of user agreement determines how legally binding the terms of use are i.e. can the owner of the website accuse user of contact breach.

Generally the implementation of user agreement is done either through a clickwrap or browsewrap agreement (Dreyer & Stockton, 2013). A clickwrap agreement consists of a pop-up notification that presents the user with buttons to agree or disagree with the terms of service, which must be available for the user before making the decision(Dreyer & Stockton, 2013; Krotov & Johnson, 2023; *Specht v. Netscape Communications Corp.*, 2001; Zynda, 2004). Clickwrap agreement ensures that to access the website user must agree to the terms of service.

Browsewrap is a non-intrusive way to constitute contractual agreement of terms of service, as it does not prohibit the use of the website in any way. It is usually implemented as a hyperlink to a separate web page containing terms of service (Dreyer & Stockton, 2013; Krotov & Johnson, 2023; *Specht v. Netscape Communications Corp.*, 2001; Zynda, 2004). The difference to clickwrap is that the user is not required to agree or even view them to access the website.

4.3 CAPTCHA

The acronym CAPTCHA stands for *Completely Automated Public Turing test to tell Computers and Humans Apart*, introduced first in the early 2000s by researchers at Carnegie Mellon University (von Ahn et al., 2000). CAPTCHAs are challenge-response tests that protect websites by distinguishing between humans and automated programs (Shi et al., 2022; von Ahn et al., 2000), becoming a central security technique for restricting automated access to website content (Shi et al., 2022).

CAPTCHAs represent a variation of the Turing test (Turing, 1950), where rather than a human judge attempting to identify a computer, a computer system now generates tests that humans can pass but current computer programs cannot (von Ahn et al., 2004, p. 58). Paradoxically a CAPTCHA is a program that generates challenges that other computer programs cannot solve. Unlike the purely conversational original Turing test, CAPTCHAs can be based on a variety of human sensory and cognitive abilities (von Ahn et al., 2004).

Modern research recognises text-, image-, audio-, video-, human cognition-, and moving object-based CAPTCHAs (Saha et al., 2015; Shi et al., 2022; Xu et al., 2012). Despite numerous variations, most CAPTCHAs are hidden from users, collecting behavioural and environmental data to classify or score them (BuiltWith®, 2025; Jin et al., 2023). This chapter focuses on text-, image-, and behaviour-based CAPTCHAs, as the first two are often used alongside the latter. A brief overview of third-party CAPTCHA provision is also included.

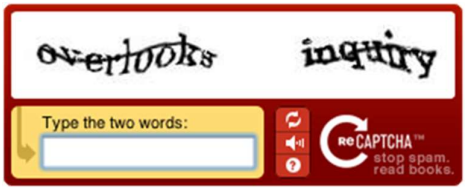
4.3.1 Text-based CAPTCHAs

Text-based CAPTCHAs were among the first implementations of a CAPTCHA, with *GIMPY* establishing early standards for the approach. *GIMPY* challenged users to recognize three out of seven distorted dictionary words presented in an image (von Ahn et al., 2004, p. 59), as illustrated in Figure 3.

Figure 3 Example of a *GIMPY* CAPTCHA (von Ahn et al., 2004, p. 58).

Since then, text-based CAPTCHAS have evolved significantly with the first major evolution being reCAPTCHA v1, which used a two-word system (Von Ahn et al., 2008) shown in Figure 4. This system presented users with two words—one known to the system and one unknown—requiring users to type both correctly to pass. The unknown words came from OCR (Optical Character Recognition) systems that had failed to confidently digitize text from books and printed materials. The approach served a dual purpose: protecting websites while simultaneously crowdsourcing the verification of text that automated OCR systems struggled to recognize due to poor print quality, unusual fonts, or damaged source materials.

Figure 4 Two-word system used in reCAPTCHA v1, it also includes audio-based option for the visually impaired (von Ahn et al., 2000).



The evolution of text-based CAPTCHAs has been extensive, with Guerar et al. (2022, p. 5) providing a comprehensive taxonomy (see Appendix 1). More recent innovations include DotCHA (Suzi Kim & Sunghee Choi, 2019), a 3D scatter-type CAPTCHA where users must identify letters composed of small spheres by rotating a 3D model (example shown in Figure 5). To succeed user must rotate the 3D model multiple times as each letter is twisted around a horizontal axis ensuring that all of them are not readable from the same rotation angle.

Figure 5 Example of a DotCHA CAPTCHA, each letter must be individually identified (Suzi & Sunghee, 2019).

4.3.2 Image-based CAPTCHAs

Already in 2014 Google revealed that 99.8 % of distorted text variants could be solved using AI technology (Shet, 2014a), and the rapid development of AI capabilities has since then only grown. This has resulted in shift of focus to image-based designs instead of text-based with the assumption that visual challenges are harder for computers that character recognition (Dinh & Hoang, 2023). Generally, the challenge includes a written text describing a task that needs visual recognition capabilities to e.g. complete an image classification task (Guerar et al., 2022, p. 7). The interaction with the challenge varies depending on the design, which have been classified by Guerar et al. (2022, p. 8-10) to six different types (see Appendix 2).

Google's reCAPTCHA is the predominant technology for website protection, accounting for approximately 89% of all identified CAPTCHA implementations (BuiltWith®, 2025). As a result of this, selection-based designs are most used in image-based CAPTCHAs, as it is the design that Google's *No captcha reCAPTCHA* utilizes alongside behaviour analysis (Shet, 2014b). If the risk analysis cannot predict confidently that the user is human, it will prompt the user with a visual recognition task. The first implementation (shown in Figure 6) presented the user a sample image and nine candidate images, from which the user then had to select the ones similar to the sample (Guerar et al., 2022, p. 11).

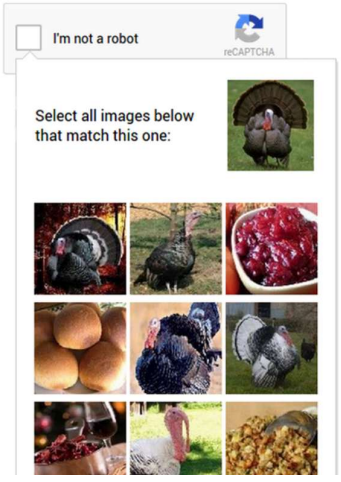


Figure 6 First implementation of selection-based CAPTCHA in *No captcha reCAPTCHA* (Shet, 2014b).

Since then, the implementation has shifted focus from recognising similar images to label recognition. In this version the CAPTCHA prompts user with a label and nine candidate images, from which user must select the ones matching the label. This implementation has some variations that are illustrated in Figure 7. One of the key differences between these implementations is that the modern version replaces the selected images with new ones.

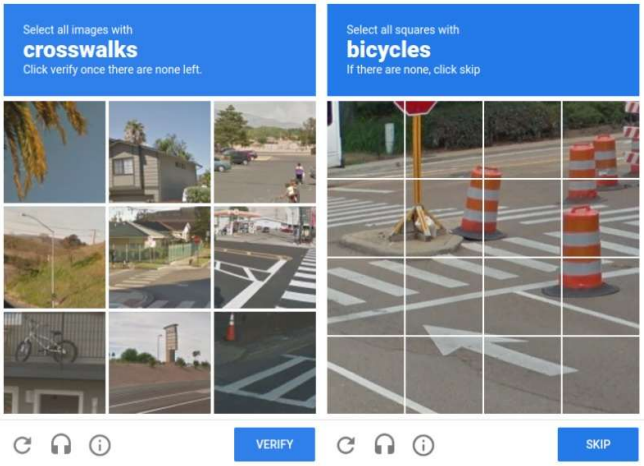


Figure 7 Different variations of a selection-based CAPTCHA (NopeCHA, 2025).

4.3.3 Behaviour-based CAPTCHAs

Due to reCAPTCHA and hCaptcha being the two predominant CAPTCHA providers for websites, with reCAPTCHA covering 89% and hCaptcha 7% of CAPTCHA usage (see BuiltWith®, 2025), the usage of other than behaviour-based CAPTCHAs in modern internet is quite limited.

Google's reCAPTCHA has had multiple iterations throughout its lifetime, but in recent years the focus has shifted from traditional designs to a behaviour-based design. The second iteration of it, *No captcha reCAPTCHA* (later known as reCAPTCHAv2), required user to interact with a checkbox to invoke the risk analysis which then decided if further security measures were needed (see Chapter 4.3.2). Alongside this a different variation was rolled out, known as reCAPTCHAv2 Invisible, which does not require a checkbox to

be presented to the user (Jin et al., 2023). In this variation the risk analysis is either invoked programmatically or the activation is bonded to a clickable element on the website. The latest version, reCAPTCHAv3, differs from the previous version in that rather than simply classifying user as human or computer it gives user a score (Google, 2024b).

The only other 'big player' in the CAPTCHA market is hCaptcha, which is very similar to reCAPTCHA. The main difference being that hCaptcha addresses some of the privacy concerns of reCAPTCHA (Cloudfare, 2020b). Other notable difference is that hCaptcha is the only provider allowing website admin to set the difficulty of passing the CAPTCHA manually (Jin et al., 2023).

4.3.4 Integrating third-party CAPTCHA systems

Jin et al. (2023, p. 4) suggests that most websites use third-party CAPTCHA providers instead of self-developed CAPTCHAs. Usage distribution data of CAPTCHAs from BuiltWith® (2025) supports the suggestion by Jin et al., thus resulting the scope of this thesis to only cover the integration framework of third-party-provided CAPTCHAs. Specifically, the focus will be on a general framework introduced by Jin et al. (2023) and the request flows of the two largest CAPTCHA providers – reCAPTCHA and hCaptcha.

The framework presented by Jin et al. (2023, p.4-5) describes the process how third-party CAPTCHA systems work as follows:

1. Website registers with CAPTCHA provider to get public and private keys.
2. User visits a webpage containing a CAPTCHA.
3. User action invokes request to load CAPTCHA content, public site key is used to invoke the CAPTCHA.
4. CAPTCHA content loads from third-party provider.'
5. User solves the CAPTCHA.
6. Provider validates solution and generates a token.
7. Token is sent to website server via the user's client.
8. Website server verifies the token with the provider using its private key.

This sequence is also illustrated in Figure 8 to further clarify the communication between systems. Step one of the sequence is assumed and not shown in the figure.

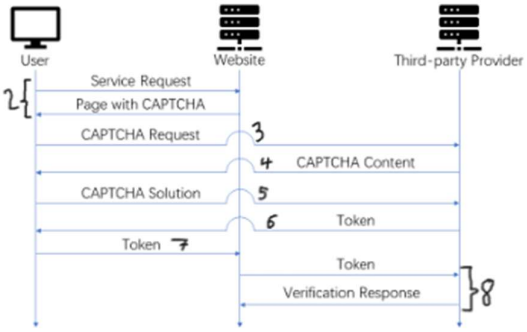


Figure 8 General framework for third-party CAPTCHAs (modified from Jin et al. 2023, p. 5.)

While both reCAPTCHA and hCaptcha follow the general third-party CAPTCHA framework, they implement it with distinct operational nuances. Google's reCAPTCHA has evolved through multiple versions, each with significant changes to how the framework is implemented.

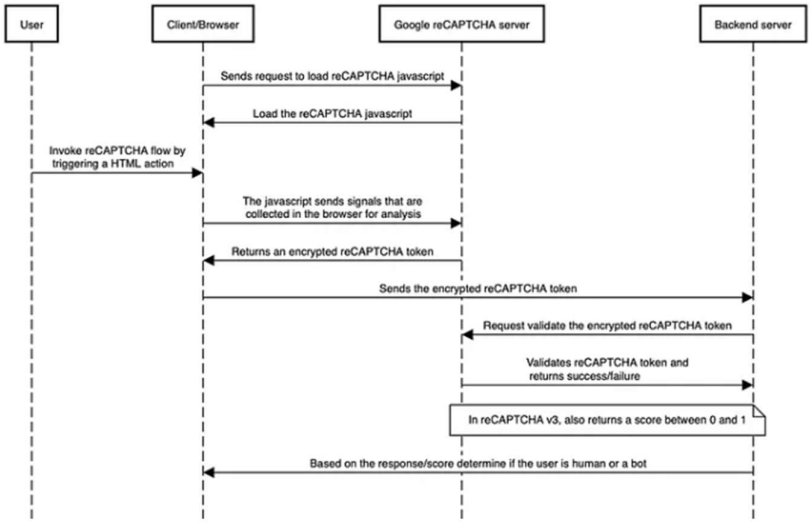


Figure 9 reCAPTCHA request flow diagram showing signal collection and reCAPTCHA v3 scoring system (Pathum, 2023).

As illustrated in Figure 9, reCAPTCHA's approach has shifted dramatically from early versions to the latest implementations. reCAPTCHA v1 used distorted text challenges (see Chapter 4.3.1) but was discontinued in 2018 (Google, 2024a) as AI reached 99.8% accuracy in solving them (Shet, 2014a). reCAPTCHA v2 introduced the familiar "I'm not a robot" checkbox that secretly analyses mouse movements—humans move their cursors in "wiggly and imperfect ways" compared to bots (Pathum, 2023). The invisible variant of v2 eliminated the checkbox requirement by binding challenges to existing buttons. Most significantly, reCAPTCHA v3 transforms the framework by removing explicit challenges entirely, instead analysing user behaviour in the background and providing a score between 0-1 to indicate human likelihood, fundamentally changing steps 3-5 of the general framework.

In contrast, hCaptcha adheres more closely to the general framework due to them maintaining a more traditional challenge-response model, rather than invisible behavioural analysis. The hCaptcha system distinctly separates its Client API from its Siteverify component, with the Client API handling challenge delivery and the Siteverify service managing server-side verification (see Figure 10).

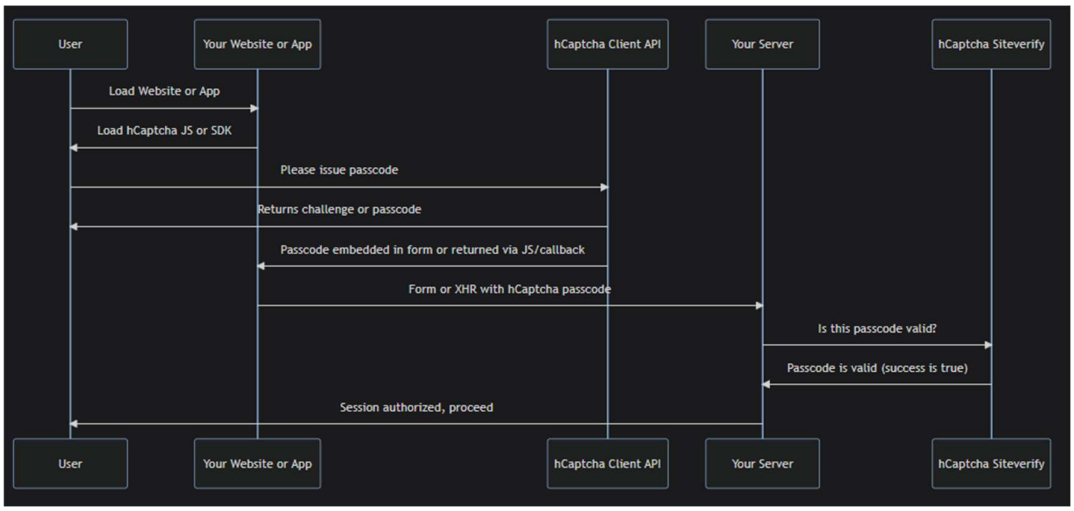


Figure 10 hCaptcha request flow diagram showing passcode generation and verification process (hCaptcha, 2025).

When a user encounters an hCaptcha widget, they receive a challenge or puzzle that, once solved, generates what hCaptcha terms a "passcode" rather than a token (hCaptcha, 2025). This passcode can be delivered through multiple technical pathways – either embedded directly within HTML forms or returned via JavaScript callbacks – providing developers with integration options suited to different application architectures. One notable technical distinction is hCaptcha's explicit support for XHR (XMLHttpRequest) submissions alongside traditional form submissions (hCaptcha, 2025), allowing for more dynamic frontend implementations while maintaining the security principles of the general framework.

Both service providers retain the critical security architecture of the general framework – keeping verification keys private on the server side – while adapting the user interaction flow to their specific technical ecosystems. The key difference between these examples is in that reCAPTCHA variations illustrated in Figure 9 invisibly assess and score user, while the highlighted version of hCaptcha utilises traditional challenge-response design. Both providers have multiple variations of CAPTCHA systems available to customers, but for example hCaptcha’s No-CAPTCHA mode is only available for their enterprise service customers (hCaptcha, 2025).

4.4 IP-based access control

In addition to application-level mechanisms such as the ones mentioned in previous chapters, websites can also enforce restrictions at the network level. Two commonly employed measures in this category are IP address blocking and IP-based rate limiting, both of which control access on the user’s IP address.

These mechanisms require monitoring server access logs, which can be done either manually or through various firewall applications (Gheorghe et al., 2018, p. 68). Every device connected to the internet is assigned a unique IP address, which allows it to communicate with other devices across networks regardless of physical location (Jyväskylän

Yliopisto & Maanpuolustuskoulutusyhdistys, 2024). This address serves as a key identifier that network-level controls can act upon.

One of the most widely used implementations of IP-based access control is the use of access-control lists (ACLs) within network firewalls. These lists define specific IP addresses or ranges that are either allowed or denied access to particular services. A commonly recommended best practice is to adopt a “deny by default” posture, allowing only explicitly permitted IPs to access the system (Scarfone & Hoffman, 2009, p. 4-1). This approach ensures that unauthorized or unexpected traffic is blocked unless a rule explicitly allows it. In practical terms, IP-based access control operates by matching packet header fields against static or dynamic allowlists and blocklists.

While IP blocking is a binary control—either permitting or denying traffic—rate limiting introduces a more nuanced approach by regulating the frequency or volume of requests from each IP address. This method is particularly effective in mitigating abusive or excessive behaviour, such as denial-of-service attempts or automated scraping (42crunch, 2019; Cloudfare, 2020a; radware, 2023).

A widely adopted mechanism or implementing rate limiting is the token bucket algorithm (Raghavan et al., 2007, p. 339), which deploys rate limiting through “metaphorical bucket of tokens that refills at a constant rate” (Manoharan, 2024, p. 1791).

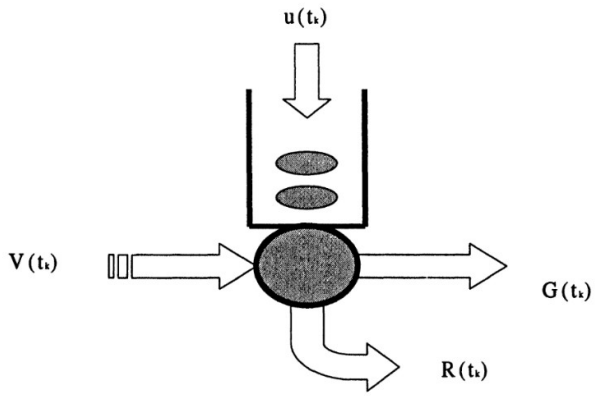


Figure 11 Dynamic model for token bucket algorithm (Ahmed et al., 2002, p. 267).

A practical implementation of this concept is the dynamic model proposed by Ahmed et al. (2002, p. 267), shown above. Notations used in the figure are as follows:

- $u(t_k)$: the number of tokens added during the time interval $[t_{k-1}, t_k]$ (i.e. refill rate).
- $V(t_k)$: size of the arriving packet at time t_k.
- $G(t_k)$: conforming traffic.
- $R(t_k)$: non-conforming traffic.

The logic of this model can be simplified to the following sequence:

1. Packet size of $V(t_k)$ arrives at the token bucket.
2. If the bucket contains enough tokens, the packet is conforming, and the equivalent number of tokens is removed.
3. If the bucket lacks enough tokens, the packet is nonconforming:

– It may be delayed until tokens are available (then reclassified as conforming), or

– It may be rejected outright, depending on the system's policy.

In essence, the token bucket algorithm allows traffic to pass freely as long as sufficient tokens are available, thus supporting short bursts of activity while limiting long-term average rate (Manoharan, 2024, p. 1791). Once the bucket is empty, the system enforces limits by slowing or dropping incoming requests. Traditionally a token bucket is implemented at a single choke point such as firewall or web server.

5 Literature review

The preceding chapters have established the foundational elements of web scraping (Chapter 2), the relevant European Union legislative frameworks (Chapter 3), and the technical implementation of various access control mechanisms (Chapter 4). Building upon this foundation, this chapter synthesises academic perspectives to address the central research questions posed in Chapter 1:

- What are the primary legal challenges associated with web scraping under current legislative frameworks and judicial interpretations, and how do these challenges affect the permissibility of automated data collection?
- What are the general techniques used to control automated access to website data? And what legal implications constitute from circumventing them?

While the previous chapters provided descriptive analyses of legislation and technical implementations in isolation, this literature review examines their interrelationship through the lens of scholarly discourse. Rather than organising this synthesis by individual sources, the chapter employs a thematic approach, examining how various legal frameworks interact with technical controls to create a complex regulatory landscape for web scraping activities.

The first section explores the legal challenges identified in existing literature, categorising and evaluating scholarly perspectives on issues such as the legal status of web scraping, unauthorised access, the enforceability of Terms of Service, and legal protections for website content.

The second section examines how the literature addresses the access control mechanisms discussed in Chapter 4, with particular emphasis on their legal implications and effectiveness. This thematic structure facilitates a comprehensive understanding of the regulatory environment governing web scraping activities across multiple jurisdictions, with particular attention to the European context established in Chapter 3.

5.1 Legal Challenges acknowledged in literature

5.1.1 Legal status of Web Scraping

Recent literature reveals a complex and evolving legal landscape for web scraping. As Possler (2019, p. 3903) highlights, the legal framework is often unspecified which characterises the framework as a "grey area". Fontana (2025, p. 205) came to the same conclusion after examining the jurisprudence and legal doctrines of web scraping. In his research he found that "there is no specific legislation in the United States, Europe or Asia that explicitly prohibits web scraping." This has resulted in the fragmentation of legal approaches across jurisdictions, which is widely acknowledged in literature (Brown et al., 2024; Fontana, 2025; Krotov & Johnson, 2023; Pagallo & Ciani Sciolla, 2023). Krotov & Johnson (2023, p. 11) suggest that the legal status of web scraping activities should be considered on a case-by-case basis, which is also backed by Brown et al. (2024, p. 18).

The fragmentation of legal approaches is highlighted by Pagallo & Ciani Sciolla (2023, p. 11-13), as they imply an inconsistency in judicial interpretations of Database Protection laws across EU member states citing contradictory Ryanair cases:

- The Spanish Supreme Court ruled that Ryanair's system was a software generating information, not a database.
- The Italian Supreme Court and Regional Court of Hamburg have declared that Ryanair does in fact have a database under Article 1(2).

As Possler (2019, p. 3903) concludes: "these legal uncertainties render ethical considerations on the part of researchers all the more relevant for scraping". The literature research conducted in this thesis validates this relevancy as there were multiple findings of academic papers addressing ethical considerations of web scraping (e.g. Brown et al., 2024; Gold & Latonero, 2017; Krotov & Johnson, 2023; Logos et al., 2023; Luscombe et al., 2022; Pagallo & Ciani Sciolla, 2023; Paige, 2024; Thelwall & Stuart, 2006; Xiao, n.d.).

5.1.2 Unauthorised access

Literature addresses multiple perspectives on whether scraping constitutes unlawful access or not. Brown et al. (2024, p. 11) suggest that “scraping that involves breaking into online spaces that are not otherwise available to the public will create higher legal risks than scraping only publicly accessible spaces.” Breaking into online spaces can be interpreted as not respecting access control mechanisms employed by the website in question. This interpretation seems to be valid as Fontana (2025, p. 203) considers the possibility of having either ToS or robots.txt file to be enough in preventing the TDM exceptions covered in Chapter 3.2. He also addresses how the cited case law suggests that even the presence of a protective measure against scraping could be sufficient enough to deem the scraping of a website unauthorised. This would consequently imply that circumventing, programmatically solving, or otherwise bypassing a CAPTCHA system could be regarded as unauthorised access to the website content.

Krotov & Johnson (2023, p. 10-11) highlight how the Digital Services Act (2022) of European Union requires “certain ‘very large online platforms’ and ‘very large search engines’ to” allow researchers to access publicly available – potentially via scraping. They also highlight that the DSA might allow “vetted researchers” to access the private data of these certain online entities.

5.1.3 Enforceability of Terms of Service

Literature identifies ToS agreements as a central mechanism for establishing protection against unwanted web scraping (Brown et al., 2024; Fontana, 2025; Krotov & Johnson, 2023; Logos et al., 2023). At the same time, it is acknowledged in literature that certain contractual agreements over ToS are more likely to be legally binding than other – specifically highlighting the distinction between legal enforceability of clickwrap versus browsewrap.

Dreyer & Stockton (2013) argue, through references to court cases, that clickwrap agreements are generally legally enforceable as it requires the user to read the terms of use, by noticing about them in a pop-up text box that requires agreeing to proceed to the website. On the other hand, browsewrap agreements do not require the user to read the terms of use, as the website notices of them usually by placing a hyperlink on the website, which leads to more subjectivity on its enforceability. Tarra Zynda (2004, p. 507) points out, in article about case *Ticketmaster Corp. v. Tickets.com, Inc.*, that "so far, courts have held browsewrap agreements enforceable if the website provides sufficient notice of the license". Continuing this it is also brought up in the article that the few courts, that have examined the validity of browsewrap agreements, have established criteria for sufficient notice requiring the terms of use to be on the landing page of the website, visible without scrolling to the bottom of the page, and presented clearly as a hyperlink. Frolova and Berman (2024) continue from this and present a list of concrete recommendations for improving legal enforceability of ToS agreements:

1. Visibility and Clarity:
    a. Terms must be placed in a conspicuous location on the website or app (e.g., login or checkout page).
2. Unambiguous Consent Mechanism:
    a. The way users accept the terms must be unmistakable, such as a checkbox or clearly labeled action button.
3. Explicit Statement of Legal Effect:
    a. The agreement should state that by accepting the terms, the user becomes a party to a legally binding agreement.
4. Consent at Registration:
    a. If site use requires registration, consent should be obtained at the registration stage before the account is created.
5. Notification of Changes:
    a. If any terms are changed—even minor ones—each user should be informed of the changes explicitly.

Even though the enforceability of ToS can be strengthened by properly implementing contractual agreements it is highlighted by Fontana (2025, p. 205), that it is a fragile instrument as enforcing ToS can be “difficult, costly and sometimes only ensures a low likelihood of success.”

5.1.4 Legal Protection of Websites and their Content

In addition to the contractual protections offered by properly implemented Terms of Service, website content may also benefit from legal protection under European Union law. As discussed in Chapter 3, where a website qualifies as a database, its content can be shielded by multiple layers of protection:

- Protection of the website as a database through copyright and/or *sui generis* right.
- Protection of individual website content through copyright.

This layered protection framework is particularly relevant when addressing web scraping, which may infringe copyright or database rights depending on the nature of the website and its contents.

5.1.4.1 Copyright infringement

Scraping can lead to copyright infringement in two ways: first, if the target website is protected as a copyrighted database; and second, if the individual content on the site is itself copyrighted. Krotov et al. (2020) highlight the role of a robots.txt file in constituting copyright for website content. They argue through relevant court law that the failure to include a robots.txt file with sufficient instructions could create implied license to use website data. On the other hand, Pagallo & Ciani Sciolla (2023, p. 10-11) place more emphasis on how scraping copyrighted content may be lawful under a set of exceptions, but highlight the pessimism related to this. They note how some interpret current legislation in a way that lawful scraping in Europe is only possible for research or cultural organisation for research purpose. This perspective is challenging as the interpretations of law suggest commercial uses to face stricter scrutiny (see e.g. chapter 3.2.), yet as

Pagallo & Ciani Sciolla (2023, p. 10) also point out there are cases where scraping has been considered lawful.

5.1.4.2 *Sui generis* infringement

Alongside copyright databases are protected by the *sui generis* right (see Chapter 3.1). As Pagallo & Ciani Sciolla (2023, p. 11) note there are two distinctive legal thresholds that trigger the *sui generis* protection for a database: substantiality of the investment; and substantiality of the extraction, which are both evaluated quantitatively and/or qualitatively.

Across different authors (Fontana, 2025; Oesch et al., 2017; Pagallo & Ciani Sciolla, 2023) the legal concept of substantial investment is acknowledged to be very ambiguous, with Oesch et al. describing it as very unclear, and Pagallo & Ciani Sciolla highlighting how there are different interpretations across EU national courts. Fontana also cites relevant court cases as examples of different reasonings for excluding *sui generis* right. Despite the ambiguity of interpretations, Oesch et al. (2017, p. 74-77, 112-118) synthesise from CJEU cases that the concept of substantial investment should be understood as specifically targeting the formation and compilation of the database itself, rather than the creation of the original data contained within it. In all cited cases the rulings of the CJEU were similar, and along the lines of "the investment aimed at the contents of the database can't be taken into consideration when evaluating the substantiality of the investment". Resulting in that only investments targeted in the collection, verification or presenting of the contents are considered.

The second threshold of *sui generis* protection – substantiality of extraction – is also ambiguous. Pagallo & Ciani Sciolla (2023, p. 12-13) suggest that the substantiality of the extraction must be determined by courts on a case-by-case basis. Their interpretation of legal perspectives is that the extraction of limited or partial data is considered admissible in any case, but that so called 'diachronic extraction' should be considered based on how it affects the target. As an example of unjust harm caused by diachronic extraction, they

present case QVC Inc. v. Resultly in which the extraction was considered substantial as it crashed the target website for two days.

5.2 Technical Access Control Mechanisms

5.2.1 Ethical implications of robots.txt

While not being designed as a security measure for websites against scraping (see Koster, 1994) the increased adoption rate of REP (Chang & He, 2025, p. 1124) has resulted in the perception that honouring robots.txt is a matter of ethical web behaviour, establishing an informal but widely respected norm against unauthorized or aggressive data collection (see e.g. Chang & He, 2025, pp. 4, 13; Gold & Latonero, 2017, p. 281; HTTP-ProxyOkeyProxy, 2024; Krotov et al., 2020).

5.2.2 Current Challenges of CAPTCHA systems

Existing literature identifies text- and image-based CAPTCHAs as the most widely adopted designs (Gutub & Kheshaifaty, 2023; Saha et al., 2015; Shi et al., 2022; Xu et al., 2012). However, studies also show that the traditional CAPTCHA models might not be sufficient to combat the rapid evolvement of AI (e.g. Guerar et al., 2022; Plesner et al., 2024). In response, leading third-party providers such as reCAPTCHA transitioned towards behaviour-based CAPTCHA systems (see Chapter 4.3.3); however, these solutions continue to lag behind the evolving capabilities of CAPTCHA solvers. This claim is supported from both academic analyses and the development of solver technologies, as relevant studies and publications have demonstrated the consistent success of solvers in overcoming most CAPTCHA systems (e.g. Jin et al., 2023; Plesner et al., 2024; Sivakorn et al., 2016; xHossein, 2021). The security test conducted by Jin et al. (2023, p. 15) on four third-party CAPTCHA providers: Google reCAPTCHA, Geetest, Arkose Labs, and hCaptcha revealed worrying results against both human solver relay- and automated attacks. A 20-year survey of CAPTCHA technologies by Guerar et al. (2022, pp. 25–26) concluded that, while none of the popular conventional and behavior-based designs had been extensively broken at the time, emerging solver capabilities posed a significant threat. But

even the ones mentioned to not yet have been broken (e.g. reCAPTCHA v2 Invisible) have since then been broken, thus proving their prediction that:

> Invisible reCAPTCHA and other academic proposals have not been broken yet, however with the advent of the fourth-generation bots that rotate through thousands of different IP addresses and mimic accurately the human behaviour, it would be difficult to design a secure CAPTCHA based solely on the user behaviour data that can be gathered in a normal (i.e., with no additional sensors or special hardware) environment.

Although sensor- and behaviour-based CAPTCHA systems offer promising avenues to counter solver advancements, they introduce significant challenges, particularly regarding user privacy, that require careful consideration. Both Dinh and Hoang (2023) and Guerar et al. (2022) emphasize the privacy concerns raised by these CAPTCHA designs, particularly regarding behavioural and sensor data of user, as well as extraction of demographic attributes.

5.2.3 Role of IP-based access control in preventing web scraping

Web scraping literature consistently identifies IP-based access control mechanisms as significant technical barriers for researchers and developers. Luscombe et al. (2022, p. 1034) categorize both IP blocking and rate limiting IP requests among their comprehensive inventory of defensive strategies employed to prevent automated data extraction. This is corroborated by other researchers, such as Shelar (2024, p. 1636) and Tabaku (2021, p. 2) similarly acknowledging these mechanisms as established methods that website administrators deploy to restrict scraping activities.

Luscombe et al. (2022, p. 1035) also address technical countermeasures, detailing workarounds for the IP-based access control mechanisms, such as randomising the used IP address to mask the scraping activity as multiple unique users accessing the website. Significantly they don’t address the workarounds as mere implementation details but

place them within broader legal and ethical framework emphasising that circumventing intentionally placed access control mechanisms requires navigation of legal ambiguities and ethical considerations. Legal implications of utilising such a workaround are similar to ignoring contents of robots.txt and/or ToS, and circumvention of a CAPTCHA (see Chapter 5.1.2) in that it could constitute to unauthorised access, thus deeming the scraping activity as unlawful.

6 Conclusions

This literature review has examined the complex intersection of legal frameworks and technical access control mechanisms related to web scraping. The analysis reveals a fragmented legal landscape where interpretation varies across jurisdictions, creating significant uncertainty for researchers, businesses, and website operators. Ambiguity of the relevant law has resulted in uncertainty regarding the legal implications of circumventing both administrative and technical access control mechanisms.

6.1 Summary of Key Findings and Research Contribution

The research identified two key dimensions that define the regulatory environment for web scraping: legal challenges and access control mechanisms. These dimensions are summarised in Tables 1 and 2. This research has highlighted the interrelationship between legal frameworks and technical implementations. The effectiveness of technical measures is often enhanced by how courts interpret attempts to circumvent them, while the legal enforceability of administrative measures like Terms of Service depends on its technical implementation (e.g., clickwrap vs. browsewrap).

Table 1 Key Legal Challenges identified in Web Scraping.

Legal Challenge	**Description**	**Key Implications**
Ambiguous Legal Status	Web scraping exists in a "grey area" with no specific legislation explicitly regulating it in major jurisdictions.	Case-by-case assessment of lawfulness is required. Ethical considerations are especially relevant.
Database Protection	EU's *sui generis* right protects databases representing substantial investment, independently of copyright protection.	Two thresholds determine protection: substantiality of investment and substantiality of extraction.
Contract Infringement	Terms of Service can establish contractual restrictions on web scraping, with varying degrees of enforceability.	Clickwrap agreements generally considered more enforceable than browse-wrap agreements, though enforcement remains challenging.
Unauthorised Access	Circumventing access control mechanisms could potentially constitute unauthorized access under various legal frameworks.	Breaking into protected spaces creates significantly higher legal risks than scraping publicly accessible content.
Copyright Infringement	Web scraping may infringe copyright if it involves copying protected content or database structures.	TDM exceptions may apply but are limited, particularly for commercial purposes, and can be overridden by technical protection measures.

Table 2 Key Access Control Mechanisms for preventing Web Scraping.

Mechanism	**Type**	**General Implementation Method(s)**	**Legal/Ethical Implications**
Robots Exclusion Protocol	Administrative	robots.txt file specifying allowed/disallowed paths and user-agents	Honouring it is widely considered as part of ethical web behaviour; ignoring may influence legal interpretations.
Terms of Service	Administrative	Contractual agreements; most commonly either clickwrap or browsewrap.	Ignoring may constitute contract breach; clickwrap agreements more legally enforceable than browsewrap.
CAPTCHA systems	Technical	Challenge-response tests and/or invisible behavioural analysis.	Circumvention may constitute unauthorized access; privacy concerns with behavioural data collection
IP-based Access Controls	Technical	IP blocking and request rate limiting	Workarounds may be viewed as unauthorised access.

6.2 Future Considerations

The regulatory landscape for web scraping continues to evolve, with several areas warranting further research and attention:

- Implications of the EU (Data Act, 2024): Set to apply from September 2025, the Act may increase legal certainty around data reuse, address contractual imbalances that restrict scraping, and trigger reforms to the Database Directive, potentially altering how database protection, specifically *sui generis* right, applies to scraping.
- Evolution of Technical Measures: As demonstrated by the rapid advancement of AI capabilities to overcome CAPTCHA systems, technical protection measures are engaged in an ongoing arms race with scraping technologies. Future research should monitor how this dynamic affects the balance between data accessibility and protection.
- Ethical Frameworks: Given the legal ambiguities, the development of robust ethical frameworks for web scraping becomes increasingly important. Krotov et al.'s work (Krotov et al., 2020; Krotov & Johnson, 2023) could be developed further to provide a more comprehensive cross-jurisdictional framework.

References

42crunch. (2019). *Rate limiting by IP address*. Retrieved 1 May 2025 from https://docs.42crunch.com/latest/content/extras/protection_rate_limiting_ip.html

Ahmed, N. U., Wang, Q., & Barbosa, L. O. (2002). Systems approach to modeling the Token Bucket algorithm in computer networks. *Mathematical Problems in Engineering*, *8*(3), 591831. https://doi.org/10.1080/10241230215282

Brown, M. A., Gruen, A., Maldoff, G., Messing, S., Sanderson, Z., & Zimmer, M. (2024). *Web Scraping for Research: Legal, Ethical, Institutional, and Scientific Considerations* (arXiv:2410.23432). arXiv. https://doi.org/10.48550/arXiv.2410.23432

BuiltWith®. (2025, April 21). *CAPTCHA Usage Distribution on the Entire Internet*. Retrieved 24 April 2025 from https://trends.builtwith.com/widgets/captcha/traffic/Entire-Internet

Case C-30/14 (15 January 2015). Ryanair Ltd v. PR Aviation BV. Retrieved 8 April 2025 from https://curia.europa.eu/juris/document/document.jsf?text=&docid=161388&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=333767

Case C-444/02 (9 November 2004). Fixtures Marketing Ltd v. Organismos Prognostikon Agonon Podosfairou AE (OPAP). Retrieved 8 April 2025 from https://curia.europa.eu/juris/showPdf.jsf;jsessionid=B7FE966C88DA4389CB38F5F7B90C159A?text=&docid=49635&pageIndex=0&doclang=en&mode=lst&dir=&occ=first&part=1&cid=316016

Chang, C., & He, X. (2025). *The Liabilities of Robots.txt* (arXiv:2503.06035). arXiv. https://doi.org/10.48550/arXiv.2503.06035

Cloudfare. (2020a). *What is rate limiting? | Rate limiting and bots*. Retrieved 1 June 2025 from https://www.cloudflare.com/learning/bots/what-is-rate-limiting/

Cloudfare. (2020b, April 8). *Moving from reCAPTCHA to hCaptcha*. Retrieved 24 April 2025 from https://blog.cloudflare.com/moving-from-recaptcha-to-hcaptcha/

Curry, D. (2025, January 22). *Travel App Revenue and Usage Statistics (2025)*. Business of Apps. Retrieved 3 February 2025 from https://www.businessofapps.com/data/travel-app-market/

Data Act (2024). Retrieved 1 May 2025 from https://digital-strategy.ec.europa.eu/en/policies/data-act

Digital Services Act (2022). Retrieved 26 April 2025 from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2065

Dinh, N. T., & Hoang, V. T. (2023). Recent advances of Captcha security analysis: A short literature review. *Procedia Computer Science*, *218*, 2550–2562. https://doi.org/10.1016/j.procs.2023.01.229

Directive 96/9 of the European Parliament and of the Council of 11 March 1996 on the Legal Protection of Databases. Retrieved 28 March 2025 from http://data.europa.eu/eli/dir/1996/9/oj/eng

Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9 and 2001/29. Retrieved 28 March 2025 from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L0790

Dreyer, A. J., & Stockton, J. (2013). *A Primer for Counseling Clients*. Retrieved 14 February 2025 from https://www.skadden.com/-/media/files/publications/2014/01/070071319-skadden.pdf

Fontana, Avv. G. (2025). Web scraping: Jurisprudence and legal doctrines. *The Journal of World Intellectual Property*, *28*(1), 197–212. https://doi.org/10.1111/jwip.12331

Frolova, E. E., & Berman, A. M. (2024). Expression of the Parties' Will in Context of Digital Transformation: Current Trends in Law Enforcement. *Law Journal of the Higher School of Economics*, *3*, 57–83. https://doi.org/10.17323/2072-8166.2024.3.57.83

Gheorghe, M., Mihai, F.-C., & Dârdal, M. (2018). *Modern techniques of web scraping for data scientists*. Retrieved 28 April 2025 from https://rochi.utcluj.ro/rrioc/articole/RRIOC-11-1-Gheorghe.pdf

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2014). Web scraping technologies in an API world. *Briefings in Bioinformatics*, *15*(5), 788–797. https://doi.org/10.1093/bib/bbt026

Gold, Z., & Latonero, M. (2017). Robots Welcome: Ethical and Legal Considerations for Web Crawling and Scraping. *Washington Journal of Law, Technology & Arts*, *13*(4), 275–

312. Retrieved 28 March 2025 from https://heinonline.org/HOL/Page?handle=hein.journals/washjolta13&id=283&div=&collection=

Google. (2024a, July 10). *Choosing the type of reCAPTCHA | Google for Developers*. Retrieved 25 April 2025 from https://developers.google.com/recaptcha/docs/versions

Google. (2024b, July 10). *reCAPTCHA v3*. Google for Developers. Retrieved 24 April 2025 from https://developers.google.com/recaptcha/docs/v3

Guerar, M., Verderame, L., Migliardi, M., Palmieri, F., & Merlo, A. (2022). Gotta CAPTCHA 'Em All: A Survey of 20 Years of the Human-or-computer Dilemma. *ACM Computing Surveys*, *54*(9), 1–33. https://doi.org/10.1145/3477142

Gutub, A., & Kheshaifaty, N. (2023). Practicality analysis of utilizing text-based CAPTCHA vs. Graphic-based CAPTCHA authentication. *Multimedia Tools and Applications*, *82*(30), 46577–46609. https://doi.org/10.1007/s11042-023-15586-5

hCaptcha. (2025). *Developer Guide | hCaptcha*. https://docs.hcaptcha.com/

HTTPProxyOkeyProxy. (2024, August 9). *The Ethical Implications of Robots.txt in Web Scraping—Okey proxy*. Retrieved 28 April 2025 from https://mirror.xyz/0xC82a668EBF772623a441eEC2f817B482634a26eb/ZgIsU7VZy4LuxLlo7algTMrhtU0XOsx0S46qByzJkJE?utm_source=chatgpt.com

Jin, R., Huang, L., Duan, J., Zhao, W., Liao, Y., & Zhou, P. (2023). *How Secure is Your Website? A Comprehensive Investigation on CAPTCHA Providers and Solving Services* (arXiv:2306.07543). arXiv. https://doi.org/10.48550/arXiv.2306.07543

Juntunen, M., & and Lehenkari, M. (2021). A narrative literature review process for an academic business research thesis. *Studies in Higher Education*, *46*(2), 330–342. https://doi.org/10.1080/03075079.2019.1630813

Jyväskylän Yliopisto & Maanpuolustuskoulutusyhdistys. (2024, April 4). *2.2 Verkon osoitteet*. Kansalaisen kyberturvallisuus -verkkokurssi. Retrieved 1 May 2025 from https://m3.jyu.fi/jyumv/ohjelmat/it/panu/kansalaisen-kyberturvallisuus-1/tekstiosuudet-aanena/recording-23-10-2021-14.44

Khder, M. (2021). Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *International Journal of Advances in Soft Computing and Its Applications*, *13*(3), 145–168. https://doi.org/10.15849/IJASCA.211128.11

Koster, M. (1994, July 30). *A Standard for Robot Exclusion*. Retrieved 14 February 2025 from https://www.robotstxt.org/orig.html#author

Koster, M., Illyes, G., Zeller, H., & Sassman, L. (2022). *Robots Exclusion Protocol* (Request for Comments RFC 9309). Internet Engineering Task Force. https://doi.org/10.17487/RFC9309

Krotov, V., & Johnson, L. (2023). Big web data: Challenges related to data, technology, legality, and ethics. *Business Horizons*, *66*(4), 481–491. https://doi.org/10.1016/j.bushor.2022.10.001

Krotov, V., Johnson, L., Murray State University, & Silva, L. (2020). Legality and Ethics of Web Scraping. *Communications of the Association for Information Systems*, *47*, 539–563. https://doi.org/10.17705/1CAIS.04724

Logos, K., Brewer, R., Langos, C., & Westlake, B. (2023). Establishing a framework for the ethical and legal use of web scrapers by cybercrime and cybersecurity researchers: Learnings from a systematic review of Australian research. *International Journal of Law and Information Technology*, *31*(3), 186–212. https://doi.org/10.1093/ijlit/eaad023

Luscombe, A., Dick, K., & Walby, K. (2022). Algorithmic thinking in the public interest: Navigating technical, legal, and ethical hurdles to web scraping in the social sciences. *Quality & Quantity*, *56*(3), 1023–1044. https://doi.org/10.1007/s11135-021-01164-0

Manoharan, M. (2024). API Rate Limiting Mechanisms in SaaS Applications: A Systematic Analysis of DDoS Protection Strategies. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *10*(6), Article 6. https://doi.org/10.32628/CSEIT241061223

NopeCHA. (2025). *Google reCAPTCHA V3*. Retrieved 24 April 2025 from https://developers.nopecha.com/token/recaptcha3/

Oesch, R., Eloranta, M., & Heino, M. (2017). *Immateriaalioikeudet ja yleinen etu*. Alma Talent Oy.

Pagallo, U., & Ciani Sciolla, J. (2023). Anatomy of web data scraping: Ethics, standards, and the troubles of the law. *European Journal of Privacy Law & Technologies*, *2*, 1–19. https://doi.org/10.57230/ejplt232PS

Paige, J. (2024). The Legality and Ethics of Web Scraping in Archaeology. *Advances in Archaeological Practice*, *12*(2), 98–106. https://doi.org/10.1017/aap.2023.42

Pathum, U. (2023, March 26). reCAPTCHA: How it works. *Medium*. Retrieved 25 April 2025 from https://medium.com/@hwupathum/recaptcha-how-it-works-4031eae74a8b

Plesner, A., Vontobel, T., & Wattenhofer, R. (2024). Breaking reCAPTCHAv2. *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, 1047–1056. https://doi.org/10.1109/COMPSAC61105.2024.00142

Possler, D., Bruns, S., & Niemann-Lenz, J. (2019). Data Is the New Oil—But How Do We Drill It? Pathways to Access and Acquire Large Data Sets in Communication Science. *International Journal of Communication*, *13*. Retrieved 16 April 2025 from https://ijoc.org/index.php/ijoc/article/viewFile/10737/2763

radware. (2023). *What is rate limiting and how does it work? | Radware*. Retrieved 1 May 2025 from https://www.radware.com/cyberpedia/bot-management/rate-limiting/

Raghavan, B., Vishwanath, K., Ramabhadran, S., Yocum, K., & Snoeren, A. C. (2007). Cloud control with distributed rate limiting. *ACM SIGCOMM Computer Communication Review*, *37*(4), 337–348. https://doi.org/10.1145/1282427.1282419

Saha, S. K., Nag, A. K., & Dasgupta, D. (2015). Human-Cognition-Based CAPTCHAs. *IT Professional*, *17*(5), 42–48. https://doi.org/10.1109/MITP.2015.79

Shet, V. (2014a, April 16). Street View and reCAPTCHA technology just got smarter. *Google Online Security Blog*. Retrieved 16 April 2025 from https://security.google-blog.com/2014/04/street-view-and-recaptcha-technology.html

Shet, V. (2014b, December 3). Are you a robot? Introducing “No CAPTCHA reCAPTCHA”. *Google Online Security Blog*. Retrieved 24 April 2025 from https://security.google-blog.com/2014/12/are-you-robot-introducing-no-captcha.html

Shi, C., Xu, X., Ji, S., Bu, K., Chen, J., Beyah, R., & Wang, T. (2022). Adversarial CAPTCHAs. *IEEE Transactions on Cybernetics*, *52*(7), 6095–6108. https://doi.org/10.1109/TCYB.2021.3071395

Singrodia, V., Mitra, A., & Paul, S. (2019). A Review on Web Scrapping and its Applications. *2019 International Conference on Computer Communication and Informatics (ICCCI)*, 1–6. https://doi.org/10.1109/ICCCI.2019.8821809

Sivakorn, S., Polakis, J., & Keromytis, A. D. (2016). *I'm not a human: Breaking the Google reCAPTCHA*. Black Hat Asia 2016. Retrieved 16 April 2025 from https://www.blackhat.com/docs/asia-16/materials/asia-16-Sivakorn-Im-Not-a-Human-Breaking-the-Google-reCAPTCHA-wp.pdf

Specht v. Netscape Communications Corp. (5 July 2001). Retrieved 16 April 2025 from https://law.justia.com/cases/federal/district-courts/FSupp2/150/585/2468233/

Stankevich, A. (2017). Explaining the Consumer Decision-Making Process: Critical Literature Review. *JOURNAL OF INTERNATIONAL BUSINESS RESEARCH AND MARKETING*, *2*(6), 7–14. https://doi.org/10.18775/jibrm.1849-8558.2015.26.3001

Thelwall, M., & Stuart, D. (2006). Web crawling ethics revisited: Cost, privacy, and denial of service. *Journal of the American Society for Information Science and Technology*, *57*(13), 1771–1779. https://doi.org/10.1002/asi.20388

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *59*(236), 433–460. Retrieved 16 April 2025 from https://www.jstor.org/stable/2251299

von Ahn, L., Blum, M., Hopper, N., & Langford, J. (2000). *The Official CAPTCHA Site*. Retrieved 16 April 2025 from http://www.captcha.net/

von Ahn, L., Blum, M., & Langford, J. (2004, February). Telling Humans and Computers Apart Automatically. *Communications of the ACM*, *47*(2), 57–60. Retrieved 16 April 2025 from http://www.captcha.net/captcha_cacm.pdf

Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science*, *321*(5895), 1465–1468. https://doi.org/10.1126/science.1160379

xHossein. (2021). *PyPasser* [Python]. Retrieved 28 April 2025 from https://github.com/xHossein/PyPasser

Xiao, G. (2021). Bad Bots: Regulating the Scraping of Public Personal Information. *Harvard Journal of Law & Technology*, *34*(2). Retrieved 26 April 2025 from

https://jolt.law.harvard.edu/assets/articlePDFs/v34/6.-Xiao-Bad-Bots-Regulating-the-Scraping-of-Public-Personal-Information-edit.pdf

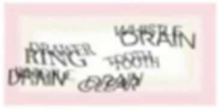
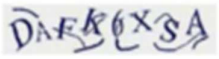
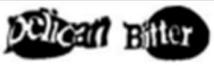
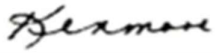
Xu, Y., Reynaga, G., Chiasson, S., Frahm, J.-M., Monrose, F., & van Oorschot, P. (2012). *Security and Usability Challenges of Moving-Object CAPTCHAs: Decoding Codewords in Motion*. https://www.usenix.org/system/files/conference/usenixsecurity12/sec12-final118.pdf

Zhao, B. (2017). Web Scraping. In L. A. Schintler & C. L. McNeely (Eds.), *Encyclopedia of Big Data* (pp. 326–328). Springer International Publishing. https://doi.org/10.1007/978-3-319-32001-4_6-1

Zynda, T. (2004). Ticketmaster Corp. V. Tickets.com, Inc. - Preserving Minimum Requirements of Contract on the Internet. *Berkeley Technology Law Journal*, *19*(1), 495–518. https://doi.org/10.15779/Z38Q965

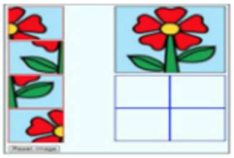
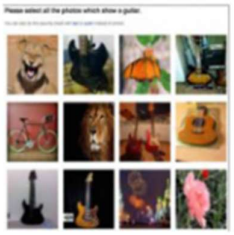
Appendices

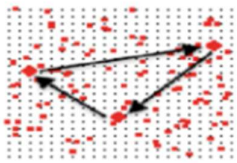
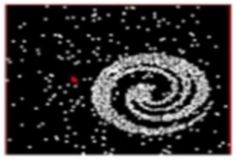
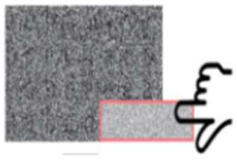
Appendix 1. Taxonomy of text-based CAPTCHAs (Guerar et al., 2022)

Type	Scheme	Sample	Year	Challenge Description
2D	GIMPY [129]		2000	Recognize three words out of of seven selected randomly from a dictionary
	EZ-GIMPY [129]		2000	Recognize one English word in a distorted image
	BaffelText [23]		2003	Recognize a pronounceable string of characters with difference masking applied
	Microsoft (MSN) [137]		2002	Recognize eight distorted characters presented with random arcs as clutters
	Google (Gmail) [137]		2006	Recognize characters which are crowded together
	Yahoo [137]		2008	Recognize a string of characters connected by intersecting random lines
	Megaupload		2010	Recognize four overlapped characters with negative intersection areas
	ReCAPTCHA V1 [128]		2008	Recognize distorted text scanned from old books
	Clickable CAPTCHA [25]		2008	Identify English words among non-English words
	Handwritten [106]		2004	Recognize a distorted handwritten text (e.g., city name)
3D	Teabag 3D [95]		2006	Recognize a sequence of characters that appears on a grid in 3D space
	3DCAPTCHA [93]		2006	Recognize a sequence of 3D characters
	Super CAPTCHA [131]		2013	Recognize a sequence of 3D characters
	DotCHA [72]		2019	Drag and rotate the model to identify each letter, then type the answer
Animated	Dracon CAPTCHA [31]		2006	Recognize five characters which fade and blur at various times over the animation frames
	KillBot Professional [90]			Recognize five moving characters among a noisy foreground and/or background
	Atlantis CAPTCHA [90]			Recognize six moving characters among other continuously changing their color
	HelloCAPTCHA [101]		2010	Recognize a sequence of six characters displayed in an animated GIF image
	NuCaptcha [94]		2008	Type the last three red moving characters

Appendix 2. Taxonomy of image-based CAPTCHAs (Guerar et al., 2022)

Type	Scheme	Sample	Year	Challenge Description
Click	Implicit CAPTCHA [6]		2005	Click on a specific area of an image (e.g., mountain top)
	SACaptcha [121]		2018	Click on some regions in the image that have a specific shape mentioned in the challenge description
Sliding	WHAT'S UP CAPTCHA [48]		2009	Move the slider to adjust at least three randomly rotated images to their upright orientation
	MintEye CAPTCHA [99]		2012	Move the slider until undistorted version of the image appears
	Tencent (Tencent.com)			Drag the slider until two puzzle pieces match
Drag and Drop	Garb CAPTCHA [132]		2013	Drag and drop the puzzle pieces to their correct position to reconstruct the original image
	Capy CAPTCHA [17]		2012	Drag a puzzle piece to complete a jigsaw
	KeyCAPTCHA [71]		2010	Drag three puzzle pieces to assemble the image
	Gao et al [44]		2010	Identify the two misplaced pieces and swap them

Selection	Hamid Ali et al [60]		2014	Drag and drop four images to an empty grid following the same order in the reference image
	Asirra [35]		2007	Select cats from a set of 12 images of cats and dogs
	No CAPTCHA reCAPTCHA [111]		2014	Select images that have the same content described in the challenge with a sample image
			2014	Select all images with street signs, cars, bridges or some specific object
	Facebook image CAPTCHA			Select the images that correspond to a hint from twelve images with different content
	HumanAuth [89]		2006	Select images with natural contents
	SEMAGE [124]		2011	Select semantically related images from a set of images
	AVATAR [33]		2012	Select avatar faces from a set of 12 images composed of a mix of human and avatar faces

	FR-CAPTCHA [50]		2014	Select real human faces distorted among nonhuman face images
	FaceDCAPTCHA [49]		2014	Select two face images of the same person
Drawing	VAPTCHA [139]		2018	Draw an resemblant trajectory to match the reference trajectory
	Drawing CAPTCHA [113]		2006	Connect a specific dots to each other
	MotionCAPTCHA [1]		2011	Draw a shape displayed in a box
Interactive	CAPTCHAStar		2015	Move the cursor until forming a recognizable shape
	Cursor CAPTCHA [122]		2013	Overlap the cursor on the identical object placed in a random generated image
	Noise CAPTCHA [97]		2012	Move a small noisy image over a large noisy image until a hidden message or object appears