Web Data
Extraction Summit

Powered by zyte

Step-by-Step Guide to Compliant Web Scraping

Real world examples to assess compliance

Use Cases

- **Product Data Extraction for Competitive Intelligence** - scrape e-commerce product pages for competitive pricing intelligence
- **Data of business executives to create a database** - scrape data from business pages about executives to create database
- **Image Extraction to Train LLM for GenAI** - scrape images to put in an LLM that will be used to build an AI image generator
- **Large scale data extraction to train LLM** - scrape data from a myriad of websites to train LLM for GenAI chat system

*I'm a lawyer, but I am not your lawyer. Please seek independent legal advice to fully assess your web scraping needs.

Potential Risk Areas

Personal Data

Any data that identifies a living human being. This is broad – you need to identify whether personal data is in scope and the applicable laws.

Website Terms / Login

Any time you have to explicitly agree to a website's terms (including mobile apps) or login – you need to comply fully with the terms you agree to.

Copyright

A tangible, original piece of work – you must identify whether copyrighted data is in scope and if your use is acceptable.

External Data Use

Web data is mostly used for internal business purposes – if your use goes beyond this you need to identify and mitigate the risks.

AI

AI litigation and legislation is cropping up all over the world. If you are using web data to train LLMs or build GenAI applications you need to ensure compliance.

Sensitive Data

Health, biometric, or other more sensitive data has extra risk – you must identify when this data may be in scope.

Where to Start?

Risk Category	Yes	No
Non-Public Data		
Agreement to Terms		
AI		
Copyrighted Data		
Personal Data		
External Use of Data		
Other Sensitive Data		

Product Data

Details

- E-commerce data
- Competitive pricing intelligence
- Fields include product name, description, price, specs, star rating, reviews, reviewer name
- No login required for most products
- Login required for in-cart scraping

Risk Category	Yes	No	Notes
Non-Public Data	✓		In-cart behind login
Agreement to Terms	✓		Accept terms to scrape in cart
AI		✓	
Copyrighted Data		✓	
Personal Data	✓		Reviewer name
External Use of Data		✓	
Other Sensitive Data		✓	

Personal Data Analysis

Reviewer Name

Is there personal data?	Jurisdiction	Lawful Basis	Data Minimization	Descope
Yes. A reviewer's name, even if a screen name and not their real name is considered personal data.	Scraping from global e-commerce sites with a focus on EU.	Is there a lawful basis to collect the personal data? Do you have consent, contractual agreement, or legitimate interest? Only option is LI, but does that exist here?	Only collect the personal data you really need. Do you absolutely need the review name to conduct your competitive intelligence project? Probably not	In this case, there is no legitimate interest because they don't really need the review name to achieve their desired outcome. As such, descope that field of data and continue with project safely.

In-Cart and Login Analysis

In-Cart Scraping

Do you need to add to cart?	Login to add to cart?	Read and Abide by ToS	Respect the Website	What is interference?
In some cases you can only see stock numbers or special pricing if you add the item to cart.	On some sites you need to login to add an item to cart. If so, you need to consider what terms you are agreeing to in order to login.	Every website has its own terms of service you must agree to when logging in. When you click and agree to the terms you create a binding contract. If the terms say no scraping, then do not scrape behind login.	If you don't have to login or the terms don't prohibit scraping, you might be able to proceed. But you want to ensure adding items to cart does not interfere with the website's operations.	In order not to interfere with the website operations, take steps to ensure you don't put items with low stock in cart, only keep in cart briefly, and check to see if having it in cart prevents others from purchasing.

Business Exec Database

Details

- Scrape names, email, phone numbers, and address for business executives across the US and EU
- Data to be used to create a database of executives with contact information

Risk Category	Yes	No	Notes
Non-Public Data		✓	
Agreement to Terms		✓	
AI		✓	
Copyrighted Data		✓	
Personal Data	✓		Name, email, phone
External Use of Data		✓	
Other Sensitive Data		✓	

Personal Data Analysis

Business Execs

Is there personal data?	Jurisdiction	Lawful Basis	Public Data	Result
Yes and No Name = yes Direct Email = yes Generic Email = no Direct Phone = yes General Bus Phone = no Business Address = no	Scraping US and EU executives so US states laws will apply and GDPR. Note that state laws all have different thresholds for when they apply, so they may not always be applicable.	Is there a lawful basis under GDPR? Potentially legitimate interest and should conduct an LIA.	In US, many state laws have an exception to allow collection of public personal data. “Information that a business has a reasonable basis to believe is lawfully made available to the general public by the consumer or from widely distributed media.” –CCPA	GDPR: Need legitimate interest analysis, notification, DSAR. If not able to do this, should not scrape. US: Likely considered public data and ok to proceed.

AI Image Generator

Details

- Public images
- Input into company's LLM to train image generator
- System that generates images from the images it is trained on
- Images are copyrighted by the original author
- Images may contain images of people, which is personal data

Risk Category	Yes	No	Notes
Non-Public Data		✓	
Agreement to Terms		✓	
AI	✓		Image Gen AI
Copyrighted Data	✓		Images
Personal Data	✓		People within images
External Use of Data		✓	
Other Sensitive Data		✓	

Large Scale Scraping for LLM

Details

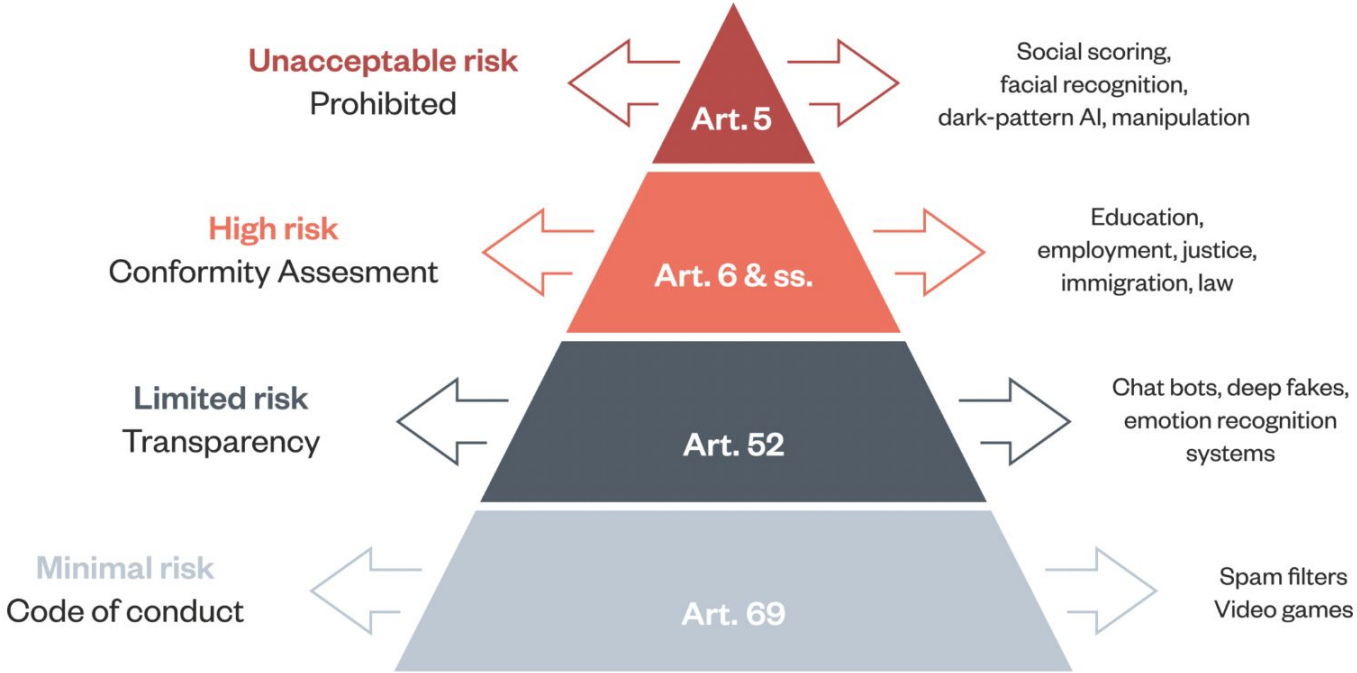
- Public data across the web, includes articles, books, research, etc
- Data use to train LLM for GenAI chat bot
- Many of the works scraped are copyrighted
- Personal data on a large scale may also be collected

Risk Category	Yes	No	Notes
Non-Public Data		✓	
Agreement to Terms		✓	
AI	✓		Create AI chatbot
Copyrighted Data	✓		Articles, books, etc
Personal Data	✓		Data within the items scraped
External Use of Data		✓	
Other Sensitive Data		✓	

GenAI
General Questions

Are copyrighted works in scope?	Is personal data being collected?	Use Case	Will the AI Act apply?	Case law ramifications?
Yes, images, articles, books, any tangible unique piece of work is copyrightable.	Yes, many images contain clear pictures of people and articles and books contain substantial personal data as well.	What you are using the GenAI for matters. Is it to compete with the source, make employment decisions, law enforcement. Context matters a lot with AI.	The AI Act will come into effect in the EU in 2024 and applies to all companies providing AI in the EU. The AI Act takes a risk based approach to compliance, with more compliance required for riskier AI.	There are multiple lawsuits in the US against GenAI providers and we won't know the results for some time. But we will assess based on initial court rulings and prior law.

EU AI Act
Risk Assessment



EU AI Act
High Risk Systems

Are copyrighted works in scope?	Do you have permission?	Is your use “fair use”?	Will the AI Act apply?	Case law ramifications?
Yes, images, articles, books, any tangible unique piece of work is copyrightable.	Some companies training their LLMs with web scraped data do have permission from the target websites, this will alleviate copyright concerns.	Ensure the images or responses generated are distinct enough from the originals to create transformative works. Court noted in Anderson v Stability AI that images generated aren’t substantially similar to original works.	If you are providing services in the EU, the AI Act will apply. However, your level of compliance will depend on the risk level of your use of the AI.	There are various lawsuits in the US looking at whether training LLMs with copyrighted materials for GenAI is a copyright violation. This is yet to be determined. Including class actions against OpenAI, Meta, Microsoft, and Google.

GenAI
Personal Data

Personal Data	Sensitive Personal Data	Lawful Basis	Data Minimization	Notification Requirements
Yes, many images contain clear pictures of people and articles and books contain substantial personal data as well.	Images and articles could include data about minors, gender, race, age, and other biometric data. This data comes with heightened standards and you want to avoid where you can.	Absent consent or contractual agreement, the only possible lawful basis for personal data going into LLMs is legitimate interest or task in public interest. We are yet to see whether this is successful, but for now consider an LIA and DPIA.	For images, descope any obvious tags that may be personal data. For example, kids, families, people, etc. Where possible, descope names and other personal data from being included in the training set.	Place clear notice on your website regarding the personal data you are collecting and using to train the LLM, place limitations on use of your GenAI apps, be clear when you are using GenAI, and ensure a proper DSAR process.

Ethics

- GenAI companies are currently being sued all over the world and we won't know the results for some time.
- Legislation is currently developing and similarly we won't have precise guidance until that is fully formed. China has enacted an AI law, the EU AI Act is pending, and the US is likely putting forth an AI Executive Order soon, and the US is looking at legislation on deep fakes.
- HOWEVER, we don't need case law and legislation to discuss ethics.
- Some ethical considerations include:
  - Respect No-AI tags. If someone has specifically said they do not want AI used on their copyrighted work, this should be respected where possible.
  - Consider biases that are created in AI systems. Examples include image generators that have racial and gender biases, job descriptions that are skewed to a particular gender, facial recognition used in law enforcement.
  - Be wary of spreading misinformation and fake content.

Compliance Risk Assessment

Our Process

Customer Intake Form	Project Review	Risk Assessment	Project Adjustments	Additional Reviews
Customer completes an initial compliance intake form which includes: ▪ Company Info ▪ Websites ▪ Data Fields ▪ Risks Areas ▪ Use Case	Zyte reviews the customer intake form and reverts to the customer with any questions or clarifications required.	Zyte provides the customer with a risk assessment to identify any compliance risks and provide customer with information on best next steps.	Work with customer on any adjustments or preparatory work required to ensure compliance.	As customer expands its projects, Zyte will work with the customer to continue to assess risk.


ewdci
CERTIFIED

Questions

Sanaea Daruwalla
Chief Legal Officer
sanaea@zyte.com
legal@zyte.com

Victoria Vlahoyiannis
Legal Counsel

Callum Henry
Legal Counsel