GPAI IP Expert

Guidelines for Scraping or Collecting Publicly Accessible Data

November 2022

GPAI / THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

This report was developed by Experts and Specialists involved in the Global Partnership on Artificial Intelligence's project Protecting AI innovation, Intellectual Property (IP). The report reflects the personal opinions of the GPAI Experts and Specialists involved and does not necessarily reflect the views of the Experts' organizations, GPAI, or GPAI Members. GPAI is a separate entity from the OECD and accordingly, the opinions expressed and arguments employed therein do not reflect the views of the OECD or its Members.

Acknowledgements

This report was developed in the context of the *Protecting AI innovation, Intellectual Property (IP)* project, with the steering of the project Co-Leads and the guidance of the Project Advisory Group, supported by the GPAI Innovation & Commercialization Working Group. The GPAI Innovation & Commercialization Working Group agreed to declassify this report and make it publicly available.

Co-Leads:
Lee Tiedrich[*], Duke University
Yann Dietrich[**], Atos

Project Advisory Group:
Teki Akuetteh[*], Africa Digital Rights' Hub
Carolyn Blankenship[†], BigBear.ai
Neeraj Gupta[**], Formulate IP
Matthieu Dhenne[**], IPSILON Dispute & Resolution
Yann Dietrich[**], Atos
Josef Drexl[*], Max Planck Institute for Innovation and Competition
Antoine Dupont[**], CEA
Robert Kroplewski[*], Poland's Minister of Digitalization in Information Society Affairs, Chancellery of Prime Minister
Andreas Liebl[*], appliedAI
Hisao Shiomi[†], Tsukuba University
Rajesh Sreenivasan[†], Technology Media and Telecoms Law Practice, Rajah & Tann Singapore LLP
Lee Tiedrich[*], Duke University
Ulrike Till[†], WIPO
Kat Walsh[†], Creative Commons

Research Assistants:
Marina Bojarski‡, GPAI Junior Investigator - under the direction of Yann Dietrich[**]
Daria Kim‡, under the direction of Josef Drexl[*]
Annmarie Messing‡, under the direction of Lee Tiedrich[*]
Joe O'Brien‡, under the direction of Lee Tiedrich[*]

Yusuke Yogoro‡, under the direction of Hisao Shiomi[†]

The report was written by: **Yann Dietrich**[**], Atos; **Carolyn Blankenship**[†], BigBear.ai; **Marina Bojarski**‡, GPAI Junior Investigator with the significant contribution and involvement of the following individuals in the Project Advisory: **Hisao Shiomi**[†], Tsukuba University; **Rajesh Sreenivasan**[†]. Finally, GPAI would like to acknowledge the efforts of colleagues at the Paris Centre of Expertise at Inria. We are particularly grateful for the support of Kaitlyn Bove, Laetitia Cuignet, Isabelle Herlin, and Catherine Pacherie-Simeral from Inria, and for the dedication of the Working Group Co-Chairs Laurence Liew and Françoise Soulié-Fogelman

* Expert of GPAI’s Innovation & Commercialization Working Group
(n.b. Teki Akuetteh and Josef Drexl are Experts in GPAI’s Data Governance Working Group)
** Observer at GPAI’s Innovation & Commercialization Working Group
† Invited Specialist
‡ Contracted parties by Inria

Citation

GPAI 2022. Protecting AI innovation, Intellectual Property (IP): GPAI IP Expert: Guidelines for Scraping or Collecting Publicly Accessible Data, Report, November 2022, Global Partnership on AI.

Disclaimer

These guidelines are intended for those with basic knowledge in intellectual property.

As these are global guidelines, if the specific issue that you are facing is not addressed, it is recommended to contact a lawyer.

Please note that the guidelines are mainly addressing intellectual property issues and do not contend with personal data issues.

The European commission in the Directive on copyright and related rights in the Digital Single Market is referring to data mining. For clarity, we have chosen to use the technical term data scraping.

Introduction

In a data-driven world, data access is key to develop digital products. The IP Committee of the GPAI Innovation & Commercialization Working Group has initiated this work to identify global recommendations applicable regardless of the country.

Data scraping is the process of extracting content from a website and importing it on a computer. The content can be used to then be analyzed or fed into an artificial intelligence algorithm. In certain instances, content scraped from public-facing websites may be protected by copyright and will require a license or an exception.

Given that there is no international exception to copyright for data scraping, jurisdictions have very different approaches to the matter. As examples, data scraping in the United States can be allowed under fair use, provided it meets the criteria. Additionally, the European Union has introduced new text and data mining exceptions to copyright[1]. Japan allows data scraping for computerized technical analysis only.

There is quite the uncertainty on whether data scraping is subjected to the authorization of the right holder in other jurisdictions.

Taking into account the diversity of applicable laws, these guidelines were designed to provide general recommendations for data scraping.

So, the first graphic intends to be thought of as global guidelines to explain what one should and should not do, in order to avoid any intellectual property (IP) issues while web scraping and training AI.

The second graphics are thought to present different exceptions that may be applicable depending on the jurisdiction.

[1] The European Commission has introduced two exceptions to copyright for text and data mining.
The DSM Directive creates a mandatory exception for the reproduction of copyrighted content and the extraction from the databases for research organizations the purpose of scientific research. No prior authorization must be requested from the copyright owners, who cannot impose any compensation for the use of content.
The second exception is applicable to any other entity. An entity can mine data provided that the right holder has not expressed its "opt-out" in an appropriate manner.

KEY ASPECTS TO CONSIDER
WHEN YOU WANT TO SCRAPE DATA ON THE INTERNET

👎 👍

You should not scrape personal data (e.g. name, e-mail, employment info, biometric data, …). Unless you have a lawful reason …

If you want to scrape personal data, you need lawful reasons such as a consent, a contract, compliance needs, vital interest or legitimate interest

PERSONAL DATA

PUBLIC DOMAIN

You should look for public domain content.

You can find open datasets on some government websites or international non-governmetal organizations.

You should not use copyrightable content without the authorization of the copyright owner, unless it is under an open source licence, or public domain
Ex: music, images, articles, …

COPYRIGHTED CONTENT

NOT COPYRIGHTABLE

You should look for not coyrightable content

It can be numbers, meteorological insights, judicial opinions, statutes, …

You should not circumvent websites or content that is password protected or has a paywall.

UNAUTHORIZED ACCESS

OPEN SOURCE

You should look for open source licences or creative commons.
For example, CC0, MIT, BSD, …
Be careful, check that your use is authorized.

You should not use data under a licence that doesn't authorize your use.
For example, if you want to use a research purpose only licence, you shouldn't use the data for a commercial product.
Some websites might also specifically deny you the right to use their data for Machine learning. Mind the restrictions of the website.

UNAUTHORIZED USE

AUTHORIZED USE

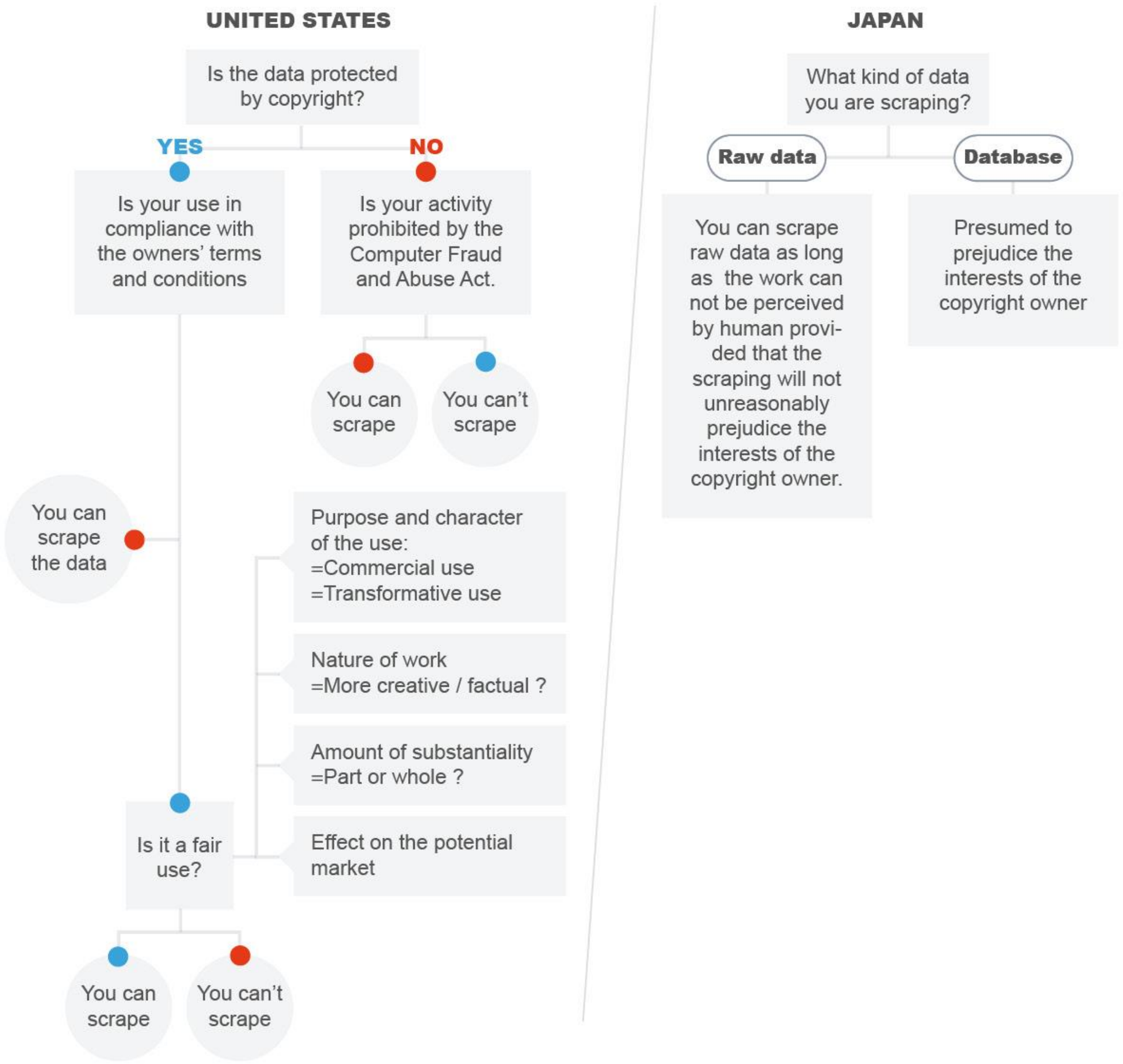
You should use a licence that allows you what you want to do.
Be careful, you may need to obtain a commercial licence.

Example : iStockphoto denies you the right to us their data for ML purposes, unless you pay for an upgraded licence.

DISCLAIMER :
PERSONAL DATA
NOT ADRESSED

EXCEPTIONS

WHICH LAW IS APPLICABLE?



UNITED STATES

Is the data protected by copyright?

YES

Is your use in compliance with the owners' terms and conditions

You can scrape the data

Is it a fair use?

Purpose and character of the use:
=Commercial use
=Transformative use

Nature of work
=More creative / factual ?

Amount of substantiality
=Part or whole ?

Effect on the potential market

You can scrape

You can't scrape

NO

Is your activity prohibited by the Computer Fraud and Abuse Act.

You can scrape

You can't scrape

JAPAN

What kind of data you are scraping?

Raw data

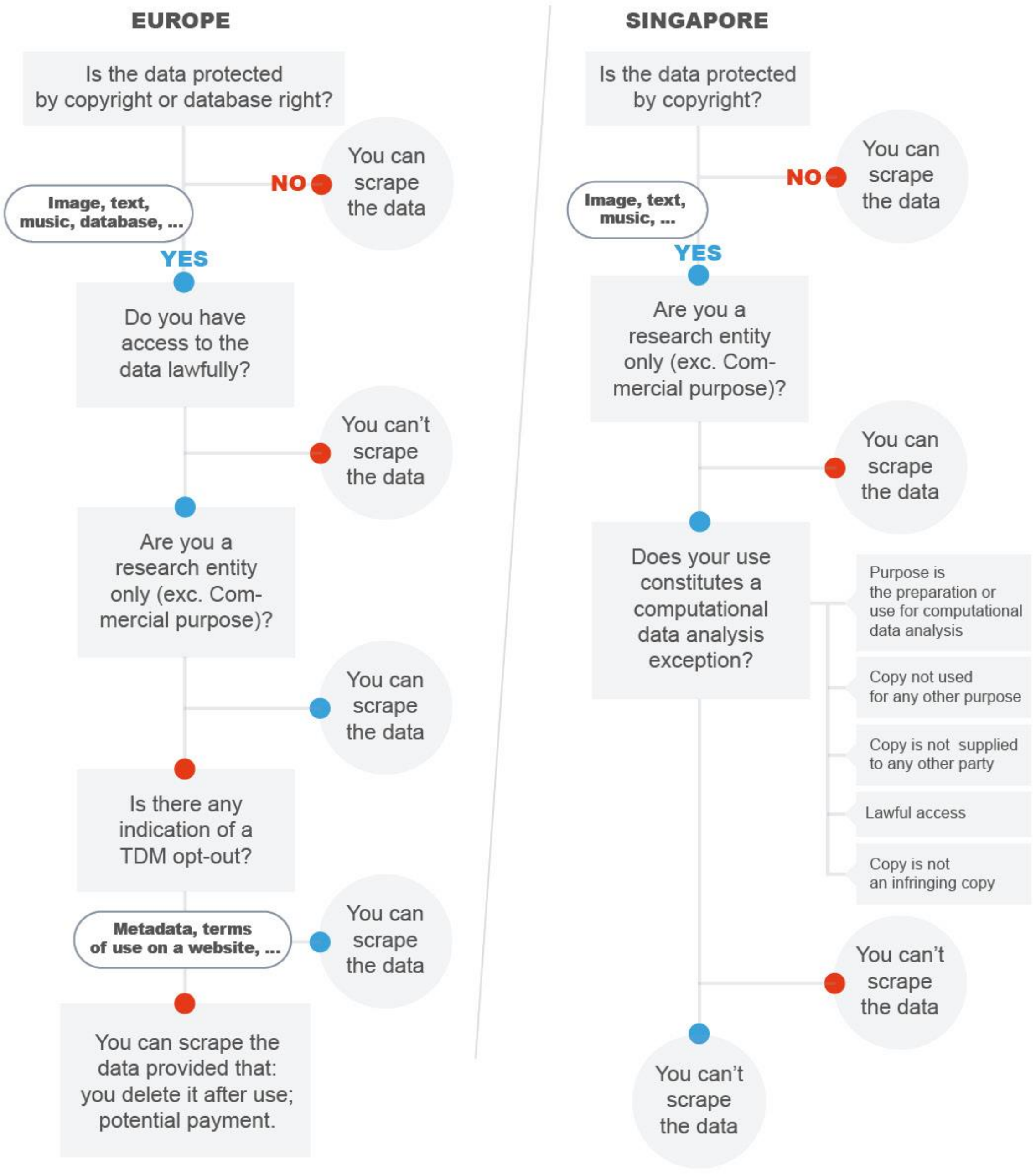
You can scrape raw data as long as the work can not be perceived by human provided that the scraping will not unreasonably prejudice the interests of the copyright owner.

Database

Presumed to prejudice the interests of the copyright owner

EXCEPTIONS

WHICH LAW IS APPLICABLE?



EUROPE

Is the data protected by copyright or database right?

Image, text, music, database, ...

NO — You can scrape the data

YES

Do you have access to the data lawfully?

You can't scrape the data

Are you a research entity only (exc. Commercial purpose)?

You can scrape the data

Is there any indication of a TDM opt-out?

Metadata, terms of use on a website, ...

You can scrape the data

You can scrape the data provided that: you delete it after use; potential payment.

SINGAPORE

Is the data protected by copyright?

Image, text, music, ...

NO — You can scrape the data

YES

Are you a research entity only (exc. Commercial purpose)?

You can scrape the data

Does your use constitutes a computational data analysis exception?

- Purpose is the preparation or use for computational data analysis
- Copy not used for any other purpose
- Copy is not supplied to any other party
- Lawful access
- Copy is not an infringing copy

You can't scrape the data

You can't scrape the data