


Ethical and Legal Aspects in CSS

SICSS

 CREST

 INSTITUT
POLYTECHNIQUE
DE PARIS

Introduction

Ethics, legal aspects: why should we care ?

Two main reasons (according to Matt Salganik)

Introduction

Ethics, legal aspects: why should we care ?

Two main reasons (according to Matt Salganik)

1) Fear-based

- At the individual & at the collective level, we have an incentive to avoid trouble.
- And there can be plenty

Introduction

Ethics, legal aspects: why should we care ?

Two main reasons (according to Matt Salganik)

1) Fear-based

2) **Hope-based**

- We are currently limiting ourselves, while some innovative research could be carried out

Introduction

Ethics, legal aspects: why should we care ?

Two main reasons (according to Matt Salganik)

- 1) Fear-based
- 2) Hope-based
- 3) Overall, we have no choice

Introduction

So what should we do ? This is where the problem start

Today: a quick introduction to the legal & ethical aspects you should be paying attention to.

Watch out:

- I am not a lawyer
- Ethical norms evolve (stay tuned)

Introduction

So what should we do ? This is where the problem start

- Strict legal rules, lot of grey area
- Disagreement on ethical rules

Introduction

2 dimensions: scraping / data management

x

2 aspects: legal aspects / ethical aspects

I. Data scraping

Rule #1 is that scraping a website is often **ILLEGAL** (yes, it's illegal).

Several issues with scraping

- Infringement of database rights
- Copyright

I. Data scraping

Rule #1 is that scraping a website is often **ILLEGAL**

Websites enforce these rules differently

- Some will let you collect the information you are seeking
- Some will try to limit your impact
 - Positively, by offering an API
 - Negatively, by placing some constraints on your actions

Some websites will actively go after you

I. Data scraping

Rule #1 is that scraping a website is often **ILLEGAL**

Remember

it is not because you can do it that it is legal

I. Data scraping

Rule #1 is that scraping a website is often **ILLEGAL**
BUT

There are 'research exceptions' to these rules

In particular, there exists a European directive that specifies the perimeter of the exception for 'text and data mining'

Here is the text of the directive, the French transposition from 2021, and some plain text explanations.

I. Data scraping

Some (vague) details about this exception

In France (~Europe), it is possible to carry out text and data mining on any single resource on which you have access to legally. No limitation in volume nor in time, but conditions.

I. Data scraping

Some (vague) details about this exception

Includes: copyrighted material you have purchased (archive of a newspaper, database), data you collected via a subscription (newspaper databases paid for by a university library) ; content available on the internet **on the condition the acquisition was licit.**

I. Data scraping

Some (vague) details about this exception

Conditions

- Access to the data was legal.
- No theft, no cheating

Data protection is key: Taking measures to ensure that the data is stored in a secure location and that measures have been taken to avoid any leak/loss.

I. Data scraping

Ethics

Even if it is allowed, don't forget to apply good practices

- Is there an API ? If yes, please use it
- Introduce yourself (give an email address)
- Be kind (add some pauses)
- Be sober, download only once

II. Data management

The problem with Human and social sciences is that they deal with **Human Subjects**.

This make all actions fall under the remit of the GDPR (General Data Protection Regulation)

II. Data management

GDPR is a vast, complex code.

To make things worse, its interpretation often depends on who reads it.

Rule #1: any data that includes either **identifying** or **sensitive information** is subject to GDPR provisions.

II. Data management

GDPR is a vast, complex code.

Personal data: is any information that can lead (directly or with other information) to the identification of a living natural person (GDPR 4, Recital 27).

II. Data management

GDPR is a vast, complex code.

Personal data: is any information that can lead (directly or with other information) to the identification of a living natural person (GDPR 4, Recital 27).

- A name, an e-mail address
- A username or handle – and everything (tweets, forum posts, etc.) connected to this username!
- An IP-address (if not leading to a VPN server)
- Images (of a face) and voice recordings
- All kinds of contextual information that can lead to identifying a person: "I am a woman, a sociologist, working at ENSAE. I am also originally from Italy, and..."
- Watch out : data can be merged and become identifying (email leading to a cell phone)

II. Data management

GDPR is a vast, complex code.

Personal data: is any information that can lead (directly or with other information) to the identification of a living natural person (GDPR 4, Recital 27).

Personal data should be processed for "**specified, explicit and legitimate purposes**" (GDPR 5)

II. Data management

Personal data

From here on, we enter into complex territory, since you can process this data in the following conditions.

- Consent
- Fulfilling a contract or legal obligation
- Protecting the vital interests of the data subject
- Carrying out a task in the public interest ← **Go for this one**
- The legitimate interests of the controller (e.g. a landlord will need to collect some personal data of tenants in order to be able to rent out apartments).

II. Data management

Identifying information

The GDPR identifies special categories of personal data (GDPR:9) as follows:

- Racial or ethnic origin
- Political opinions and religious or philosophical beliefs
- Trade union membership
- Information about health, sex life and sexual orientation
- Genetic and biometric data that can lead to the identification of a living person (but not images in general).

Difference:

Processing personal data is allowed "if...", processing special categories is prohibited "unless..."

II. Data management

Yet, avoid avoiding doing research because of this.

Dealing with Human Subjects

5 (non-legally) sufficient advice

Dealing with Human Subjects

5 (non-legally) sufficient advice

1. Store your data adequately

- Are your files on a secure server?
- Is your disk encrypted?
- Do you protect the name of the people you interview?
- Where do you store copies of your material?

Dealing with Human Subjects

5 (non-legally) sufficient advice

2. Ask yourself : are you affecting your subjects ?

Non/Obstrusive methods

Is it worth it ? Is it scientifically legitimate ?

Dealing with Human Subjects

5 (non-legally) sufficient advice

3. Do you anonymize properly ?

- Change /Erase all names

But : Is this enough ? Probably not

Reason #1: anonymity != confidentiality

Reason #2: Re-identification can happen with limited data (Sweeney, 2000)

Dealing with Human Subjects

5 (non-legally) sufficient advice

4. 'But the data is already public' is not a good excuse

Dealing with Human Subjects

5 (non-legally) sufficient advice

5. **Should you release the data?**

A growing demand (for replication, for cumulativity)

All good, but should not happen at the expense of your subjects.

Dealing with Human Subjects

5 (non-legally) sufficient advice

⇒ **Tread carefully, and borrow from ethnographers**

Conclusion

- Even if it can be annoying, don't forget that privacy rules are (in principle) made to protect privacy.
- Data can stay for long, or even return at unexpected moments in your life



IT CAME
FROM THE
DATA LAKE

See [this talk](#) by M. Ceglowski