Web Data Collection with R

Session 4: Ethical and Legal Considerations

Sascha Göbel
sascha.goebel@uni-konstanz.de

November 25th, 2020

Today: Ethics and Legality

- ethical and legal challenges
- polite web data collection with R
- some general advice

Ethical Considerations

The problem

- uncertainty about what can/should and cannot/should not be done
- neither strict rules-based nor ad-hoc approach suited
- norms, rules, and laws lag behind capabilities for massive web data collection
- there is no risk-free web data collection

A principles-based approach (Salganik 2018)

- our focus in this seminar is on passive observation
- but the skills you obtain can be used for intervention, too
- evaluate existing rules *and* general ethical principles

Ethical Considerations

The problem

- uncertainty about what can/should and cannot/should not be done
- neither strict rules-based nor ad-hoc approach suited
- norms, rules, and laws lag behind capabilities for massive web data collection
- there is no risk-free web data collection

A principles-based approach (Salganik 2018)

- our focus in this seminar is on passive observation
- but the skills you obtain can be used for intervention, too
- evaluate existing rules *and* general ethical principles

(1) Respect for persons

(1) individuals should be treated as autonomous

(2) individuals with diminished autonomy should be entitled to additional protections

- *unseen seer*
- when possible, get informed consent
  - participants not researchers get to decide
  - "some form of consent for most research"
  - consent given through ToS?
  - minors on Twitter?

(2) Beneficence

(1) do not harm

(2) maximize possible benefits and minimize possible harms

- conduct risk-benefit analysis
- what is the probability of adverse events?
- impact on study participants and social contexts
- open and reproducible research
- mind information risk:
  - unanticipated secondary use
  - *database of ruin*
  - anonymize
  - comply with ToS
  - regulate data access

(3) Justice

- protection and access
- disadvantaged groups should benefit from knowledge
- allow participation of all groups

(4) Respect for law and public interest

- include law in your considerations
- compliance with local laws and website ToS
- special circumstances may justify violating ToS
- explain your decisions openly ...
- at the same time balance legal risks

(3) Justice

- protection and access
- disadvantaged groups should benefit from knowledge
- allow participation of all groups

(4) Respect for law and public interest

- include law in your considerations
- compliance with local laws and website ToS
- special circumstances may justify violating ToS
- explain your decisions openly ...
- at the same time balance legal risks

Always ask yourself

- Am I comfortable with publishing my research including details on data collection?
- Would I expect (my) information to be (re-)used for this purpose?
- How will others react?

Ask others

- Discuss your ideas in class
- Make use of office hours

Always ask yourself

- Am I comfortable with publishing my research including details on data collection?
- Would I expect (my) information to be (re-)used for this purpose?
- How will others react?

Ask others

- Discuss your ideas in class
- Make use of office hours

Table 6.2: The "Five Safes" are Principles for Designing and Executing a Data Protection Plan (Desai, Ritchie, and Welpton 2016)

Safe	Action
Safe projects	Limits projects with data to those that are ethical
Safe people	Access is restricted to people who can be trusted with data (e.g., people who have undergone ethical training)
Safe data	Data are de-identified and aggregated to the extent possible
Safe settings	Data are stored in computers with appropriate physical (e.g., locked room) and software (e.g., password protection, encrypted) protection
Safe output	Research output is reviewed to prevent accidental privacy breaches

Source: Salganik 2018

What about IRB approval?

- no direct equivalent at German universities
- ethics boards may reject review on "no human subjects research" grounds
- still, try if deemed necessary, but know that its not enough

Polite Web Data Collection

Polite behavior

- If available, use official/authorized sources
- identify yourself and be responsive
- ask for permission
- query sparsely

Rude behavior can have consequences, not just for you

- IP address (field) blocked
- API access credentials blacklisted
- Providers move content to APIs, rate limit, restrict content
- challenge-response tests (CAPTCHAs)
- frequent changes in HTML source code

Polite Web Data Collection

Polite behavior

- If available, use official/authorized sources
- identify yourself and be responsive
- ask for permission
- query sparsely

Rude behavior can have consequences, not just for you

- IP address (field) blocked
- API access credentials blacklisted
- Providers move content to APIs, rate limit, restrict content
- challenge-response tests (CAPTCHAs)
- frequent changes in HTML source code

If there is no authorized data access

Identify yourself

- don't hide behind your IP address and common user-agents
- state your intentions
- website owners should be able to get in touch with you
- let's adjust your user-agent in R

Ask for permission

- robots.txt (Robots Exclusion Protocol)
- informal rules, not binding by law, no technical barrier
- website owners want to keep server traffic in check
- tells which information may be scraped
- if access is disallowed, contact website owner

If there is no authorized data access

Identify yourself

- don't hide behind your IP address and common user-agents
- state your intentions
- website owners should be able to get in touch with you
- let's adjust your user-agent in R

Ask for permission

- robots.txt (Robots Exclusion Protocol)
- informal rules, not binding by law, no technical barrier
- website owners want to keep server traffic in check
- tells which information may be scraped
- if access is disallowed, contact website owner

robots.txt

- text file stored in root directory of a website
- "User-agent" specifies who, * = everyone
- "Disallow" and "Allow" specify what (path)
- general ban: User-Agent: * Disallow: /
- Crawl-delay: pause between requests for a certain number of seconds
- let's have a look at robots.txt of our running examples

Query sparsely

- accessing data causes server traffic
- always download the target HTML site first, then parse and extract
- when downloading several pages, use pauses in between downloading HTML files
- use `Sys.sleep()` in R

Legal Consideration

Disclaimer

- neither the content of this seminar nor what I say constitute legal advice.
- if you are uncertain about the legality of your web data collection project, consult a legal expert.

But is this stuff legal?

- web data collection is not per se illegal
- local laws, copyrights, ToS, bandwith usage, your technical approach, etc. matter
- often no clear criteria about what is allowed or not
- most prominent legal cases involve commercial interests

Legal Consideration

Disclaimer

- neither the content of this seminar nor what I say constitute legal advice.
- if you are uncertain about the legality of your web data collection project, consult a legal expert.

But is this stuff legal?

- web data collection is not per se illegal
- local laws, copyrights, ToS, bandwith usage, your technical approach, etc. matter
- often no clear criteria about what is allowed or not
- most prominent legal cases involve commercial interests

General Advice

Violations and gray areas

- is it okay to use questionable means?
- ultimately, that is your decision to make and you are responsible for your actions!

Best practices

- look for official ways to collect data
- obey robots.txt and API rate limiting
- stay identifiable
- don't overburden servers
- document the sources of data at all times
- heed copyrights
- do not repurpose content commercially

General Advice

Violations and gray areas

- is it okay to use questionable means?
- ultimately, that is your decision to make and you are responsible for your actions!

Best practices

- look for official ways to collect data
- obey robots.txt and API rate limiting
- stay identifiable
- don't overburden servers
- document the sources of data at all times
- heed copyrights
- do not repurpose content commercially