

Mining Software Repositories While Respecting Privacy

Jesus M. Gonzalez-Barahona

Universidad Rey Juan Carlos

@jgbarah

<https://jgbarah.github.io/presentations>

2nd International Conference on Applied Technologies
Latacunga (Ecuador), December 2nd 2020
Materials from: Educational Track MSR June 29th 2020

The plan

- 1 Should we worry?
- 2 Analysis: EU-funded research
- 3 Definitions (GDPR)
- 4 Some guidelines
- 5 Case: git data
- 6 Case: Anonymizing public datasets
- 7 Important details
- 8 Call for action
- 9 To probe further

Disclaimer

I'm not a lawyer.

I don't have any
formal training in law.

This is not legal advice.



MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

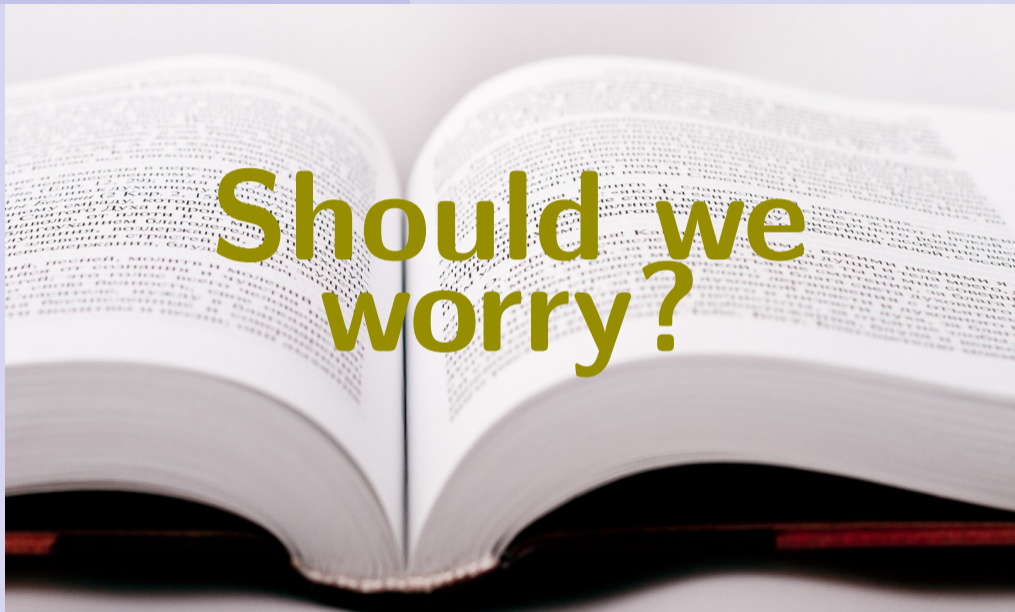
Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action



Should we worry?

Mining software repositories

*Should we worry
(about data privacy)
when we research
based on data
extracted from
software development repositories?*

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Let's analyze from two (interrelated)
points of view:

- legal requirements
- ethical requirements

Ethics

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

- Ethics recommendations in our field
- Ethics rules by our funders
- Ethics committees in our institutions

Ethics: ICT research principles

- stakeholders perspectives
- respect for persons
- beneficence
- justice: fairness and equity
- respect for law and public interest

“The Menlo Report: Ethical Principles Guiding Information and
Communication Technology Research”,
Dittrich & Kenneally (2012)

Ethics: SE research core concepts

- informed consent
- scientific value
- beneficence
- confidentiality

“Ethical Issues in Empirical Studies of Software Engineering”
Singer & Vinson (2002)

Ethics: MSR

Software repositories always contain personal information or identifiers that can be mapped to individuals. Given that repositories are usually publicly available, even supposedly anonymised datasets usually contain sufficient information to allow mapping of the anonymised data to individual developers. Therefore, one usually has to assume that research using an MSR dataset can affect human subjects, requiring careful consideration of ethics implications. It is an often occurring misunderstanding that analysis of publicly available data is free from the requirement of ethics consideration. Particular problems in MSR research are the considerations for informed consent, risks to the subjects, and compliance.

“Ethical Mining: A Case Study on MSR Mining Challenges”,
Gold & Krinke, MSR (2020)

Reminder:
This presentation
will focus on
privacy &
data protection



Applicable law

In the European Union (and elsewhere):

- **General Data Protection Regulation** (GDPR)
Affects all of EU, and rights of EU citizens
- Specific law in member states
- Recommendations by national data protection agencies

Becoming influential in other countries

Applicable law

Similar law in other jurisdictions:

- California Consumer Privacy Act
- Lei Geral de Proteção de Dados (LGPD) (Brazil)
- Act on Protection of Personal Information (Japan)
- Personal Information Protection Act (South Korea)

(Plus international agreements, eg EU-Japan)

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Analysis: EU-funded research

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

What should we have into account,
if we get funded
by the European Commission?

In the process, we'll review one
of the more complete set of requirements
on data privacy

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

While the EU's ethics review process is primarily concerned with ethics issues, your project must demonstrate compliance with the GDPR. However, the fact that your research is legally permissible does not necessarily mean that it will be deemed *ethical*.

Crucially, if your research proposal involves the processing of any personal data, whatever method is used, you – and all of your partners, collaborators and service providers – must, if called upon, be able to demonstrate compliance with both legal and ethical requirements. Such requests could come from data subjects, funding agencies or data protection supervisory authorities.

H2020 document on Ethics and Data Protection, by EC

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Research aspects

Processing

Publication

Processing

Processing of personal data should be:

- lawful,
- fair,
- and transparent.

GDPR, Preface (39)

Lawful?

- explicit consent: difficult when mining repositories
- task in the public interest (public institutions)
- legitimate interest (non-public institutions)

GDPR and Research: An Overview for
Researchers, UK Research and Innovation

Fair, transparent?

Fair, transparent:

Processing for [...] research purposes, shall be subject to appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject.

GDPR, Article 89

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
researchDefinitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Those safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation. Those measures may include pseudonymisation [...]. Where those purposes can be fulfilled by further processing which does not permit [...] the identification of data subjects, those purposes shall be fulfilled in that manner.

GDPR, Article 89

Publication

Publication of personal data:

Sharing personal data should be through managed processes, with access and usage controls to protect from re-identification

Potential problem for publication of results, when they involve personal data (eg: datasets)

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
researchDefinitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

It is highly likely that if your project involves any data about identifiable persons, even if they are not directly participating in the research, you are processing personal data and must comply with EU and national law. Only data that have been fully and irreversibly anonymised are exempt from these requirements. Importantly, while **pseudonymisation** can provide individual data subjects with a degree of protection and anonymity, pseudonymised data still fall within the scope of personal data because it is possible to re-identify the data subject (see below).

Even if your project is using only **anonymised data**, the origin or acquisition of the data may still raise significant ethics issues.

H2020 document on Ethics and Data Protection, by EC

MSR & Privacy

“Higher risks” related to GDPR in research

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Types of personal data	<ul style="list-style-type: none"> * racial or ethnic origin * political opinions, religious or philosophical beliefs * genetic, biometric or health data * sex life or sexual orientation * trade union membership
Data subjects	<ul style="list-style-type: none"> * children * vulnerable people * people who have not given their explicit consent to participate in the project
Scale or complexity of data processing	<ul style="list-style-type: none"> * large-scale processing of personal data * systematic monitoring of a publicly accessible area on a large scale * involvement of multiple datasets and/or service providers, or the combination and analysis of different datasets (i.e. big data)
Data-collection or processing techniques	<ul style="list-style-type: none"> * privacy-invasive methods or technologies (e.g. the covert observation, surveillance, tracking or deception of individuals) * using camera systems to monitor behaviour or record sensitive information * data mining (including data collected from social media networks), ‘web crawling’ or social network analysis * profiling individuals or groups (particularly behavioural or psychological profiling) * using artificial intelligence to analyse personal data * using automated decision-making that has a significant impact on the data subject(s)
Involvement of non-EU countries	<ul style="list-style-type: none"> * transfer of personal data to non-EU countries * collection of personal data outside the EU

H2020 document on Ethics and Data Protection, by EC

MSR & Privacy

Jesus M. Gonzalez-Barahona

Should we worry?

Analysis: EU-funded research

Definitions (GDPR)

Some guidelines

Case: git data

Case: Anonymizing public datasets

Important details

Call for action

Types of personal data	<ul style="list-style-type: none"> * racial or ethnic origin * political opinions, religious or philosophical beliefs * genetic, biometric or health data * sex life or sexual orientation * trade union membership
Data subjects	<ul style="list-style-type: none"> * children * vulnerable people * people who have not given their explicit consent to participate in the project
Scale or complexity of data processing	<ul style="list-style-type: none"> * large-scale processing of personal data * systematic monitoring of a publicly accessible area on a large scale * involvement of multiple datasets and/or service providers, or the combination and analysis of different datasets (i.e. big data)
Data-collection or processing techniques	<ul style="list-style-type: none"> * privacy-invasive methods or technologies (e.g. the covert observation, surveillance, tracking or deception of individuals) * using camera systems to monitor behaviour or record sensitive information * data mining (including data collected from social media networks), 'web crawling' or social network analysis * profiling individuals or groups (particularly behavioural or psychological profiling) * using artificial intelligence to analyse personal data * using automated decision-making that has a significant impact on the data subject(s)
Involvement of non-EU countries	<ul style="list-style-type: none"> * transfer of personal data to non-EU countries * collection of personal data outside the EU

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Data subjects

* children

* vulnerable people

* people who have not given their explicit consent to participate in the project

Children & vulnerable people

Maybe we don't know...

...but they may be in our dataset

- they are subject to special protection
- even when we usually cannot tell who they are...
- ...others could

This situation is very difficult to deal with

No explicit consent

- We collect data from services...
- ...which didn't get explicit consent for our cases
- Even if they got, they don't guarantee that for us
- In summary: usually, no explicit consent

Can we avoid this case?

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

**Scale or complexity
of data processing**

large-scale processing of personal data

* systematic monitoring of a publicly accessible area on a large scale

* involvement of multiple datasets and/or service providers, or the combination and analysis of different datasets (i.e. big data)

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Large-scale

- Usually, several data sources
- The more data, the better
- If we can combine datasets, we do

The better the research, the riskier

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
researchDefinitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Data-collection or processing techniques	<ul style="list-style-type: none"> privacy-invasive methods or technologies (e.g. the covert observation, surveillance, tracking or deception of individuals) * using camera systems to monitor behaviour or record sensitive information * data mining (including data collected from social media networks, web crawling' or social network analysis) profiling individuals or groups (particularly behavioural or psychological profiling) * using artificial intelligence to analyse personal data * using automated decision-making that has a significant impact on the data subject(s)
---	---

Privacy invasive methods

- Example study: who is working off-hours
- Methodology: tracking individual activity in all available data sources
- Risk: tagging specific people

You can learn working hours, days off, vacation...

Data mining from social media

- Our data source are social media
- Of course we mine data from them

Data mining is the core of our business

Profiling individuals or groups

- Example: activities by newcomers
- Methodology: tracking individual activity in all available data sources
- Risk: tagging specific people

You can show specific activity of persons

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Using AI to analyze personal data

- Example: find out experts
- Methodology: analyze activity to find out experts in some languages, using AI
- Risk: singling out specific persons

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Involvement
of
non-EU countries

transfer of personal data to non-EU countries

collection of personal data outside the EU

Transfer of data across EU border

- Collecting data from non EU data sources
- Sharing data with non-EU researchers

Can we avoid these scenarios?

MSR & Privacy

Jesus M. Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded research

Definitions (GDPR)

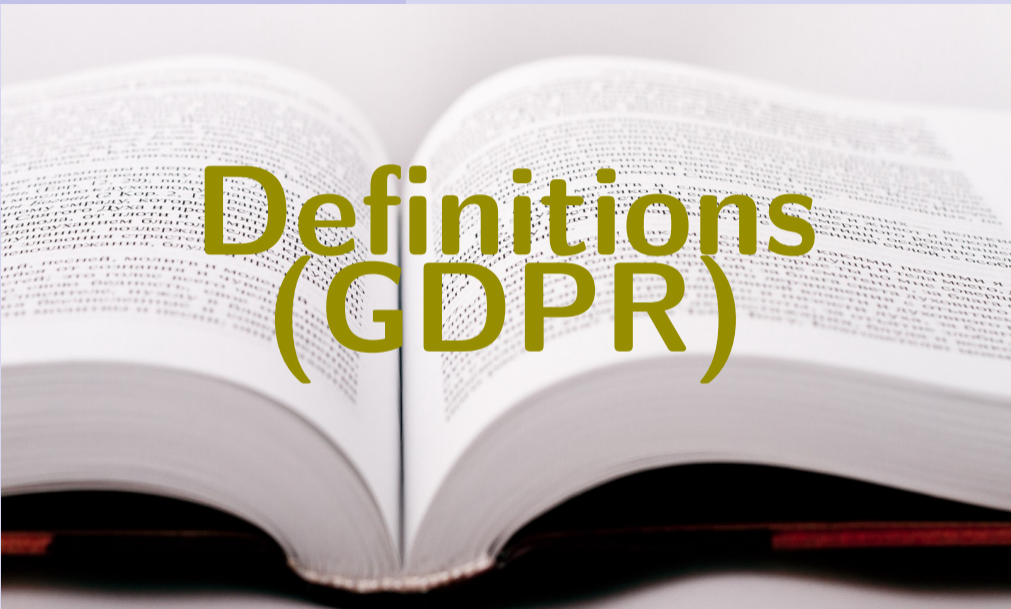
Some guidelines

Case: git data

Case: Anonymizing public datasets

Important details

Call for action



Definitions (GDPR)

Data processing

(2) Data processing [includes] any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.

GDPR, Article 4.

Actors subject to GDPR

- Data controller: determines the purposes and means of the processing of personal data
- Data processor: processes personal data on behalf of the controller

Data protection office (DPO)

- Appointed by controllers and processors (eg: usually one in each University)
- Involved in all issues related to protection of personal data
- Data subjects may contact DPO directly
- Good, local point to seek advice for a researcher

DPIA

Data Protection Impact Assessment:

Process designed to assess the data-protection impacts [...] and, [...] to ensure that remedial actions are taken as necessary to correct, avoid or minimise the potential negative impacts on the data subjects.

DPIA likely required

Should we worry?

Analysis:
EU-funded
researchDefinitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

A company systematically monitoring its employees' activities, including the monitoring of the employees' work station, internet activity, <i>etc.</i>	<ul style="list-style-type: none"> - Systematic monitoring - Data concerning vulnerable data subjects
The gathering of public social media data for generating profiles.	<ul style="list-style-type: none"> - Evaluation or scoring - Data processed on a large scale - Matching or combining of datasets - Sensitive data or data of a highly personal nature

H2020 document on Ethics and Data Protection,
by European Commission

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action



Some guidelines

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Guidelines based on the
European Commission case,
but useful even when they
are not mandatory

Detailed analysis

Ethics issues raised by our methodology:

- data collection and processing operations
- ethics issues that these raise
- mitigation of these issues in practice.

Submission to the Research Ethics Committee.

Detailed analysis

- Mandatory for EC-funded research proposals
- Important: involve the DPO
(Data Protection Officer)
- Maybe: requirement to conduct a DPIA
(Data Protection Impact Assessment)

GDPR approach

Data protection by design (DPbD):

Data controllers are required to implement appropriate technical and organisational measures to give effect to the core data-protection principles of GDPR.

GDPR, Articles 5 and 25

DPbD in research

Data protection by design:

- Anonymization / pseudonymization
- data minimization
- cryptography (hashing, encrypting)
- data protection focused service providers & storage
- procedures for exercising fundamental rights (access, consent)

Data minimization

(1) Data processing must be lawful, fair and transparent. It should involve only data that are necessary and proportionate to achieve the specific task or purpose for which they were collected

GDPR, Article 5

Data minimization

- Collect minimal personal data
- Anonymize and pseudonymize
- Store data securely
- Dispose data when no longer needed
- Limit access to data

Mitigation

Anonymization, pseudonymization

...but even in this case, ethics issues:

- origin of the data
- potential misuse of methodology or findings
- potential for deanonymization

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action



Case: git data

The problem

You have a collection of
all commits fulfilling some properties
from a large set of public repositories.

You analyze number of committers per time
period.

How can you publish the dataset
in a reproduction package?

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

The problem

Let's assume we have
a lawful basis
for the data processing

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

```
commit 491f9205c36fdf54b4bbb7f25ba83b6cb99874b9
Author: Jesus M. Gonzalez-Barahona <jgb@gsyc.es>
Date:   Mon May 13 13:33:04 2019 +0200
```

Add notice about GNOME Extension needed
for indications.

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

```
commit 491f9205c36fdf54b4bbb7f25ba83b6cb99874b9
Author: Jesus M. Gonzalez-Barahona <jgb@gsync.es>
Date:   Mon May 13 13:33:04 2019 +0200
```

Add notice about GNOME Extension needed
for indications.

Personal data?

Personal data

(1) 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;

GDPR, Art. 4

Personal data

Personal data include data such as internet protocol (IP) addresses (unique identifiers that can be used to identify the owner of devices connected to the internet) and data from 'smart meters' monitoring energy usage by addresses linked to identifiable persons.

H2020 document on Ethics and Data Protection,
by European Commission

Certainly, it includes names & email identifiers.

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Personal data

Jesus M. Gonzalez-Barahona <jgb@gsyc.es>

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Personal data

But... wait!!!!

Our data is public data!!

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Open source data

[Box 4] Using 'open source' data

The fact that some data are publicly available does not mean that there are no limits to their use.

On the contrary, **if you take 'open source' personal data about identifiable persons and create new records or files/profiles, you are processing personal data about them** and must have a lawful/legitimate basis for doing so.

You must ensure that the data processing is fair to the data subject and that their fundamental rights are respected.

H2020 document on **Ethics and Data Protection**,
by European Commission

Open source data

If your research project uses **data from social media networks** and you do not intend to seek the data subjects' explicit consent to the use of their data, you must assess whether those persons actually intended to make their information public (e.g. in the light of the privacy settings or limited audience to which the data were made available).

It is not enough that the data be accessible; they must have been made public to the extent that the data subjects do not have any **reasonable expectation of privacy**. **You must also ensure that your intended use of the data complies with any terms and conditions published by the data controller.**

If you are in any doubt as to what you can and cannot do with this kind of data, you should seek advice from your DPO or a suitably qualified expert and include their opinion in your proposal.

H2020 document on **Ethics and Data Protection**,
by European Commission

How to fix the problem

- Anonymize: Strip all personal data (but still... more on this later)
- Pseudonymize personal data (the dataset will be much richer, but still...)

Pseudonymizing

(5) 'pseudonymisation' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;

GDPR, Art. 4

Pseudonymization

[Box 2] Pseudonymisation and anonymisation: understanding the difference

Pseudonymisation entails substituting personally identifiable information (such as an individual's name) with a unique identifier that is not connected to their real-world identity, using techniques such as coding or hashing. However, if it is possible to re-identify the individual data subjects by reversing the pseudonymisation process, data protection obligations still apply. They cease to apply only when the data are fully and irreversibly anonymised.

Anonymisation involves techniques that can be used to convert personal data into anonymised data. Anonymisation is increasingly challenging because of the potential for re-identification.

Re-identification is the process of turning pseudonymised or anonymised data back into personal data by means of data matching or similar techniques.

H2020 document on Ethics and Data Protection,
by EC

Pseudonymizing 1

```
echo -n "<jgb@gsync.es>" | sha256sum  
3fdffb4a435cc3a5bab7d96b3cc2cefea90ca879f7fba0341
```

```
commit 491f9205c36fdf54b4bbb7f25ba83b6cb99874b9  
Author: 3fdffb4a435cc3a5bab7d96b3cc2cefea90ca879f7fba0341  
Date: Mon May 13 13:33:04 2019 +0200
```

```
Add notice about GNOME Extension needed...
```

Pseudonymizing 1

Not good enough:

Data returned for an Email Append

Input (Encrypted Email)	Recovered Email	First Name	Last Name	Address	City	State	Zip	Phone
cbf05329de4e57e4cba09471448ddb98	joe.smith@gmail.com							
5469703a9c26d5e8be4e46bef4596e2f836088c0	commonme@yahoo.com	Don	Johnson	478 19TH PL W	Redmond	WA	98052	4255557892

In addition to reversing hashed email addresses, Datafinder also provides personal information including name, address and phone number associated with an email address.

Four cents to deanonymize: Companies reverse hashed
email addresses, by Gunes Acar

Pseudonymizing 1

For the curious:

- Estimated: 5 billion email addresses (2018)
- Amazon EC2: 450 billion hashes/sec.
- Lists of (targeted) email lists for sale
- Data breaches leaking billions of addresses

There is a whole business ecosystems around email addresses

Four cents to deanonymize: Companies reverse hashed

Pseudonymizing 2

- Hash “name + address”

Better, but still subject to attack if you harvested addresses

- Salt the hash, use different algorithm
- Non-hash functions (eg, sequential code)
- Encryption instead of hash

Better, but you need to disclose details if you want others to merge with your dataset

Pseudonymizing 3

A possibility emerges: Encryption / coding

- Datasets: public
- Key / coding table:
only to researchers asking for it

(variant: reference them in the paper)

Is this good enough for “legitimate use”?

Pseudonymizing 3

```
commit 491f9205c36fdf54b4bbb7f25ba83b6cb99874b9
```

```
Author: 4334345
```

```
Date: Mon May 13 13:33:04 2019 +0200
```

```
Add notice about GNOME Extension needed...
```

Separate table:

```
4334345, Jesus M. Gonzalez-Barahona <jgb@gsync.es>
```


MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

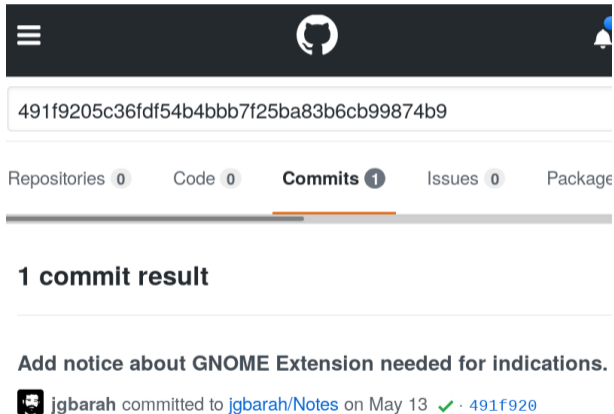
Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

We still have a problem



The screenshot shows the GitHub interface for a repository. At the top, there is a navigation bar with a hamburger menu, the GitHub logo, and a notification bell. Below this is a search bar containing the commit hash `491f9205c36fdf54b4bbb7f25ba83b6cb99874b9`. Underneath the search bar, there are statistics for the repository: Repositories (0), Code (0), Commits (1), Issues (0), and Package (0). The 'Commits' section is highlighted with an orange bar. Below the statistics, the heading '1 commit result' is displayed. The commit details show a message: 'Add notice about GNOME Extension needed for indications.' The commit was made by user 'jgbarah' to the repository 'jgbarah/Notes' on May 13, and it is marked as successful with a green checkmark and the commit hash `491f920`.

We still have a problem

```
curl https://archive.softwareheritage.org/api/1/1/491f9205c36fdf54b4bbb7f25ba83b6cb99874b9/
```

```
{"author":  
  {"name": "Jesus M. Gonzalez-Barahona",  
    "fullname": "Jesus M. Gonzalez-Barahona <jgb@g",  
    "email": "jgb@gsyc.es"}  
}  
...  
}
```

What can we do?

Pseudonymize the hash too:

```
commit 6777888876876
```

```
Author: 4334345
```

```
Date: Mon May 13 13:33:04 2019 +0200
```

Add notice about GNOME Extension needed...

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

We still have a problem (2)

The screenshot shows a GitHub repository interface. At the top, there is a navigation bar with a hamburger menu, the GitHub logo, and a notification bell. Below this is a search bar containing the text "Add notice about GNOME Extension needed for indications.". Under the search bar, there are statistics for the repository: "Repositories 0", "Code 4K+", "Commits 10+", "Issues 2", and "Packages 0". A horizontal progress bar is shown below these statistics, with the "Commits" section highlighted in orange. The main content area displays "Showing 10 available commit results" with a help icon. Below this, the same commit title "Add notice about GNOME Extension needed for indications." is shown. The commit was made by user "jgbarah" to the repository "jgbarah/Notes" on May 13, with a green checkmark and the commit hash "491f920".

We still have a problem (2)

- Commit comment can be used to deanonymize author.
- Date can be used to deanonymize author.

```
commit 6777888876876
```

```
Author: 4334345
```

```
Date: 5353453453
```

```
Comment: 4343434334
```

...and separate coding tables

Why we need tables

- For reproduction: commits per time period
- for reuse: identity merging
- for reuse: relationship between message and time of the day
- for reuse: link to issues
- ...

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

When data is public...

This is a problem for anyone
having access to the data
that allows deanonymization

But also for anyone letting others deanonymize

Big trouble if it is publicly available data

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

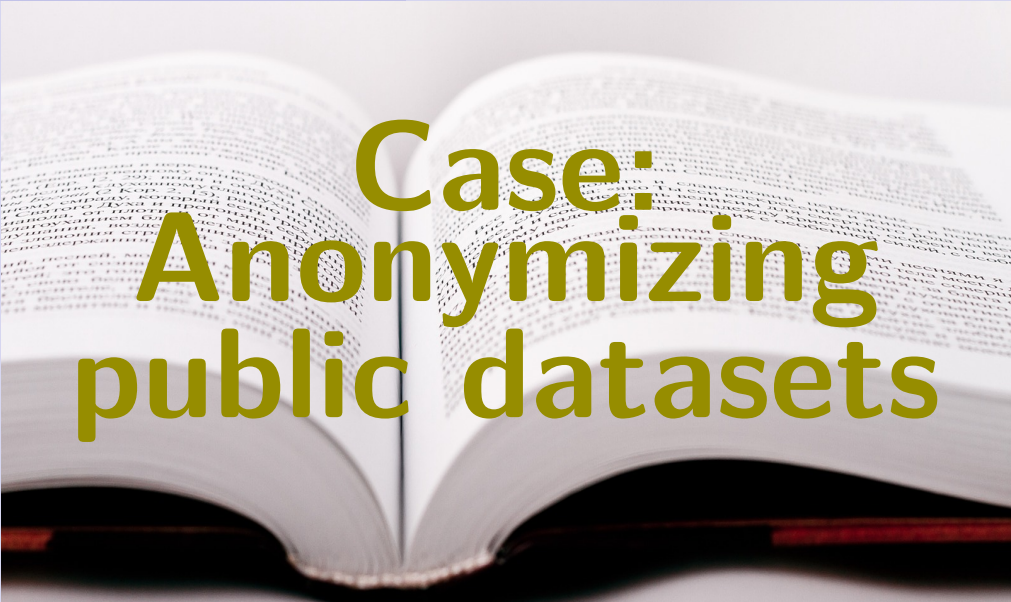
Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action



Case: Anonymizing public datasets

Sharing a dataset

- Use Perceval to collect whatever the data source offers
- Analyze fields in each JSON document with strings that can be used to identify persons
- Extra step in the pipeline: anonymize those strings
- Store the resulting JSON documents

Is this good enough?

How to anonymize

- Hashing is not good enough (easy to break if you have collections of identities)
- Alternatives: coding, hashing with salt, HMAC

Each option has its own characteristics

Attack models

- **Single target:**
Given an identity, find all items linked to it
- **Trawl fishing:**
Given a large set of identities, find items linked to any of them in the dataset

Attack models

- **Public exploitation:**
Find a public, un-anonimized item, compare with same item in dataset, de-anonymize all anonymous identities linked to it

Public exploitation

- If you can find an item in the original data source, you can deanonymize identities in it
- Avoidance: different anonymizations for each item.
- For example: $HMAC(id, item_id)$

Problem: now you cannot group items per person

Re-identification by authorized parties

Something to allow:

- A FOSS Foundation may have consent from its developers
- A company may have consent from its employees (eg, incentives program)
- An organization may get consent on a one-by-one basis

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Re-identification by authorized parties

Can we use the anonymized data
re-identifying it?

Yes, we can (but there are issues)

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
researchDefinitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Re-identification by authorized parties (A)

- Anonymize by hashing, never published
- Making dataset public: $HMAC(hash)$
- Deliver the HMAC key to trusted parties
- Trusted parties hash identities, $HMAC(hash)$, re-assign identities

Re-identification by authorized parties (A)

- If you don't have the key, you cannot HMAC
- If you don't have the identity, you cannot hash

Problems:

- *Store* stores hashes (far from ideal)
- If you're not that trusted,
you can try deanonymizing all identities

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Re-identification by authorized parties (B)

- Anonymize with HMAC, *KeyStore*, make dataset public
- Get identities from trusted parties, HMAC with *KeyStore*, deanonymize in dataset and send securely to authorized party

Re-identification by authorized parties (B)

Problems:

- You need to trust *Store*
- Deanonimized dataset could be compromised (HMAC it with Key_{Party})

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Re-identification by authorized parties (C)

Could we have the best of (A) and (B)?

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Important details



Timing of anonymization

- Collection (no personal data is processed)
Example: web form, no browser tracking
Example: anonymous dataset from 3rd party
- Later than collection:
Raw data is not anonymized
(needs special protection)

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Informed consent

Explain to participants what your research is about, what their participation in your project will entail and any risks that may be involved. When they have fully understood, see and obtain their express permission.

GDPR, Article 4 (11), Article 7

Managing consent

You need to document and archive consent

You need to be able of producing evidence

Consent management applications:

- ethically robust, secure
- model consent processes
- manage, document, evidence

Secondary use

Usually, we perform secondary use
(we don't collect data directly from persons)

Important: check original consent

Explain how data was obtained,
justify use (legitimate use)
ensure processing is fair to data subjects

Secondary use

(47) the legitimate interests of a controller [...] or of a third party, may provide a legal basis for processing, provided that the interests or the fundamental rights and freedoms of the data subject are not overriding, taking into consideration the reasonable expectations of data subjects based on their relationship with the controller

Data security

- Appropriate technical & organisational measures
- Level of security commensurate to risks faced by the data subjects
- Prevention of unauthorized access, disclosure, accidental deletion

Data security

Do

- ✓ use GDPR-compliant tools to collect, process and store research subjects' personal data;
- ✓ take communications security seriously, and devise and implement dedicated protocols for your project as necessary;
- ✓ check the terms and conditions of all of the service providers you use (software, applications, storage, etc.) to process personal data within your project, in order to identify and mitigate risks to the data subjects;
- ✓ encrypt your research data and/or the devices on which they are stored, and ensure that keys/passwords are appropriately protected; and
- ✓ consult your DPO or a suitably qualified expert for advice on how to achieve a level of data security that is commensurate to the risks to your data subjects.

H2020 document on Ethics and Data Protection,
by EC

Data security

- ✘ collect data on a personal device such as a smartphone without ensuring that they are properly protected (e.g. consider the implications of automatic back-ups to the cloud, and the device's security features);
- ✘ use free services that may use your participants' data for their own purposes in lieu of payment, or collect data or communicate with research participants via social media platforms without first assessing the data protection implications;
- ✘ use unencrypted email, SMS or insecure 'voice over IP' platforms to communicate with vulnerable participants or those who may be subject to state surveillance;
- ✘ expose personal data to unauthorised access or use when accessing them remotely (e.g. by using insecure wifi connections) or travelling to countries where your devices may be inspected or seized; and
- ✘ assume that your research partners, collaborators or service providers have appropriate information security and data protection policies without checking that this is the case.

H2020 document on Ethics and Data Protection,
hv FC.

Transfer outside EU

Possible (GPDR, Chapter 5)::

- to countries with “adequacy determination”
- when explicit consent is obtained
- when compliant data-transfer agreements are in place

Beware of third-party services!!!

Non-EU countries with adequacy determination

Transfer outside EU

- ✓ ensure that any international data transfers fulfil at least one of the relevant conditions in Chapter V GDPR;
- ✓ check that any third-party services you intend to use (e.g. survey tools, data analytics, cloud storage, etc.) are incorporated in an EU Member State or legally represented in the EU in accordance with the GDPR;
- ✓ adopt legally binding and enforceable agreements with partners or service providers prior to data transfers;
- ✓ prohibit the onward transfer of personal data by members of your consortium and any other recipients outside the framework of such agreements; and
- ✓ implement appropriate organisational and technical measures to ensure that personal data are transferred securely.

H2020 document on Ethics and Data Protection,
by EC

Archiving

Archive personal data
only as needed for research purposes.

Delete it as soon as possible
(define a maximum retention period).

Beware of backups and data in cloud storage.

H2020 document on Ethics and Data Protection,
by EC

Archiving

But you can keep personal data indefinitely if you ensure it is only for:

- archiving purposes in the public interest
- scientific or historical research purposes
- statistical purposes

Archiving

Open issue: archiving for reproduction

In principle, covered by research exemption,
but it is not absolute

Collection outside EU

Subject to GDPR

even if data comes from outside EU

if data controller is based in EU

Of course, compliance with
the law of the country of collection

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

Children

If your research project involves collecting data from children, you must follow the “EC Guidance note on informed consent”, in particular the provisions on obtaining the consent of a parent/legal representative and, where appropriate, the assent of the child.

H2020 document on Ethics and Data Protection,
European Commission

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

The GDPR establishes special safeguards for children in relation to “information society services”, a broad term covering all internet service providers, including social media platforms. These include a requirement for verified parental consent in respect of information society services offered directly to children aged under 16. Individual Member States may provide for this threshold to be lowered to 13.

H2020 document on Ethics and Data Protection,
European Commission

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action



Call for action

Researchers & developers

Maybe we researchers need to work
with developers
to learn what is legitimate use for them?

Maybe developers should specify
what is legitimate use for them,
in their repositories?

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

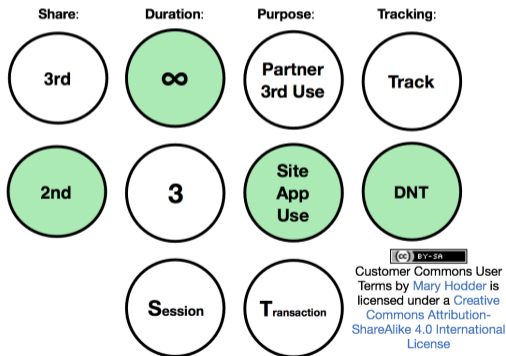
Case: Anonymizing
public datasets

Important details

Call for action

An inspiration

MY TERMS: Icon format and structure



NOTE: I'm the first party. My terms are: 2nd-∞-SU-DNT

User Submitted Terms, by
Mary Hooder
Customer Commons and
User Submitted Terms

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action

A related IEEE WG

Working Group Details

Working Group MRPT_WG - Machine Readable Privacy Terms Working Group

Sponsor Committee SSIT/SC - Social Implications of Technology Standards Committee

Society **SSIT - IEEE Society on Social Implications of Technology**

Machine Readable Privacy Terms Working Group

FOSS is open for a reason

Having the source code available
is a conscious decision

It could be extended to all data
related to software development
(better understanding of the project)

Clarification of intent would help to define
legitimate interest and prove ethics compliance

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

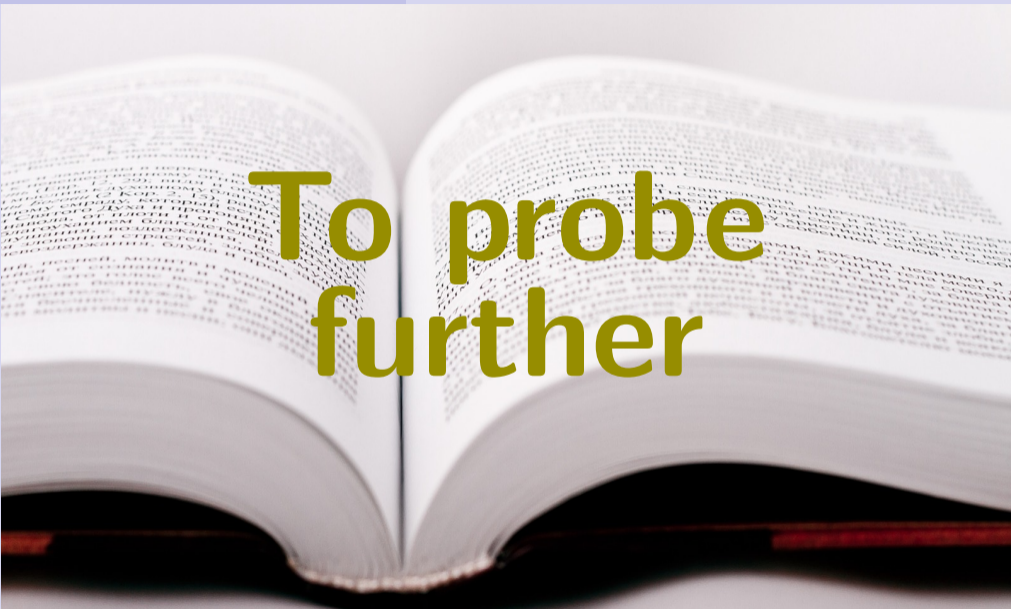
Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action



To probe further

References

- **GDPR portal** by EC
(includes full text of GDPR)
- **H2020 document on Ethics and Data Protection**, European Commission
- **GDPR and Research: An Overview for Researchers**, UK Research and Innovation

Credits



Book, by NikolayFrolochkin, Pixabay.

License: Creative Commons CC0



Caution, Warning. Precaucion, Alerta, by Ehecatl1138, OpenclipArt.

License: Creative Commons CC0

MSR & Privacy

Jesus M.
Gonzalez-Barahona

Should we worry?

Analysis:
EU-funded
research

Definitions
(GDPR)

Some guidelines

Case: git data

Case: Anonymizing
public datasets

Important details

Call for action



©2019-2020 Jesus M. Gonzalez-Barahona.

Some rights reserved. This document is distributed under the terms
of the Creative Commons License “Attribution-ShareAlike 4.0”,
available in

<http://creativecommons.org/licenses/by-sa/4.0/>

This document (including source) is available from
<https://jgbarah.github.io/presentations>