GDPR compliant guidelines for processing personal data in legal documents

Noora Arajärvi and Livia Holden (with the assistance of Anna Ziliotto and Joshua Bishay)

1. Aim and scope of the guidelines

CULTEXP, the main output created by EURO-EXPERT and the first multilingual and cross-jurisdictional database on cultural expertise, contains judgments and expert reports that have been collected from the following sources: OPEN ACCESS databases, law court archives, and experts. The law on the processing of personal data in the European Union is set out in the General Data Protection Regulation 2016/679 (the GDPR). At the time of writing these guidelines, interpretations, and practices regarding the processing of personal data in legal documents vary greatly. At EURO-EXPERT we have developed specific GDPR compliant guidelines for the treatment of personal data contained in the documents archived on CULTEXP. In developing these guidelines, we have referred to the GDPR, to other regulations and case law, and to the expert opinion of Prof. Jougleux (European University Cyprus) who was appointed by EURO-EXPERT to address the purpose of CULTEXP and the guiding principles of personal data processing in legal documents.

2. What is anonymisation, and why anonymise?

The concept of personal data under the GDPR[1] is very broad. Personal data is not only "direct personal data" that allows immediate identification of the data subject (such as the person's name, or "Prime Minister of France") but also "indirect personal data" (e.g. 45-year old Greek female professor living in Cambridge). The concept of indirect personal data embraces any data that could lead to identification using reasonable means.

Anonymisation is a technique of data processing that can be applied to personal data to achieve irreversible de-identification. One of the advantages of anonymisation is that it allows research that would not otherwise be possible due to privacy concerns. This means removing direct identifiers such as names, birth dates and addresses,

[1] Complete Guide to GDPR Compliance: https://gdpr.eu/.

although this alone may not be sufficient to avoid identification of individuals. Therefore, each case must be considered in its entirety to ensure that no individuals are identifiable.

Effective anonymisation prevents singling out an individual in a dataset by linking several records within the same dataset (or between several separate datasets) or inferring any information from such dataset(s). The aim of anonymisation is to ensure that the data no longer relates to identifiable persons, which means it is no longer considered "personal data". Therefore, properly anonymised data falls outside the scope of data protection rules. Once a dataset is truly anonymised and individuals are no longer identifiable, the GDPR no longer applies. In 2018, the Court of Justice of the European Union introduced measures it was adopting to comply with the GDPR in the publication of its decisions.[2] National legislation regarding privacy and data management is also relevant in this regard.

Pseudonymisation consists of replacing one attribute with another. It is defined in the GDPR as "*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.*"[3] Most EU legal systems use a two-step technique to achieve the most effective outcome: First, the names of the parties to the litigation are changed to their initials. Then, personal identifiers such as ID numbers or addresses are replaced with "x" characters. Pseudonymisation, however, may require further redaction to comply with the GDPR. Even when the name of a party is completely replaced with initials, if the decision describes a situation that is unique or that is very well known, the decision still contains personal data. Similarly, if, by using techniques of data mining, users could find cross-references leading to specific identifiers of persons involved, the data may still be within the scope of the GDPR and national legislation. It is important to note that member states have implemented and applied the GDPR somewhat differently across the EU. Hence, an *ad hoc* assessment of the balance between the risk of disclosing personal data and a legitimate purpose of personal data processing may be required in particular cases.

Generalisation is another anonymisation technique. This approach consists of generalising or diluting the attributes of data subjects by modifying the relevant scale or order of magnitude (e.g. using a region rather than a city, a month rather than a week, etc.). Whilst generalisation can be effective to prevent singling out of individuals, and may be useful as a supplemental tool, it does not allow for effective anonymisation in all cases.

Processing personal data for research purposes occupies a privileged position within the GDPR.[4]
While the re-publication of case law, in the case of CULTEXP, falls within the 'legitimate interest' of facilitating access to justice on topics related to cultural expertise,[5] the EURO-EXPERT data collectors are nonetheless instructed to

[2] Court of Justice of the European Union, 'From 1 July 2018, requests for preliminary rulings involving natural persons will be anonymised', Press release No 96/18, Luxembourg, 29 June 2018, https://curia.europa.eu/jcms/upload/docs/application/pdf/2018-06/cp180096en.pdf.
[3] Article 4 (5) GDPR.
[4] Art. 89 GDPR on the safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes. Also, G Maldoff, 'How GDPR changes the rules for research': https://iapp.org/news/a/how-gdpr-changes-the-rules-for-research.
[5] With precise, specific, and detailed legitimate interest, in the sense described in the Working Party Opinion 06/2014, "Notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC", accessible at

remove all personal data from their data sets. EURO-EXPERT has adopted the following combination of anonymisation and pseudonymisation, which is explained in further detail below:

1) The names of the parties are reduced to initials or removed.
2) All personal identifiers such as ID numbers, addresses, social security numbers and so on are removed or redacted.

CULTEXP will add a third step of monitoring to verify that no direct or indirect personal data have been missed by previous pseudonymisation/ anonymisation.

3. What to anonymise?

Anonymisation techniques render data subjects unidentifiable. As a general rule, people professionally involved with a court case are not anonymised; for example judges, clerks, lawyers, bailiffs, court experts, interpreters and custodians[6], but in situations with a high risk of reprisals (e.g., terrorism and organised crime cases), they may also need to have their identities hidden[7]. The data protection authorities of the countries participating in EURO-EXPERT have different views as to what information should be included in court files that are accessible to the public.[8] At this stage, EURO-EXPERT will also **anonymise or pseudonymise the names of people professionally involved in a court case**, unless the individual provides a specific, informed and written consent to the inclusion of their name (as may be the case with some expert reports). However, a careful assessment should be made as to whether the disclosure of the name of the expert might lead to the inadvertent disclosure of the names of the parties, in which case the name of the expert should be removed. As the guidelines for the processing of personal data in the publication of judgements in legal databases evolve in future, EURO-EXPERT may relax this precautionary approach and decide not to anonymise the names of the judges, experts, lawyers, and court staff. Data collectors who are trained to anonymise and redact legal documents are advised to use their expert discretion in using anonymisation or pseudonymisation and always to err on the side of caution if in doubt.

EURO-EXPERT requires anonymisation (or pseudonymisation) of:

1. Names of **all** individuals: Defendants, plaintiffs, claimants, applicants, appellants, respondents, victims, witnesses, other parties, court personnel (judges, prosecutors, lawyers, tribunal officials), experts and consultants.
    a. NB. At this stage, the names of experts must also be redacted in expert reports (as well as in case files) UNLESS 1) the report was provided directly by the expert themselves, AND 2) they have given

https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf. See also Art. 89 GDPR.

[6] M van Opijnen *et al.*, 'On-line Publication of Court Decisions in the EU: Report of the Policy Group of the Project 'Building on the European Case Law Identifier'', 15 February 2017, https://www.bo-ecli.eu/uploads/deliverables/Deliverable%20WS0-D1.pdf, p. 24

[7] T Allard, L. Béziaud & S. Gambs, 'Online publication of court records: circumventing the privacy-transparency trade-off', arXiv:2007.01688 [cs.CR], https://arxiv.org/abs/2007.01688.

[8] E.g. The Belgian Data Protection Authority has published an opinion stating that the names of lawyers, judges and clerks must also be anonymised (2012) but according to the report of the Commission of the Modernisation of the Judiciary (2014), the Data Protection Authority changed its opinion with regard to the anonymisation of people professionally involved. In the Romanian ROLII website the names of judges, clerks and others professionally involved are also anonymised.

their specific, individual, informed and written consent for their name to appear on the report, which will be published in an open database. It is important to consider whether the disclosure of the name of the expert might lead to the identification of the applicant, in which case the name of the expert should be reduced to initials or removed.

2. Names of places: Addresses (home address and other); specific locations (villages, towns, and cities, depending on the risk that individuals might be identifiable with this information – consider generalisation); specific places (hotels, shopping centres, shops, farms, churches, accommodation, hospitals, buildings or monuments, boats, etc).

3. Dates: Birthdates (consider generalisation: a year instead of a date); other dates or years which could make a person identifiable (dates suggesting specific and recognisable events in the life of subject, such as the date of marriage); dates of other decisions, orders and notifications (asylum application, decision by an administrative authority, etc).

   a. In some cases, redacting only part of the date may be sufficient if the timeline is relevant for understanding the case. Consider whether the month/year may be retained, for instance, in cases where the age of the individual is relevant. For example:

   "It is claimed they were born 30 Jan 2021, whereas other documents say they were born 12 Dec 2021"

   There are two options; the first option provides more information:

   ➡ It is claimed they were born [Redacted] Jan 2021, whereas other documents say they were born [Redacted] Dec 2021

   OR

   ➡ It is claimed they were born [Redacted] 2021, whereas other documents say they were born [Redacted] 2021

   b. In other cases, the specific dates can be removed without losing key information. For example:

   The asylum seeker claims: "sono in Italia dal 30 agosto 2015 e vengo dal Senegal, Kolda. Ho 28 anni, non sono sposato e non ho figli. Ho lasciato il Senegal nel maggio 2013. Sono andato in Mali, poi in Niger, in Libia dove sono stato nel 2014 per un anno."

   ➡ sono in Italia dal [removed] e vengo dal Senegal, [removed]. Ho 28 anni, non sono sposato e non ho figli. Ho lasciato il Senegal nel [removed]. Sono andato in Mali, poi in Niger, in Libia dove sono stato nel [removed] per un anno.]

4. Numbers: National identification numbers, passport numbers, fiscal codes, social security numbers, bank accounts, vehicle registration numbers, identification numbers (e.g. personal identification under the

asylum procedure), police report numbers (codes of recordings, codes for intercepts), telephone numbers, postal codes, etc.

5. Signatures: In some jurisdictions case files include signatures. Note also signatures written in the margins of the documents (e.g. Italy). Many expert reports include signatures, which should be removed or redacted.

6. Brands, labels, corporate logos (including those of legal database services) if they provide information that could lead to the identification of individuals.
    a. In a case of terrorism, the brand of the shirt of the accused was redacted because the eyewitness identified the accused by the branding of their clothes.

    b. In the UK cases, many decisions were downloaded from Westlaw, and their logo has been redacted.

7. In some cases, names of private associations, institutions, and commercial enterprises, if those can be linked to an individual and lead to the identification of a natural person. This is addressed in detail in Annex 1. Consider the context, scope and level of detail: The CEO of a named small local company is identifiable whereas a salesperson working for Carrefour or Tesco in a capital city may not be.

    a. If names of companies are redacted, "nesting" corporate entities may be necessary to retain readability of the case. For example:

    "Marvel Studios is a subsidiary of Disney Corp, which purchased it from Marvel Entertainment Group."

    ➡ [Corp 1.1] is a subsidiary of [Corp 1], which purchased it from [Corp 2]

8. In some cases, the case numbers: If there is no imminent risk or access to sensitive information via the case number, this does not have to be redacted. An assessment should be made balancing the legitimate purpose of the database and the indirect risk of disclosure of personal data; the case number might be disclosed for the purpose of legal citations. In some jurisdictions, anyone can order the case from a court with a case number and could access the personal data that way. In the Italian context, the case numbers of decisions of the Commissione Territoriale reveal the social security number of the applicant and should be removed or redacted.

    a. In other Italian judgments, some cases are identified as follows:
    "Cass. Sez. 6^, 20 ottobre 1999, Bajrami" (or with the name of judge).

    ➡ "Cass. Sez. 6^, 20 ottobre 1999, B[removed]"

4. What not to anonymise?

Do not remove or redact, unless the data can lead to the identification of individuals:

1. Names of historical figures, famous persons, public figures (but if a public figure is a party to a case, they may have a right to privacy; considered also issues such as libel suits).
    a. In some cases, well-known public figures may have been indirectly involved in circumstances surrounding the case. In such instances, consider if it may be useful to not only redact their name but also their role and title. For example:

    "Emmanuel Macron was president of France when he signed the contract with John Doe."

    ➡ [EM] was [Redacted] when he signed the contract with [JD]

2. Names of criminal organizations (Cosa Nostra, Islamic State, Black AXE, etc).

3. Names of provinces or states (unless that information could lead to the identification of the person(s) – e.g. asylum seeker from a small province in Eritrea living in a specified area in north-west Portugal).

4. URLs.

5. Bibliographical references (authors, titles).

6. Public and international associations or organisations (United Nations, Red Cross, World Health Organization, etc).

7. Names and other data related to legal persons (companies, associations, etc.), if those cannot be linked to an individual and do not lead to the identification of a natural person. Note that, under the GDPR, only natural persons have personal data. This is addressed in detail in Annex 1.

8. Legal precedents that contain the names of the parties (unless these lead to identification of the individuals involved in the case that you are anonymising).

    a. Consider redacting names when the court refers to names of individuals in discussing the facts of the precedent. For example:

    "In this case of Appellant we rely on the *Parker* case. In that case Mr Parker was driving down the street when he got in an altercation with the driver of another vehicle."

    ➡ In this case of Appellant we rely on the *Parker* case. In that case [Redacted] was driving down the street when he got in an altercation with the driver of another vehicle.

5. How to anonymise?

1. Keep two folders on the Shared Drive: One for the original documents (for reference and if there is a need to check details, but to be deleted at the end) and one for the anonymised documents. Avoid keeping any original documents in locations (physical or digital) which could be accessed by others or could be at any risk of a data breach.

2. When saving the anonymised document, make sure the file name does not contain any names or other personal data. The same applies to its 'document properties' or metadata.

3. Depending on your dataset, you may consider different options for anonymisation or pseudonymisation of names of individuals:
    a. Use initials representing the first name and surname.
    b. If using initials might reveal the identity of the person, use letters or numbers but try to be consistent throughout the case, as long as that will not reveal the identity:
        i. A1, A2 etc for claimants/plaintiffs and B1, B2 etc for respondents/defendants)
        ii. For experts, use E1, E2, etc.
        iii. For witnesses, use W1, W2, etc.
        iv. For judges, use J1, J2, etc.
        v. For lawyers, counsel, attorneys and other legal representatives of parties, use L1, L2, etc.
    b. If there is still a risk of identifying individuals, use -- or xx for names of all or some of the individuals involved in the case.

4. If the document makes one passing reference to any person, you can redact, use -- or xx. These can be used when the role of that individual is irrelevant to the narrative as a whole.

5. When going through the case, you may find it useful to keep a pen and paper at hand to write down your "code", e.g. "Dr Carol Smith = E1". Make sure to properly destroy these notes afterwards.

6. In Word, under "Edit", you will find "Find" under which there is the function "Replace". In some versions of Word you will find this function under "Home" > "Editing". Use this with caution. See below for details.

7. If you work with PDF documents, using Adobe may be the only secure option. Remember to sanitise the document at the end!

8. Do not use any online PDF-Word converters for documents containing personal data or sensitive information. Adobe Pro includes a PDF-Word converter. If you have problems, contact the EURO-EXPERT team.

6. Challenges and tips

Removing direct identifiers is not enough to ensure that identification of the data subject is no longer possible; pseudonymised data is not the same as anonymised data!

Simply altering the main identifier (usually the name) does not prevent someone from identifying a data subject if quasi-identifiers (such as places, dates, age and so on) remain in the dataset. Extra steps should be taken, such as

removing and generalising attributes or deleting the original data, or at least bringing them to a highly aggregated level.

- Example: if a person is described as 'a man', the anonymity set size is three and a half billion, but if he is described as 'a middle-aged Dutchman with a beard' the anonymity set size is reduced to, say, half a million. If he is described as 'a middle-aged Dutchman with a beard who lives near Cambridge' the anonymity set size is only three or four and consequently allows for a high probability of identification.

Make sure you:

- Read and understand the case and the roles of persons involved.

- Clarify whether the data will be anonymised (the link to the data subject will be destroyed) or pseudonymised (the data could be reversible).

- Do not rely on the "find-replace" function:
    - Documents may be partly corrupt and not all references are identifiable by the search function.
    - Documents may contain inaccuracies, misprints or mistakes.
    - The same name, place, or other identifier may appear with several spellings in the same document (e.g. Mohammed, Mohamed, Muhammad, Muhammed etc).
    - In certain languages (e.g. Finnish) suffixes may alter the core of the word, rendering it undiscoverable by the search function (e.g. Mäki - Mäen).
    - The find-replace function can be a useful supplemental tool in the final check of the document to see if any identifiers remain.

- Remember to check the case name/citation for any identifiers!

- Remember to check also footnotes for any personal data, if applicable.

- If some data will not be anonymised, explain why you cannot anonymise the data.

- If the data will be coded, describe the coding system, and who will have access to it; and confirm that it cannot be traced back to individuals unless essential for the study.

- Assess whether the remaining data combined with additional information might permit the identification of an individual.

- Prepare a short statement disclosing your anonymisation technique or the mix of techniques that you have used, and highlight any caveats, challenges, or problems. Doing so will help to show that your work is conscientious and reflective and will serve to maintain the quality and legal compliance of the outcomes of the project.

Some further resources

This document is one of the primary outputs of EURO-EXPERT, European Research Council funded project led by Livia Holden, and it is protected by copyright laws. Please cite as follows: Arajärvi, Noora and Livia Holden (with the assistance of Anna Ziliotto and Joshua Bishay) 2021 "GDPR compliant guidelines for processing personal data in legal documents", available at https://culturalexpertise.net/wp-content/uploads/2021/06/gdprcompliantguidelinesforprocessingpersonaldata.pdf

P Jougleux, Legal Opinion on the OA database for research and dissemination purposes, EURO-EXPERT, 16 March 2021.

EUI, Guide on Good Data Protection Practice in Research (2019), https://www.eui.eu/Documents/AboutEUI/Organization/DataProtection/DPO-Good-Data-Protection-Practice.pdf.

Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques, 10 April 2014, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.

Annex 1

1. Introduction

There are several reasons why for anonymising or redacting the name of a legal person (e.g. a corporation) to protect the identity and personal data of a natural person (an individual). This memorandum outlines the two most common examples of when to anonymise or redact the legal entity's name: First, if redaction or anonymisation allows readability without sacrificing the strict standards which we apply for GDPR compliance, and second, to erase context which may reveal information about unrelated parties.
The need to anonymise or redact the name of a legal person arises most commonly in proceedings where the actions of natural persons are key to litigation between legal persons, for example, when the CEO of a company is accused of defrauding another company's corporate officers. In these situations, we have two options: (1) to anonymise or redact the natural person's name and relationship with the legal person, or (2) to anonymise or redact the natural person's name and the legal person's name.

2. Maintaining comprehension and readability

In Excerpt 2.1, it is clear that the natural person is the founder of a corporate entity. The two options are shown below to fully anonymise the paragraph:
Excerpt 2.1:

> 2 The First Defendant, VAK ("VAK "/"D"), is a Russian citizen and the founder of the JFC Group. The Second Defendant, JFC Group Holding (BVI) Limited

In Excerpt 2.1(a), the natural person's data redacted:
Excerpt 2.1(a):

> 2 The First Defendant, VAK ("VAK "/"D"), is a Russian citizen and the founder of the JFC Group. The Second Defendant, JFC Group Holding (BVI) Limited

In Excerpt 2.1(b), the legal person's data redacted:
Excerpt 2.1(b):

> 2 The First Defendant, VAK ('VAK "/"D"), is a Russian citizen and the founder of the JFC Group. The Second Defendant, JFC Group Holding (BVI) Limited

In excerpt 2.1(a) the relationship that VAK has with the company is lost, although we know the name of the company. However, with Excerpt 2.1(b) the pseudonymised name of the legal person is "Corp1" (Corporate 1), so we know that VAK is the founder of "Corp1" and more information is preserved to improve readability. Furthermore (in Excerpt 2.1(b)), leaving "the founder" unredacted does not risk disclosure of personal data but maintains the importance of this person's position in the company as a measure of their decision-making power, or to contextualise the later evidence discussed in the case. For example, with excerpt 2.2:

Excerpt 2.2:

> e. At the suggestion of Mrs YZ , Mr VAK agreed that I should take over as General Director of JFC Russia from April 2011 so that Mrs YZ could concentrate her efforts on finding alternative sources of funding.

In Excerpt 2.2(a) below, the testimony of the witness would be useless unless one is allowed to know that YZ, who was CEO at the time, convinced the founder of "Corp1" that the witness should become GD of Corp1's subsidiary "Corp1.1" as below would maintain:

Excerpt 2.2(a):

> e. At the suggestion of Mrs YZ , Mr VAK agreed that I should take over as General Director of JFC Russia from April 2011 so that Mrs YZ could concentrate her efforts on finding alternative sources of funding.

Another example (Excerpt 2.4, below) of this arises in relation to real property (buildings, land, apartments, etc.). In the example below, there is a question of whether Loc3 (Location 3) is owned by Corp10.1 or VA. By redacting the precise location and the company name, this context can be maintained so that the reader understands the dispute between Corp10's subsidiary (Corp10.1) and VA.

Excerpt 2.3:

> (2) A residential development at Solnechnoye, to the north of St Petersburg and near Loc3 on the Gulf of Finland, owned by Corp10.1, a subsidiary of Corp10 which Dr VA says was owned by him;

3. Protecting the identity of third parties uninvolved in litigation

In Excerpt 3.1, disclosing the legal entity's name would reveal the involvement of a natural person who is entirely uninvolved in the case:

Except 3.1

> 9 Between June and early November 2016, Corp4 was engaged by Corp5 ("Corp5 "), a consultancy based in Washington DC, providing research, strategic intelligence and due diligence services to clients. Mr CS and Corp4 had developed a prior working relationship with Corp5 over a number of years. Corp5 engaged Corp4 to prepare a series of confidential memoranda based on intelligence concerning
>
> DJT . The parties describe these memoranda as the and

In Excerpt 3.1, redacting the names of two companies is necessary to protect the identity of CS, but it also conceals further information about individual DJT, who is not a party or witness in the case. By redacting the names of the companies, we do not have to redact "had developed a prior working relationship with Corp5 over a number of years".