

Annotation guidelines

1. Introduction

This document contains annotation guidelines for document anonymization within the ANONYMIZED project. Anonymisation is the process of removing textual elements that may lead to an individual being identified, either directly (because of e.g. a name or personal number) or indirectly by combining so-called "quasi-identifiers" together with background knowledge (e.g. nationality, country of residence and profession in combination).

2. Annotation steps

The anonymization process consists of two steps: (i) identifying the words to anonymise, and (ii) suggesting a replacement.

2.1. Identifying the information to anonymise

Any term or group of terms that contain *personal data*, i.e. identifying information, should be annotated by adding an XML start (<a>) and end tag () around the term(s), as the example below shows:

(1) <a>Edvard Munch was a Norwegian painter.

Personal data includes, among others, any "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 [e.g. religious beliefs, ethnicity] or social identity" which can directly or indirectly (in combination) identify a person (<https://gdpr-info.eu/art-4-gdpr/>).

Span of annotations: consecutive groups of tokens relating to the same sensitive information should be enclosed within the same XML tag, as example (1) containing a two-token name shows.

2.2. Suggesting replacements

In this first stage of annotation, providing replacements is not required.

3. Data utility

To preserve the usefulness of the anonymised documents, only the necessary information outlined above should be changed. In other words, the anonymisation should balance two objectives: on one hand ensure that the identity of the person is no longer disclosed, but on the other hand retain as much semantic content as possible.

References

European Parliament and the Council of the EU. (2016). General Data Protection Regulation - Regulation (EU) 2016/679. Official Journal of the European Union (OJ), 59(1-88), 294. <<https://gdpr-info.eu/>>