Qualitative and Quantitative Methods in Libraries (QQML) 9,3: 441-460,2020

Text and Data Mining for the National Library of Greece in consideration of Internet Security and GDPR

Marinos Papadopoulos,[1] Michalis Gerolimos,[2] Konstantinos Vavousis[3] and Christos Xenakis[4]

[1] PhD, Attorney-at-Law, Legal Counsel of the National Library of Greece
[2] PhD, e-Resources & Systems Librarian at the National Library of Greece
[3] PhD cand IT Security Professional in the private sector (TRUST-IT Ltd.); secnews.gr Editor-in-Chief
[4] Professor, University of Piraeus, Department of Digital Systems; System Security Laboratory

Abstract: Text and Data Mining (TDM) as a technological option is usually leveraged upon by large libraries worldwide in the technologically enhanced processes of web-harvesting and web-archiving with the aim to collect, download, archive, and preserve content and works that are found available on the Internet. TDM is used to index, analyze, evaluate and interpret mass quantities of works including texts, sounds, images or data through an automated "tracking and pulling" process of online material. Access to the web content and works available online are subject to restrictions by legislation, especially to laws pertaining to Copyright, Industrial Property Rights and Data Privacy. As far as Data Privacy is concerned, the application of the General Data Protection Regulation (GDPR) is considered as an issue of vital importance for the smooth operation of TDM service offered by national libraries mostly in the EU Member States, which among other requirements mandates the adoption of privacy-by-design and advanced security techniques. This article focuses on the TDM deployed by National Library of Greece (NLG) and considerations for applied Internet Security solutions taking into account GDPR requirements. NLG has deployed TDM as of February 2017 in consideration of the provision of art.4(4)(b) of Law 4452/2017, as well as of the provisions of Regulation 2016/679/EU (GDPR). Art.4(4)(b) of law 4452/2017 sets the TDM activity in Greece under the responsibility of NLG, appointed as the organization to undertake, allocate and coordinate the action of archiving the Hellenic web, i.e. as the organization responsible for text and data analysis at national level in Greece. The deployment of TDM by NLG, presented by the authors, caters for a framework of technical and legal considerations, so that the electronic service enabled based on the TDM operation complies with the data protection requirements set by the new EU legislation. While the presentation elaborates upon minimum set of technical Internet Security means considered by NLG for achieving GDPR compliance, the paper (to-be-

Received: 19.5.2020 Accepted: 9.9.2020 ISSN 2241-1925

published) focuses on TDM and GDPR issues specifically in relation to art.89 of GDPR titled "*Safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes*" that is a key-tern ruling for the operation of NLG in compliance with GDPR.

Keywords: web harvesting, web archiving, data analysis, text & data mining, TDM, text mining, content mining, computational text analysis, text and data analysis, web scraping, archiving, copyright law, methods and applications, policies, TDM on databases, reproduction, Optimal Infrastructure, Strong Security Mechanism, GDPR

1. Text & Data Mining and GDPR issues for the National Library of Greece

Text and Data Mining (hereinafter, TDM) activity may involve the processing of personal data. This processing, though, of personal data is processing for archiving purposes in the public interest, or processing for scientific or historical research or statistical purposes. In many cases it is processing that combines more than one of the above-mentioned purposes.

Under the General Data Protection Regulation (Regulation 2016/679/EU, hereinafter GDPR),[1] the data protection principles set out the main responsibilities for organizations. These principles are applicable in the case of organizations which benefit from the TDM exception, of course. The principles are similar to those described in Directive 95/46/EC (the Data Protection Directive)[2] of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data; however, GDPR has added detail at certain points and a new accountability requirement. The most significant addition is the accountability principle. The GDPR requires from a data processor to show how it complies with the data protection principles, for example by documenting the decisions it takes about a processing activity.

The data protection principles are described in article 5 of GDPR.[3] [4] Article 5 of GDPR lays down all the key principles for the protection of personal and special

[1] *Regulation 2016/679/EU* of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), available at URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679 [last check, April 30, 2020].

[2] The Data Protection Directive, officially Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, is a European Union directive adopted in 1995 which regulated the processing of personal data within the European Union (EU). The General Data Protection Regulation has superseded the Data Protection Directive and came into force as of May 25, 2018.

[3] See, also, Data Protection Directive, art.4 titled "*Principles relating to processing of personal data*"; see, also, Recital 39 according to which *Any processing of personal data should be lawful and fair. It should be transparent to natural persons that personal data concerning them are collected, used, consulted or otherwise processed and to what*

categories data, i.e. the lawfulness, fairness, transparency, purpose-limitation, data-minimization, accuracy, storage-limitation, integrity and confidentiality, and accountability. According to the provision of article 5 of GDPR:

1. *Personal data shall be:*
    a. processed lawfully,[5] *fairly*[6] *and in a transparent*[7] *manner in relation to individuals;*[8]

---

extent the personal data are or will be processed. The principle of transparency requires that any information and communication relating to the processing of those personal data be easily accessible and easy to understand, and that clear and plain language be used. That principle concerns, in particular, information to the data subjects on the identity of the controller and the purposes of the processing and further information to ensure fair and transparent processing in respect of the natural persons concerned and their right to obtain confirmation and communication of personal data concerning them which are being processed. Natural persons should be made aware of risks, rules, safeguards and rights in relation to the processing of personal data and how to exercise their rights in relation to such processing. In particular, the specific purposes for which personal data are processed should be explicit and legitimate and determined at the time of the collection of the personal data. The personal data should be adequate, relevant and limited to what is necessary for the purposes for which they are processed. This requires, in particular, ensuring that the period for which the personal data are stored is limited to a strict minimum. Personal data should be processed only if the purpose of the processing could not reasonably be fulfilled by other means. In order to ensure that the personal data are not kept longer than necessary, time limits should be established by the controller for erasure or for a periodic review. Every reasonable step should be taken to ensure that personal data which are inaccurate are rectified or deleted. Personal data should be processed in a manner that ensures appropriate security and confidentiality of the personal data, including for preventing unauthorized access to or use of personal data and the equipment used for the processing.

[4] See, also, Directive 2016/680/EU art.4 for principles relating to processing of personal data, and Recitals 26-28 in this Directive.

[5] The lawfulness of the processing is described in art.6(1) of GDPR; relevant to the lawfulness of the processing are Recitals 40-49. The conditions of data subject's consent are described in art.7 of GDPR. Regarding a child's consent in the case of offering of Information Society services, art.8 of GDPR applies. Lawfulness of the processing means that personal data processing respects all applicable requirements; personal data processing should be considered as lawful if processing is in accordance with law, pursues a legitimate purpose, and is necessary and proportionate in a democratic society in order to achieve that purpose.

[6] Fair processing implies that personal data or special categories data have not been obtained or otherwise processed through unfair means, by deception or without the knowledge of data subject.

[7] Transparency means that it should be clear to natural persons that personal data concerning them are collected, used, consulted or otherwise processed. Relevant is Recital 39 which explains transparency and sets requirements for the quality of information to be given to data subjects: it should be easily accessible and easy to understand, and it should also include information through which natural persons should be made aware of risks and safeguards in relation to the processing of their personal data.

b. collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall not be considered to be incompatible with the initial purposes;[9]

c. adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed;[10]

d. accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay;[11]

e. kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes subject to implementation of the appropriate technical and organizational measures required by the GDPR in order to safeguard the rights and freedoms of individuals;[12]

f. processed in a manner that ensures appropriate security of the personal data, including protection against unauthorized or unlawful processing and against accidental loss, destruction or

---

See, also, art.12 of GDPR; relevant to the transparency of the processing are Recitals 58-59.

[8] See, also, art.8 of Directive 2016/680/EU.

[9] The purpose-limitation principle requires data to be processed for specified, explicit, and legitimate purposes (the purpose-dimension of this principle) and not further processed in a manner that is incompatible with those purposes (the compatible-dimension of this principle). Both dimensions of the purpose-limitation principle should occur at the time of collection of the personal data, i.e. at the beginning of the processing of personal data and/or special categories data. Relevant to the purpose-limitation principle is the provision of art.6(4) of GDPR. There are only two cases for exception of the purpose-limitation principle: 1) if the data subject consents to a new, incompatible purpose for his/her data processing, and 2) if the processing is based on EU or Member-State law. Aside from these two cases, GDPR considers as a priori compatible with the initial purpose of data processing the cases of processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes.

[10] The data-minimization principle pertains to both the quantity and quality of data which should only be processed only if the purposes aimed cannot be fulfilled by other means. Recital 39 is relevant to the minimization-principle.

[11] The accuracy principle. See, also, art.7(2) of Directive 2016/680/EU.

[12] The storage-limitation principle. See, also, art.25 of GDPR and art.20 of Directive 2016/680/EU. Relevant are art.4(1)(e) and art.5 of Directive 2016/680/EU.

> *damage, using appropriate technical or organizational measures.*[13]
>
> 2. *The controller shall be responsible for, and be able to demonstrate compliance with, paragraph 1 ('accountability').*[14]

The processing of personal data through the TDM activity is inevitable. TDM involves processing of text and data which may include any information relating to an identified or identifiable natural person, a.k.a. personal data.[15] Essential to the concept of personal data is the linkability of information to an individual allowing his/her identification. Regarding TDM and personal data protection there is concern that sets of correlated data that could be considered insignificant or even trivial can provide intimate knowledge about data subjects where TDM is applied (Hargreaves, I., et al., 2014). Any information that allows for identification of natural person by reasonable means may constitute personal data. Truly anonymous data do not constitute personal data, as is stated in Recital 26 of GDPR: "[...]*The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable*."

TDM constitutes processing of personal data, in the sense that it involves 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, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.[16] In most cases TDM works in the following manner (Botti, M., et al., 2019b):

1. It identifies input materials to be analyzed, such as works, or data individually collected or organized in a pre-existing database;
2. It copies substantial quantities of materials—which encompasses:
   a. pre-processing materials by turning them into a machine-readable format compatible with the technology to be deployed for the TDM so that structured data can be extracted. Pre-processing typically encompasses the following tasks:

[13] The integrity and confidentiality principle. Art.32-34 of GDPR are relevant to this principle. Also, art.4(1)(f) and art.39-31 of Directive 2016/680/EU focus on the integrity and confidentiality principle.

[14] The accountability principle means that the controller must be able to demonstrate that the processing is in compliance with the legal applicable rules. Relevant is art.24 of GDPR.

[15] Art.4(1) of GDPR.

[16] Art.4(2) of GDPR.

i. *Tokenization*: this is typically the first step in a natural language processing solution and it refers to splitting the text into meaningful character sequences/self-contained semantic units, e.g. words or sentences.

ii. *Normalization*: this involves removing morphological variations from words such as capitalization, plural number or tenses, in order to grasp similarities between them (e.g., the same word in singular and plural), obviously with a loss of information. Two types of techniques are used regarding normalization. These are *stemming* and *lemmatization*. In the former, language specific patterns are recognized, using for example the rules for converting words from singular to plural or verb tenses. This technique is simple, fast and applicable for large volumes of text. Lemmatization involves using a dictionary (such as WordNet that is both a dictionary and a thesaurus) to extract the roots of common words. This approach can be more accurate compared to stemming, but it is more resource intensive and dictionaries may be incomplete for certain languages. The two methods can complement each other and they are often used in conjunction.

iii. *Parsing*: this involves a group of functions that are used after term isolation and document cleanup, i.e., after normalization and parsing, which facilitate working in higher abstraction layers. Typically, parsing includes morphological and syntactical analysis of tokens in order to identify their role within sentences (e.g. noun, verb, adjective or object-verb-subject), which is referred to as *Part-of-Speech (POS) tagging*.

b. possibly, but not necessarily, uploading the pre-processed materials on a platform, depending on the TDM technique to be deployed;

1. It extracts the data; and
2. It recombines data to identify patterns into the final output.

Therefore, to undertake TDM a researcher must access and make a copy of the work/data in order to apply the necessary algorithms for the extraction of new knowledge. This necessary copying of the work/data in the process of the application of TDM has led to considerations of the necessity for an open norm in the European Copyright legal framework which could be similar to the open norm of the "*fair use*" doctrine in the American Copyright Act. Unfortunately, there's no room for such an open norm doctrine in the EU Copyright law, for the time being (Botti, M., et al., 2019a; Botti, M., et al., 2019b).

It has also led to considerations regarding data protection which have become more vivid taking into account the "*straightjacket*" of the GDPR. TDM is restricted by GDPR. It is not an activity that can be lawfully executed without restrictions. The application of GDPR rules on TDM restricts the later through the principles of processing, the legal grounds for the processing, and informatory obligations of the data subjects for the processing, at least (Caspers,

M., et al., 2016). Thus, the collection and processing of personal data in the framework of TDM activity undertaken for scientific research purposes is subject to the safeguards imposed by GDPR principles, such as the necessity of having a legitimate ground to process such data, the obligation to collect data only as far as it is necessary in order to achieve the specified and legitimate purpose (principle of finality/the purpose limitation principle of art.5(1)(b) pf GDPR);[17] the prohibition against collecting more data and to keep them for a longer period than is necessary for the purposes for which they are collected and/or further processed (the 'data minimization' principle).[18] Also, the organization which deploys TDM is bound by the principle of accountability which means that said organization must be able to demonstrate that it has appropriate processes to ensure that it only collects and holds the personal data that it needs in order to achieve the scientific purpose for which TDM was deployed. Besides, said organization must bear in mind that GDPR says individuals have the right to complete any incomplete data which is inadequate for the organization's purpose, under the right to rectification.[19] They also have right to force the organization which processed their data to delete any data that is not necessary for the organization's purpose, under the right to erasure (right to be forgotten).[20]

2. Article 89 of GDPR

In the case of TDM activity undertaken by the National Library of Greece (hereinafter, NLG), article 89 of GDPR is applicable. According to article 89 of

[17] According to art.5(1)(b) of GDPR personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes. In practice, the purpose limitation principle means that the organization which deployed TDM must: (1) be clear from the outset why it is collecting personal data that may be included in the text or data aimed to be mined and what it intends to do with it; (2) comply with said organization's documentation obligations to specify the purposes for TDM which may involve the collection of personal data; (3) comply with the organization's transparency obligations to inform individuals about its purposes regarding TDM and the processing of personal data mined; and (4) ensure that if the organization plans to use or disclose personal data for any purpose that is additional to or different from the originally specified purpose, the new use is fair, lawful and transparent.

[18] According to art.5(1)(c) of GDPR personal data shall adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed (data minimization). This means that the organization which deploys TDM activity must identify the minimum amount of personal data that it needs to fulfil the purpose of scientific research through TDM. Once this minimum amount of personal data is identified, the organization should hold no more personal data than what was identified as necessary.

[19] Art.16 of GDPR.

[20] Art.17 of GDPR.

GDPR, titled "*Safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes*" (emphasis through underscore added by the authors):

1. Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards, in accordance with this Regulation (Regulation 2016/679/EU), for the rights and freedoms of the data subject. Those safeguards shall ensure that technical and organizational measures are in place in particular in order to ensure respect for the principle of data minimization. Those measures may include pseudonymization provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner.
2. Where personal data are processed for scientific or historical research purposes or statistical purposes, Union or Member State law may provide for derogations from the rights referred to in Articles 15,[21] *16,*[22] *18*[23] *and 21*[24] *subject to the conditions and safeguards referred to in paragraph 1 of this Article in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes.*
3. Where personal data are processed for archiving purposes in the public interest, Union or Member State law may provide for derogations from the rights referred to in Articles 15, 16, 18, 19,[25] *20*[26] *and 21 subject to the conditions and safeguards referred to in paragraph 1 of this Article in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes.*
4. Where processing referred to in paragraphs 2 and 3 serves at the same time another purpose, the derogations shall apply only to processing for the purposes referred to in those paragraphs.

Article 89(1) of GDPR repeats the principles[27] of data minimization,[28] purpose limitation,[29] and storage limitation.[30] The whole article 89 encompasses the

[21] Right of access by the data subject
[22] Right to rectification
[23] Right to restriction of processing
[24] Right to object
[25] Notification obligation regarding rectification or erasure of personal data or restriction of processing
[26] Right to data portability
[27] See article 4 of Directive 2016/680/EU regarding the principles relating to the processing of personal data. See, also, Recital 26 of the same Directive, according to which *Any processing of personal data must be lawful, fair and transparent in relation to the natural persons concerned, and only processed for specific purposes laid down by law. … Natural persons should be made aware of risks, rules, safeguards and rights in*

processing of both personal data of article 6 of GDPR as well as special categories data of article 9 of GDPR.[31] Article 89 does not describe a legal basis for the processing of personal data or of special categories data; the legal bases for the processing of this data are described strictly in article 6 of GDPR. Article 6(1) of GDPR exhaustively stipulates what may constitute a legal basis for data processing. Therefore, the processing of data through the TDM application in a library setting could be lawful only if either one of the options described in article 6(1) of GDPR is applicable. Most likely, processing of data in the framework of TDM complies with processing that is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and

---

relation to the processing of their personal data and how to exercise their rights in relation to the processing. In particular, the specific purposes for which the personal data are processed should be explicit and legitimate and determined at the time of the collection of the personal data. The personal data should be adequate and relevant for the purposes for which they are processed. It should, in particular, be ensured that the personal data collected are not excessive and not kept longer than is necessary for the purpose for which they are processed. Personal data should be processed only if the purpose of the processing could not reasonably be fulfilled by other means. In order to ensure that the data are not kept longer than necessary, time limits should be established by the controller for erasure or for a periodic review. Member States should lay down appropriate safeguards for personal data stored for longer periods for archiving in the public interest, scientific, statistical or historical use.

[28] Art.5(1)(c) of GDPR. The principle of data minimization is a specification of the general principle of proportionality. The principle of data minimization posits that the collection of personal data shall be *adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimization').*

[29] Art.5(1)(b) of GDPR. According to this principle, personal data may be *collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation')*.

[30] Art.5(1)(e) of GDPR. According to this principle, personal data may be *kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) subject to implementation of the appropriate technical and organizational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject ('storage limitation')*; Upon expiration of that period, data must be deleted or anonymized.

[31] See art.9(2)(j) of GDPR according to which *processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject.*

freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.[32]

Paragraphs 2 and 3 of article 89 entitle the Member States to provide for derogations from certain rights of the data subjects. The scope of article 89(2) is limited to processing for scientific or historical research purposes and statistical purposes. Said paragraph of article 89 provides for derogations from the right of access,[33] the right of rectification,[34] the right of restriction of processing,[35] and the right of objection.[36] However, such derogations are still subject to the conditions and safeguards referred to in article 89(1) of GDPR, which means firstly that appropriate safeguards must be in place to protect the rights and freedoms of data subjects even when derogations apply;[37] [38] [39] secondly, that the

[32] Art.6(1)(f) of GDPR.
[33] Art.15 of GDPR.
[34] Art.16 of GDPR.
[35] Art.18 of GDPR.
[36] Art.21 of GDPR.
[37] See Recital 156 according to which *The processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes should be subject to appropriate safeguards for the rights and freedoms of the data subject pursuant to this Regulation. 2Those safeguards should ensure that technical and organizational measures are in place in order to ensure, in particular, the principle of data minimization. The further processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes is to be carried out when the controller has assessed the feasibility to fulfil those purposes by processing data which do not permit or no longer permit the identification of data subjects, provided that appropriate safeguards exist (such as, for instance, pseudonymization of the data). Member States should provide for appropriate safeguards for the processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes. Member States should be authorized to provide, under specific conditions and subject to appropriate safeguards for data subjects, specifications and derogations with regard to the information requirements and rights to rectification, to erasure, to be forgotten, to restriction of processing, to data portability, and to object when processing personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes. The conditions and safeguards in question may entail specific procedures for data subjects to exercise those rights if this is appropriate in the light of the purposes sought by the specific processing along with technical and organizational measures aimed at minimizing the processing of personal data in pursuance of the proportionality and necessity principles.*
[38] See Recital 157 according to which *In order to facilitate scientific research, personal data can be processed for scientific research purposes, subject to appropriate conditions and safeguards set out in Union or Member State law.*
[39] See Recital 162 according to which *Union or Member State law should, within the limits of this Regulation, determine statistical content, control of access, specifications for the processing of personal data for statistical purposes and appropriate measures to safeguard the rights and freedoms of the data subject and for ensuring statistical confidentiality.*

use of the rights from which derogations are given must be likely to render impossible or seriously impair the achievements of the specific purposes, and such derogations are necessary for the fulfilment of those purposes;[40] and thirdly, that derogations only apply to the purposes mentioned in the respective paragraphs of article 89.

Recital 159 states that "*the processing of personal data for scientific research purposes should be interpreted in a broad manner including for example technological development and demonstration, fundamental research, applied research and privately funded research.*" National legislation of Member-States may include a definition of the term "*scientific research purposes*"; the intention of the EU legislator is to include under the definition of "*scientific research purposes*" the broadest possible meaning and allow scientific research purposes to be pursued at least to the extent possible under the Data Protection Directive. In the Greek legal framework there's no specific definition of the term "*scientific research purposes*", though there is law 3653/2008 that pertains to the enhancement of scientific research and technology in Greece. The individual right to scientific research and teaching is recognized in the wording of article 16(1)(a) of the Greek Constitution according to which "*Art and science, research and teaching shall be free and their development and promotion shall be an obligation of the State.*" However, so far there are very few Greek Court decisions elaborating upon this individual right and the meaning of "*scientific research purposes*", with the most notable being the decision of Council of State No.1043/1989 (Papadopoulos, M., *Scientific Research, Web Harvesting and Text & Data Mining*, in Zachou, V., ed., 2020).

The term "*historical research*" is not defined in the GDPR. However, Recital 160 makes clear that under the term "*historical research*" fit both "*historical research and research for genealogical purposes*".

---

[40] See art.14(5)(b) of GDPR, according to which the provision of information where personal data have not been obtained from the data subject shall not apply insofar as *the provision of such information proves impossible or would involve a disproportionate effort, in particular for processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, subject to the conditions and safeguards referred to in Article 89(1) or in so far as the obligation referred to in paragraph 1 of this Article is likely to render impossible or seriously impair the achievement of the objectives of that processing. In such cases the controller shall take appropriate measures to protect the data subject's rights and freedoms and legitimate interests, including making the information publicly available.* See, also, art.17(3)(d) of GDPR, according to which the provisions of paragraphs 1 and 2 of article 17 that pertains to the right to erasure (the right to be forgotten) shall not apply to the extent that processing is necessary *for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) in so far as the right referred to in paragraph 1 is likely to render impossible or seriously impair the achievement of the objectives of that processing.*

Recital 162 defines the meaning "*statistical purposes*" as "*any operation of collection and the processing of personal data necessary for statistical surveys or for the production of statistical results*". Further, the statistical purpose in the processing of data implies that the result of the processing is not personal data, but aggregate data, and that this result of the processing of personal data may not be used in support of measures or decisions regarding any particular natural person (data subject). Special regulation may be applicable in case of processing of data for statistical purposes, such as article 338(2) of the Treaty for the Functioning of the European Union (TFEU)[41] or the Regulation EC 223/2009 on European Statistics.[42]

Article 89 does not distinguish between research pursuing public interests and research done for private and/or purely commercial purposes. Thus, it applies in research pursued through TDM or other means either for public or for private interest, either for commercial or for non-commercial purpose.

Article 89(3) applies to archiving purposes in the public interest. Not every archive falls under the scope of article 89(3), but only those that have a legal obligation to maintain records in the scope of the public interest. According to Recital 158, "*Public authorities or public or private bodies that hold records of public interest should be services which, pursuant to Union or Member State law, have a legal obligation to acquire, preserve, appraise, arrange, describe, communicate, promote, disseminate and provide access to records of enduring value for general public interest.*" This means that archives which do not fit in the public interest scope, are not covered by article 89. NLG is entitled to hold records of public interest and has a legal obligation to acquire, preserve, appraise, arrange, describe, communicate, promote, disseminate and provide access to records of enduring value for general public interest according to Greek law 3149/2003 as amended through law 4452/2017.

Under article 89(3) of GDPR where personal data are processed for archiving purposes in the public interest, EU law of Member-State law may provide for derogations from the right of access by the data subject,[43] the right of

---

[41] According to art.338(2) of TFEU *The production of Union statistics shall conform to impartiality, reliability, objectivity, scientific independence, cost-effectiveness and statistical confidentiality; it shall not entail excessive burdens on economic operators.*

[42] Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics and repealing Regulation (EC, Euratom) No 1101/2008 of the European Parliament and of the Council on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities, Council Regulation (EC) No 322/97 on Community Statistics, and Council Decision 89/382/EEC, Euratom establishing a Committee on the Statistical Programmes of the European Communities (Text with relevance for the EEA and for Switzerland), OJ L 87, 31/03/2009, p.164–173.

[43] Art.15 of GDPR.

rectification,[44] the right to restriction of processing,[45] the notification obligation,[46] the right to data portability,[47] and the right to object.[48] However, such derogations in the case of article 89(3) are still subject to the conditions of article 89(1) of GDPR.

In the case of processing of special categories of personal data, i.e. personal data of article 9 of GDPR, either for archiving purposes in the public interest or for scientific and historical research purposes or for statistical purposes national law may stipulate conditions for the lawfulness of the processing, according to article 9(2)(j) of GDPR. Article 22 of Greek law 4624/2019 caters for the lawfulness of the processing of special categories personal data; said article of law 4624/2019 is applicable in the case of processing of special categories personal data through the TDM process that is considered lawful according to article 9(2)(a) of law 4624/2019, i.e. on the basis of (absolutely) necessary processing for the purpose of public interest.

In addition to the public interest purpose, TDM activities by the National Library of Greece may also be undertaken for historical[49] or scientific[50] research purposes, at least. According to Recital 159 of GDPR in order to meet the specificities of processing personal data for scientific research purposes, specific conditions must apply in particular as regards the publication or otherwise disclosure of personal data in the context of scientific research purposes. If the result of scientific research gives reason for further measures in the interest of the data subject, the general rules of Regulation 2016/679/EU should apply in view of those measures. Thus, given that the output of TDM deployed by the National Library of Greece (NLG) includes personal data found in the works harvested from the Web, NLG must apply specific conditions regarding the publication or otherwise disclosure—in copyright terms this is deemed to be

---

[44] Art.16 of GDPR.
[45] Art.16 of GDPR.
[46] Art.19 of GDPR.
[47] Art.20 of GDPR.
[48] Art.20 of GDPR.
[49] See Recital 160 of GDPR according to which *Where personal data are processed for historical research purposes, this Regulation should also apply to that processing. This should also include historical research and research for genealogical purposes, bearing in mind that this Regulation should not apply to deceased persons.*
[50] The 'scientific research purpose' is meant widely for the application of GDPR. According to Recital 159, *For the purposes of this Regulation, the processing of personal data for scientific research purposes should be interpreted in a broad manner including for example technological development and demonstration, fundamental research, applied research and privately funded research. In addition, it should take into account the Union's objective under Article 179(1) TFEU of achieving a European Research Area. Scientific research purposes should also include studies conducted in the public interest in the area of public health.*

relevant to the presentation/communication to the public right of copyright[51]—of the output of the TDM that includes personal data of persons who have not deceased. These specific conditions could pertain to the intranet or the extranet through which the TDM output is accessible to a certain public that is narrower than the general public.

Regarding TDM activities undertaken for statistical purposes, NLG must cater for the following requirements in consideration of Recital 162 of GDPR: the result of the processing of personal data or special categories data whenever is undertaken either in the framework of TDM or other means for statistical purposes cannot be personal data, but only aggregate data; the result of the processing for statistical purposes cannot be used in support of measures or decisions regarding any particular data subject (natural person). In consideration of Recital 162 statistical results may further be used for purposes other than scientific research purposes. The term '*statistical purposes*' is meant as "*any operation of collection and the processing of personal data necessary for statistical surveys or for the production of statistical results*".[52] According to Recital 162 of GDPR "*The statistical purpose implies that the result of processing for statistical purposes is not personal data, but aggregate data, and that this result or the personal data are not used in support of measures or decisions regarding any particular natural person.*" Thus, there're two conditions which both must be met in the case of processing of personal data or special categories data for statistical purposes: a) the result of the processing of personal data or special categories data must not be personal data but should be aggregate data; b) the result of the processing of personal data or special categories data must not be used in support of measures or decisions regarding any particular natural person.

According to art.5(1)(b) of GDPR, personal data processed during TDM activities undertaken by the NLG for historical or scientific purposes may further be processed for archiving purposes in the public interest or for statistical purposes, in accordance with art.89(1), without being considered to be incompatible with the initial purposes ('*purpose limitation*'). Regarding the '*storage limitation*' requirement, art.5(1)(e) of GDPR rules that personal data processed during TDM activity may be stored for longer periods insofar as the personal data are processed solely for archiving purposes[53] in the public interest, scientific or historical research purposes or statistical purposes in accordance

[51] See art.3(1)(h) of Greek Copyright Law 2121/1993, which pertains to the communication to the public right of the works of copyright-holders, by wire or wireless means or by any other means, including the making available to the public of their works in such a way that members of the public may access these works from a place and at a time individually chosen by them.

[52] Recital 162 of GDPR.

[53] According to Recital 158, the 'archiving purpose' includes in particular "*providing specific information related to the political behavior under former totalitarian state regimes, genocide, crimes against humanity, in particular the Holocaust or war crimes.*"

with art.89(1) and are subject to implementation of appropriate technical and organizational measures required by Regulation 2016/679/EU in order to safeguard the rights and freedoms of the data subject.

Regarding the processing of personal data for archiving purposes through the TDM process, data referring to deceased persons do not constitute personal data,[54] thus there's no conflict of the processing of deceased persons with the provisions of GDPR. In the same vein, further processing of personal data for archiving purposes, for example with a view to providing specific information related to the political behavior under former totalitarian state regimes, genocide, crimes against humanity, in particular the Holocaust, or war crimes is not in conflict with the provisions of GDPR.[55]

Art.89(1) makes a reference to '*pseudonymization*'; '*pseudonymization*' is referred as a technical measure to ensure respect of the principle of data minimization. Pseudonymization is meant 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 organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person*".[56] Art.89(1) refers to '*pseudonymization*', but not to '*anonymization*'. However, the reference in article 89(1) to "*further processing which does not permit or no longer permits the identification of data subjects*" could be interpreted as including anonymization. This interpretation is also inferred from Recital 156, according to which "*The further processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes is to be carried out when the controller has assessed the feasibility to fulfil those purposes by processing data which do not permit or no longer permit the identification of data subjects, provided that appropriate safeguards exist (such as, for instance, pseudonymization of the data)*". The list of safeguards mentioned in article 89(1) is not exhaustive, thus both anonymization and pseudonymization could be favored under article 89 of GDPR. The legal distinction between anonymized and pseudonymized data is its categorization as personal data. Pseudonymous data still allows for some form of re-identification (even indirect and remote), while anonymous data cannot be re-identified.[57]

---

[54] See art.4 and Recital 158 of GDPR.

[55] See Recital 158 of GDPR, according to which "*Member States should also be authorized to provide for the further processing of personal data for archiving purposes, for example with a view to providing specific information related to the political behavior under former totalitarian state regimes, genocide, crimes against humanity, in particular the Holocaust, or war crimes.*"

[56] Art.4, No.5 of GDPR.

[57] See Recital 26 of GDPR according to which *"…Personal data which have undergone pseudonymization, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person…..*

Anonymized data do not subject to data protection obligations through the application of GDPR or relevant legislation. Anonymous data cannot be linked back to identifiable data subjects; also, anonymous data is useless for almost anything but almost high-level data aggregation and analysis.[58]

In contrast to anonymized data, pseudonymized data retains some statistical utility relative to the level of pseudonymization. For this reason, data pseudonymization is preferable for statistical analysis. That is why art.89(1) of GDPR makes reference to '*pseudonymization*' rather than to '*anonymization*' as a technical measure to ensure data minimization. Pseudonymization techniques differ from anonymization techniques. With anonymization, the data is scrubbed for any information that may serve as an identifier of a data subject. Pseudonymization does not remove all identifying information from the data but merely reduces the linkability of a dataset with the original identity of an individual (e.g., via an encryption scheme). Both pseudonymization and anonymization are encouraged in the GDPR and enable its constraints to be met. These techniques should therefore be generalized and recurring. Those in possession of personal data should implement one or other of these techniques to minimize risk, and automation can reduce the cost of compliance.

Pseudonymization is referred to GDPR as a method—a technical means—that can be used for demonstrating GDPR-compliance in more than one article or recital of the Regulation. Article 25(1) of GDPR makes a reference to pseudonymization as an appropriate technical measure which is designed to implement data-protection principles. Article 32(1)(a) of GDPR names pseudonymization and encryption of data as a technical means to ensure a level of security appropriate to the risk, thus pseudonymization is advocated as a risk-based approach to data security. Also, Recital 78 of GDPR reports pseudonymization of personal data as soon as possible as a measure which meets the principles of data protection by design and data protection by default.

There are multiple methods for pseudonymization such as data masking, encryption and tokenization. Encryption entails the use of a key to encode or protect a data set. Consequently, encryption is mathematically reversible and is subject to the complexities of key management. Tokenization by comparison, involves replacing identifying or sensitive data with a mathematically unrelated value. Therefore, the tokens cannot be mathematically reversed. Both encryption

---

The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes."

[58] See Working Party of Article 29, *Opinion 05/2014 on Anonymization Techniques*, WP 216, April 10, 2014, available at URL: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf [last check, April 30, 2020].

and tokenization can be format-preserving and tokens may optionally include elements of the original value for data-processing purposes. Tokenization is used in the TDM process. Data masking is a process for obfuscating data that is typically accomplished via encryption. Using masking, data can be de-identified and de-sensitized so that personal information remains anonymous in the context of support, analytics, testing, or outsourcing.

The most suitable method of pseudonymization depends on the specific use case and needs of an organization, although it's worth noting that from a compliance standpoint, tokenization via a cloud-based tokenization provider is the only method that enables an organization to completely remove sensitive or identifying data from its systems. This is a significant differentiator from both a compliance and a data security perspective. As is already stated above, tokenization is part of the TDM process.

TDM activities involve both web harvesting and web archiving processing of subject matter. Regarding the National Library of Greece (NLG) which is a public law entity according to art.1(1) of law 3149/2003, thus is a legal entity which aims at serving the general public interest, Recital 158(2) of GDPR is of interest. NLG as a public body that holds records of public interest—these include the output of TDM records as well as works in the general or specific collections of works archived and made available to the public through NLG—is an organization that provides services which, pursuant to Greek law, has a legal obligation to acquire, preserve, appraise, arrange, describe, communicate, promote, disseminate and provide access to records of enduring value for the general public interest.

Also, in consideration of Recitals 156 and 158 of GDPR, NLG which is empowered to deploy TDM for archiving and research purposes for the public interest must cater for the application of appropriate safeguards for the rights and freedoms of the data subjects pursuant to GDPR. Those safeguards should ensure that technical and organizational measures are in place in order to ensure, in particular, the principle of personal data minimization or the principle of storage limitation. The further processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes is to be carried out when the Controller—it might be a different entity than the NLG which is the Processor definitely—has assessed the feasibility to fulfil those purposes by processing data which do not permit or no longer permit the identification of data subjects, provided that appropriate safeguards exist (such as, for instance, pseudonymization of the data).

Article 89(2) of GDPR allows for derogations from rights referred to in article 15 (the right of access), article 16 (right of rectification), article 18 (right of restriction of processing), and article 21 (right of objection) of GDPR when personal data are processed for scientific or historical research or statistical purposes in so far as such rights are likely to render impossible or seriously

impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes. Said GDPR article also posits that EU or Member State law may provide for derogations from the rights of the aforesaid articles. Regarding the Greek law, article 29 and article 30 of law 4624/2019 are applicable in the case of NLG.

Derogations from the right to object (art.21 of GDPR) where personal data are processed for scientific or historical purposes or statistical purposes pursuant to article 89(1) of GDPR are provisioned in article 21(6) of GDPR, too; article 21(6) allows for derogations from the right to object in cases of processing of personal data which is necessary for the performance of a task carried out for reasons of public interest.

Regarding the obligation to inform the data subject when personal data processed through the TDM process have not been obtained from the data subject,[59] article 23(1) of GDPR allows EU or Member State law to provide restrictions to the obligation to inform. As far as the Greek law is concerned, article 32(1)(a)(aa) of law 4624/2019 is applicable in the case of NLG and its obligation to inform the data subjects for the processing of their personal data through the TDM process. Thus, NLG is not obliged to inform the data subjects for the processing of their personal data through the deployment of TDM by NLG that serves important objectives of general public interest.

Finally, regarding the right to erasure—the right to be forgotten—article 17(3)(d) of GDPR caters for a restriction to it where processing of personal data takes place for archiving purposes in the public interest, scientific or historical research purposes of statistical purposes in accordance with article 89(1) of GDPR in so far as the right to be forgotten is likely to render impossible or seriously impair the achievement of the objectives of the processing.

3. Epilogue

The General Data Protection Regulation (EU) 2016/679 offers a digital environment for companies and organizations where they can better trace, secure and handle data within the IT infrastructure and beyond. In the same vein, GDPR requires strong security mechanisms to be in place in order to safeguard the data under consideration. Hence, powerful security mechanisms should be adopted for the adequate protection of sensitive and private data stored and in order to comply with GDPR. The latest trends in cyber security have embedded technologies with enhanced mechanisms for better results, including machine learning and big data analytics on network security

[59] See art.14 of GDPR that pertains to information to be provided to the data subject where personal data are processed without being obtained from the data subject.

solutions.[60] Regarding these mechanisms and network security solutions planning and/or deployment by NLG there will be further analysis in a forthcoming paper published, soon.

Selected bibliography:

Article 29 Working Party, (2014), *Opinion 05/2014 on Anonymization Techniques, WP216*, April 10, 2014.

Botti, M., Papadopoulos, M., Zambakolas, C., & Ganatsiou, P., (2019a), *On the Eve of Web-Harvesting and Web-Archiving for Libraries in Greece*, Erasmus Law Review, 12, p.178-189.

Botti, M., Papadopoulos, M., Zampakolas, C., & Ganatsiou, P., (2019b), *Text and Data Mining in Directive 2019/790/EU. Enhancing Web-Harvesting and Web-Archiving in Libraries and Archives*, Open Journal of Philosophy (OJPP), 9, p.369-395.

Caspers, M., Guibault, L., McNeice, K., Piperidis, S., Pouli, K., Eskevich, M., & Gavriilidou, M., (2016), *Reducing Barriers and Increasing Uptake of Text and Data Mining for Research Environments Using a Collaborative Knowledge and Open Information Approach*, Baseline Report of Policies and Barriers of TDM in Europe.

Hargreaves, I., Guibault, L., Handke, C., Martens, B., Lynch, R., & Filippov, S., (2014), *Standardisation in the Area of Innovation and Technological Development, Notably in the Field of Text and Data Mining*, Report from the Expert Group, European Union.

Kantarcioglu, M., Xi, B., (2016), *Adversarial Data Mining: Big Data Meets Cyber Security*, CCS '16: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, p.1866–1867, October 2016.

Zachou, V., ed., 2020, Archives and Cultural Readings, Ocelotos Publications.

Additional bibliography:

Botti, M., Papadopoulos, M., Zampakolas, C., & Ganatsiou, P., (2019c), *Text and Data Mining in the EU 'Acquis Communautaire'. Tinkering with TDM & Digital Legal Deposit*, Erasmus Law Review, 12, p.190-208.

Botti, M., Papadopoulos, M., Zampakolas, C., & Ganatsiou, P., (2018), *Legal and Technical Issues for Text and Data Mining in Greece*, in Computer Ethics—Philosophical Enquiry (CEPE) Proceedings.

De Wolf & Partners, (2014), *Study on the Legal Framework of Text and Data Mining (TDM)*, European Union.

European Commission & COM 192 Final, (2015), Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee, and the Committee of the Regions, *A Digital Single Market Strategy for Europe*, Document: 52015DC0192, Brussels.

European Commission & SWD 301 Final Part 1/3, (2016), Commission Staff Working Document, *Impact Assessment on the Modernization of EU Copyright Rules*, Brussels.

European Commission & SWD 301 Final Part 2/3, (2016), Commission Staff Working Document, *Impact Assessment on the Modernization of EU Copyright Rules*, Brussels.

European Commission & SWD 302 Final, (2016), Commission Staff Working Document, *Executive Summary of the Impact Assessment, on the Modernization of EU Copyright Rules*, Document: 52016SC0302, Brussels.

---

[60] Kantarcioglu, M., Xi, B., (2016), *Adversarial Data Mining: Big Data Meets Cyber Security*, CCS '16: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, p.1866–1867, October 2016.

European Copyright Society, (2017), *General Opinion on the EU Copyright Reform Package*.
Feiler, L., Forgó, N., Weigl, M., (2018), *The EU General Data Protection Regulation (GDPR): A Commentary*, German Law Publishers.
Geiger, C., Frosio, G., & Bulayenko, O., (2018), *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market-Legal Aspects*, Centre for International Intellectual Property Studies (CEIPI) Research Paper.
IFLA, (2013), *IFLA Statement on Text and Data Mining*.
Kuner, C., Bygrave, L., Docksey, C., Drechsler, L., eds., (2020), *The EU General Data Protection Regulation (GDPR) – A Commentary*, Oxford University Press.
Linder, A., ed., (2016), *European Data Protection Law – General Data Protection Regulation 2016*, CreateSpace Independent Publishing Platform.
Markham, K., (2020), *A Practical Guide to the General Data Protection Regulation (GDPR): 2nd Edition*, Law Brief Publishing.
Papadopoulos, M., Botti, M., Ganatsiou, P., & Zampakolas, C., (2020), Emporical Research on Web Harvesting in the process of Text and Data Mining in National Libraries of EU Member States, Open Journal of Philosophy (OJPP), 10, 369-395.
Sag, M., (2019), *The New Legal Landscape for Text Mining and Machine Learning*, Journal of the Copyright Society of the USA, 66.
Voigt, P., Brussche, A. v.d., (2017), *The EU General Data Protection Regulation (GDPR): A Practical Guide*, Springer.