Online Journal of Applied Knowledge Management

A Publication of the International Institute for Applied Knowledge Management

Volume 8, Issue 1, 2020

Personal data protection and academia: GDPR issues and multi-modal data-collections "in the wild"

Ingo Siegert, Mobile Dialog Systems, Otto von Guericke University Magdeburg, Germany, ingo.siegert@ovgu.de

Vered Silber Varod, Open Media and Information Lab, The Open University Israel, Israel, vereds@openu.ac.il

Nehoray Carmi, Open Media and Information Lab, The Open University Israel, Israel, nehorayc@gmail.com

Pawel Kamocki, Leibniz Institut für Deutsche Sprache, Mannheim, Germany, pawel.kamocki@gmail.com

Abstract

The European Union (EU) General Data Protection Regulations (GDPR) has a direct impact on research activities, as it raises the awareness of personal rights not only among the scientists but also among the data-subjects scientists process information from. This paper presents the dilemma related to the privacy of audio and video data, compliance with the EU GDPR, and techniques to anonymize and pseudonymize such data. We further discuss issues of "in the wild" personal data collection by focusing on multi-modal collections, mainly of audio, video via these channels. Throughout this paper we define relevant core issues and highlight two challenges of "in the wild" data collection: Internet crawling and public data collecting. In the last section, some exemplary use cases are demonstrating the raised issues, illuminating how GDPR affects the collection of publicly available data; how privacy concerns influence participant behavior, and which de-anonymization levels can be reached with what kind of data. The key point we present is that the identity of the participants is revealed in the voice or video signal, while the latter is at the same time the object of the research. One implication is that the research community has to actively disconnect the data from the personal information on the participants. Hence the importance of a process of anonymity or omission of data for research activity. This entail the development of an infrastructure for data access control to enable data sharing among researchers.

Keywords: General Data Protection Regulations (GDPR), data collection, interaction resources, personal data, academia, "in the wild".

Introduction

As of May 25, 2018, enforcement of compliance with the *European Union (EU)* General Data Protection Regulation (GDPR) 2016/679 began. The GDPR is designed to unify all data privacy laws across Europe while protecting the information of all EU residents against information leakage and privacy breaches. The EU GDPR has a direct impact on research activities, as it raises

the awareness of personal rights not only among the scientists but also among the *experimental subjects* (i.e., *Subjects*, *Respondents*, *Informants* or *Participants*, see Morse, 1991) scientists process information from. Moreover, it affects the way data can be collected, stored, and processed (analyzed, exchanged, etc., (Sveningsson Elm, 2008)). This entails questions such as what constitutes "personal data" and how can it be collected together with the data under investigation, generally regarded here as the subject's behavior. In addition, how can data, especially data that reveal the subject's identity, such as his/her speech signal or his/her face, be efficiently anonymized? and are there alternatives to the use of such data? Up until now, scientists tried to deal with it on their own and to help their peers by publishing documents and papers on ethical issues. Batliner and Schuller (2014), for example, listed crucial ethical issues, including the challenge to guarantee the consent and the privacy of the subjects and the need to encode the data to guarantee this privacy.

Furthermore, although the GDPR, under certain conditions, allows for the processing of non-anonymized personal data for research purposes, it is not clear which lines should not be crossed. Additionally, in times of big-data and deep-learning methods, de-anonymization seems to be easier than ever, and thus, the use of enriched data is under question. At the same time, especially in the field of human-human interaction, the use of deep learning technologies on the one hand, and the scientific desire to study naturalistic behavior (of subjects) on the other hand, raises the demand for massive collections of human interaction (The term human-human interaction denotes conversations between a human and another human, as opposed to human-machine, human-device, or human-robot interaction, (see Krämer et al., 2012 for an overview of different theories on human-human interactions). This implies to avoid the constraints of the lab environment and to collect unsupervised spontaneous interactions while still ensuring high data quality. Simply speaking, academia is aiming for "in the wild" data collections, meaning, to process information of people even when they are not aware of it. This entails the use of data enriched with additional metadata such as age, sex, profession, socio-demographic information (Dudzik et al., 2019), or specific personality traits (similar to Big Data studies that use freely available data found on the Internet). We consider this type of data (e.g., sex, age, language, proficiency, personality, etc.) as crucial to behavior analyses (Truong et al., 2007). These needs are challenging the demands of the GDPR and in particular the principle of data minimization leading to a huge uncertainty in research data collection and sharing.

Due to these contradictory requirements, the anonymization of data is a crucial issue. On the one hand, anonymized data are not subject to the GDPR and as such can in principle be freely processed; on the other hand, anonymization is a highly complicated process which requires considerable efforts and should be periodically reviewed, which leads to the question if it is possible to de-anonymize data and how can publicly available "anonymous" data be used in light of the above? Unfortunately, anonymization is an alternation of the collected data, which is not desirable from the methodological point of view, as many analyses demand non-alternated data, for example, facial analyses need information from the eye-region. Similarly, acoustic analyses need information from the full acoustic spectrum. Both extradite the individual's identity.

Another particular issue of interest is the obligation of information, which, under GDPR's Art. 12 to 14, applies regardless of whether the data were obtained from the 'data subject' (see a definition

below) directly or indirectly (e.g., collected from publicly available sources), and can only be derogated from in very special circumstances. The information should be provided to the 'data subject' before the processing, which raises questions about its impact on future research, as informed individuals may subconsciously behave in an altered manner.

This contribution aims to discuss the above issues by focusing on the effect on multi-modal data collections, mainly data collection of audio, video, and human behavior via these channels (e.g., HCI, personal interviews, etc.). In the next section we define relevant core issues. We then highlight two challenges of "in the wild" Data collection: internet crawling and public data collecting. In the last section, some exemplary use cases are demonstrating the raised issues, illuminating how GDPR affects the collection of publicly available data; how privacy concerns influence participant behavior, and which de-anonymization levels can be reached with what kind of data.

Definitions of Core Issues

Concept of Personal Data in the GDPR

An effective definition of the term *personal data* depends on the jurisdiction and the purposes for which the term is being used. Art. 4, 1. of the GDPR defines 'personal data' as 'any information relating to an identified or identifiable natural person ('data subject'). This definition is essentially the same as in the 1995 Personal Data Directive (95/46/EC), which was replaced by the GDPR. The GDPR's concept of 'personal data' is significantly broader than the concept of 'personally identifiable information' (PII) used e.g. in the US. Personal data definition comprises four elements. Firstly, 'any information' can constitute personal data, regardless of its form (digital/analog, text/image/sound recording, etc.) and even regardless of whether it's a fact or an opinion, or even whether it's true or false. Secondly, in order to constitute personal data, the information has to 'relate' to a person. According to Article 29 Data Protection Working Party (WP136, 2007), information can relate to a person by its *content* (i.e., it says something about the person), by its *purpose* (i.e., it can be used to evaluate the person and treat his/her in a certain way) or by its *result* (i.e., it is likely to have an impact on the person's rights and freedoms). Thirdly, the person that the data relate to should be 'identified or identifiable'. A person is identified when it's singled out from a group, typically by a sufficiently unique name-surname combination, but other identifiers (e.g., username or ID number, or in a certain context – a photograph) can also be taken into account. Moreover, a person is 'identifiable' if it can be singled out from a group by any means reasonably likely to be used (such as cross-referencing with data from social networks). Fourthly, the information has to relate to a 'natural person', i.e., a living individual (and not e.g., a legal entity). Information related to deceased persons, however, may also be regarded as personal data inasmuch as it relates to living individuals (typically the descendants or other family members). The person that the data relate to is referred to as 'data subject' (Art. 4, 1. of the GDPR).

Anonymization and Pseudonymization Under the GDPR

As mentioned above, a person is regarded as identifiable if it can be identified (singled out) by any means 'reasonably likely to be used' (recital 26 of the GDPR). The process of 'breaking the relation' between the information and the person is called *anonymization*. Anonymized data are

no longer considered personal data and therefore the GDPR does not apply to them. The quite broad definition of personal data within the GDPR and the character ('sensitivity' and accessibility) complicates the anonymization process, as assessing what identification means are 'reasonably likely to be used' have to be constantly evaluated (e.g., banking details are subject to stricter anonymization standards than the profession of the person). Moreover, the standard for anonymization changes dynamically over time: some advanced technological means (such as biometric identification) may not be reasonably likely to be used in most contexts today, but they are likely to become commonplace in the years to come. The Article 29 Data Protection Working Party divided anonymization techniques into two groups: randomization (e.g., noise addition) and generalization (e.g., k-anonymity and t-closeness), (WP216, 2014).

The concept of pseudonymization should be distinguished from anonymization. It is defined (in Art. 4, 5. of 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 organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person'. Pseudonymized data are still considered personal data. Their processing is subject to the GDPR; pseudonymization is merely considered an additional safeguard allowing to implement the principle of integrity and confidentiality or to meet the criteria for a research exemption of Art. 89 of the GDPR.

Data Minimization

According to Art. 5.1, c) of the GDPR, personal data should be 'adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed'. This principle (which existed also under the 1995 Data Protection Directive), referred to as 'data minimization', is arguably the biggest hurdle for data-intensive research and technology. The GDPR does not allow any derogations from the principle for research purposes — quite the contrary, processing for research purposes, in order to qualify for some exemptions, should be accompanied with safeguards to ensure respect of data minimization. This implies that, even for research purposes, data that are not relevant and necessary should be deleted or anonymized — hence the importance of anonymization or data omission for research activities (European Data Protection Supervisor, 2020).

Anonymization and De-anonymization Techniques

For an efficient anonymization, two factors have to be taken into account: the privacy-preserving capabilities and the preserved intelligibility (Korshunov et al., 2012). Whilst the first factor reflects the ability to mask particular information of the subject to prevent his/her identification, the second factor describes the ability to still distinguish actual events or features of interest. Thus, anonymization presents a trade-off between privacy and action/event recognize-ability. Most recent approaches performing privacy preservation use some form of transformation on the data, by, e.g., reducing the granularity of the representation. Most common techniques are randomization, k-anonymity or l-diversity, distributed privacy preservation, and downgrading application effectiveness (see Aggarwal & Yu, 2008 for an overview). The actual successor for data samples is differential privacy, where it can be guaranteed that one specific record only has a

very small impact on the result of an analysis. For most purposes of *textual data,* it is sufficient to remove the name, address, and full postal code, or perform generalization and perturbation for relational data. Ensuring a higher degree of privacy l-diversity or differential privacy methods should be applied. For anonymize *facial data* mostly naive techniques such as blurring and pixelation are applied (Newton et al., 2005). More sophisticated approaches, preserving the privacy of the recorded persons are developed. Letournel et al. (2015) presented an anonymization technique based on visual landmarks with varied adaptive filtering that even preserves the important face clues supporting further behavior and emotion analyses.

As we move to adopt voice-controlled systems, transferring and storing users' voice will become a greater privacy concern. For voice anonymization two layers of personal data have to be distinguished, the content containing personal information and the voice itself. Both layers are able to identify the speaker and therefore to be anonymized. Advanced *voice-based speaker anonymization* aims to suppress the speaker identity and is mostly based on voice transformation techniques, changing source, or filter characteristics of the speech (Pobar & Ipšić, 2014). Recent research suggested to use x-vector speaker representations to suppress the timbre of a speaker, and thus hindering the speaker identification (Fang et al., 2019). Although these techniques can preserve the textual context and may sound natural, they cannot sustain the emotional expression of a speaker, as they have never aimed that. Thus, these voice anonymization techniques are not applicable for data aimed at investigating certain speaker states (emotions, dispositions, engagement, etc.) as masking to protect one's privacy means that the features that differentiate one's inner state are also disguised. Naive speech content anonymization mostly utilizes blanking or silencing to remove personal information from the voice stream. But to guarantee anonymity, this has to be done manually and is mostly used for recorded research data stored locally. Winkler and Buchmann (2018) discussed a breakdown of the voice interaction and suggest the parts of the voice that should be masked. Their study provides a prototype technique to anonymize voice called a Dummy-Based Anonymization Technique. This approach was developed to obfuscate the speaker's identity for online services is presented in Winkler and Buchmann (2018). The idea is to feed the voice assistant with voice samples from an internal database containing dummy requests. Dummy requests in that context mean that requests similar to real requests are made but with changed privacy information, for a calendar task obfuscation this would lead to dummy requests that would request Alexa to read, add and remove calendar entries at different times, by different voices, without having an impact on the correctness of the service.

This approach allows to obscure private application data enabling an identification of the user. As also a person’s habits or activity times can be seen as personal data (WP136, 2007). One disadvantage of this approach is that the personal information is transmitted nevertheless and could be “filtered” out of the stream of dummy data (sound volume, direction, intonation, background noise). Furthermore, regarding conversational analyses, this approach prevents to study human-machine interaction, as every “dialog” between the voice assistant and the anonymization device is just a coincidental sequence of random meaningless voice samples.

A possible solution for this issue would be to apply the principle of data sparsity (closely related to data minimization required by the GDPR): To process only the kind of data that is necessary for a specific use case. This means that the dataset owner holds the full acoustic data but only extracts

and shares certain features for all subjects with enriched data so that an identification is avoided. This can be done by using broader data classifications. For example, Silber-Varod et al. (2019) used only the acoustic signal and speaker's sex, as the data was the property of a private industrial company.

Data de-anonymization, on the other hand, is the method of re-identifying the individual from anonymous data together with publicly available information. Especially as the amount of publicly available information is growing this is a major privacy concern. For re-identification based on textual data not much information is needed generally. The combination of zip codes, birth date, and sex from anonymized data together with voter databases is enough to identify individuals (Narayanan & Shmatikov, 2008; Ohm, 2009).

Besides personal information also face and voice data can be used to identify a person. In the domains of speech communication, human-machine interaction, or speech technologies for healthcare, the speech signal is the core material scientists are exploring as it carries acoustic and word-based cues that are used for diagnosis and behavioral patterns. Moreover, for multi-modal corpora, facial patterns are analyzed as well and this information has to be treated with care as it carries personal data that can be used for biometric identification systems (Kinnunen & Li, 2010; Li & Jain, 2011). Especially naive techniques have to be avoided. Either, because they can be subverted (see parrot recognition (Newton et al., 2005)) or they do not maintain sufficient speaker characteristics needed for scientific analyses as these algorithms only aim to preserve a good intelligibility of the speech content while enabling the best anonymity. The intelligibility of emotional or dispositional factors is not focused on so far. Anonymization techniques need to find a good trade-off between a good disguise of privacy-sensitive information and the possibility of analyzing scientifically the interaction in terms of emotional content and attitude.

Challenges of "In the Wild" Data Collection

Internet-Crawling

With the development and popularity of the World Wide Web (WWW), information has never been more accessible and at the same time, the internet is the largest unstructured database known to us (Patel & Bhatt, 2014). Data mining is the extraction of relevant information from a database (structured or unstructured), using parameters or predictive algorithms (Patel & Bhatt, 2014). With the aid of a crawler software, an agent capable of analyzing and extracting information from databases, researchers may narrow down the amount of information they receive from the internet to manageable amounts. For example, organizations mine social networks to learn about their workers' relationships and to identify knowledge hubs (Fire & Puzis, 2016); epidemiologists had mined over 79,000 articles on cancer to find the connection between female parity and cancer (Tourassi et al., 2015). Data mining methodologies and tools are not limited to web pages only but also apply to other unstructured databases such as YouTube and data collected from publicly available sources.

Public Data Collection

Nowadays, many researchers wanting to collect "natural" interactions create captivating experiments at great expense to let the participants forget about the observation (Rösner et al.,

2017), or install technical devices in public spaces to engage people to interact with it. Alternatively, for the collection of public interaction data, researchers usually conduct lab-experiments where participants are invited to intact with a technical device. Thereby the device can either be acting autonomously or driven by a Wizard of Oz technique. Whilst the first option requires a lot of development work beforehand, the latter option requires extensive support during the execution of the experiment. Furthermore, knowing that they are part of an experiment changes the behavior of the participants: This effect is known as the *Observer Effect*. Thereby participants change aspects of their behavior in response to their awareness of being observed (Parsons, 1974). This change in the participants' behavior contradicts, however, an investigation of the undisturbed "natural" interaction and public data collection studies are therefore conducted.

The main purpose of these public data recordings is to understand unconstrained conversations, analyze feedback mechanisms, and dialog repair strategies.

The main challenges include:

A. Getting people to interact with a machine (robot) – just having an interactive device does not mean people will interact with it. Appropriate conditions must, therefore, be taken to stimulate the interaction, through an appropriate environment (exhibition) or unique design (Cuteness).

B. Provide a useful interaction – getting individuals to interact with a talking system, with no prior experience, requires an appealing design or convincing interaction.

C. Enable the technical system to cope with unexpected behavior – as the individuals are not observed, they tend to explore the capabilities of the system.

D. Avoid collecting background data – as the setting within public spaces cannot be completely controlled, the system should only record individuals interacting with it.

Use Cases

YouTube Crawler

Data collection process is still tedious and time-consuming, especially when one needs human subjects or biological data such as voice or bio-signals. Some tools were designed to bypass the problem by constructing databases of biodata such as Physionet (Goldberger et al., 2000), for complex bio samples, or Vocalid (Bunnell & Patel, 2014), for voice samples. However, such databases may not be up to date or may not exactly suit the researcher's needs. The solution for this problem came from web crawlers, *bots*, whose purpose is to systematically browse the internet and gather requested data (for example, Ben-David, 2019).

Many studies use YouTube data mining to gain information from billions of users. For example, Severyn et al. (2016) gathered users' opinions on global events from multi-lingual YouTube comments; Khosla (2016) demonstrated how data generated from YouTube can be mined and utilized to make targeted, real-time and informed decisions. YouTube became popular for research probably due to a variety of reasons: YouTube videos can replace questioners via dedicated channels; users' private meta-data can be used, it enables an analysis of view time, and division of data by country of origin.

YouTube is a vast source of information in audio and video-based research. According to YouTube press statistics, YouTube has over a billion registered users; 300 hours of video are uploaded to YouTube every minute; almost five billion videos are watched on YouTube every single day; and YouTube gets over 30 million visitors per day. Moreover, the platform contains valuable meta-data, which is collected during the video lifetime. Furthermore, the YouTube API V3 was released by Google in 2014, as part of the Google developer project. It allows developers to access video statistics and YouTube channels' data (note that the API also allows users to manipulate videos and channel programmatically) (Mosconi et al., 2019).

In the following, we will present the YouTube Metadata Information Retriever (YMIR) (Carmi, 2018), as a use case to demonstrate that collecting data from YouTube for research purposes does not necessarily entail breaking the GDPR. The YMIR was designed to browse and capture metadata of YouTube videos (Carmi, 2018). The crawler is a minimized version of existing crawlers (e.g., Khosla, 2016), dedicated for YouTube research. By limiting the scope of the crawler, Carmi (2018) increased the run time response and improved the relevance of the results. Moreover, by creating a simplified crawler for data collection, a database may be rebuild or updated with ease, expediting the above process, and bypassing the need for manual labor. Although the code is not extensive, the main challenge was to create a simple tool that can be altered to suit digital humanities and social sciences' needs. YMIR can retrieve metadata from a given list of YouTube videos, a YouTube channel name, or a search query. For each URL, it collects the video's title, description, number of comments, number of views, Likes and Dislikes, the date it was published, and its duration. The tool also produces a separate file with comments metadata: for each comment, the crawler collects its content, viewer's rating, Likes, and more.

The design of YMIR consists of three main parts:

1. Initialization (GUI and YSP (YouTube Service provider);

2. Search (Video results thread and Comment results thread);

3. Save (CSV builder and Comment Indexer).

Initialization – input: the user provides a link to his/her YouTube account which enables to access statistics from his/her own channel or from search queries he makes. With the project key for account verification the YSP connects and sends all queries to the YouTube servers. User input can be either a list of YouTube videos URL or a search phrase (Figure 1).

Search: acquires database parameters. Search client – generates and receives results from the service provider, either for videos or comments.

Save – output: Variables indexer in a CSV format. The data received from the API is divided into public and private data. Public data (Views, Upload date, Likes, Dislikes, Comments, etc.) can be accessed by anyone with a key and may be viewed directly on the video webpage. Private data (Views per country, Percentage of video watched (as users may not watch the whole video), Average View Duration, The number of time the video was added to a playlist, google analytics, etc.) is available only to the account associated with the video and contain further analytics and information for channel managers. One of the advantages of this tool is that the same database may be recreated using the same topics repeatedly.

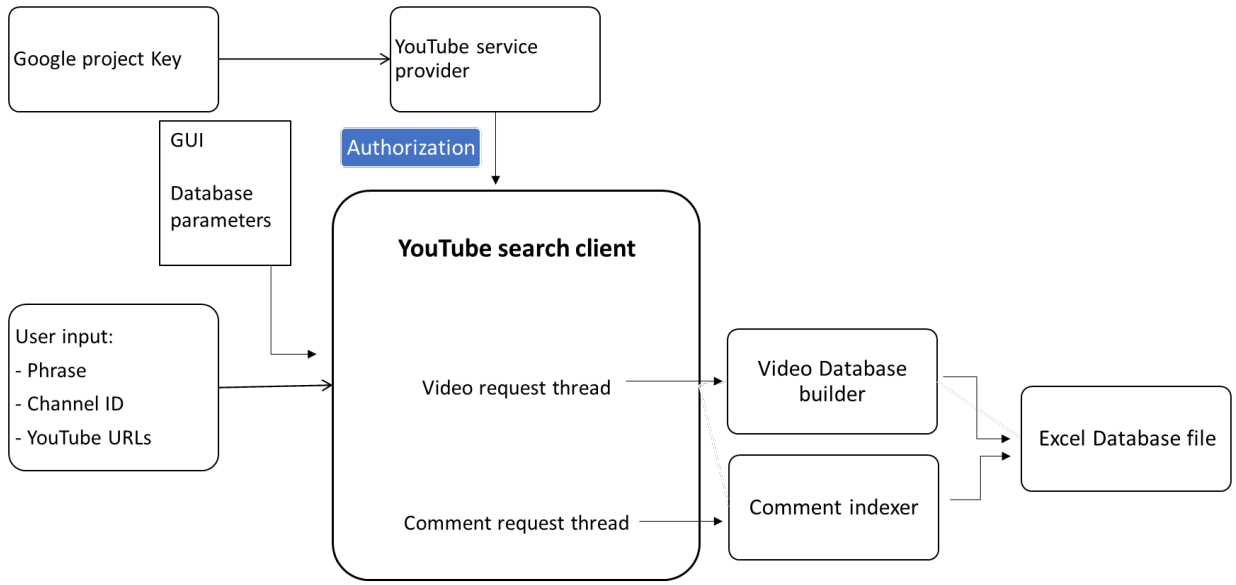


Figure 1. The design of YouTube Metadata Information Retriever (for details see Carmi, 2018)

Public Interaction Recording

A very first approach to public interaction recordings is described in Gustafson and Bell (2000), where an experimental Swedish dialog system having an animated talking agent, was exposed to the general public for about six months. Another public platform, the robot Herme, was installed for three months at a science fair in Dublin in 2011 (Han et al., 2012). In that study, visitors volunteered to talk with a robot, thus providing the researchers with audio-visual samples of informal, chatty dialogues. Recent examples are of Lopatovska and Oropeza (2018), who featured voice assistants of Amazon's Alexa in a public academic space for 24 days; of Bernstein (2016) and Moore et al. (2017) who utilized a public interaction experiment at an art museum in Philadelphia; and of Siegert (2020) who collected interactions with Amazon's Alexa at a science fair in Germany and Austria for about 6 months.

All of these systems have used different strategies to solve the first challenge – getting people to interact. The Gustafson system use an animated agent that encourages visitors to interact with the system by its human-like appearance and personality. The Herme robot uses a different strategy and relies on its unique design (cuteness). Lopatovska and Oropeza (2018) installed Amazon's Alexa in the main public hall of the Pratt Institute School of Information, where they expected graduate students to cross by. They assumed that young students are more likely to interact with a modern voice interface. Other systems make use of appropriate conditions of the surrounding and space, to stimulate the interaction (a museum (Moore et al., 2017) & a science exhibition (Siegert, 2020)). Using an interactive voice assistant alone is not sufficient. Moore, Pan, and Engineer (2017) pointed out that people were not aware of the technology and therefore did not engage with it. Additionally, some visitors, both individually and in a group, were too shy to ask Alexa questions. Therefore, a special skill was developed, while the names and titles of non-English artists and artwork are trained to Alexa. The system in Siegert (2020), although a generic Alexa

system without a special skill, could be engaged with several visitors, as the system actively committed the shortcomings of the Alexa system. It was presented as a part of a science fair, and the Alexa system was part of an interactive quiz game to demonstrate that Alexa is not able to be used as an answer system for a general knowledge quiz.

This directly brings us to the second challenge – to provide an interaction with users having no prior experience with talking systems. To this end, a design of an appealing object or a convincing interaction was found effective (Moore et al., 2017). For example, Gustafson and Bell (2000)'s system attracted visitors due to its animated character, while the Herme robot had a quite cute look, which was especially aimed at children, generally more adventurous.

The third challenge is to enable the technical system to cope with unexpected behavior. It was found out that when users are not observed, they tend to explore the capabilities of the system. This could be either solved by employing a system update (recognition lexicon, semantic analyzer, system output) after several months of usage by analyzing the so far recorded interactions and by an additional simplification of the conversation. For example, by reacting to keywords linked to specific dialog contexts, as done in the Gustafson and Bell (2000)'s system. Another even more pragmatic solution is implemented in the Herme robot: Due to the extremely noisy environment no attempt was made to incorporate any traditional form of speech recognition in the dialogue interface. Instead, the robot waited "an appropriate amount of time" to take the next step through the dialogue to establish and maintain the illusion of attention and back-channeling (Han et al., 2012).

In terms of the GDPR, the last challenge is the most severe one. Hereby two factors have to be taken into account. First, only collect data of users who are interacting with the system. Second, sharing the recorded data should be possible. For the first factor, most researchers rely on a system activation that is initiated by the participant. Therefore, it avoids recording other conversations (Gustafson & Bell, 2000; Lopatovska & Oropeza, 2018; Moore et al., 2017; Siegert, 2020). Sometimes, an information sign was placed next to the system, to inform the visitors about the ongoing study (Lopatovska & Oropeza, 2018; Siegert, 2020). Han et al. (2012) explicitly asked the visitors to sign a consent form, but they also recorded the video of the visitors. As we have stated in the definitions, anonymized data do not fall under the GDPR, and public recorded data, with no additional information about the participants, allegedly are a good source for anonymous data. Nonetheless, voice and facial information can be used to identify the participants and a good anonymization technique that will prevent de-anonymization and will still allow conversational analyses does not yet exist. Therefore, the data is either not shared, which is an unsatisfactory situation for the research community, or seldom available for collaborative research, where the data provider holds full control over the data to assure the anonymity of the recorded participants. Only a few studies reported about approval from an Ethics Committee or a data security officer (Han et al., 2012; Siegert, 2020), which furthermore explains the strict data sharing policy of researchers.

Discussion

This paper presents the dilemma related to the privacy of audio and video data, compliance with the EU GDPR. GDPR and other EU's Ethics Guidelines (such as the Ethics Guidelines for

Trustworthy AI (European Commission, 2018)) set up strategies to reach privacy and data governance. In this paper, we wanted to highlight the relevance of the new regulations to the academia with regard to voice and video "in the wild" data collection. We focused on the implications of two issues: (1) individuals will eventually have full control over their own data, and (2) Individuals' data will be anonymized or at least pseudonymized.

The new data protection rules in the EU impose several steps to be taken by anyone who processes personal data, including for research purposes. According to the Privacy by Design principle (Article 25 of the GDPR), the processing, already at the conception phase, has to be designed in such a way as to respect the data protection principles set forth in Article 5 of the GDPR (such as lawfulness, data minimization, purpose and storage limitation, integrity and confidentiality). Most importantly, data subjects have to be provided with information about the processing in a concise, transparent, intelligible and easily accessible form, using clear and plain language (cf. Articles 12-14 of the GDPR). Other rights of data subjects, such as access (Article 15 of the GDPR) and erasure (Article 17 of the GDPR), also need to be observed. Furthermore, several instruments such as a record of data processing activities (Article 30 of the GDPR) or a record of data breaches (Article 33(5) of the GDPR) need to be implemented in research institutions.

The GDPR does not apply to anonymized data, but the standard for anonymization is high: all the means reasonably likely to be used to identify the data subjects should be considered, and the process should be irreversible. Although a lot of effort has been conducted in the anonymization of data, speech data despite speaker identity has been rarely investigated.

Pseudonymization, which, by definition, is reversible, does not place the data beyond the scope of the GDPR, but can be regarded as a safeguard for rights and freedoms of data subjects. The specific rules applicable to research may still evolve, as the practice of national data protection authorities clarifies, and as international organizations adopt new guidelines (such as those adopted recently by the European Data Protection Supervisor (2020)). A Code of conduct, such as the one proposed for the language community (Kamocki et al., 2018), could also harmonize and simplify the current practice in the concerned sectors.

We mentioned above recent attempts to obfuscate voice assistants regarding personal information (Winkler & Buchmann, 2018). More promising is the use of privacy-preserving based on Blockchain technology, for which studies showed that it can prevent the disclose of training data and model (Weng et al., 2019). Still, the applicability for voice data has yet to be demonstrated.

Conclusion

Although academia can guarantee that data concerning the individuals will not be used to harm or discriminate against them, supporting full control of the personal data by both parties, is a goal that academia should be prepared to. Regulations are strict and can cause organizations to reinvent strategy and invest more in managing, storing and processing information. Unlike information security standards, the GDPR is not a standard but a statement by the organization that it complies with the required regulations. Should academic institution authorities acquire new research management skills? The current situation is that despite not sharing personal data or not collecting at all, at least in voice interaction the solution is yet to be found. Thus, research especially in the

term of speaker identity anonymization while preserving speaker states (emotions, dispositions, engagement, etc.) is still a challenge. To fulfill both aims, keeping all the information in the voice or visual signal about the speaker to allow scientific analyses on one hand while not revealing the speaker's identity on the other hand, we suggest a combined strategy from the legal and academic perspectives. In this sense, it has to be determined by the researcher's scientific goals to which extent anonymization techniques still maintain specific speaker states or whether end-to-end learning or auto-encoders can preserve the speaker state information while obfuscating the speaker's identity. Similarly, the evolvement of new rules for pseudonimization or a code of conduct for data recordings could lead to a harmonized practice for academia determining the prerequisites for GDPR-compliant data recordings, storage and exchange.

References

Aggarwal C. C., & Yu P. S. (2008). A general survey of privacy-preserving data mining models and algorithms. In C. C. Aggarwal & P. S. Yu (Eds), *Privacy-preserving data mining. Advances in Database Systems* (vol. 34, pp. 11-52). Springer. https://doi.org/10.1007/978-0-387-70992-5_2

Batliner, A., & Schuller, B. (2014). More than fifty years of speech and language processing-the rise of computational paralinguistics and ethical demands. https://www5.informatik.uni-erlangen.de/Forschung/Publikationen/2014/Batliner14-MTF.pdf

Ben-David, A. (2019). 2014 not found: A cross-platform approach to retrospective web archiving. *Internet Histories*, *3*(3-4), 316-342.

Bernstein, S. (2016). What can Alexa teach us about the Barnes? [Blog post]. https://medium.com/barnes-foundation/what-can-alexa-teach-us-about-the-barnes-21154d68700c#.v08qjcf4k

Bunnell, H. T., & Patel, R. (2014). VocaliD: Personal voices for augmented communicators. *The Journal of the Acoustical Society of America*, *135*(4), 2390-2390. https://doi.org/10.1121/1.4877902

Carmi, N. (2018). YMIR - YouTube metadata information. https://github.com/nehorayc/YMIR

Dudzik, B., Jansen M., Burger F., Kaptein, F., Broekens, J., Heylen. D. K. J., et al. (2019). Context in human emotion perception for automatic affect detection: A survey of audiovisual databases. *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction*, 206-212. https://doi.org/10.1109/ACII.2019.8925446

European Commission (2018). Ethics guidelines for trustworthy AI. https://ec.europa.eu/futurium/en/ai-alliance-consultation

European Data Protection Supervisor (2020). A preliminary opinion on data protection and scientific research. https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf

Fang, F., Wang, X., Yamagishi, J., Echizen, I., Todisco, M., Evans, N., et al. (2019). Speaker anonymization using x-vector and neural waveform models. *Proceedings of the 10th ISCA Speech Synthesis Workshop*, 20-22. https://doi.org/10.21437/SSW.2019-28

Fire, M., & Puzis, R. (2016). Organization mining using online social networks. *Networks and Spatial Economics*, *16*(2), 545-578. https://doi.org/10.1007/s11067-015-9288-4

Goldberger, A. L., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *circulation*, *101*(23), e215-e220. https://doi.org/10.1161/01.CIR.101.23.e215

Gustafson, J., & Bell, L. (2000). Speech technology on trial: Experiences from the August system. *Natural Language Engineering*, *6*(3-4), 273-286. https://doi.org/10.1017/S13513249000 02485

Han, J. G., Gilmartin, E., De Looze, C. Vaughan, B., & Campbell, N. (2012). Speech & multimodal resources: The Herme database of spontaneous multimodal human-robot dialogues. *Proceedings of the 8th Language Resources and Evaluation Conference,* 1328-1331.

Kamocki, P., Ketzan, E., Wildgans, J., & Witt, A. (2018). Toward a CLARIN data protection code of conduct, In I. Skadina & M. Eskevich (Eds), *CLARIN annual conference 2018 proceedings,* 49-52.

Khosla, C. (2016). Youtube data analysis using hadoop [Unpublished Doctoral dissertation]. California State University, Sacramento. https://bit.ly/2Wb8EQz

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, *52*(1), 12-40. https://doi.org/10.1016/ j.specom.2009.08.009

Korshunov, P., Araimo, C., De Simone, F., Velardo, C., Dugelay, J. L., & Ebrahimi, T. (2012). Subjective study of privacy filters in video surveillance. *Proceedings of the 14th International Workshop on Multimedia Signal Processing,* 378-382. https://doi.org/ 10.1109/MMSP.2012.6343472

Krämer, N. C., von der Pütten, A., & Eimler, S. (2012). Human-agent and human-robot interaction theory: Similarities to and differences from human-human interaction. In M. Zacarias & J. V. de Oliveira (Eds), *Human-computer interaction: The agency perspective* (Vol. 396, pp. 215-240). Studies in Computational Intelligence. Springer. https://doi.org/10.1007/978-3-642-25691-2_9

Letournel, G., Bugeau, A., Ta, V. T., & Domenger, J. P. (2015). Face de-identification with expressions preservation. *Proceedings of the International Conference on Image Processing,* 4366-4370. https://doi.org/10.1109/ICIP.2015.7351631

Li, S. Z., & Jain, A. K. (2011). *Handbook of face recognition* (Vol. 1). Springer.

Lopatovska, I., & Oropeza, H. (2018). User interactions with "Alexa" in public academic space. In L. Freund (Ed.), *Proceedings of the association for information science and technology*, *55*(1), 309-318. Wiley. https://doi.org/10.1002/pra2.2018.14505501034

Moore, S., Pan, D., & Engineer, M. (2017). A case study on using voice technology to assist the museum visitor. *MW17 Museums and the Web 2017.*

Morse, J. M. (1991). Subjects, respondents, informants, and participants? *Qualitative Health Research*, *1*(4), 403-406. https://doi.org/10.1177/104973239100100401

Mosconi, G., Li, Q., Randall, D., Karasti, H., Tolmie, P, Barutzky, J., et al. (2019). Three gaps in opening science. *Computer Supported Cooperative Work*, *28*, 749-789. https://doi.org/10.1007/s10606-019-09354-z

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *Proceedings of the IEEE Symposium on Security and Privacy,* 111-125. https://doi.org/10.1109/SP.2008.33

Newton, E. M., Sweeney, L., & Malin, B. (2005). Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering*, *17*(2), 232-243. https://doi.org10.1109/TKDE.2005.32

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, *57*, 1701. https://ssrn.com/abstract=1450006

Parsons, H. M. (1974). What happened at hawthorne? *Science*, *183*(4128), 922-932. https://doi.org/10.1126/science.183.4128.922

Patel, R., & Bhatt, P. (2014). A survey on semantic focused web crawler for information discovery using data mining technique. *International Journal for Innovative Research in Science and Technology*, *1*(7), 168-170.

Pobar, M., & Ipšić, I. (2014). Online speaker de-identification using voice transformation. *Proceedings of the 37th International Convention on Information and Communication Technology, Electronics and Microelectronics,* 1264-1267. https://doi.org/10.1109/MIPRO.2014.6859761

Rösner, D., Frommer, J., Wendemuth, A., Bauer, T., Günther, S., Haase, M., et al. (2017). The LAST MINUTE corpus as a research resource: From signal processing to behavioral analyses in user-companion interactions. In S. Biundo & A. Wendemuth (eds), *Companion technology* (pp. 277-299). *Cognitive Technologies series*. Springer. https://doi.org/10.1007/978-3-319-43665-4_14

Severyn, A., Moschitti, A., Uryupina, O., Plank, B., & Filippova, K. (2016). Multi-lingual opinion mining on YouTube. *Information Processing & Management*, *52*(1), 46-60. https://doi.org/10.1016/j.ipm.2015.03.002

Siegert, I. (2020). Alexa "in the wild" – Collecting unconstrained conversations with a modern voice assistant in a public environment. Proceedings of the 12th Language Resources and Evaluation Conference, 608-612.

Silber-Varod, V., Lerner, A., Carmi, N., Amit, D., Guttel, Y., Orlob, C., et al. (2019). Computational modelling of speech data integration to assess interactions in b2b sales calls. *Proceedings of the IEEE 5th International Conference on Big Data Intelligence and Computing,* 152-157. https://doi.org/10.1109/DataCom.2019.00031

Sveningsson Elm, M. (2008). How do various notions of privacy influence decisions in qualitative internet research? In A. N. Markham & N. K. Baym (Eds), *Internet inquiry: Conversations about method* (pp. 69-87). Sage Publications. https://doi.org/10.4135/9781483329086.n7

Tourassi, G., Yoon, H. J., Xu, S., & Han, X. (2015). The utility of web mining for epidemiological research: studying the association between parity and cancer risk. *Journal of the American Medical Informatics Association*, *23*(3), 588-595. https://doi.org/10.1093/jamia/ocv141

Truong, K. P., van Leeuwen, D. A., & Neerincx, M. A. (2007). Unobtrusive multimodal emotion detection in adaptive interfaces: Speech and facial expressions. In D. D. Schmorrow, & L. M. Reeves (Eds), *Foundations of Augmented Cognition* (Vol. 4565, pp. 354-363). Springer. https://doi.org/10.1007/978-3-540-73216-7_40

Weng, J., Weng, J., Zhang, J., Li, M., Zhang, Y., & Luo, W. (2019). Deepchain: Auditable and privacy-preserving deep learning with blockchain-based incentive. *IEEE Transactions on Dependable and Secure Computing*. https://doi.org/10.1109/TDSC.2019.2952332

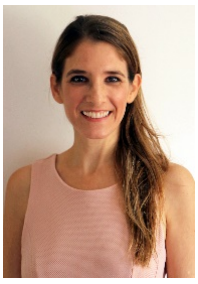
Winkler, K., & Buchmann, E. (2018). Dummy-based anonymization for voice-controlled IoT devices. *Proceedings of the 12th International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies,* 1-8. https://www.thinkmind.org/index.php?view=article&articleid=ubicomm_2018_1_10_10043

WP136. (2007). Article 29 data protection working party, Opinion 4/2007 on the concept of personal data, adopted on 20 June 2007 (WP136).

WP216. (2014). Article 29 data protection working party, Opinion 05/2014 on anonymization techniques, adopted on 10 April 2014 (WP216).

Authors Biographies

Ingo Siegert, Ph.D. Assistant professor for Mobile Dialog Systems at the Otto von Guericke University Magdeburg. Research interests and publications focus on signal-based analyses and interdisciplinary investigations of human-computer interaction in terms of addressee detection and the utilization of further interaction patterns, such as filled pauses or discourse particles. He has published 90+ peer reviewed papers on several conferences and various journals and is co-organizer of several workshops and conferences.



Vered Silber-Varod, Ph.D. Director of the Open Media and Information Lab (OMILab), The Open University of Israel. Former Research Fellow at The Research Center for Innovation in Learning Technologies, The Open University of Israel. Research interests and publications focus on various aspects of speech sciences, with expertise in speech prosody, acoustic phonetics, speech communication and text analytics. Honored to be part of ISCA'S WomenNSpeech list. Currently the treasurer-Secretary, The Haiim B. Rosen Israeli Linguistic Society.

Nehoray Carmi, M.Sc. Data scientist at the Open Media and Information Lab (OMILab). B.Sc. in Electrical Engineering from Jerusalem collage of technology and M.Sc. in Computer Science from the Open University of Israel. Adjunct instructor at the Open University of Israel, and R&D advisor in the private market. His research explores the connection between machine learning models and human psychology.

Pawel Kamocki, Dr. iur Researcher at the Leibniz Institut for Deutsche Sprache, trained in IT law and linguistics. His research interests include legal issues in machine translation, access and control of language data, and anonymization/pseudonymization of language resources.