ON THE CHANGING REGULATIONS OF PRIVACY AND PERSONAL INFORMATION IN MIR

Pierre Saurel
Université Paris-Sorbonne
pierre.saurel
@paris-sorbonne.fr

Francis Rousseaux
IRCAM
francis.rousseaux
@ircam.fr

Marc Danger
ADAMI
mdanger
@adami.fr

ABSTRACT

In recent years, MIR research has continued to focus more and more on user feedback, human subjects data, and other forms of personal information. Concurrently, the European Union has adopted new, stringent regulations to take effect in the coming years regarding how such information can be collected, stored and manipulated, with equally strict penalties for being found in violation of the law.

Here, we provide a summary of these changes, consider how they relate to our data sources and research practices, and identify promising methodologies that may serve researchers well, both in order to be in compliance with the law and conduct more subject-friendly research. We additionally provide a case study of how such changes might affect a recent human subjects project on the topic of style, and conclude with a few recommendations for the near future.

This paper is not intended to be legal advice: our personal legal interpretations are strictly mentioned for illustration purpose, and reader should seek proper legal counsel.

1. INTRODUCTION

The International Society for Music Information Retrieval addresses a wide range of scientific, technical and social challenges, dealing with processing, searching, organizing and accessing music-related data and digital sounds through many aspects, considering real scale use-cases and designing innovative applications, exceeding its academic-only initiatory aims.

Some recent Music Information Retrieval tools and algorithms aim to attribute authorship and to characterize the structure of style, to reproduce the user's style and to manipulate one's style as a content [8], [1]. They deal for instance with active listening, authoring or personalised reflexive feedback. These tools will allow identification of users in the big data: authors, listeners, performers.

As the emerging MIR scientific community leads to industrial applications of interest to the international business (start-up, Majors, content providers, platforms) and to experimentations involving many users in living labs (for MIR teaching, for multicultural emotion comparisons, or for MIR user requirement purposes) the identification of legal issues becomes essential or strategic.

Legal issues related to copyright and Intellectual Property have already been identified and expressed into Digital Rights Management by the MIR community [2], [7], when those related to security, business models and right to access have been expressed by Information Access [4], [11]. Privacy is another important legal issue. To address it properly one needs first to classify the personal data and processes. A naive classification appears when you quickly look at the kind of personal data MIR deals with:

- User's comments, evaluation, annotation and music recommendations are obvious personal data as long as they are published under their name or pseudo;
- Addresses allowing identification of a device or an instrument and Media Access Control addresses are linked to personal data;
- Any information allowing identification of a natural person, as some MIR processes do, shall be qualified as personal data and processing of personal data.

But the legal professionals do not unanimously approve this classification. For instance the Court of Appeal in Paris judged in two decisions (2007/04/27 and 2007/05/15) that the Internet Protocol address is not a personal data.

2. WHAT ARE PROCESSES OF PERSONAL DATA AND HOW THEY ARE REGULATED

A careful consideration of the applicable law of personal data is necessary to elaborate a proper classification of MIR personal data processes taking the different international regulations into account.

2.1 Europe *vs.* United States: two legal approaches

Europe regulates data protection through one of the highest State Regulations in the world [3], [9] when the United States lets contractors organize data protection through agreements supported by consideration and entered into voluntarily by the parties. These two approaches are deeply divergent. United States lets companies specify their own rules with their consumers while Europe enforces a unique regulated framework on all companies providing services to European citizens. For instance any company in the United States can define how long they keep the personal data, when the regulations in Europe would specify a maximum length of time the personal data is to be stored. And this applies to any company offering the same service.

A prohibition is at the heart of the European Commission's Directive on Data Protection (95/46/CE – The Directive) [3]. The transfer of personal data to non-European Union countries that do not meet the European Union adequacy standard for privacy protection is strictly forbidden [3, article 25][1]. The divergent legal approaches and this prohibition alone would outlaw the proposal by American companies of many of their IT services to European citizens. In response the U.S. Department of Commerce and the European Commission developed the Safe Harbor Framework (SHF) [6], [14]. Any non-European organization is free to self-certify with the SHF and join.

A new Proposal for a Regulation on the protection of individuals with regard to the processing of personal data was adopted the 12 March 2014 by the European Parliament [9]. The Directive allows adjustments from one European country to another and therefore diversity of implementation in Europe when the regulation is directly enforceable and should therefore be implemented directly and in the same way in all countries of the European Union. This regulation should apply in 2016. This regulation enhances data protection and sanctions to anyone who does not comply with the obligations laid down in the Regulation. For instance [9, article 79] the supervisory authority will impose, as a possible sanction, a fine of up to one hundred million Euros or up to 5% of the annual worldwide turnover in case of an enterprise.

2.2 Data protection applies to any information concerning an identifiable natural person

Until French law applied the 95/46/CE European Directive, personal data was only defined considering sets of data containing the name of a natural person. This definition has been extended; the 95/46/CE European Directive (ED) defines 'personal data' [3, article 2] as: "*any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity*".

For instance the identification of an author through the structure of his style as depending on his mental, cultural or social identity is a process that must comply with the European data privacy principles.

2.3 Safe Harbor is the Framework ISMIR affiliates need not to pay a fine up to hundreds million Euros

Complying with Safe Harbor is the easiest way for an organization using MIR processing to fulfill the high level European standard about personal data, to operate worldwide and to avoid prosecution regarding personal data. As explained below any non-European organization may enter the US – EU SHF's requirement and publicly declare that they do so. In that case the organization must develop a data privacy policy that conforms to the seven Safe Harbor Principles (SHP) [14].

First of all organizations must identify personal data and personal data processes. Then they apply the SHP to these data and processes. By joining the SHF, organizations must implement procedures and modify their own information system whether paper or electronic.

Organizations must notify (P1) individuals about the purposes for which they collect and use information about them, to whom the information can be disclosed and the choices and means offered for limiting its disclosure. Organizations must explain how they can be contacted with any complaints. Individuals should have the choice (P2) (opt out) whether their personal information is disclosed or not to a third party. In case of sensitive information explicit choice (opt in) must be given. A transfer to a third party (P3) is only possible if the individual made a choice and if the third party subscribed to the SHP or was subject to any adequacy finding regarding to the ED. Individuals must have access (P4) to personal information about them and be able to correct, amend or delete this information. Organizations must take reasonable precautions (P5) to prevent loss, misuse, disclosure, alteration or destruction of the personal information. Personal information collected must be relevant (P6: data integrity) for the purpose for which it is to be used. Sanctions (P7 enforcement) ensure compliance by the organization. There must be a procedure for verifying the implementation of the SHP and the obligation to remedy problems arising out of a failure to comply with the SHP.

3. CLASSIFICATION FOR MIR PERSONAL DATA PROCESSING

Considering the legal definition of personal data we can now propose a less naive classification of MIR processes and data into three sets: (i) nominative data, (ii) data leading to an easy identification of a natural person and (iii) data leading indirectly to the identification of a natural person through a complex process.

3.1 Nominative data and data leading easily to the identification of a natural person

The first set of processes deals with all the situations giving the name of a natural person directly. The second set deals with the cases of a direct or an indirect identification easily done for instance through devices.

In these two sets we find that the most obvious set of data concerns the "Personal Music Libraries" and "recommendations". Looking at the topics that characterize

[1] Argentina, Australia, Canada, State of Israel, New Zealand, United States – Transfer of Air Passenger Name Record (PNR) Data, United States – Safe Harbor, Eastern Republic of Uruguay are, to date, the only non-European third countries ensuring an adequate level of protection: http://ec.europa.eu/justice/data-protection/document/international-transfers/adequacy/index_en.htm

ISMIR papers from year 2000 to 2013, we find more than 30 papers and posters dealing with those topics as their main topic. Can one recommend music to a user or analyze their personal library without tackling privacy?

3.2 Data leading to the identification of a natural person through a complex process

The third set of personal data deals with cases when a natural person is indirectly identifiable using a complex process, like some of the MIR processes.

Can one work on "Classification" or "Learning", producing 130 publications (accepted contributions at ISMIR from year 2000 to year 2013) without considering users throughout their tastes or style? The processes used under these headings belong for the most part to this third set. Looking directly at the data without any sophisticated tool does not allow any identification of the natural person. On the contrary, using some MIR algorithms or machine learning can lead to indirect identifications [12].

Most of the time these non-linear methods use inputs to build new data which are outputs or data stored inside the algorithm, like weights for instance in a neural net.

3.3 The legal criteria of the costs and the amount of time required for identification

This third set of personal data is not as homogeneous as it seems to be at first glance. Can we compare sets of data that lead to an identification of a natural person through a complex process?

The European Proposal for a Regulation specifies the concept of "identifiability". It tries to define legal criteria to decide if an identifiable set of data is or is not personal data. It considers the identification process [9, recital 23] as a relative one depending on the means used for that identification: "*To determine whether a person is identifiable, account should be taken of all the means reasonably likely to be used either by the controller or by any other person to identify or single out the individual directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the individual, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration both available technology at the time of the processing and technological development.*"

But under what criteria should we, as MIR practitioners, specify when a set of data allows an easy identification and belongs to the second set or, on the contrary, is too complex or reaches a too uncertain identification so that we would not legally say that these are personal data? To answer these questions, we must be able to compare MIR processes with new criteria.

4. MANAGING THE TWO FIRST SETS

On an example chosen to be problematic (but increasingly common in the industry), we show how to manage personal data in case of a simple direct or indirect identification process.

4.1 Trends in terms of use and innovative technology

Databases of personal data are no more clearly identified. We can view the situation as combining five aspects, which lead to new scientific problems concerning MIR personal data processing.

Data Sources Explosion. The number of databases for retrieving information is growing dramatically. Applications are also data sources. Spotify for instance provides a live flow of music consumption information from millions of users. Data from billions of sensors will soon be added. This profusion of data does not mean quality. Accessible does not mean legal or acceptable for a user. Those considerations are essential to build reliable and sustainable systems.

Crossing & Reconciling Data. Data sources are no longer isolated islands. Once the user can be identified (cookie, email, customer id), it is possible to match, aggregate and remix data that was previously isolated.

Time Dimension. The web has a good memory that humans are generally not familiar with. Data can be public one day and be considered as very private 3 years later. Many users forget they posted a picture after a student party. And the picture has the misfortune to crop up again when you apply for a job. And it is not only a question of human memory: Minute traces collected one day can be exploited later and provide real information.

Permanent Changes. The general instability of the data sources, technical formats and flows, applications and use is another strong characteristic of the situation. The impact on personal data is very likely. If the architecture of the systems changes a lot and frequently, the social norms also change. Users today publicly share information that they would have considered totally private a few years earlier. And the opposite could be the case.

User Understandability and Control. Because of the complexity of changing systems and complex interactions users will less and less control over their information. This lack of control is caused by the characteristics of the systems and by the mistakes and the misunderstandings of human users. The affair of the private Facebook messages appearing suddenly on timeline (Sept. 2012) is significant. Facebook indicates that there was no bug. Those messages were old wall posts that are now more visible with the new interface. This is a combination of bad user understanding and fast moving systems.

4.2 The case of an Apache Hadoop File System (AHFS) on which some machine learning is applied

Everyone produces data and personal data without being always aware that they provide data revealing their identification. When a user tags / rates musical items [13], he gives personal information. If a music recommender exploits this user data without integrating privacy concepts, he faces legal issues and strong discontent from the users.

The data volume has increased faster than "Moore's law": This is what is meant by "Big Data". New data is generally unstructured and traditional database systems such as Relational Database Management Systems cannot handle the volume of data produced by users & machines & sensors. This challenge was the main drive for Google to define a new technology: the Apache Hadoop File System (AHFS). Within this framework, data and computational activities are distributed on a very large number of servers. Data is not loaded for computation, nor the results stored. Here, the algorithm is close to the data. This situation leads to the epistemological problem of separability into the field of MIR personal data processing: are all MIR algorithms (and for instance the authorship attribution algorithms) separable into data and processes? An answer to this question is required for any algorithm to be able to identify the set of personal data it deals with.

Now, let us consider a machine learning classifier/recommender trained on user data. In this sense, the algorithm is inseparable from the data it uses to function. And, if the machine is internalizing identifiable information from a set of users in a certain state (let say EU), it is then in violation to share the resulting function in a non-adequate country (let say Brazil) the EU if it was trained in, say, the US.

4.3 Analyzing the multinational AHFS case

Regarding to the European regulation rules [3, art. 25], you may not transfer personal data collected in Europe to a non-adequate State (*see* list of adequate countries above). If you build a multinational AHFS system, you may collect data in Europe and in US depending on the way you localized the AHFS servers. The European data may not be transferred to Brazil. Even the classifier would not legally be used in Brazil as long as it internalizes some identifiable European personal information.

In practice one should then localize the AHFS files and machine-learning processes to make sure no identifiable data will be transferred from one country with a specific regulation to another with another regulation about personal data. We call these systems "heterarchical" due to the blended situation of a hierarchical system (the global AHFS management) and the need of a heterogeneous local regulation.

To manage properly the global AHFS system we need a first analysis of the system dispatching the different files on the right legal places. Privacy by Design (PbD) is a useful methodology to do so.

4.4 Foundations Principals of Privacy by Design

PbD was first developed by Ontario's Information and Privacy Commissioner, Dr. Ann Cavoukian, in the 1990s, at the very birth of the future big data phenomenon. This solution has gained widespread international recognition, and was recently recognized as a global privacy standard.

According to its Canadian inventor[1], is PbD based on seven Foundation Principles (FP): PbD *"is an approach to protect privacy by embedding it into the design specifications of technologies, business practices, and physical infrastructures. That means building in privacy up front – right into the design specifications and architecture of new systems and processes. PbD is predicated on the idea that, at the outset, technology is inherently neutral. As much as it can be used to chip away at privacy, it can also be enlisted to protect privacy. The same is true of processes and physical infrastructure":*

- Proactive not Reactive (FP1): the PbD approach is based on proactive measures anticipating and preventing privacy invasive events before they occur;
- Privacy as the Default Setting (FP2): the default rules seek to deliver the maximum degree of privacy;
- Privacy embedded into Design (FP3): Privacy is embedded into the architecture of IT systems and business practices;
- Full Functionality – Positive Sum, not Zero-Sum (FP4): PbD seeks to accommodate all legitimate interests and objectives (security, etc.) in a "win-win" manner;
- End-to-End Security – Full Lifecycle Protection (FP5): security measures are essential to privacy, from start to finish;
- Visibility and Transparency — Keep it Open (FP6): PbD is subject to independent verification. Its component parts and operations remain visible and transparent, to users and providers alike;
- Respect for User Privacy — Keep it User-Centric (FP7): PbD requires architects and operators to keep the interests of the individual uppermost.

At the time of digital data exchange through networks, PbD is a key-concept in legacy [10]. In Europe, where this domain has been directly inspired by the Canadian experience, the EU[2] affirms: "*PbD means that privacy and data protection are embedded throughout the entire life cycle of technologies, from the early design stage to their deployment, use and ultimate disposal*".

4.5 Prospects for a MIR Privacy by Design

PbD is a reference for designing systems and processing involving personal data, enforced by the new European proposal for a Regulation [9, art. 23]. It becomes a method for these designs whereby it includes signal analysis methods and may interest MIR developers.

This proposal leads to new questions, such as the following: Is PbD a universal methodological solution about personal data for all MIR projects? Most of ISMIR contributions are still research oriented which doesn't mean

[1] http://www.ipc.on.ca/images/Resources/7foundationalprinciples.pdf

[2] "Safeguarding Privacy in a Connected World – A European Data Protection Framework for the 21st Century" COM (2012) 9 final.

that they fulfill the two specific exceptions [9, art. 83][1]. To say more about that intersection, we need to survey the ISMIR scientific production, throughout the main FPs. FP6 (transparency) and FP7 (user-centric) are usually respected among the MIR community as source code and processing are often (i) delivered under GNU like licensing allowing audit and traceability (ii) user-friendly. However, as long as PbD is not embedded, FP3 cannot be fulfilled and accordingly FP2 (default setting), FP5 (end-to-end), FP4 (full functionality) and FP1 (proactive) cannot be fulfilled even. Without any PbD embedded into Design, there are no default settings (FP2), you cannot follow an end-to-end approach (FP5), you cannot define full functionality regarding to personal data (FP4) nor be proactive. Principle of pro-activity (FP1) is the key. Fulfilling FP1 you define the default settings (FP2), be fully functional (FP4) and define an end-to-end process (FP5).

In brief is PbD useful to MIR developers even if it is not the definitive martingale!

5. EXPLORING THE THIRD SET

“Identifiability” is the potentiality of a set of data to lead to the identification of its source. A set of data should be qualified as being personal data if the cost and the amount of time required for identification are reasonable. These new criteria are a step forward since the qualification is not an absolute one anymore and depends specifically on the state of the art.

5.1 Available technology and technological development to take into account at this present moment

Changes in Information Technology lead to a shift in the approach of data management: from computational to data exploration. The main question is “What to look for?” Many companies build new tools to “make the data speak”. This is the case considering the trend of personalized marketing. Engineers using big data build systems that produce new personal dataflow.

Is it possible to stabilize these changes through standardization of metadata? Is it possible to develop a standardization of metadata which could ease the classification of MIR processing of personal data into identifying and non-identifying processes.

Many of the MIR methods are stochastic, probabilistic or designed to cost and more generally non-deterministic. On the contrary the European legal criteria [9, recital 23] (*see* above § 3.3) to decide whether a data is personal or not (the third set) seem to be much to deterministic to fit the effective new practices about machine learning on personal data.

This situation leads to a new scientific problem: Is there an absolute criterion about the identifiability of personal data extracted from a set of data with a MIR process? What characterizes a maximal subset from the big data that could not ever be computed by any Turing machine to identify a natural person with any algorithm?

5.2 What about the foundational separation in computer science between data and process?

Computer science is based on a strict separation between data and process (dual as these two categories are interchangeable at any time; data can be activated as a process and a process can be treated as a data).

We may wonder about the possibility of maintaining the data/process separation paradigm if i) the data stick to the process and ii) the legal regulation leads to a location of the data in the legal system in which those data were produced.

6. CONCLUSION

6.1 When some process lead to direct or indirect personal data identification

Methodological Recommendations. MIR researchers could first audit their algorithm and data, and check if they are able to identify a natural person (two first sets of our classification). If so they could use the SHF which could already be an industrial challenge for instance regarding Cyber Security (P5). Using the PbD methodology certainly leads to operational solutions in these situations.

6.2 When some process may lead to indirect personal data identification through some complex process

In many circumstances, the MIR community develops new personal data on the fly, using the whole available range of data analysis and data building algorithm. Then researchers could apply the PbD methodology, to insure that no personal data is lost during the system design.

Here PbD is not a universal solution because the time when data (on the one hand) and processing (on the other hand) were functionally independent, formally and semantically separated, has ended. Nowadays, MIR researchers currently use algorithms that support effective decision, supervised or not, without introducing ‘pure’ data or ‘pure’ processing, but building up acceptable solutions together with machine learning [5] or heuristic knowledge that cannot be reduced to data or processing: The third set of personal data may appear, and raise theoretical scientific problems.

Political Opportunities. The MIR community has a political role to play in the data privacy domain, by explaining to lawyers —joining expert groups in the US, UE or elsewhere— what we are doing and how we overlap with the tradition in style description, turning it into a computed style genetic, which radically questions the analysis of data privacy traditions, cultures and tools.

[1] (i) these processing cannot be fulfilled otherwise and (ii) data permitting the identification are kept separately from the other information, or when the bodies conducting these data respect three conditions: (i) consent of the data subject, (ii) publication of personal data is necessary and (iii) data are made public

Future Scientific Works. In addition to methodological and political ones, we face purely scientific challenges, which constitute our research program for future works. Under what criteria should we, as MIR practitioners, specify when a set of data allows an easy identification and belongs to the second set or on the contrary is too complex or allows a too uncertain identification so that we would say that these are not personal data? What characterizes a maximal subset from the big data that could not ever be computed by any Turing machine to identify a natural person with any algorithm?

7. REFERENCES

[1] S. Argamon, K. Burns, S. Dubnov (Eds): The Structure of Style, Springer-Verlag, 2010.

[2] C. Barlas: "Beating Babel - Identification, Metadata and Rights", Invited Talk, *Proceedings of the International Symposium on Music Information Retrieval*, 2002.

[3] Directive (95/46/EC) of 24 October 1995 *Official Journal L 281, 23/11/1995 P. 0031 - 0050* : http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML

[4] J.S. Downie, J. Futrelle, D. Tcheng: "The International Music Information Retrieval Systems Evaluation Laboratory: Governance, Access and Security", *Proceedings of the International Symposium on Music Information Retrieval*, 2004.

[5] A. Gkoulalas-Divanis, Y. Saygin, Vassilios S. Verykios: "Special Issue on Privacy and Security Issues in Data Mining and Machine Learning", *Transactions on Data Privacy*, Vol. 4, Issue 3, pp. 127-187, December 2011.

[6] D. Greer: "Safe Harbor - A Framework that Works", *International Data Privacy Law*, Vol.1, Issue 3, pp. 143-148, 2011.

[7] M. Levering: "Intellectual Property Rights in Musical Works: Overview, Digital Library Issues and Related Initiatives", Invited Talk, *Proceedings of the International Symposium on Music Information Retrieval*, 2000.

[8] F. Pachet, P. Roy: "Hit Song Science is Not Yet a Science", *Proceedings of the International Symposium on Music Information Retrieval*, 2008.

[9] Proposal for a Regulation on the protection of individuals with regard to the processing of personal data was adopted the 12 March 2014 by the European Parliament: http://www.europarl.europa.eu/sides/getDoc.do?type=TA&reference=P7-TA-2014-0212&language=EN

[10] V. Reding: "The European Data Protection Framework for the Twenty-first century", *International Data Privacy Law*, volume 2, issue 3, pp.119-129, 2012.

[11] A. Seeger: "I Found It, How Can I Use It? - Dealing With the Ethical and Legal Constraints of Information Access", *Proceedings of the International Symposium on Music Information Retrieval*, 2003.

[12] A.B. Slavkovic, A. Smith: "Special Issue on Statistical and Learning-Theoretic Challenges in Data Privacy", *Journal of Privacy and Confidentiality*, Vol. 4, Issue 1, pp. 1-243, 2012.

[13] P. Symeonidis, M. Ruxanda, A. Nanopoulos, Y. Manolopoulos: "Ternary Semantic Analysis of Social Tags for Personalized Music Recommendation", *Proceedings of the International Symposium on Music Information Retrieval*, 2008.

[14] U.S. – EU Safe Harbor: http://www.export.gov/safeharbor/eu/eg_main_018365.asp