How Should Digital Humanities Pioneers Manage Their Data Privacy Challenges?

Francis Rousseaux, Pierre Saurel

How Should Digital Humanities Pioneers Manage their Data Privacy Challenges?

Francis Rousseaux[1], Pierre Saurel[2]

[1] Institut de Recherche et de Coordination Acoustique Musique, Paris, France
francis.rousseaux@ircam.fr
[2] Paris-Sorbonne University, Paris, France
pierre.saurel@paris-sorbonne.fr

Abstract. Since Digital Humanities researchers and developers are regularly creating somehow industrial applications concerning international business, it is time for those communities to be aware and make the most of legacy constraints and opportunities.
For instance, let us consider the Computer Music state of the art, and particularly the Music Information Retrieval community and the wonderful algorithms it produces around authorship attribution and style recognition: even if some music style or authorship is finally attributed to some persons, this attribution may not result from a set of computable data somewhere reportable, the information being typically learned (in the sense of Machine Learning, more or less supervised) from dislocated data throughout the big data or the global database, and disseminated in the global programming system. Is this legal?
In Europe and worldwide, Privacy by Design (PbD) is the actual response to protect the fundamental right to data protection and to guarantee the free movement of personal data between business stakeholders or Member States.

Keywords: Digital Humanities; Computer Music; Music Information Retrieval; Machine Learning; Knowledge Management; Right Management; Privacy by Design; Big Data; Authorship Attribution; Style Recognition.

1 Introduction

Since the Digital Humanities researchers and developers are often about to create somehow industrial applications potentially concerning international business [12, 28, 43], it is time for those communities to be aware and make the most of legacy constraints and opportunities.

For instance, let us consider the Computer Music state of the art, and particularly the Music Information Retrieval (MIR) community, as it is typically structured and organized by the International Society of Music Information Retrieval (ISMIR, see http://www.ismir.net/).

ISMIR is now fifteen years old, and getting out of adolescence. After a fast-growing childhood, bottle-fed by the best IT algorithms and the most vitamin-rich

signal analysis methods, the International Society for Music Information Retrieval is now addressing 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, overflowing its academic-only initiatory aims.

As the emerging MIR scientific community reaches its disciplinary maturity and leads to potential industrial applications of interest to the international business (start-up, Majors, content providers, download or exchange platforms) and to large scale 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 even strategic.

Among legal issues, those related to copyright and Intellectual Property have already been identified and expressed into Digital Right Management subsystems by the MIR community [8, 27, 33], when those related to security, business models and right to access have been understood and expressed by Information Access [17, 36]. If those domains remain islands beside the MIR continent, Privacy, as another important part of legal issues, is not even a living island in the actual MIR archipelago.

However, Privacy and personal data issues are currently addressed by many Information Technology (IT) communities, aware of powerful and efficient paradigms like Privacy by Design.

2 Privacy by design: New Challenges in Big Data

Privacy by Design (PbD) was 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 made-in-Ontario solution has gained widespread international recognition, and was recently recognized as a global privacy standard.

2.1 What Has Changed within the Big Data?

The first radical change is obviously the web. Everyone produces data and personal data. However, the user is not always aware that he provides personal data allowing his identification. For instance, as described by [41], when a user tags or rates musical items, he gives personal information about himself. If a music recommender exploits this kind of user data without integrating strong privacy concepts, he faces legal issues and strong discontent from the users.

The volume of data has been increasing faster than the "Moore's law". This evolution is known as the concept of "Big Data". New data are generally unstructured and traditional database systems such as Relational Database Management Systems cannot handle the volume of data produced by users and by machines & sensors. This challenge was the main driver for Google to define a new technology: the Apache Hadoop File System. Within this framework, data and computational activities are distributed on a very large number of servers. Data are not loaded for being computed, and the result stored. Here, the algorithm is close to the data.

Databases of personal data are no more clearly identified. We can view the situation as combining five aspects:

- **Explosion of Data Sources.** The number of databases for retrieving information is growing dramatically. Applications are also data sources. *Spotify* for instance, embedded in Facebook, 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 islands. Once the user can be identified (cookie, email, customer id), it is possible to match, aggregate and remix data that were previously technically isolated.
- **Time Dimension.** The web has generally a good memory that humans are 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 bad idea 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 they would have considered totally private 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 be able to control their information. This lack of control is caused by the characteristics of the systems and by the mistakes and the misunderstanding 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.

Changes in the 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" and usually find personal data. This is the case considering the underlying trend of heavily personalized marketing. Engineers using the big data usually deal with existing personal data and build systems that produce new personal dataflow.

2.2 Foundations of Privacy by Design

Accordind to its inventor Ann Cavoukian[1], "Privacy by Design is an approach to protect privacy by embedding it into the design specifications of technologies,

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

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".

1. Proactive not Reactive; Preventative not Remedial. The PbD approach is characterized by proactive rather than reactive measures. It anticipates and prevents privacy invasive events before they happen. PbD does not wait for privacy risks to materialize, nor does it offer remedies for resolving privacy infractions once they have occurred — it aims to prevent them from occurring. In short, PbD comes before-the-fact, not after.
2. Privacy as the Default Setting. We can all be certain of one thing — the default rules! PbD seeks to deliver the maximum degree of privacy by ensuring that personal data are automatically protected in any given IT system or business practice. If an individual does nothing, their privacy still remains intact. No action is required on the part of the individual to protect their privacy — it is built into the system, by default.
3. Privacy Embedded into Design. PbD is embedded into the architecture of IT systems and business practices. It is not bolted on as an add-on, after the fact. The result is that privacy becomes an essential component of the core functionality being delivered. Privacy is integral to the system, without diminishing functionality.
4. Full Functionality — Positive-Sum, not Zero-Sum. PbD seeks to accommodate all legitimate interests and objectives in a positive-sum "win-win" manner, not through a dated, zero-sum approach, where unnecessary trade-offs are made. PbD avoids the pretense of false dichotomies, such as privacy vs. security, demonstrating that it is possible to have both.
5. End-to-End Security — Full Lifecycle Protection. PbD, having been embedded into the system prior to the first element of information being collected, extends securely throughout the entire lifecycle of the data involved — strong security measures are essential to privacy, from start to finish. This ensures that all data are securely retained, and then securely destroyed at the end of the process, in a timely fashion. Thus, PbD ensures cradle to grave, secure lifecycle management of information, end-to-end.
6. Visibility and Transparency — Keep it Open. PbD seeks to assure all stakeholders that whatever the business practice or technology involved, it is in fact, operating according to the stated promises and objectives, subject to independent verification. Its component parts and operations remain visible and transparent, to users and providers alike. Remember, trust but verify.
7. Respect for User Privacy — Keep it User-Centric. Above all, PbD requires architects and operators to keep the interests of the individual uppermost by offering such measures as strong privacy defaults, appropriate notice, and empowering user-friendly options. Keep it user-centric.

2.3 Prospects for Privacy by Design

In Europe and worldwide [23], Privacy by Design is considered as the best current operational response to both protect the fundamental right to data protection and guarantee the free flow of personal data between business stakeholders or Member States. Thus, at the time of digital data massive exchange through networks, privacy by design is a key-concept in legacy [32, 37, 44, 47].

For instance in Europe, where this domain has been directly inspired by the Canadian experience, the European Community[1] affirms that "Privacy by Design 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".

Privacy by Design becomes a reference for designing new systems and processing involving personal data. It becomes even an essential tool and constraint for these designs whereby it includes signal analysis methods as long as these analyses integrate or produce personal data.

Concerning the scientific community, we can recall two main points:

- Processing relative to historical, statistical and *scientific research* purposes, falls under specific conditions defined by article 83 of the "Safeguarding Privacy in a Connected World" European law that facilitates the use of personal data in certain cases. This article defines two specific exceptions, i.e. when: (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;
- Penalties in case of non-compliance are severe. As long as processing is not compliant, these penalties are the same whether the algorithms and the processing used in real business are issued from the research community or not. The supervisory authority "shall impose a fine up to €1,000,000 or, in case of a company up to 2 % of its annual worldwide turnover".

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

Europe regulates data protection through one of the highest State Regulations in the world [16, 31] 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.

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

A prohibition is at the heart of the European Commission's Directive on Data Protection (95/46/CE – The Directive) [16]. 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 [16, 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) [23, 42]. 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 has been adopted on 12th March 2014 by the European Parliament [31]. 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 [31, 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.

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' [16, 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.

3 The Way ISMIR Supports Legal Issues

Let us look ahead ISMIR works from the point of view of pro-activity about data and especially about the legal and personal data.

3.1 ISMIR Works regarding Privacy issues

Is it possible to get a clear view of ISMIR evolution regarding the legal themes and especially privacy from the year 2000 to the year 2013 without going into technical

[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

details? Let us begin by some 2000 to 2013 sessions raw comparisons of 1° Topics mentioned within the official calls for contributions and papers, 2° Topics that characterize the accepted papers, posters and interventions.

In 2000, one category out of the eleven offered, concerned legal aspects: "Intellectual property rights issues". And in the forthcoming edition there is one category "social, legal, ethical and business issues" out of a total of thirty-two. Looking simply at those figures, we can note that personal data and rights issues occupy a secondary place on ISMIR's agenda. But, as we see below, those questions are very transversal to many research topics.

Let us now have a look at the topics in the ISMIR titles of accepted papers, posters and interventions.

Table 1. A word-cloud fed with ISMIR 2000 & 2001 scientific papers

Table 2. The corresponding 2011 & 2012 word-cloud

At first glance, there is no significant change in the scope of legal, right or personal data: only tree interventions mention the term 'legal' or 'right' in their title, no publication mentions explicitly "Personal Data" or "Privacy". However, behind many themes of the different ISMIR works, personal data are involved in one way or another. The most obvious one concerns the "Personal music libraries" and "recommendation": more than 30 papers and posters deal with those topics as being their main topic. How to recommend music to a user or analyze their personal library without tackling privacy? And how to work on "Classification" or "Learning", producing 130 publications without considering users throughout their tastes and their style?

3.2 Why the lack of pro-activity regarding legal issues can lead to failure?

We may agree that the MIR scientific community, as noticeable through ISMIR successive publications, is deepening its scientific objects and sub-domains, creating powerful algorithms, features and metadata, considering research as its main activity.

But business is not far away, and will completely reorganize the traditional stakeholders' models, requiring user involvement to design recommendation systems and to extract knowledge and evaluation. For having ignored the necessity to address legal issues, and particularly privacy issues, other IT innovative areas have already collapsed.

For instance, the Digital Right Management for digital music was an attempt from the producers to recover their intellectual property on music already largely shared by users on the web. It is clear that the legal aspects were not integrated "by design" by the different stakeholders in the context of music stored in digital lossless files, exchangeable worldwide, easy to be copied.

In the context of personal data, the principle of "End to security" is regularly in the limelight. For instance in 2011 millions of PlayStation accounts where hacked. The consequences for Sony have been considerable and, two years later, prosecutions are still in progress. In Great Britain, Sony was condemned in 2013, and during the trial, the authorities mentioned that: "polls conducted after the breach suggested a greater awareness of the risks in handing over personal data". Once again, the effort here is conducted after the issue.

Those two simple examples only show the beginnings of the problems that will arise if private data management is not proactive. The new capabilities of the Information Technology announce a much more complicated world in terms of personal data.

3.3 Actual MIR Practices are PbD-Compatible but not PbD-Compliant

Just as music/sound-based system design is not the MIR core target, constraints related to architecture design (technical constraints or related to user interfaces) are not in the core focus of MIR researchers either. That is why, even though this notion is more than twenty years old, Privacy by Design — as an intersection between IT content and method — has not directly involved the ISMIR contributors yet, no more than the International Computer Music Conference ones.

Furthermore, most of ISMIR contributions are still research oriented, in the sense of Article 83 of the "Safeguarding Privacy in a Connected World". To say more about that intersection, we need to enter into a cross survey of the ISMIR scientific ten-years production, throughout the main PbD Foundational Principles (FP).

FP6 (transparency) and FP7 (user-centric) are most of the time fully 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 into Design, FP3 cannot be fulfilled and accordingly FP2 (default setting), FP5 (end-to-end), FP4 (full functionality) and FP1 (proactive) cannot be fulfilled either. Without any PbD embedded into Design, there

are no default settings (FP2), you can not follow and end-to-end approach (FP5), you can not define full functionality regarding to personal data (FP4) and you can even less be proactive. Principle of pro-activity (FP1) is the key principle. If you fulfill FP1 you can define the default settings (FP2), be full functional (FP4) and define an end-to-end process (FP5).

Actual MIR Practices claim to be relatively neutral to data privacy and are compatible with PbD. The MIR Practices could be compliant to PbD as long as they would fulfill the FP1 principle of pro-activity.

4 How ISMIR Builds New Kind of Personal Data

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

The Gamelan Project is a study case where machines produce new data and new personal data from inputs.

4.1 A Case Study: The Gamelan Project

Gamelan (see http://projet-gamelan.fr/) was an Industrial Research category research project coordinated by IRCAM, gathering INA, EMI Music France and UTC, and supported by the French *Agence Nationale pour la Recherche*. The project began in November 2009 and lasted 48 months.

Digital studios involve a great amount of traceable processes and objects, because of the intense producer-device interactions during contents production. The important flow of these traces called for a system to support their interpretation and understanding. The Gamelan research teams has studied and developed such a system in the digital music production context, towards musical object and process reconstitution. We were aiming at combining trace engineering, knowledge modeling and knowledge engineering, based on the differential elaboration of an ontology, standard formats and common knowledge management tools.

Involving professional users, the Gamelan research teams succeeded in applying this system to several different real use cases, put forward by different kind of end users, and we are now able to discuss some hypothesis about trace-based knowledge management, digital music preservation and reconstitution, opening on to some considerations about artistic style, and to the specification of the next generation prototypes that music industry would need to develop.

Gamelan is also the name of the developed software environment, built upon the production ecosystem, to address the reconstitution issue of digital music production, by combining trace engineering, knowledge modelling and knowledge engineering. Most of the time reconstitution is relegated afterwards. Gamelan aims at reconstructing the composer-system interactions that have led to the creation of a work of art that is about to leave the production studio. The purposes of reconstruction concern long-

term preservation, repurposing, versioning and evolution of the work of art, and more generally the disclosure of the contingencies of its initial outcome.

In the music production studio, everything is about creativity [15, 19]. Until now, music tools design has mainly focused on the making of the final product, because the very first aim of the studio is to provide the creator with efficient means to make and shape the musical object he or she came in the studio for. But this requisite priority on creativity has overshadowed another need that appears later: reconstitution.

Of course, creativity empowering raises tough challenges to work out. For instance, on the conceptual side, bridging the gap between creative thinking and application interfaces remains a challenging issue [2, 11, 35], while on the technical side, the heterogeneity of tools, systems, components, protocols and interfaces keeps renewing difficulties for the management of production environment [13].

A creator finishing his or her work in a studio marks the end of the production process: the so-awaited object is finally there, thus the creator, the producer, the sound engineer and all the people involved are happy or at least relieved; the goal is reached and the story reaches its end. However, at this very moment, because the final object is there, no one wonders about its reconstitution.

But —say ten years later— when “back-catalog" teams of music companies want to edit some easy to sell Greatest Hits at up-to-date audio formats, mining the musical archives is no longer easy. Back to the reachable-recorded digital files, it may be painful to figure out which one of the bunch of files left is the one needed. File dates and file names are not trustable.

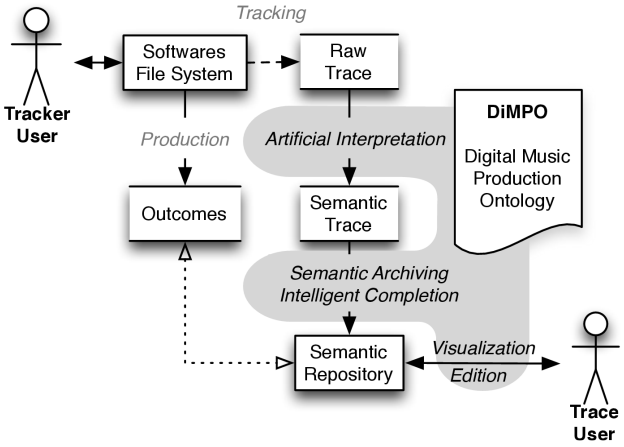

Fig. 1. Gamelan: the interaction graph

Closer in time —say two months after the production— the simple task of collecting vital information on the contributors who actually worked on the project may turn into a real problem. A musician may replace another without logging his/her name. Or a name is missing because we only have the nickname and we don't have the phone number either. There is a whole set of information on contributions (name, role, time spent, etc.) necessary to manage salaries, rights and royalties that regularly proves

hard to collect afterwards. Evidently, this kind of information would be far easier to collect directly at production time.

On the surface, nothing to do with privacy and personal data! But in fact, and it is typically the case as soon as a complex person-software device is involved, this type of project invites us to rethink classical approaches and qualifications of privacy issues.

4.2 Collecting traces to build a knowledge model in the Gamelan Project

The Gamelan project exemplifies several of the many R&D emerging questions that are raised in the digital audio processing domain.

First of all reconstitution requires to collect traces during the production process itself. Automatically-collected software traces differ from human-entered traces. The former can be seamlessly collected through automatic watching components, with interfaces traces and logs as heuristic material, while the latter inevitably requests a human contributor for information that cannot be automatically captured or inferred from automatic traces. A full-production tracking environment would resemble Living Labs, towards a Living Studio.

Secondly, these traces call for an appropriate knowledge model. To stay as little invasive as possible, such a model should provide means to determine which information is worth to ask humans during the production or not compared to the creativity disturbing cost. Without a knowledge model, it would not be possible to interpret the traces or to determine the kind of traces worth capturing. To achieve this model, professional knowledge must be identified, listed and characterized with experts, defining a digital music production Knowledge Level.

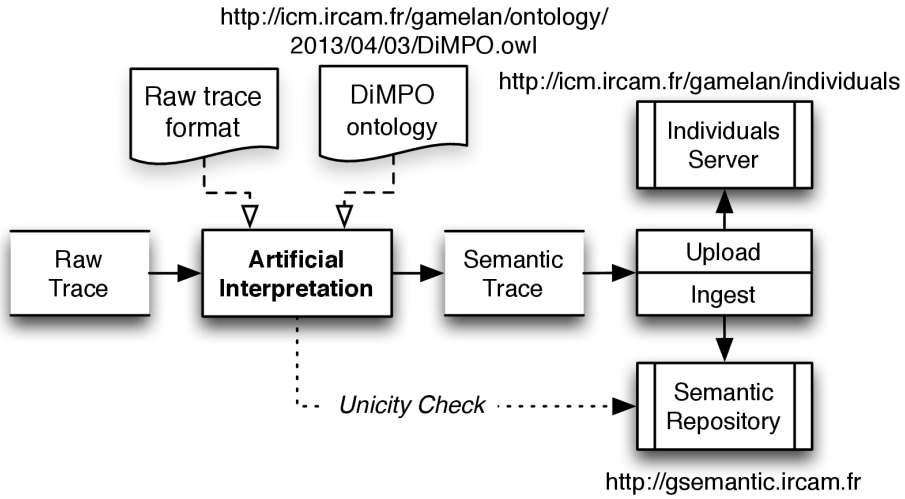

Fig. 2. Gamelan: the different functionalities

Within Gamelan, traces from used operating system and from used professional applications are extracted, semantic networks dealing with typical digital audio composition acts are involved towards some abstraction of those traces, but personal data are nowhere considered: some real time digital audio flow are involved, transformed on the fly by creative acts that sign the composer particular style and their artistic singularity.

The composer style, as part of built up personal data, often not even named, is computed to support the Gamelan reconstruction process: what is interesting to reconstitute has something to do with the abstract truth of the given piece of art and its stylistic genesis. To understand that "the composer is currently testing a sample within the whole piece framework" is more efficient than being aware of a succession of cut-paste-listen actions that has to be generalized.

Thus some personal data, like artistic style [1, 7, 13, 18] are built up on the fly, relatively to processing algorithms, knowledge bases and title repositories [9], evolving from the system experience itself, and only known by the system. The ultimate target is clearly the style-recognition [22] of the creator, viewed as the correlation between their practice and the character of their work of art.

4.3 Information retrieval produces personal data from inputs

Most of the time the ISMIR developed methods create new outputs that are personal data or include personal data. These data may be used through streaming without being stored. These data could be so evanescent that they could not be reproduced at all, depending on the specific situation of the performance processing and of the music listener. These data are personal data as they depend on the specific listener or the specific musician.

Authorship attribution creates new personal data, whatever the algorithmic way you build this attribution.

Some of these non-linear algorithms encode personal data implicitly. For instance Kohonen maps, Neural Nets, Hidden Markov Models (HMM), Bayesian Maps create and store data [21], which are the weights of the network (especially the weights of the hidden layer) or a set of probability [20]. In that case these data are personal data integrated inside the processing and encoding for instance the author attribution. Moreover this technical way to encode these new personal data converges thanks to the big data.

Fast Fourier Transform, Wavelets, and more generally algorithms that consist in projecting data on a specific basis produce personal data in the same way. Results of the projection are new data, which are personal as long as these projections depend on the listener or the musician.

This is still the case when a MIR research team uses some algorithms to support heuristic reasoning and decision support, dealing for example with authorship attribution or style recognition. In this case, music could be considered as a sequence that can be randomly walked or, better, turned into some semi-predictive program solution by taking advantage of (let's say) local HMM amnesia to (let's say) converge towards a relevant future. Then a musical sequence can be considered as a solution to some

Constraints Satisfaction Problem [6, 30], and no one is henceforth able to separate data from processing, a fortiori personal data from processing.

5 MIR communities: an emblem of Digital Humanities ones?

5.1 Style identification algorithms in actual MIR: a classification

Among the algorithms that identify artists creating a style of their own, we can distinguish two meta-heuristics groups. The first one is based on interconnections of databases and more generally on the big data. The second one is based solely on the intrinsic properties of music data.

In this section we propose a classification of algorithms and heuristics allowing the identification of the creators of musical data exclusively on the basis of this second approach.

It is usual to consider the style of an artist is available at three levels: patterns, meanings and feelings (cf. [3], p. xii). For each of these levels, the algorithms implemented enter the structured list given below which can be used as read gate to analyze style recognition algorithms.

Symbolic approaches

Shape grammars, rules production and combination

Shape grammars are examples of symbolic approach to style production (Mitchell 1990). As a reverse problem it deals with style identification. Shape grammars were first introduced [38] for painting and sculpture. Later on it has been used for producing new designs in architecture [39]. Even useable pieces of software were developed in this field [29].

Shape grammars are a kind of results obtained by rules production and combination. In this same class of algorithms we can put L-systems [24].

As reverse problem let's consider the results of the rules production system and let's try to find the kind of rules that could produce the same shapes. Find theses rules and you find the style of the production system.

Combination of patterns

Another way to deal within this symbolic approach is to combine rules but defined short patterns. This is the kind of algorithms produced by David Cope with the so-called EMI project [14].

Optimization methods

Some of the algorithms used to categorize features are dealing with optimization techniques. Among those we find Support Vector Machines [43] that is used do define a *hyper plane* separating two sets of examples (positive and negatives ones). This *hyper plane* is the set of points that do maximize the distance between the two sets of examples.

Yang and Pedersen [46] proposed a large comparison between those technical optimization methods.

Algorithms based on topology

Some other algorithms use topology and connectivity to categorize features and style. Among those we can consider Kohonen cards so-called Self Organizing Maps [26].

For instance Jupp and Gero [24] do apply these algorithms to categorize styles.

Non deterministic algorithms

Frequency and statistical methods

One of the most intuitive technique to define an artist's style is to build a dictionary of the specific « atoms » of creation he uses. These atoms can be words, music notes, etc. The algorithm consists on building this dictionary and storing the specific personal frequencies of uses of the atoms by the artist.

Argamon and Koppel [4] solve the authorship verification problem by using this key algorithm.

J. Karlgren [25] explains how these algorithms, defined to detect sylitic variations based on different low-level features can not directly be used for variations of a higher level of abstraction [10].

Markov models and especially Hidden Markov Models

In the simplest case, Markov Models consist on a set of states and transition probabilities from one state to the other ([3], Chapter 7 & Chapter 10). More sophisticated Markov Models, also called Hidden Markov Models (HMM) consist on Markov Models with hidden states. Hidden states are kind of internal states that cannot be directly perceived.

These models were widely used to model and categorize styles. For instance they were used for instance to identify and distinguish Beethoven sonatas and Mozart sonatas [2].

Assayag, Bloch, Cont and Dubnov [5] apply these techniques to Anticipatory Learning, which mixes these HMM and Q-learning [40].

5.2 From MIR R&D communities to Digital Humanities ones

In the current legal frameworks, Personal Data are still considered as a particular kind of data, as opposed to processing according to the classical Information Technology paradigm. By seriously considering the PbD methods and technologies for mastering and appropriation, the MIR community will probably rediscover that the paradigmatic data/processing separation has been finally overcome, as soon as many MIR algorithms raise their results on the fly from digital music flows, often in real-time: thus the MIR community will naturally join and lead the conducting of the PbD contemporary concept towards more advanced concepts, able to take note that sometimes, per-

sonal data can be outputs from some authorship attribution artificial system, made of complex person-machine interactions and act accordingly.

The time where 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 heuristic knowledge that cannot be reduced to data or processing.

This means that new research fields do not separate data and processing anymore. This can be done in different ways. In many circumstances, the MIR community develops new personal data using the whole range of data analysis and data building algorithms. The MIR community is especially well positioned to identify the new personal data produced through these algorithms.

From this respect, MIR is a good emblem of what is currently happening within the whole Digital Humanities R&D communities.

6 Conclusion

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

Methodological Recommendations. Digital Humanities researchers and developers 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 Safe Harbor Framework which could already be an industrial challenge for instance regarding Cyber Security (P5). Using the Privacy by Design 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 Digital Humanities researchers and developers 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 Privacy by Design 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, Digital Humanities researchers and developers 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 [21] 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 Digital Humanities 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.

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 Digital Humanities 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?

References

1. R. Ackoff. From Data to Wisdom. Journal of applied systems analysis, 16(1):3–9, 1989.
2. C. Alamkan, W.P. Birmingham, M.H. Simoni. Stylistic Structures: An initial Investigation of the Stochastic Generation of Tonal Music. University of Michigan, Computer Science and Engineering Division, Department of Electrical Engineering and Computer Science, 1999.
3. S. Argamon, K. Burns, S. Dubnov (Eds): The Structure of Style, Springer-Verlag, 2010.
4. S. Argamon and M. Koppel. The Rest of the Story: Finding Meaning in Stylistic Variations, in S. Argamon, K. Burns, S. Dubnov (Eds): The Structure of Style, Springer Verlag, 2010.
5. G. Assayag, G. Bloch, A. Cont and S. Dubnov. Interaction with Machine Improvisation, in S. Argamon, K. Burns, S. Dubnov (Eds): The Structure of Style, Springer Verlag, 2010.
6. J-J. Aucouturier, F. Pachet, P. Roy, A. Beurivé: “Signal + Context = Better Classification”, Proceedings of the International Symposium on Music Information Retrieval, 2007.
7. H. Barki and J. Hartwick. Measuring user participation, user involvement, and user attitude. Mis Quarterly, pages 59–82, 1994.
8. C. Barlas: “Beating Babel - Identification, Metadata and Rights”, Invited Talk, Proceedings of the International Symposium on Music Information Retrieval, 2002.
9. T. Bertin-Mahieux, D.P. Ellis, B. Whitman, P. Lamere: “The Million Song Dataset”, Proceedings of the International Symposium on Music Information Retrieval, 2011.
10. D. Biber. A typology of English texts. Linguistics 27:3-43, 1989.
11. N. Bonnardel and E. Marmèche. Evocation processes by novice and expert designers: Towards stimulating analogical thinking. Creativity and Innovation Management, 13(3):176–186, 2004.
12. N. Bonnardel and F. Zenasni. The impact of technology on creativity in design: An enhancement? Creativity and Innovation Management, 19(2):180–191, 2010.
13. J.D. Carney. The style theory of art. Pacific Philosophical Quarterly, 72(4):272–289, 1991.
14. D. Cope. "Experiments in Music Intelligence." In Proceedings of the International Computer Music Conference, San Francisco, 1987.
15. D. Cope, Computer Models of Musical Creativity. Cambridge, MA: MIT Press, 2006.

16. Directive (95/46/EC) of 24 Oct. 1995, *Official Journal L281, 23/11/1995,* http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML, *p.31-50*
17. 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.
18. S. Dubnov, G. Assayag, O. Lartillot, and G. Bejerano. Using machine-learning methods for musical style modeling. Computer, 36(10):73–80, 2003.
19. E.A. Edmonds, A. Weakley, L. Candy, M. Fell, R. Knott, and S. Pauletto. The studio as laboratory: combining creative practice and digital technology research. International Journal of Human-Computer Studies, 63(4):452–481, 2005.
20. S.E. Fienberg: "The Relevance or Irrelevance of Weights for Confidentiality and Statistical Analyses", Journal of Privacy and Confidentiality, Vol. 1, Issue 2, pp. 183-195, 2009.
21. 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.
22. M. Grachten, G. Widmer: "Who Is Who in the End? Recognizing Pianists by Their Final Ritardandi", Proceedings of the International Symposium on Music Information Retrieval, 2009.
23. D. Greer: "Safe Harbor - A Framework that Works", International Data Privacy Law, Vol.1, Issue 3, pp. 143-148, 2011.
24. J. Jupp and J. Gero. Let's Look at Style: Visual and Spatial Representation and Reasoning Design, in S. Argamon, K. Burns, S. Dubnov (Eds): The Structure of Style, Springer Verlag, 2010.
25. J. Kalgren. Textual Stylistic Variation: Choices, Genres and Individuals, in S. Argamon, K. Burns, S. Dubnov (Eds): The Structure of Style, Springer Verlag, 2010.
26. T. Kohonen, Self-Organizing Maps, vol. 30, Springer Verlag, 1995.
27. 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.
28. T. Lubart. How can computers be partners in the creative process: classification and commentary on the special issue. International Journal of Human-Computer Studies, 63(4):365–369, 2005.
29. A. McKay, S.C. Chase, K. Shea, K., H.H. Chau. Spatial grammar implementation: From theory to useable software. AI EDAM (Artificial Intelligence for Engineering Design, Analysis and Manufacturing) 26(02), 143-159, 2012.
30. F. Pachet, P. Roy: "Hit Song Science is Not Yet a Science", Proceedings of the International Symposium on Music Information Retrieval, 2008.
31. 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
32. J.D. Reiss, M. Sandler: "Audio Issues in MIR Evaluation", Proceedings of the International Symposium on Music Information Retrieval, 2004.

33. V. Reding: “The European Data Protection Framework for the Twenty-first century”, *International Data Privacy Law,* volume 2, issue 3, pp.119-129, 2012.
34. G. Rozenberg and A. Salomaa. The Mathematical Theory of L-Systems (Academic Press, New York, 1980
35. D. A. Schön. The reflective practitioner: How professionals think in action, volume 5126. Basic books, 1983.
36. 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.
37. 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.
38. G. Stiny, J. Gips. Shape grammars and the generative specification of painting and sculpture. In Information Processing 71, 1460–1465. North-Holland Publishing Company, 1972.
39. G. Stiny. Introduction to shape and shape grammars. Environment and Planning B: Planning and Design 7(3), 343-351, 1980.
40. R. S. Sutton and A.G. Barto. Reinforcement learning: an introduction. MIT Press, Cambridge, MA. 1998.
41. 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.
42. U.S. – EU Safe Harbor: http://www.export.gov/safeharbor/eu/eg_main_018365.asp
43. V. Vapnik. The Nature of Statistical Learning Theory, Springer-Verlag, 1999
44. D. Wright, K. Wadhwa: “Introducing a Privacy Impact Assessment Policy in the EU Member States”, International Data Privacy Law, Vol. 3, Issue 1, pp. 13-28, 2012.
45. F.Y. Wang. Is culture computable? Intelligent Systems, IEEE, 24(2):2–3, 2009.
46. Y. Yang, J.P. Pedersen. A Comparative Study on Feature Selection in Text Categorization, Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), pp412-420, 1997.
47. A.M. Zaslavsky: “Privacy and the Statistician: What Do We Need to Know to Certify Nondisclosure?”, Journal of Privacy and Confidentiality, Vol. 3, Issue 2, pp. 83-90, 2011.