CoMaFeDS

Consent Management for Federated Data Sources

Max-R. Ulbricht (✉), Frank Pallas
Information Systems Engineering Research Group
Technische Universität Berlin
Berlin, Germany
{mu, fp}@ise.tu-berlin.de

***Abstract*—This paper highlights the critical role of individually given consent for ensuring privacy in big data scenarios employing federated sources of personal data. It derives the need for consent management to be implemented in respective systems and identifies the most important requirements that have to be met. Based on already existing approaches like Hippocratic databases, sticky policies or distributed usage control, it develops an own approach of consent management for federated data sources.**

Keywords—consent management; federated systems; privacy preserving data integration; informational self-determination; data protection

I. INTRODUCTION

Questions of privacy have always played a major role in the context of cloud computing and big data. Even if often misunderstood as merely being about terms like data / information security, data minimization or anonymity, privacy covers many further aspects which must also be taken into account in the design of information systems used for the collection, processing and use of personal data. In this regard, the concept of consent is of utmost importance.

Emanating from the established understanding of privacy being about "informational self-determination", any access to and use of data referring to a certain person shall be subject to the individually provided consent of that person. As privacy theorist Alan Westin put it already in 1967, "Privacy is the claim of individuals, groups, or institutions to determine for themselves when, and how, and to what extent information about them is communicated to others." [1, p. 7] . This understanding was picked up by early European privacy legislation and is still authoritative in the current process of a European data protection reform.

As provided by art. 7 of the still valid European data protection directive [2], for example, personal data may only be processed if "the data subject has unambiguously given his consent" or if the processing is legitimated on other grounds like the necessity for carrying out a contract with the data subject or an existing legal obligation. As such other grounds do not apply for most potential applications of big data technologies, individual consent is often the only available basis for making big data applications employing personal data viable at all.

Such individual consent must, however, meet further requirements: According to art. 2, lit h) of the European data protection directive [2], individual consent must, in order to constitute a legitimate basis for processing personal data, be "freely given", "specific", and "informed". In order to be specific and informed, in turn, consent must refer to specific personal data, be given for a specific purpose and be declared with regard to a specific "utilizer" processing the personal data for that purpose.

For current and future applications of big data, this requirement proves increasingly challenging. In particular, much of the value proposed by big data technologies arises from the approach of interrelating different and already existing datasets with each other and combinedly processing them for purposes that could not have been foreseen at the moment of data collection. From the perspective of European data protection law as well as in terms of the above-mentioned, more general understanding introduced by Westin, such processings of personal data for novel purposes would always require a separate consent from respective data subjects. While unquestionably serving the intended goal of achieving the data subjects' "informational self-determination", this would render a multitude of big data applications impossible.

On the other hand, many applications of big data utilizing already existing sets of personal data can provide significant societal value (e.g. in the field of research) and would therefore also find the consent of the persons that the respective data refers to. Given the above-mentioned consent-related requirements, however, obtaining and providing consent all too often turns out to raise prohibitive efforts for data utilizers and data subjects, thus rendering valuable usages of existing data impossible. We therefore see significant potential in technical mechanisms for the easy and low-effort provision of "specific" and "informed" consent to the processing and use of already existing sets of personal data for novel purposes.

This paper therefore delineates the general challenges arising with regard to the provision of consent in the context of federated sources of personal data in section II. Section III then presents existing technical approaches for managing consent in some more detail and discusses their suitability for use cases with federated data sources and with purposes not known at the moment of initial data collection. In section IV, we then present our own approach of "Consent Management for Federated Data Sources (CoMaFeDS)", which picks up several

concepts from the aforementioned approaches and complements them with functionalities specifically tailored to scenarios of federated data sources. Section V provides an outlook and concludes.

II. Illustrative Scenario & Challenges

The challenges that are to be addressed can be illustrated with a simple scenario. We assume a research institute trying to heighten traffic safety. For this purpose, it wants to discover to what extent traffic density and wheather conditions influence car drivers' stress level and how these two factors interrelate with each other. Instead of collecting respective data in a controlled and separate empirical study – which would cause significant costs and efforts – researchers discuss an alternative approach of employing already existing car motion data from data source *A* (e.g. a navigation system maufacturer) and data on pulse-rates, skin moisture, and skin resistance collected by smart watches or other wearable devices and stored in data source *B*. Furthermore, wheather conditions are to be determined on the basis of data collected by personal wheather stations and stored in the respective manufacturer's data source *C*. All data are assumed to carry either explicit or implicit personal relations to natural persons and thus to be personal data.

Making this kind of research possible raises certain requirements: First, it must be possible to search existing data sources for data matching certain criteria like data types, geo-location etc. It must be possible that these data sources are autonomously maintained by different parties and of largely heterogeneous nature. Second, a mechanism is needed that allows for the utilization of these data based on data subjects' consent meeting the criteria laid out above. This, in turn, requires mechanisms allowing data subjects to state their consent for specific purposes of data utilization to be conducted by specific parties. As neither potentially relevant purposes nor the respective utilizers can be assumed to be known in advance, all mechanisms must, third, allow for the dynamic and ad-hoc addition of novel purposes and utilizers.

Further requirements will of course emerge but at least these main functionalities are indispensable for facilitating scenarios of analyzing existing personal data from different data sources for novel purposes. Establishing information systems that meet these requirements for sustainable consent management would thus allow for a multitude of societally valuable big data applications to be realized in a way that is compatible with established conceptions of privacy (as, in particular, embodied in European data protection law) und thus to become viable at all. In the past, several technological approaches and concepts have been proposed in this regard. These shall be presented and discussed in the following section.

III. Approaches & Concepts for Technical Solutions

As the problem of privacy preserving data transfers and/or integration is far from being a new one, we will first present and analyze some concepts that claim to address the outlined requirements. Thereby we try to figure out what advantages and disadvantages each concept has and evaluate whether the concept or part of it is suitable for the intended solution of integrated consent management for heterogenous and autonomous data sources.

A. Hippocratic databases

Hippocratic databases are a technical approach to preserve privacy and implement data protection directly into a database system (privacy by design approach). The idea is to use supplementary mechanisms and additional database tables to ensure that only authorized recipients with valid purposes can access particular attributes within database tables. [3]

In traditional relational database systems, all data pertaining to a given data subject are stored in different tables which contain values for all given attributes of a data subject. To get access to these values, a user or application performs queries against the database engine in a specific query language like SQL. The design of the mentioned database systems allows every user or application with sufficient access privileges to the system itself also to access the information that's stored in the database through appropriate query statements without any restrictions. If the database tables contain personal information, this behavior of this kind of database systems can lead to fundamental privacy harms when unauthorized persons get access to potentially sensitive informations about individuals.

To prevent unauthorized access, hippocratic databases implement, in the figurative sense, a "guard" between the database tables and the user or application that requests the desired information from the database. This guard implements different mechanisms to manipulate requests that originate from outside the system. These so called query modifications use additional database tables to decide whether an access to the requested attribute is allowed or not. [4]

The mentioned tables can codify both the privacy policy of the provider/operator of the database system as well as the privacy preferences of the individual that is represented by the respective data. Within these privacy policies and preferences, it is specified which recipients are allowed to access what data for which well-defined purpose [3]. If implemented correctly, this implies that only queries with the right combination of recipient and purpose can "pass the guard" and get access to the requested data. To automate the generation of the respective policy- and preference-tables, there exist some mechanisms to directly import privacy policies following the P3P standard [5][6].

Hippocratic databases are thus capable to foster informational self-determination by ensuring that data can only be processed when an individual has stated her consent to exactly the purpose of the respective processing. If the implementation is done right, even the administrator of the database system cannot access data that is not explicitly approved for her and the purpose she is stating. Concepts from the field of hippocratic databases can thus prove valuable with regard to the provision of specific consent and – to a certain extent – the enforcement thereof.

One of the downsides of this concept is the nearly complete absence of real world implementations. Besides some academic demonstrations [7][4], it is hard to find productive systems

implementing hippocratic database principles to preserve data privacy. In addition, in a scenario with federated data sources all of these data sources must be structured in a hippocratic way for full benefit of the concept for the complete system.

B. *Sticky Policies*

The basic idea of Sticky Policies is a dualistic approach to assure informational self-determination based on given consent to different purposes of data processing. The first component is a data subject's privacy preference sticked to a dataset that contains personally identifiable information relating to her, while a second part uses strong encryption to regulate the access to this dataset.

The privacy preferences of an individual are represented by a policy wherein simple rules determine the circumstances under which the personal information inside the dataset may be accessed. These circumstances can consist of specific technical environment variables that determine that e.g. the data should only be processed on a system with predefined security mechanisms or only inside a given subnet of an institution's computer network. Furthermore, policies can also contain specific purposes like "research" or "contract fulfillment" for which the processing of the dataset is permitted. Other requirements like expiration dates or trustworthy institutions that are allowed to process the data without any further negotiations are possibly thinkable.

These policies are – in conjunction with encryption mechanisms – used to ensure informational self-determination as follows: At the moment of the collection of personal information, the data subject has to specify for what purposes and under which conditions this information can be accessed. The system of the data holder converts these preferences into a standardized privacy policy. All collected data are then encrypted by the system and the key for decrypting them is transfered to a third trustworthy institution, a so called trusted authority. Then the privacy policy that belongs to the encrypted dataset is enhanced with the information which trusted authority holds the decryption key for using the dataset. This enhanced policy is then sticked to the encrypted dataset, resulting in a "Sticky Policy" and a dataset that is not usable without its associated decryption key. [8]

If another institution is interessested in using and processing the dataset, it can send a request to the data holder. As an answer the encrypted dataset with the sticky policy is transfered to the potential utilizer. With the information inside the policy, a request can be transmitted to the announced trusted authority, containing a request for the dataset itself as well as a commitment to all rules stated in the policy. At this stage, the trusted authority can perform some provisional actions to validate the technical environment of the desired processing ("remote software verification"), log the request and commitment, and if all rules inside the policy are fulfilled release the key necessary to decrypt and process the requested dataset. [8][9]

As outlined above, this concept ensures that datasets cannot be processed without the interaction with a trusted authority. That third instance can log all accesses to the data and makes sure that data is only processed under circumstances that satisfy the preferences of the individual the respective dataset refers to. Like concepts from the field of hippocratic databases, sticky policies can therefore prove valuable in the context of stating consent specific to certain purposes, utilizers and processing conditions. The strong dependence on the trusted authority and, thus, on a third institution is, however, a significant drawback of this concept. Also, the additional steps required during the collection of data (encryption of the dataset, transferring the keys to trusted authorities, ...) generate extra costs and require some modifications on the data holders' systems that collect the data in the first place. Another downside of the presented concept is the fact that only the access to the data is checkable. A violation of the rules codified within the Sticky Policies by a disobliging utilizer after the access can not be detected or prevented.

C. *Distributed Usage Control*

The concept of distributed usage control enhances the Sticky Policies approach by the additional aspect of ensuring that a data utilizer's violations of stated rules can be prevented or observed even after initial access. For this purpose, the policies that codify the preferences of a data subject are splited into two documents, called provisions and obligations.

Of these, provisions contain the rules that regulate the access to a dataset. Comparably to the Sticky Policies outlined above, provisions prescribe the requirements that must be met in order to obtain a requested dataset. Examples of these requirements can be proofs about the technical security infrastructure (e.g. processing systems controlled by means of trusted computing [10]) or the existence of certain organizational security measures (for example a security management system according to ISO 27001 [11]) to be present for the environment of the planned data processing. Obligations, in contrast, specify rules and regulations regarding the handling of the data after the access is potentially granted. They represent some sort of contracts that, besides rules like "delete data after 30 days" or "only 3 times copying allowed", also define compensations that come into force if particular rules or parts of them are violated. [12]

If a prospective utilizer wants to process a dataset, she requests the access from the data holder. The data holder can then demand provisional actions like proofs about the technical systems used for the processing and/or commitment to all rules stated in the provisions that can not be verified in a technical way. In case all conditions are fulfilled, a negotiation about the obligation is initiated. If the inquiring utilizer commits to all given rules and accepts the suggested compensations for cases of noncompliance, the requested dataset is transferred.

To be able to control the further usage of transferred datasets, some mechanisms have to be installed on the utilizers processing systems. These are divided into two different sorts of components. Control mechanisms are agents that prevent misuse or undesired handling of transferred data through technical instruments. Some of these agents can prohibit unauthorized copying, some suppress a transmission to other parties and others again disable the printing function of the system [13]. The second sort of mechanisms is responsible for things that can not be practically prevented technologically.

These mechanisms observe the processing of the moot dataset. If parts of the processing infringe a rule from the obligations, the observer mechanism activates a signal mechanism that informs the data holder about the breach so that the negotiated compensations can be triggered. [12]

Having in mind our desired consent management for federated data sources, this approach to access and usage control might particularly support the enforcement of specifically stated consent. It can ensure the compliance with an individual's consent for a specific purpose and specific utilizers in both ways, ex-ante and ex-post. Some rules can be enforced by technical artifacts and the violation of others which can not be prevented are reported and punished afterwards. Enabling these features would, however, provoke significant costs for implementing the control-, observe- and signal-components on all participating systems. For this, a high level of cooperation is needed and potential utilizers have to give up significant parts of their autonomy.

D. Dynamic consent

The approach of "dynamic consent" is specifically focused on the insight also outlined in our illustrative scenario that already existing data can be highly useful for future research but that as soon as personal data is considered, the need for purpose- and utilizer-specific consent often hinders such applications. Dynamic consent is not a self-contained system but rather a concept that can be seen as enhancement of existing systems/approaches.

The concept originates from biomedical research using so-called biobanks and the idea to tackle the outlined problem is to provide a personalized interface that enables uncomplicated communication between researcher and participant. This interface should be capable to use different communication channels like text messages, email or even letters. [14] With the help of this interface, it should be possible for researcher to directly request new consent from a participant whose data already are available from an older or ongoing project in case new research ideas emerge. For the participant – or, more generally, the data subject – that already gave her consent to a previous data processing, the interface can provide a simple way to alter her decisions or enhance them because of changed circumstances.

Thinking this concept consequently out leads to information systems that provide different kinds of interfaces for the dynamic adjustment of given or new consent to variable sorts of data processing purposes conducted by variable utilizers. Depending on the preferences, needs and circumstances of the data subject, the interface could be implemented as web service, application for smartphones and tablet computers, or as a cheap standalone device with the only function to provide access to the dynamic consent system.

As drawbacks of this model, time and efforts for the implementation and integration into existing systems, or the extension of them, are obvious. But with one of the aims of this approach, to accelerate the renunciation from the often used crutch of "broad consent" in the context of research, this concept clearly points into the right direction of a new way for handling consent in a world with lots of data that can be used for meaningful new applications well beyond the initial purpose of collection and processing. Especially in matters of making our approach dynamically adaptive to novel and unforeseen uses of personal data as well as to also novel and unforeseen utilizers, we thus see the concept of dynamic consent as a highly promising one.

IV. Our Concept: CoMaFeDS

In the following sections we introduce our concept for a consent management platform for data mining of heterogeneous, autonomous and distributed data sources, called CoMaFeDS. The goal is to provide a platform that enables data mining on existing datasets from different autonomous sources with the possibility of ensuring the informational self-determination of data subjects represented by the processed datasets. As building blocks for this platform, we borrow several approaches from the concepts analyzed above. In particular, our platform is intended to appropriately address the identified drawbacks of these approaches.

A. Prerequisites

Given the intended dynamic and multi-party nature of the settings to be covered by our platform, privacy preferences, especially the consent to different processing purposes, should – like in the concept of Sticky Policies – be located together with the respective datasets. Within these policies, a data subject should be able to give her consent to various data processing purposes performed by different recipients in advance.

To simplify these decisions, potential utilizers as well as possible processing purposes should be categorized. This allows to give consents like "My data can be processed by independent research institutes for the purpose of demographic development evaluation, but not from governmental agencies for tax estimations." Such policies must be represented in well defined format that allows for concise specification of categories of purposes and recipients. The chosen format must permit arbitrary levels of details in the contemplated policies, so that definitions of numerous subcategories are possible.

Furthermore, our concept is dependent on the knowledge where to find specific datasets. To solve this problem, a data collector or holder that is interested to participate in data mining contexts should generate a description or specification of his database, holding details about provided datasets, and specify the internal structure of the database.

For every potential data source, a machine-readable document specifying where to find which kind of data should thus exist. Furthermore, for every dataset detailed preferences regarding to various processing purposes and recipients should also be available.

B. Architecture

As our consent management system should be as flexible and universally applicable as possible, we design CoMaFeDS as a platform in a broader sense. That means it can be a hosted service in the cloud just as it can be a (stand alone) software

component that may be used to enhance an existing data mining toolkit.

Figure 1 shows the general architecture of a data mining system using CoMaFeDS for consent management. The concept should be deployed as a connector between data mining applications and the data sources they want to analyze. For this purpose, CoMaFeDS has standardized (API-like) interfaces in both directions.

If a data holder with an interesting database decides to open up its datasets for big data analytics performed by external organizations, it can simply connect to the CoMaFeDS platform. During the connection process, the description of the datasets and the specifications of the internal structure of the database as well as the corresponding privacy policies are transfered to the platform.

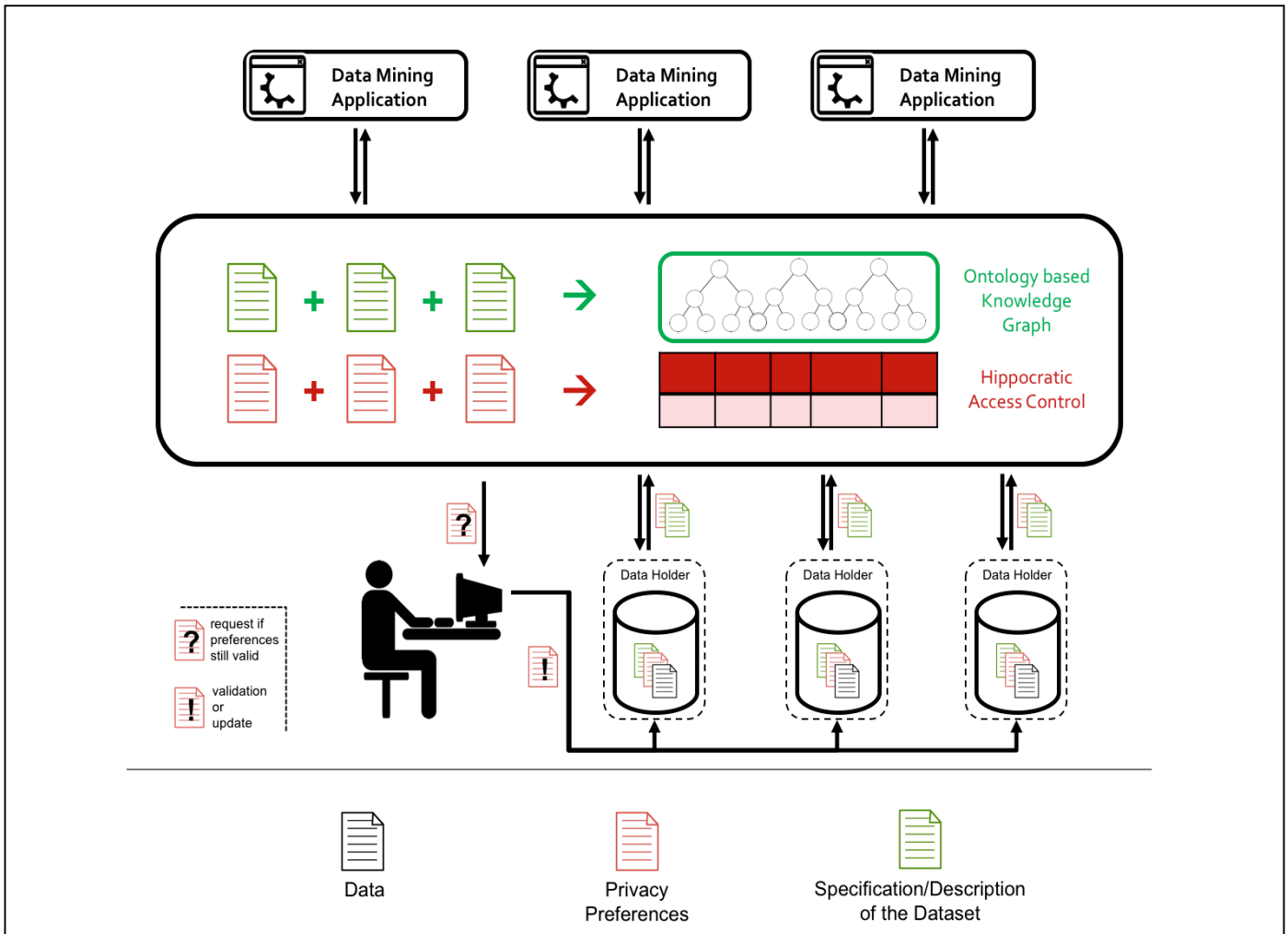

Fig. 1. Architecture of a system using CoMaFeDS

Based on these documents, CoMaFeDS performs some internal conversions. All database- and dataset-related information and specifications are used to develop an ontology-based knowledge graph. This graph codifies the knowledge about the location and the possibilities to access specific datasets ("where to find what kind of data and how to reach them") and can be easily processed to obtain that internal knowledge.

The privacy policies, in turn, are used to create an internal hippocratic integration model. This can be realized based on dedicated tables or other kinds of storage structures that are able to hold the information which attributes of a specific dataset are accessable for which recipient and for what processing purpose. Similar to standalone hippocratic databases, this design leads to a system that prevents all data accesses that did not match the right combination of recipient and purpose.

An organisation or institution that seeks datasets for a data mining application has the opportunity to connect to the CoMaFeDS platform and query the above mentioned knowledge graph to figure out if useful data for its use case are offered by any of the connected data sources through the platform. If desired datasets are available, the potential utilizer has to present its identity and the purpose for the required data processing. If this specific combination of recipient and purpose matches the requirements of the corresponding rules

within the hippocratic integration model, access is granted and the data mining can be performed.

As an enhancement, a mechanism allowing data subjects to provide dynamic consent is thinkable. If a potential utilizer finds suitable datasets inside the knowledge graph, but no consent for the considered data processing purpose exists, the platform should be able to ask the data subject for new or updated permissions. The data subject now has the possibility to alter her setting directly within her original privacy policy that is located inside the database where the dataset was originally generated/collected. The CoMaFeDS platform then checks the respective data source for existing updates of the privacy policy in periodic intervals. If the answer is positive, it updates the internal hippocratic integration model.

C. Discussion

A potential drawback of our briefly sketched platform could be that a mendacious potential utilizer gives bogus statements about her identity or the purpose of data processing. Such behaviour could be counteracted with an accreditation process that verifies the given identity and information in advance. If this verification is satisfactory, electronic certificates can be issued for the utilizer. As, however, such accreditation process generates efforts on all sides, it is not intended in the first iteration of the proposed design.

Another possible objection emanates from the question whether a fraudulent utilizer can exploit the platform to obtain lots of personal data after the access is granted and then copy and use them for every purpose she wants. Even though it is not obvious in the proposal, we want to make clear that there is no way for a utilizer to get direct access to the data sources. All accesses are managed by the platform and data are presented to external data mining applications as if they were stored in one big database. This approach allows it to perform data mining against the platform without the possibility to copy specific datasets.

V. Outlook & Conclusion

In this paper, we highlighted the importance of consent-provision for the viability of future big data applications based on multiple and federated sources of personal data. From the perspective of data protection law as well as from the broader view of privacy, the provision of individual consent to specific purposes and data utilizers is crucial for achieving the goal of "informational self-determination". There is thus a clear need for future information systems employing federated data sources to implement appropriate mechanisms for consent management.

We discussed and analyzed several existing approaches for technically supporting such consent management. Hippocratic databases allow to regulate data access based on individual privacy policies that can, in particular, also specify accepted purposes of data usage in relation to different potential utilizers. Sticky Policies follow a comparable approach but keep these policies sticked to the respective datasets, thus allowing for more flexibility and decentralization than hippocratic databases. On the other hand, they introduce the critical role of a "trusted authority". Distributed usage control, in turn, extends the "Sticky Policies" concept by more sophisticated mechanisms for enforcement and monitoring of data usage after initial access but has the drawback of significant implementation efforts for all participating parties. Finally, the concept of dynamic consent allows for more flexible specifications of desired purposes and utilizers and particularly offers ad-hoc modifications of given consent.

Based on these concepts, we propose a novel approach for a consent management platform specifically tailored to the increasingly relevant use cases employing multiple, federated sources of personal data for cloud-based big data analytics. Our concept shall allow for integrated queries upon multiple and autonomous data sources, taking into account individually given, purpose- and utilizer-specific, and dynamically adjustable consent provided by data subjects. Our next steps will include the prototypical implementation and more detailed delineations on the legal dimension of technically mediated consent. Later extensions might also include the incorporation of mechanisms for distributed usage control.

References

[1] A. F. Westin, *Privacy and Freedom*. New York: Atheneum, 1967.

[2] *Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data*. 1995.

[3] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Hippocratic Databases," in *Proceedings of the 28th International Conference on Very Large Data Bases*, Hong Kong, China, 2002, pp. 143–154.

[4] Y. Laura-Silva and W. Aref, "Realizing Privacy-Preserving Features in Hippocratic Databases," Purdue University, Department of Computer Science Technical Reports, Dec. 2006.

[5] L. F. Cranor, "P3P: making privacy policies more useful," *IEEE Security & Privacy*, vol. 1, no. 6, pp. 50–55, Nov. 2003.

[6] J. Reagle and L. F. Cranor, "The Platform for Privacy Preferences," *Commun. ACM*, vol. 42, no. 2, pp. 48–55, Feb. 1999.

[7] J. Azemović, "Data Privacy in SQL Server based on Hippocratic Database Principles," *The Microsoft MVP Award Program*, 2012. [Online].Available: http://blogs.msdn.com/b/mvpawardprogram/archive/2012/07/30/data-privacy-in-sql-server-based-on-hippocratic-database-principles.aspx. [Accessed: 04-Dec-2015].

[8] S. Pearson and M. C. Mont, "Sticky policies: an approach for managing privacy across multiple parties," *Computer*, vol. 44, no. 9, pp. 60–68, 2011.

[9] M. C. Mont, S. Pearson, and P. Bramhall, "Towards accountable management of identity and privacy: sticky policies and enforceable tracing services," in *14th International Workshop on Database and Expert Systems Applications, 2003. Proceedings*, 2003, pp. 377–382.

[10] E. W. Felten, "Understanding trusted computing: will its benefits outweigh its drawbacks?," *IEEE Security Privacy*, vol. 1, no. 3, pp. 60–62, May 2003.

[11] J. Brenner, "ISO 27001: Risk management and compliance," *Risk management*, vol. 54, no. 1, p. 24, 2007.

[12] A. Pretschner, M. Hilty, and D. Basin, "Distributed Usage Control," *Commun. ACM*, vol. 49, no. 9, pp. 39–44, Sep. 2006.

[13] E. Lovat and A. Pretschner, "Data-centric Multi-layer Usage Control Enforcement: A Social Network Example," in *Proceedings of the 16th ACM Symposium on Access Control Models and Technologies*, New York, NY, USA, 2011, pp. 151–152.

[14] J. Kaye, E. A. Whitley, D. Lund, M. Morrison, H. Teare, and K. Melham, "Dynamic consent: a patient interface for twenty-first century research networks," *Eur J Hum Genet*, vol. 23, no. 2, pp. 141–146, Feb. 2015.