Big Data and Analytics in the Age of the GDPR

Bonatti, Piero A.; Kirrane, Sabrina

Published: 08/07/2019

Document Version
Peer reviewed version

Big Data and Analytics in the Age of the GDPR

Piero A. Bonatti
Dep. of Electrical Eng. and Information Technologies
Università di Napoli Federico II
Naples, Italy
pab@unina.it

Sabrina Kirrane
Department of Information Systems and Operations
Vienna University of Economics and Business
Vienna, Austria
sabrina.kirrane@wu.ac.at

***Abstract*—The new European General Data Protection Regulation places stringent restrictions on the processing of personally identifiable data. The GDPR does not only affect European companies, as the regulation applies to all the organizations that track or provide services to European citizens. Free exploratory data analysis is permitted only on anonymous data, at the cost of some legal risks. We argue that for the other kinds of personal data processing, the most flexible and safe legal basis is *explicit consent*. We illustrate the approach to consent management and compliance with the GDPR being developed by the European H2020 project SPECIAL, and highlight some related big data aspects.**

***Keywords*-Personal Data Processing, Analytics, Anonymization, Usage Control Policies, GDPR**

I. Introduction

The new European General Data Protection Regulation[1] (GDPR), that has come into force on May 25, 2018, places stringent restrictions on the processing of personally identifiable data. The GDPR does not only affect European companies, as the regulation applies to *all* the organizations that track or provide services to European citizens (cf. Article 3).[2] Infringements may severely affect the reputation of the violators, and are subject to substantial administrative fines (up to 4% of the total worldwide annual turnover or 20 million Euro, whichever is higher). Therefore, the risks associated to infringements constitute a major disincentive to the abuse of personal data. Given that the collection and the analysis of personal data are paramount sources of innovation and revenue, companies are interested in maximizing personal data usage within the limits posed by the GDPR. Consequently, *data controllers* (i.e. the personal and legal entities that process personal data) are looking for methodological and technological means to comply with the regulation's requirements efficiently and safely.

The GDPR is changing the way personal data are processed. It states that by default, personal data shall not be processed, and in this way it encourages the use of *anonymous data*. They are not regarded as personal data, so anonymous data lie outside the scope of the GDPR and can be freely used. Subsequently, the regulation introduces a list of exceptions to the default prohibition. Personal data can be collected, stored, and analyzed according to the legal bases defined in Article 6 of the GDPR. Some examples of such legal bases include public interest, the vital interests of the data subject, contracts, and the legitimate interests of the data controller, just to name a few. These legal bases are constrained by a number of provisos and caveats that restrict their applicability.[3] So, in practice, the kinds of personal data processing that are most useful for data-driven applications are almost exclusively allowed by another legal basis, namely, the *explicit consent* of the data subjects.[4]

In the following, we discuss the two mainstream options offered by the GDPR, namely anonymization and explicit consent. We will argue that currently anonymization techniques cannot be extensively applied in the applications grounded on big data processing and analytics; this will lead us to focus on consent management. The approach to anonymization and consent management we propose here is being developed within the European H2020 project SPECIAL.[5] The main use cases of the project are provided by three of its industrial partners, namely Deutsche Telekom, Proximus, and Thomson Reuters.

The remainder of the paper is structured as follows: *Section* II presents the challenges associated with ensuring that data is legally anonymous. *Section* III discusses the role of consent as a legal basis for personal data processing, and the machine-understandable encoding of consent. *Section* IV examines the aspects of consent management that may be regarded as big data themselves. *Section* V points to alternative work on GDPR compliance and differentiates our work from related approaches. Finally, we present our conclusions and interesting directions for future work in *Section* VI.

[1] http://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf

[2] Moreover, the case of Cambridege Analytica and the increasing number of information abuses and related crimes are fostering discussions outside Europe about the opportunity of adopting regulations similar to the GDPR. California has recently approved a *Consumer privacy act* (CCPA) that is similar to GDPR in many respects, see for example https://iapp.org/news/a/gdpr-matchup-california-consumer-privacy-act/ for a comparison.

[3] Of particular relevance here are the data minimization principle introduced in Article 5, and the limitations to the legitimate interests of the controller rooted in Article 6.1(f).

[4] Article 6.1(a)

[5] https://www.specialprivacy.eu/

II. Anonymization

Article 4.(1) of the GDPR states that for the purpose of the regulation, "personal data" means any information relating to an *identified or identifiable* natural person (the *data subject*), where:

> an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

Moreover, recital[6] 26 explains that:

> To determine whether a natural person is identifiable, account should be taken of *all the means reasonably likely to be used* [...] either by the controller or by another person to identify the natural person directly or indirectly.

The available technologies for analytics, that leverage the available big data sources, are superb examples of such indirect identification means. Their effectiveness in identifying individuals based on their "behavioral fingerprint" is witnessed by a number of applications and cases, illustrated in detail in the extensive report by Christl and Spiekermann [1]. In this context it is difficult to assess whether any of the available anonymization methods produces data that are *legally* anonymous. In the following we discuss this issue with respect to the two main families of anonymization approaches, namely:

- k-anonymity and its evolution, such as l-diversity and t-closeness, just to name a few;
- ϵ-differential privacy and its refinements.

The reader is referred to [2] for an overview of anonymisation criteria and techniques – including the above approaches – and a discussion of their pros and cons.

The methods applying to the first family typically operate by removing information (e.g. by deleting day and month from birth dates); the methods applying to the second family typically introduce noise in query answers or directly in the data, in a controlled way (special methods are available for non-numeric data). Still there exist relationships between the two families, see for example [3], [4].

A first difficulty in avoiding the GDPR's restrictions through data anonymization is that the guarantees provided by anonymization methods and the requirements posed by the GDPR have different natures. For example, the goal of k-anonymity is having each record match at least k different individuals. Clearly, as k grows, the re-identification of data subjects does not become easier, but how can one assess, for a given k, that there is no "reasonable" mean to re-identify a data subject, as required by the regulation's definition of anonymous data? Similar arguments hold for the other methods in the same family, i.e. l-diversity and t-closeness. Differential privacy guarantees that by anonymizing a dataset (or its view), the knowledge about "which data subjects are described in the dataset" is approximately the same before and after querying the dataset's anonymized view. Such knowledge is expressed in terms of all pairs of similar databases, that differ in one record (the one that represents the considered data subject); in particular, the ratio between the probability of the two datasets – given the query result – should be bounded by $\exp(\epsilon)$, where ϵ is a given parameter. Again, there is no formal way of assessing that the legal notion of anonymity is met for a given ϵ (also due to the influence of background knowledge, discussed below).

Will the legislators eventually stipulate that the GDPR's definition of anonymity is met for some standardized values of the parameters k, l, t and ϵ? This is unlikely, given that the degree of protection ensured by any parameter choice depends on the additional data sources available to an attacker, and that the amount of such background knowledge is difficult to estimate. In more formal terms, all anonymization methods are vulnerable to attacks based on background knowledge (and, in particular, to the aggregation of different information sources), see for example [5], [6], [7], [8]. Therefore, even if the mismatch between the technical and the legal definitions were reconciled, still the difficulty of estimating the amount of available backround knowledge would be reflected on the estimate of acceptable parameter values.

A further, well known issue in data anonymization is that by removing details and introducing noise, anonymisation methods decrease data quality and – consequently – data *utility* [2], [9], [6], [10], [7]. The required anonymisation level may turn out to be incompatible with the necessary data quality in many applications. An immediate consequence of the negative effects of anonymization on data quality is that the problem of estimating the available background knowledge cannot be simply bypassed through a cautious, very "conservative" choice of the privacy parameters.

In the light of the above discussion it is clear that – as of today – anonymous data and anonymized data are two distinct concepts, and that data controllers make use of anonymization methods at their own risk. The regulation's youth and the consequent lack of court decisions make the assessment of legal risks even harder. For this reason, SPECIAL is focussing its efforts on the other mainstream approach at personal data processing, that is, consent management (that is the subject of the next section). Still, anonymization methods play an interesting role, that will be illustrated in the following.

[6] In law, a recital consists of an account or repetition of the details of some legal document, that clarifies the document's purpose and its intended interpretation.

III. Data Usage Descriptions and Consent Management

The GDPR poses at least two requirements that call for a machine-understandable representation of data usage modalities.

Article 30 states that each controller shall maintain a record of the personal data processing activities under its responsibility. The first paragraph specifies that such a ledger should describe (among other information) the following aspects of *data usage*:

P1. the *purpose* of processing;
P2. a description of the *categories of data subjects* and of the *categories of personal data*;
P3. the categories of *recipients* to whom the personal data have been or will be disclosed;
P4. *transfers* of personal data to a third country or an international organisation (since cross-border data transfer are subject to limitations);
P5. the envisaged *time limits for erasure* of the different categories of data;
P6. information about the *processing*, such as the security measures mentioned in Article 32.

Recital 42 stresses that, where processing is based on the data subject's consent, the controller should be able to demonstrate that the data subject has given consent to the processing operation. SPECIAL addresses this issue by recording consent in a *transparency ledger*. The description of consent is similar to the description of processing activities as per Article 30. While Article 6.1.(a) – that introduces consent as a legal basis for personal data processing – and Recital 42 explicitly mention only the purpose of processing, Articles 13 and 14 add the other elements P2–P6 listed above. Concerning P6 (processing), it should be specified whether any automated decision making is involved, including profiling.

Once such data usage descriptions are encoded in a machine-understandable way, several tasks, related to GDPR compliance, can be automated, including:

T1. Checking whether the processing complies with several restrictions imposed by the GDPR, such as additional requirements on the processing of sensitive data, restrictions on cross-border transfers, and compatibility of data usage with the chosen legal basis. This kind of validation requires a machine-understandable formalization of the relevant parts of the GDPR.
T2. Checking whether a specific operation is permitted by the available consent.
T3. Running ex-post auditing on the controller's activities. In SPECIAL this task is supported by logging data processing events in the transparency ledger, and comparing such events with consent.
T4. Finding the consent that justifies a specific processing (for auditing or responding to a data subject's enquiry).

The transparency ledger can also be used to provide dashbords to data subjects, that support them in monitoring the use of their data and *explaining* why their consent allowed specific operations. Such dashboards can also be used as a uniform interface to let data subjects exercise their rights (access to data, right to erasure, etc.) as specified by Articles 15–18 and 21–22.

A. Issues in Collecting Consent

The range of approaches to collecting consent from data subjects lies between two extreme approaches.

- (Purely static) Consent is requested a priori, listing in one document all the possible variants of data processing that the controller may carry out in the future. The result is a very long document that users typically do not fully read and do not fully understand.
- (Purely dynamic) Consent is requested on the fly for each specific operation being executed. The user is pestered by requests, many of which are similar.

In both cases, data subjects may be induced to deny consent, and possibly stop using the service. SPECIAL is addressing this issue in two ways:

- Experimenting with a novel, incremental dynamic approach that lies in between the above two extremes [11].
- Re-using previous consent, as explained in the following sections.

Another difficulty that currently seems to have no solution is related to exploratory data analysis and mining. This kind of processing fosters innovation by discovering new knowledge, that may suggest novel services and applications. However, by its very nature, such exploratory analyses do not have any specific pre-conceived purpose, and – unfortunately – "finding novel interesting relationships among data" is considered too vague to constitute the purpose of a valid consent request.[7] Consequently, *exploratory analyses are possible only on anonymous data*, under the difficulties mentioned in Section II.

Anonymization techniques may also be used together with consent to increase the percentage of opt-in's. In this case anonymization acts as a guarantee of protection, that may encourage data subjects to consent to data processing. The processing – at the same time – is legally permitted by consent, so the probability of re-identification, in this case, does not affect compliance with the GDPR.

B. Encoding Usage Descriptions and Consent

The common structure of the activity records and of the consent forms, consisting of properties P1–P6, is called *simple (usage) policy* in SPECIAL. In general, both the controller's activities and the consent of data subjects can

[7] Consent must be "specific" (par. 4.(11) and 6.1(a)); recital 39 insists that "purposes [...] should be [...] determined at the time of collection of the personal data". The only case where purposes might not be "fully identified" a priori is scientific research, cf. recital 33.

be described by a *set* of simple usage policies (covering different data categories and purposes), called *full (usage) policies*. Each simple policy can be specified simply by attaching to each property P_i (such as purpose, data category, recipients, etc.) a term selected from a suitable *vocabulary* (ontology).

Example 3.1: A company – call it BeFit – sells a wearable fitness appliance and wants (i) to process biometric data (stored in the EU) for sending health-related advice to its customers, and (ii) share the customer's location data with their friends. Location data are kept for a minimum of one year but no longer than 5; biometric data are kept for an unspecified amount of time. In order to do all this legally, BeFit needs consent from its customers. Consent can be represented with two simple policies, specified using SPECIAL's vocabularies:

```
{
  has_purpose: FitnessRecommendation,
  has_data: BiometricData,
  has_processing: Analytics,
  has_recipient: BeFit,
  has_storage: { has_location: EU }
}
{
  has_purpose: SocialNetworking,
  has_data: LocationData,
  has_processing: Transfer,
  has_recipient: DataSubjFriends,
  has_storage: {
      has_location: EU,
      has_duration: [1year,5year]
      }
}
```

If "HeartRate" is a subclass of "BiometricData" and "ComputeAvg" is a subclass of "Analytics", then the above consent allows BeFit to compute the average heart rate of the data subject in order to send her fitness recommendations. BeFit customers may restrict their consent, e.g. by picking a specific recommendation modality, like "recommendation via SMS only". Then the first line should be replaced with something like

```
has_purpose:{
  FitnessRecommendation,
  contact: SMS}
```

Moreover, a customer of BeFit may consent to the first or the second argument of the union, or both. Then her consent would be encoded, respectively, with the first simple policy, the second simple policy, or both. Similarly, each single process in the controller's lines of business may use only biometric data, only location data, or both. Accordingly, it may be associated to the first simple policy, the second simple policy, or both. ∎

SPECIAL's vocabularies are temporarily derived by adapting previous standardized terms introduced by initiatives related to privacy and DRM, such as P3P[8] and ODRL,[9] while more refined vocabularies are being developed through W3C's *Data Privacy Vocabularies and Controls Community Group*, (DPVCG),[10] promoted by SPECIAL and spanning over a range of stakeholders wider than the project's consortium.

As shown in Example 3.1, usage policies can be formatted with a minor extension of Jason (in particular, compound terms and policy sets require additional operators), while vocabularies can be encoded in RDFS or lightweight profiles of OWL2 such as OWL2-EL and OWL2-QL.

Internally, SPECIAL's components encode also policies and the entries of the transparency ledger with a fragment (profile) of OWL2 called $\mathcal{PL}$ (policy logic) [12]. The adoption of a logic-based description language has manifold reasons. First, it has a clean, unambiguous semantics, that is a must for policy languages. A formal approach brings the following advantages:

- strong correctness and completeness guarantees on the algorithms for permission checking and compliance checking;
- the mutual coherence of the different reasoning tasks related to policies, such as policy validation, permission checking, compliance checking, and explanations (cf. tasks T1–T4 and the subsequent paragraph);
- correct usage after data is transferred to other controllers (i.e. interoperability).

The last point is related to so-called *sticky policies* [13], that constitute a sort of a license that applies to the data released to third parties. It is essential that all parties understand the sticky policy in the same way.

Policies are modelled as OWL2 *classes*. If the policy describes a controller's activity, then its instances represent all the operations that the controller may possibly execute. If the policy describes consent, then its instances represent all the operations permitted by the data subject. A description of (part of) the controller's activity – called *business policy* in SPECIAL – *complies* with a consent policy if the former is a subclass of the latter, that is, all the possible operations described by the business policies are also permitted by the given consent.

Example 3.2: Consider again Example 3.1. The Jason-like representation used there can be directly mapped onto an OWL2 class `ObjectUnionOf(`P_1 P_2`)`, where P_2 is:

```
ObjectIntersectionOf(
  ObjectSomeValueFrom(
    has_purpose SocialNetworking )
  ObjectSomeValueFrom(
    has_data LocationData)
  ObjectSomeValueFrom(
    has_processing Transfer)
  ObjectSomeValueFrom(
    has_recipient DataSubjFriends)
  ObjectSomeValueFrom(
    has_storage ObjectIntersectionOf(
```

[8] http://www.w3.org/TR/P3P11

[9] https://www.w3.org/TR/odrl/

[10] www.w3.org/community/dpvcg/

```
      ObjectSomeValueFrom(has_location: EU)
      DataSomeValueFrom(has_duration
        DatatypeRestriction(xsd:integer
          xsd:minInclusive "365"^^xsd:integer
          xsd:maxInclusive "1825"^^xsd:integer
)))
```

(with omit P_1 due to space limitations; the reader may easily derive it by analogy with the above example).

In order to check whether a business policy BP (encoded as an OWL2 class) complies with the above policy one should check whether

$$\texttt{SubClassOf(}BP\ \texttt{ObjectUnionOf(}P_1\ P_2\texttt{))}$$

is a logical consequence of the ontology that defines SPECIAL's vocabularies. ∎

The encoding of policies as classes *facilitates the re-use of available consent* (which is the preferred option, when applicable, given the impact of repeated consent requests on user experience). The GDPR sometimes allows to process personal data for a purpose other than that for which the data has been collected, provided that the new purpose is "compatible" with the initial purpose.[11] Compatibility cannot be assessed automatically, in general, because it is not formalized in the regulation, and involves enough subtleties to need the assessment of a lawyer. However, by expressing purposes as classes, one can at least have the data subject consent upfront to a specified range of "similar" purposes. Roughly speaking, the accepted class of purposes is like an agreement – between data subjects and controllers – on which purposes are "compatible" in the given context. Also expressing the other policy properties (P2–P6) as classes is beneficial. As data subjects consent to wider classes of usage modalities, the need for additional consent requests tends to decrease, thereby improving usability and reducing the costs associated to consent requests. Consider that sometimes the difficulties involved in reaching out the data subjects, and the concern that too many requests may annoy users, make controllers decide *not* to deliver a service that requires additional consent.

C. Fast Compliance Checking Algorithms

Business policies (that describe the processing of each of the controller's processes) are not only needed to fulfil the requirements of Article 30. They can also be used to check whether a running process complies with the available consent, as a sort of access control system. Several implementation strategies are possible, depending on the controller's system architecture; to fix ideas, the reader may consider the following generic approach: Each of the controller's processes is labelled with a corresponding business policy that describes it, and before processing a piece of data, the business policy is compared with the data subject's consent to check whether the operation is permitted.

[11] See for example articles 5.1 (b) and 6.4.

In general, such compliance checks occur frequently enough to call for a scalable implementation. Consider, for example, a telecom provider that collects location information to offer location-based services. Locations cannot be stored without a legal basis, such as law requirements or consent – not even temporarily, while a batch process selects the parts that can be legally kept. So compliance checking needs to be executed on the fly. In order to estimate the amount of compliance checks involved, consider that the events produced by the provider's base stations are approximately 15000 per second; the probing records of wi-fi networks are about 850 millions per day.

In order to meet such performance requirements, SPECIAL has developed ad-hoc reasoning algorithms for $\mathcal{PL}$ [12], that leverage $\mathcal{PL}$'s simplicity to achieve unprecedented reasoning speed. Compliance checking is split into two phases: first, business policies are normalized and closed under the axioms contained in the vocabularies; in the second phase, business policies are compared with consent policies with a *structural subsumption* algorithm. We have just completed the evaluation of a sequential Java implementation of those algorithms, called PLR. We chose Java to facilitate the comparison with other engines, by exploiting the standard OWL APIs, and we refrained to apply parallelization techniques in order to assess the properties of the basic algorithms. Before discussing more performant implementation options, we report the results for PLR.

PLR can pre-compute the first phase, since the business policies are known in advance and are typically persistent. So the runtime cost is reduced to structural subsumption. In this way, on the test cases derived from SPECIAL's use cases (cf. Table I), the performance we achieve, respectively, is 150μsec and 190μsec per compliance check, using the following system:

processor:	Intel Xeon Silver 4110
cores:	8
cache:	11M
RAM:	198 GB
OS:	Ubuntu 18.4
JVM:	1.8.0_181
heap:	32 GB (actually used: less than 700 MB).

This means that PLR alone can execute about 6000 compliance checks per second and more than 518 millions per day, that is, 60% of wi-fi probing events and 40% of base station events.

In order to raise performance up to the required levels, one can re-engineer PLR using a language more performant than Java, and/or parallelize processing by means of big data architectures (discussed below). Compliance checking is particularly well suited to parallelization, since each test is independent from the others and no synchronization is required. Additionally, the investigation of parallelization within PLR's algorithms is under investigation.

	Pilot 1	Pilot 2
Ontology		
inclusions	186	186
disjoint class axioms	11	11
property range axioms	10	10
functional properties	8	8
classification hierarchy height	4	4
Business policies		
# generated policies	120	100
avg. simple pol. per full pol.	2.71	2.39
Consent policies		
# generated policies	12,000	10,000
avg. simple pol. per full pol.	3.77	3.42
Test cases		
# compliance checks	12,000	10,000

Table I: Size of the test cases derived from SPECIAL's pilots

D. Compliance Checking Architectures

SPECIAL-K is an Apache Kafka[12] distributed streaming platform that enables personal data processing compliance checking and transparency. The SPECIAL-K system components, depicted in *Figure* 1, include:

Personal Data Inventory: This component is responsible for a preliminary analysis aimed at determining: What data is collected and which data points would be classified as personal data; what is the purpose of data collection and processing; where are collected data stored; for how long are the data stored; with whom is the data shared. Such information is needed in order to configure the SPECIAL-K architecture.

Personal Data Gateway: The gateway component is responsible for enabling the personal data processing/sharing events generated by existing Line of Business applications to be intercepted by Apache Kafka.

Applications & Personal Data Processing Topics: Each application log is represented using a distinct Kafka topic, while a separate compliance topic is used to store the enriched log after compliance checks have been completed.

MongoDB & Consent and Policy Log Kafka Topic: The prevailing consent, usage policies and the respective vocabularies are stored in a Mongo database, while changes to consent and business policies are recorded in a Kafka topic called "Consent and Policy Log".

Compliance Checker: The compliance checker component uses the consent together with business policies and the application logs provided by Kafka to check that data processing and sharing complies with the relevant usage control policies. The results of this check are saved onto a new Kafka topic called "Compliance Log".

Consent, Transparency & Compliance Backends: The interaction between the various architectural components is managed by mu.semte.ch[13] an open source micro-services framework for building RDF enabled applications.

Consent, Transparency & Compliance Dashboards: Users interact with the system via the consent management user interface and the transparency and compliance dashboards. The former supports granting and revoking consent for processing and data sharing. The latter provides the data subject with transparency with respect to data processing and sharing events in a digestible manner.

Elasticsearch: As logs can be serialized using JSON-LD, it is also possible to benefit from the faceting browsing capabilities of Elasticsearch[14] and the out of the box visualization capabilities provided by Kibana[15].

In order to improve the performance of our PLR algorithm (which alone is only able to deal 60% of wi-fi probing events and 40% of base station events), we are currently in the process of preparing the SPECIAL-K benchmark. The objective of the benchmark is to evaluate the performance of the PLR algorithm based on realistic test cases derived from SPECIAL's pilots, and to further stress test the algorithm from a scalability perspective using synthetic test cases of increasing size. The goal of the former tests is to investigate the suitability of the SPECIAL-K platform to cater for realistic policies, while the latter focus on identifying potential choke points or bottlenecks.

IV. Big Data Aspects

Big data are not only the input of anonymization and analytics, or the domain to be modelled with SPECIAL's vocabularies. This section illustrates – in terms of the usual four 'v' – the aspects related to big data that arise in SPECIAL's components themselves.

Volume: The SPECIAL-K system enables data subjects, controllers, processors, and supervisory authorities to verify that personal data is processed according to consent granted by the respective data subject. It does so by recording a history of changes to consent in the Consent and Policy Log Kafka Topic, by storing all data processing and sharing events in the Personal Data Processing Topic, and by persisting the result of the compliance checking in the Compliance Log Kafka Topic. Thus in SPECIAL, Kafka plays the role of a high-performance, low latency, distributed filesystem, which is used to persist our append only logs.

Velocity: Considering the need for real time compliance checking, arising from the SPECIAL uses cases, SPECIAL-K is designed to deal with a constant flow of data events (i.e., an event stream). In the case of the location based services, discussed in section III, this roughly equates to approximately 15000 base station events per second; and 850 million wifi network probing events per day.

[12]https://kafka.apache.org/

[13]https://mu.semte.ch/

[14]https://www.elastic.co/products/elasticsearch

[15]https://www.elastic.co/products/kibana

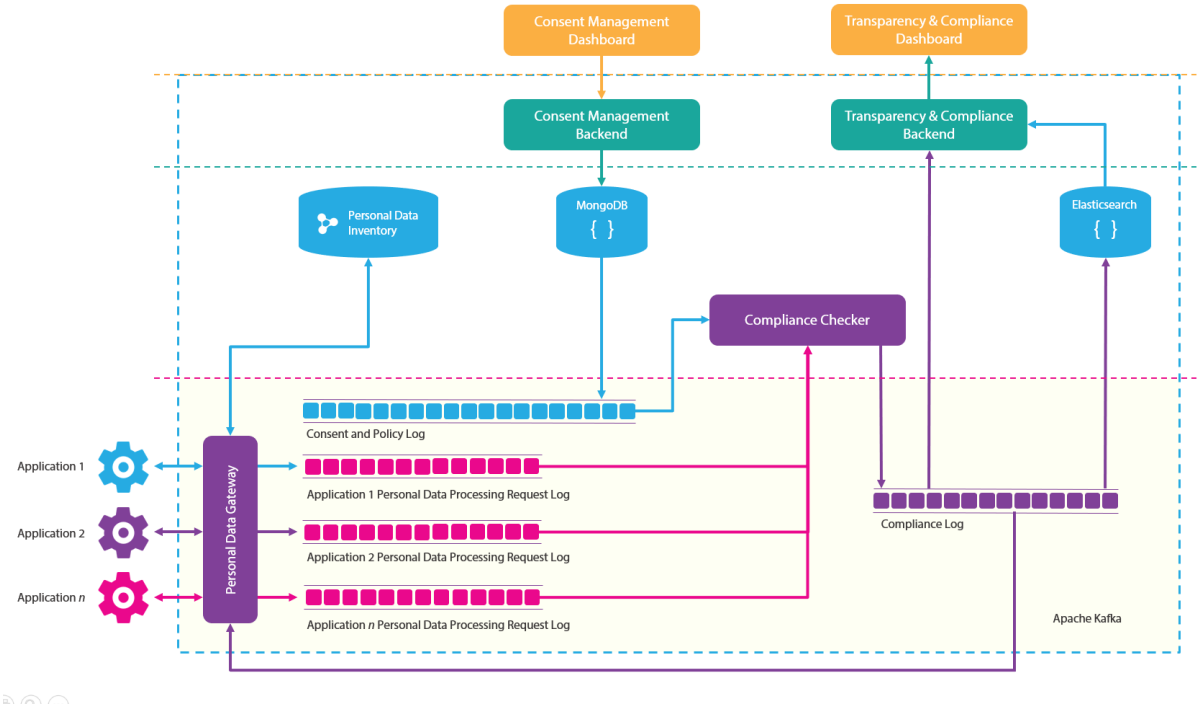

Figure 1: SPECIAL-K architecture [14]

Variety: SPECIAL's components and architecture shall be integrated in existing systems, whose application domain cannot be restricted a priori. Both the data categories and their encoding (format) may significantly vary across different controllers and (for large companies) even within the organization's borders. The dashboards that support data subjects in exercising their rights and monitoring data usage shall process such multiplicity of data, and interact with different instantiations of SPECIAL, which adds to the variety of the involved data. SPECIAL addresses variety via the personal data inventory (from an analysis and semantic representation perspective), the personal data gateway (enabling existing Line of Business applications to interface with the SPECIAK-K platform), and the DPVCG's standardization initiative (cf. Section III-B). The latter is aimed at creating a framework of general classes of common interest, that can be extended with application-specific terms. Such a framework enhances interoperability and guides the formulation of application-specific vocabularies.

Veracity: From a SPECIAL perspective, veracity is about the faithfulness of policies and log events with regard to the human-readable text in the consent requests and the actual behavior of the system. Here, there is a need to automatically construct human readable text from policy annotations, to automatically reconstruct systems workflows from data processing and sharing events, and to verify the correctness of the human readable policies and workflows with the help of human experts. Both of which are the subject of future work.

V. Related Work

Considering the well known utility versus privacy trade-off associated with the suppression and generalization techniques for anonymization, there is a new stream of research which focuses on using deep learning to synthesize data, which is representitive of real data [15], [16]. Although such personal data synthesis could potentially address the utility issue, the privacy guarantees offered by such approaches is still an open research question.

From a GDPR compliance perspective, there exist several compliance tools (cf. [17], [18], [19], [20]) that enable companies to assess the compliance of their applications and business processes via predefined questionnaires. Additionally, there is a body of work that focuses on modelling the text of the GDPR in a manner that supports legal reasoning and compliance checking [21], [22], [23], [24], [25]. Other work in this area, demonstrates how the European Legislation Identifier (ELI) ontology can be used to model the GDPR as linked data [26]. Outputs include a DCAT[16] catalog containing the official text of the GDPR and a SKOS[17] ontology defining concepts related to GDPR. De Hert et al. [27] propose a systematic interpretation of

[16]DCAT, https://www.w3.org/TR/vocab-dcat/
[17]SKOS, https://www.w3.org/TR/skos-reference/

the right to data portability from both a minimalist and an empowering perspective.

Both rule languages and OWL2 have already been used as policy languages; a non-exhaustive list is [28], [29], [30], [31], [32]. As noted in [33], the advantage of OWL2 – hence description logics – is that all the main policy-reasoning tasks are decidable (and tractable if policies can be expressed with OWL2 profiles), while compliance checking is undecidable in rule languages, or at least intractable – in the absence of recursion – because it can be reduced to datalog query containment. So an OWL2-based policy language is a natural choice in a project like SPECIAL, where policy comparison is the predominant task. Among the aforementioned languages, both Rei and Protune [31], [32] support logic program rules, which make them unsuitable to SPECIAL's purposes. KAoS [30] is based on a description logic that, in general, is not tractable, and supports role-value maps – a construct that easily makes reasoning undecidable (see [34], Chap. 5). The papers on KAoS do not discuss how to address this issue.

P3P's privacy policies – that are encoded in XML – and simple $\mathcal{PL}$ policies have a similar structure: the tag `STATEMENT` contains tags `PURPOSE`, `RECIPIENT`, `RETENTION`, and `DATA-GROUP`, that correspond to the analogous properties of SPECIAL's usage policies. Only the information on the location of data is missing. The tag `STATEMENT` is included in a larger context that adds information about the controller (tag `ENTITY`) and about the space of web resources covered by the policy (through so-called *policy reference files*). Such additional pieces of information can be directly encoded with simple $\mathcal{PL}$ concepts.

There exist several well-engineered reasoners for OWL2 and its profiles. Hermit [35] is a general reasoner for OWL2. Over the test cases inspired by SPECIAL's use cases, it takes 3.67 ms and 3.96 ms per compliance check, respectively, that is, over 20 times longer than PLR. ELK [36] is a specialized polynomial-time reasoner for the OWL2-EL profile. It does not support functional roles, nor the interval constraints used to model storage duration, therefore it cannot be used to reason on the $\mathcal{PL}$ profile. Konclude [37] is a highly optimized reasoner with "pay-as-you-go" strategies (i.e. it becomes more efficient on less complex profiles of OWL2). Konclude is designed for classification, and is currently not optimized for subsumption tests (i.e. the reasoning task underlying compliance checks). Consequently, it turns out to be slower than Hermit on our test cases.

VI. Conclusion, Further Issues, and Future Work

The GDPR allows exploratory analytics to be carried out only on anonymous data. Unfortunately, the anonymization techniques available today are not guaranteed to produce results that are anonymous in the GDPR's sense. Moreover, anonymization is known to reduce the quality and the utility of data. As a consequence, in many cases it is preferrable to adopt explicit consent as the legal basis for personal data processing.

It is profitable to adopt a semantic representation of consent, in order to ensure correct interoperability (e.g. through sticky policies), and get stronger correctness and completeness guarantees on the several algorithms that perform compliance checking, support auditing and user rights, and provide explanations about the consequences of policies and consent. Semantic representation does not mean slow processing. With suitable algorithms and big data architectures, it is possible to achieve the required performance in demanding application contexts.

The class-based consent management proposed in SPECIAL may also provide a viable means to apply exploratory analyses and mining techniques to non-anonymous data, to some extent. For instance, SPECIAL could be used in order to obtain consent for the "analysis" of "location data" for the purpose of "*improving BeFits wearable fitness appliances*" (a flexible definition, yet still more specific than the fully generic "finding novel interesting relationships among data"). Data mining could then be applied to the bundle of personal data of all customers that consent to such data usage. Of course, the company would need to ensure that the processing complies with the stated consent policy; for example, only location data should be analyzed, and the goal of their analysis should be "improving BeFits wearable fitness appliances" and nothing else. From a legality perspective, it has yet to be determined if "*improving BeFits wearable fitness appliances*" would be seen as specific enough. Additionally, in general, it is unknown at this point how to prevent the analysis from leading to other insights, not permitted by the consent policy.

The SPECIAL-K system is based on the Apache Kafka distributed streaming platform, which has already proven its effectiveness in terms of its ability to handle high volumes and velocity (cf. [38], [39], [40]). However, when it comes to compliance checking and transparency for personal data processing, there are a number of open challenges concerning variety and veracity. From a variety perspective, there is the need to semantically encode the data, purpose, processing, storage and sharing associated with existing Line of Business applications, however a necessary first step is to determine where personal data is located and how it is used within the company. The SPECIAL consortium are currently investigating the use of automated data and knowledge discovery techniques to gain insights into personal data which is usually scattered across several systems (cf. [41], [14]). Whereas from a veracity perspective, it is necessary to verify that business policies faithfully represent the data processing and sharing performed by the corresponding Line of Business application. Here, existing work on process conformance checking, such as [42], [43], provide for interesting starting points. Additionally, there is

a need to automatically construct human readable policies from SPECIAL usage policies, for instance by leveraging existing approaches for the generation of text from computer-executable code (cf. [44], [45]).

ACKNOWLEDGMENT

This research is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N. 731601. The authors are grateful to all of SPECIAL's partners; without their contribution this project and its results would not have been possible.

REFERENCES

[1] W. Christl and S. Spiekermann, *Networks of Control: A Report on Corporate Surveillance, Digital Tracking, Big Data & Privacy*. Facultas, 2016. [Online]. Available: https://books.google.it/books?id=sE45vgAACAAJ

[2] G. D. Acquisto, J. Domingo-Ferrer, P. Kikiras, V. Torra, Y. de Montjoye, and A. Bourka, "Privacy by design in big data – an overview of privacy enhancing technologies in the era of big data analytics," ENISA Report, V 1.0, 2015, https://www.enisa.europa.eu/publications/big-data-protection.

[3] N. Li, W. H. Qardaji, and D. Su, "On sampling, anonymization, and differential privacy or, *k*-anonymization meets differential privacy," in *7th ACM Symposium on Information, Compuer and Communications Security, ASIACCS '12, Seoul, Korea, May 2-4, 2012*, H. Y. Youm and Y. Won, Eds. ACM, 2012, pp. 32–33. [Online]. Available: http://doi.acm.org/10.1145/2414456.2414474

[4] S. P. Kasiviswanathan and A. D. Smith, "A note on differential privacy: Defining resistance to arbitrary side information," *CoRR*, vol. abs/0803.3946, 2008. [Online]. Available: http://arxiv.org/abs/0803.3946

[5] X. Xiao, Y. Tao, and N. Koudas, "Transparent anonymization: Thwarting adversaries who know the algorithm," *ACM Trans. Database Syst.*, vol. 35, no. 2, pp. 8:1–8:48, 2010. [Online]. Available: http://doi.acm.org/10.1145/1735886.1735887

[6] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. ACM, 2011, pp. 193–204.

[7] C. Clifton and T. Tassa, "On syntactic anonymity and differential privacy," in *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, April 2013, pp. 88–93.

[8] C. Liu, P. Mittal, and S. Chakraborty, "Dependence makes you vulnberable: Differential privacy under dependent tuples." in *NDSS*. The Internet Society, 2016. [Online]. Available: http://dblp.uni-trier.de/db/conf/ndss/ndss2016.html#LiuMC16

[9] R. Sarathy and K. Muralidhar, "Evaluating laplace noise addition to satisfy differential privacy for numeric data," *Trans. Data Privacy*, vol. 4, no. 1, pp. 1–17, 2011. [Online]. Available: http://www.tdp.cat/issues11/abs.a064a10.php

[10] J. Lee and C. Clifton, "How much is enough? choosing ϵ for differential privacy," in *Information Security, 14th International Conference, ISC 2011, Xi'an, China, October 26-29, 2011. Proceedings*, ser. Lecture Notes in Computer Science, X. Lai, J. Zhou, and H. Li, Eds., vol. 7001. Springer, 2011, pp. 325–340. [Online]. Available: https://doi.org/10.1007/978-3-642-24861-0_22

[11] E. Schlehan and R. Wenning, "Deliverable 1.6 - Legal requirements for a privacy enhancing Big Data, V2," SPECIAL, Tech. Rep., 2018, https://www.specialprivacy.eu/images/documents/SPECIAL_D26_M21_V10.pdf. Currently confidential.

[12] P. A. Bonatti, "Fast compliance checking in an OWL2 fragment," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, J. Lang, Ed. ijcai.org, 2018, pp. 1746–1752. [Online]. Available: https://doi.org/10.24963/ijcai.2018/241

[13] S. Pearson and M. C. Mont, "Sticky policies: An approach for managing privacy across multiple parties," *IEEE Computer*, vol. 44, no. 9, pp. 60–68, 2011. [Online]. Available: https://doi.org/10.1109/MC.2011.225

[14] W. Dullaert, U. Milosevic, J. Langens, A. S'Jongers, N. Szepes, V. Goossens, N. Rudavsky-Brody, W. Delabastita, S. Kirrane, and J. Fernandez, "Deliverable 3.4 - Transparency & Compliance Release," SPECIAL, Tech. Rep., 2018, https://www.specialprivacy.eu/images/documents/SPECIAL_D34_M25_V10.pdf.

[15] M. Alzantot, S. Chakraborty, and M. Srivastava, "Sensegen: A deep learning architecture for synthetic sensor data generation," in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 2017, pp. 188–193.

[16] M. Platzer, "Enabling privacy-preserving big data," Mostly.ai, Tech. Rep., 2018, https://mostly.ai/assets/Synthetic%20Data%20Engine%20-%20White%20Paper.pd.

[17] Information Commissioner's Office (ICO) UK, "Getting ready for the GDPR," 2017. [Online]. Available: https://ico.org.uk/for-organisations/resources-and-support/data-protection-self-assessment/getting-ready-for-the-gdpr/

[18] Microsoft Trust Center, "Detailed GDPR Assessment," 2017. [Online]. Available: http://aka.ms/gdprdetailedassessment

[19] Nymity, "GDPR Compliance Toolkit." [Online]. Available: https://www.nymity.com/gdpr-toolkit.aspx

[20] S. Agarwal, S. Steyskal, F. Antunovic, and S. Kirrane, "Legislative compliance assessment: Framework, model and gdpr instantiation," in *Annual Privacy Forum*. Springer, 2018, pp. 131–149.

[21] M. Robol, M. Salnitri, and P. Giorgini, "Toward gdpr-compliant socio-technical systems: modeling language and reasoning framework," in *IFIP Working Conference on The Practice of Enterprise Modeling*. Springer, 2017, pp. 236–250.

[22] M. Palmirani, M. Martoni, A. Rossi, C. Bartolini, and L. Robaldo, "Pronto: Privacy ontology for legal reasoning," in *International Conference on Electronic Government and the Information Systems Perspective*. Springer, 2018, pp. 139–152.

[23] M. Palmirani, M. MARTONI, A. ROSSI, C. BARTOLINI, and L. ROBALDO, "Legal ontology for modelling gdpr concepts and norms," in *Legal Knowledge and Information Systems: JURIX 2018: The Thirty-first Annual Conference*, vol. 313. IOS Press, 2018, p. 91.

[24] L. Elluri, A. Nagar, and K. P. Joshi, "An integrated knowledge graph to automate gdpr and pci dss compliance," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 1266–1271.

[25] L. Elluri and K. P. Joshi, "A knowledge representation of cloud data controls for eu gdpr compliance," in *2018 IEEE World Congress on Services (SERVICES)*. IEEE, 2018, pp. 45–46.

[26] H. J. Pandit, K. Fatema, D. O'Sullivan, and D. Lewis, "Gdprtext-gdpr as a linked data resource," in *European Semantic Web Conference*. Springer, 2018, pp. 481–495.

[27] P. De Hert, V. Papakonstantinou, G. Malgieri, L. Beslay, and I. Sanchez, "The right to data portability in the gdpr: Towards user-centric interoperability of digital services," *Computer Law & Security Review*, vol. 34, no. 2, pp. 193–203, 2018.

[28] T. Y. C. Woo and S. S. Lam, "Authorizations in distributed systems: A new approach," *Journal of Computer Security*, vol. 2, no. 2-3, pp. 107–136, 1993. [Online]. Available: https://doi.org/10.3233/JCS-1993-22-304

[29] S. Jajodia, P. Samarati, M. L. Sapino, and V. S. Subrahmanian, "Flexible support for multiple access control policies," *ACM Trans. Database Syst.*, vol. 26, no. 2, pp. 214–260, 2001. [Online]. Available: http://portal.acm.org/citation.cfm?id=383891.383894

[30] A. Uszok, J. M. Bradshaw, R. Jeffers, N. Suri, P. J. Hayes, M. R. Breedy, L. Bunch, M. Johnson, S. Kulkarni, and J. Lott, "KAoS policy and domain services: Towards a description-logic approach to policy representation, deconfliction, and enforcement." in *4th IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY)*. Lake Como, Italy: IEEE Computer Society, Jun. 2003, pp. 93–96.

[31] L. Kagal, T. W. Finin, and A. Joshi, "A policy language for a pervasive computing environment." in *4th IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY)*. Lake Como, Italy: IEEE Computer Society, Jun. 2003, pp. 63–.

[32] P. A. Bonatti, J. L. D. Coi, D. Olmedilla, and L. Sauro, "A rule-based trust negotiation system," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 11, pp. 1507–1520, 2010. [Online]. Available: https://doi.org/10.1109/TKDE.2010.83

[33] P. A. Bonatti, "Datalog for security, privacy and trust," in *Datalog Reloaded - First International Workshop, Datalog 2010, Oxford, UK, March 16-19, 2010. Revised Selected Papers*, ser. Lecture Notes in Computer Science, O. de Moor, G. Gottlob, T. Furche, and A. J. Sellers, Eds., vol. 6702. Springer, 2010, pp. 21–36. [Online]. Available: https://doi.org/10.1007/978-3-642-24206-9_2

[34] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, Eds., *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[35] B. Glimm, I. Horrocks, B. Motik, G. Stoilos, and Z. Wang, "Hermit: An OWL 2 reasoner," *J. Autom. Reasoning*, vol. 53, no. 3, pp. 245–269, 2014. [Online]. Available: https://doi.org/10.1007/s10817-014-9305-1

[36] Y. Kazakov, M. Krötzsch, and F. Simancik, "The incredible ELK - from polynomial procedures to efficient reasoning with EL ontologies," *J. Autom. Reasoning*, vol. 53, no. 1, pp. 1–61, 2014. [Online]. Available: https://doi.org/10.1007/s10817-013-9296-3

[37] A. Steigmiller, T. Liebig, and B. Glimm, "Konclude: System description," *J. Web Semant.*, vol. 27-28, pp. 78–85, 2014. [Online]. Available: https://doi.org/10.1016/j.websem.2014.06.003

[38] L. Engineering. (2015) Running kafka at scale. [Online]. Available: https://engineering.linkedin.com/kafka/running-kafka-scale

[39] N. Engineering. (2016) Kafka inside keynote pipeline. [Online]. Available: https://medium.com/netflix-techblog/kafka-inside-keystone-pipeline-dd5aeabaf6bb

[40] B. Svingen. (2017) Publishing with apache kafka at the new york times. [Online]. Available: https://www.confluent.io/blog/publishing-apache-kafka-new-york-times/

[41] S. Kirrane, U. Milošević, J. D. Fernández, A. Polleres, and J. Langens, "Deliverable 2.7 - Transparency Framework V2," SPECIAL, Tech. Rep., 2018, https://www.specialprivacy.eu/images/documents/SPECIAL_D27_M23_V10.pdf.

[42] A. I. Ali-Gombe, B. Saltaformaggio, D. Xu, G. G. Richard III *et al.*, "Toward a more dependable hybrid analysis of android malware using aspect-oriented programming," *computers & security*, vol. 73, pp. 235–248, 2018.

[43] J. Carmona, B. van Dongen, A. Solti, and M. Weidlich, *Conformance Checking: Relating Processes and Models*. Springer, 2018.

[44] P. W. McBurney and C. McMillan, "Automatic documentation generation via source code summarization of method context," in *Proceedings of the 22nd International Conference on Program Comprehension*. ACM, 2014, pp. 279–290.

[45] L. Moreno, J. Aponte, G. Sridhara, A. Marcus, L. Pollock, and K. Vijay-Shanker, "Automatic generation of natural language summaries for java classes," in *2013 21st International Conference on Program Comprehension (ICPC)*. IEEE, 2013, pp. 23–32.