SPECIAL

Scalable Policy-awarE Linked Data arChitecture for prIvacy, trAnsparency and compLiance

Deliverable D5.6

Report on application guidelines

Document version: V1.0

SPECIAL DELIVERABLE

Name, title and organisation of the scientific representative of the project's coordinator:

Ms Jessica Michel t: +33 4 92 38 50 89 f: +33 4 92 38 78 22 e: jessica.michel@ercim.eu

GEIE ERCIM, 2004, route des Lucioles, Sophia Antipolis, 06410 Biot, France

Project website address: http://www.specialprivacy.eu/

Project	
Grant Agreement number	731601
Project acronym:	SPECIAL
Project title:	Scalable Policy-awarE Linked Data arChitecture for prIvacy, trAnsparency and compLiance
Funding Scheme:	Research & Innovation Action (RIA)
Date of latest version of DoW against which the assessment will be made:	17/10/2016
Document	
Period covered:	M18-M36
Deliverable number:	D5.6
Deliverable title	Report on application guidelines
Contractual Date of Delivery:	31-12-2019
Actual Date of Delivery:	28-01-2020
Editor (s):	Sabrina Kirrane
Author (s):	Rigo Wenning
Reviewer (s):	Ben Whittam Smith, Martin Kurze
Participant(s):	WU, ERCIM, Refinitiv, TLABS
Work package no.:	5
Work package title:	Use case implementation and evaluation
Work package leader:	Refinitiv
Distribution:	PU
Version/Revision:	1.0
Draft/Final:	Final
Total number of pages (including cover):	50

Disclaimer

This document contains description of the SPECIAL project work and findings.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any responsibility for actions that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the SPECIAL consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 28 Member States of the Union. It is based on the European Communities and the Member States cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors (http://europa.eu/).

SPECIAL has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731601.

Contents

List of Figures

1 Introduction

When SPECIAL started, the academic and the industry partners where enthusiastic about the advances in data protection. The promise to get out of the consent swamp and leave the GDPR penalty threads behind was a great motivator.

SPECIAL had four use cases:

1. A general use case on location based services
2. A use case focusing on location based services and profiling
3. A use case focusing on corporate compliance and GDPR
4. A use case focusing on location statistics and insights on network quality

All the cases had their load of surprises and difficulties. This document tries to draw conclusions from the insights gained while overcoming the often surprising barriers of the specific use cases. Having three rather different use cases allowed SPECIAL to find the rather generic issues while implementing the SPECIAL system. The present guidelines address those generic issues. They give hints about obstacles and how to overcome them.

This document first explains the overall abstract concept behind SPECIAL. Because SPECIAL uses known technologies and extends them in specific ways, an implementer needs to understand the basic ideas behind it. This includes technical aspects, opportunistic benefiting from existing situations and hints on how to deal with certain legal requirements.

Because data protection and data processing potentially can cover every aspect of live, boundaries for the comprehensive approach have to be found. SPECIAL has chosen to use the approach via use cases. It has three industry partners and as many use cases. Two use cases concern location based services and one concerns a compliance mechanism in a highly regulated environment. SPECIAL also created a generic location based service to exemplify findings. This generic use case used the now common fitness bracelets that are monitoring your daily life.

Following the presentation of the general concept, this report takes a consulting approach by giving hints on where to start and how to prepare the grounds based on the use cases that were explored. In current systems, too many things remain unspoken or implicit. A hard part of the work to create data protection compliant interesting services is a full understanding of data life cycles, workflows, access control and especially about legacy applications, currently dealing with the data. Once the assessment finished, legal and technical considerations will help to use the SPECIAL system in order to achieve a feasible, compliant and meaningful solution for the initial goal. Those goals can be exemplified along the SPECIAL use cases. But the SPECIAL system and method is very generic and can also be made working for other use cases beyond the ones that were implemented as proof of concept.

SPECIAL focused on data protection use cases. This included the deep and insightful legal analysis of several use cases. Those allowed to use the system creatively to fulfill the requirements coming from the legal evaluation of the use cases. But SPECIAL also made significant contributions to resolve the consent paradox. Courts all over

Europe require more and more that users need to consent to data collection. There is an escalation going on. While courts want consent, the data collection industry reacted with pre-ticked boxes knowing that most of the users will not bother to interact with that configuration. In turn the legislator and the courts required more and more *informed* consent and invalidated the pre-ticked boxes. More and more information in privacy policies was required over time leading to long complex documents that nobody reads. SPECIAL developed a concept how to escape this escalation while keeping an eye on economic viability.

Finally, given the SPECIAL system, it becomes much easier to create data value chains and make them compliant in a complex legal and technical environment. While the concept of sticky policies is old, SPECIAL provided some new aspects to further advanced solutions.

Chapter 1

The SPECIAL concept

1 The basics

The basic idea behind SPECIAL is to further develop a concept that uses metadata and semantic information to help algorithms in a workflow to do the right thing, which is a very high level description of the SPECIAL system. The actual implementation of such a system faces enormous challenges. To overcome those challenges, a system needs to take into account a number of technical, but also of legal requirements and an arbitrary number of combinations thereof. It is reassuring that many opportunities to resolve challenges come from protocols and other environmental information already present at collection time. SPECIAL facilitates the intelligent re-use of that data using existing technologies.

A very good way to assess the concept is to think it as a data life cycle. The SPECIAL overall architecture depicted below also helps to understand why thinking in data streams and data lakes helps to better assess the challenges and opportunities.

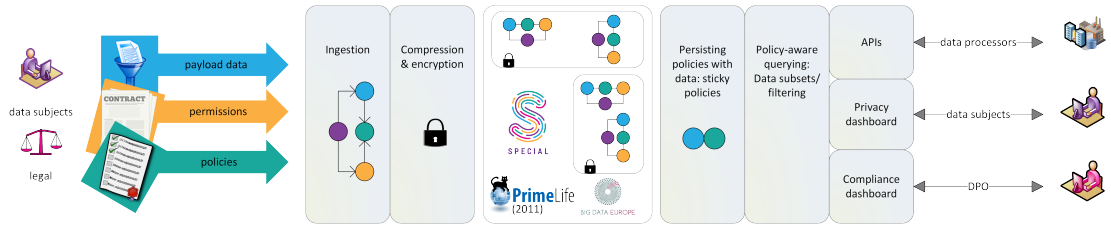

Figure 1.1: The SPECIAL Architecture

The diagram is constructed alongside a typical data life cycle. Data is created, collected or otherwise ingested into a system. Within the system, data from several sources are combined. In SPECIAL one decisive advantage is that the data ingestion includes potentially a broad range of metadata. Data and metadata are processed and evidence is created via digital ledger technologies and encryption. Data is then processed according to the rules expressed in the metadata using the SPECIAL policy language and other metadata vocabularies, e.g. provenance[9]. Finally, the result can be shared with the user in a dashboard. Or the metadata can be used to create policy aware data value chains via the use of the the sticky policy paradigm.

2 Using metadata

SPECIAL is a big data project. Implementing the SPECIAL paradigm means that one has to handle high volume data streams. Whether there is high velocity or variety will depend on the use case. The bigger volume of data originates from the fact that the environmental and policy data is stored as metadata linked to the instance data or to data categories.

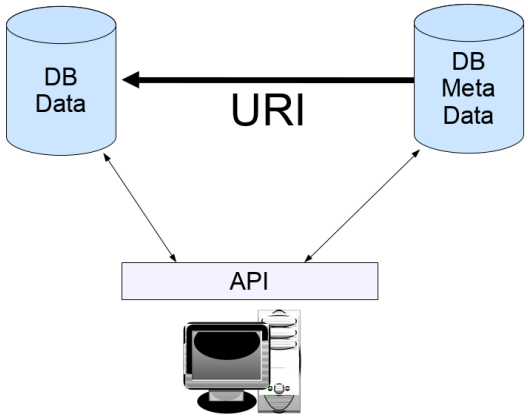

Figure 1.2: Data annotation with linked data

Having the system rely on annotations of data has multiple implications that will be detailed in later sections. First of all, such system can digest any kind of metadata. This is important for the ingestion of information. Data can now be augmented by policy rules, provenance information and data quality metadata.

From there, it becomes clear that not all information provided in the context of the data ingestion is already ready for use of the system. As SPECIAL requires data to be linked data in order for the annotation system to work, metadata and other policy data found at ingestion time has to be transformed to fit into the system. This is called semantific lifting. Semantic Lifting was developed in the framework of the Big Data Europe Project[1].

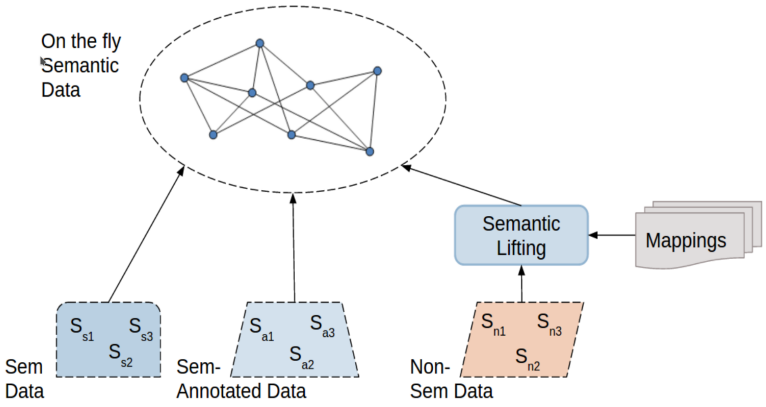

Figure 1.3: Semantic Lifting of Legacy Data

Some of the sources may already be Linked data. This is especially true if using the SPECIAL compliance mechanism and the attached event and object data formats as explained in Deliverable D2.7. But there are a lot of other sources, where data comes from line of business applications (LOB).

It is important to collect all data sources for a given or intended workflow. If the data is not yet in the Linked data format, it has to be transformed, which is called *semantic lifting*. Once this is done, it must be determined what the object of annotation will be. For SPECIAL as a privacy project, the object naturally was personal data. All other Linked data was then annotating the personal data it addressed, including policy information, access control, usage limitations and the like.

3 Creating the data lake

Once all the data is ingested, this creates a combination of data and metadata. As data and metadata is linked, a graph is created. This graph contains a wealth of relations, most of them automatically generated at ingestion and not all of them being necessarily useful or usable in the context of data protection.

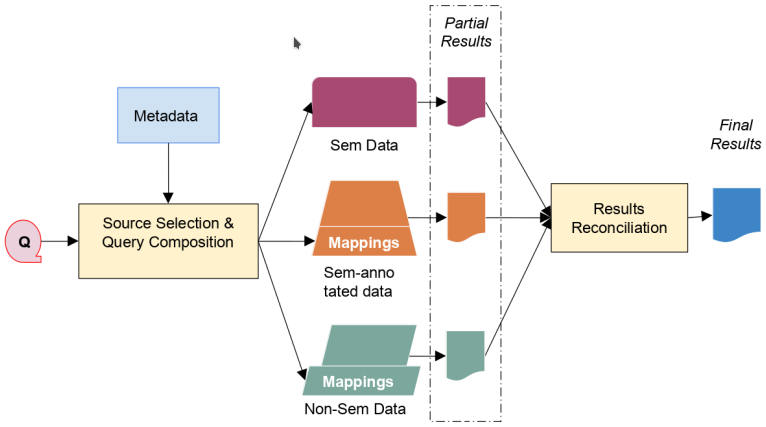

Figure 1.4: Semantic Data Lake

Further processing and algorithms can now take the wealth of relations into account and use them as conditions or further influence or change such processing.

Once in the data lake, the fact of processing certain data can be used to add further metadata by recording that the data was processed and how it was processed. Such data annotations may also be used to create a log of access control events, impose mechanisms for quality control. This way, a compliance checker may find out whether such access or processing was within the given policy boundaries.

Consequently, creating the data lake and the privacy data graph does not by itself make a system compliant. But especially for data protection rules, systems need to be context aware. Data may be usable for one purpose in one context and not be usable in another. The graph in the data lake can express that. This means the semantic data lake is a precondition for data protection aware processing, but it is not a sufficient condition to have compliance.

SPECIAL also makes compliant systems easy and agile. Of course any type of data handling can be hard-coded into the application programming of a given use case. But a slight change and the entire application needs to be re-programmed and re-deployed. For SPECIAL instead, it may be sufficient to change a line in an ontology or policy. Data and metadata are in the database or knowledge graph. From there, no need to have complex code to make compliant systems because it is only a matter of the right query that selects the usable data sets. In fact, as we want to process data, we normally know the legal constraints that come with it, either from data protection or from other legal constraints like licenses. If the metadata contains semantics about those constraints, the query will only select those data sets containing the right semantics or pointing to the right ones in a knowledge graph.

It is important to report the conclusion of the long discussions the project had about the use of linked data. Linked data is important as a mind model as it allows to link data and metadata and allows to make those relations explicit. But in the telecommunications use cases for example, the industry partner used a fine tuned specially dedicated database for real time processing. Of course nobody would ever replace such a system with a linked data engine with inferior performance. The conclusion was to use linked data as an important mind model to understand the relations between several sets of data and metadata. From there, creative data modeling was used to preserve the data model even within the constraints of the highly optimized system and be able to re-export to linked data, at least in theory. This way, data keeps being transportable, transformable and even marketable without constraining the existing line of business applications.

4 Interfacing intelligently

The SPECIAL system can now be used to create innovative interfaces. Those are not exclusively interfaces for data subjects, but may provide a new way to create policy aware data markets. This may range from processing for data analytics over data value chains to the creation of a user dashboard.

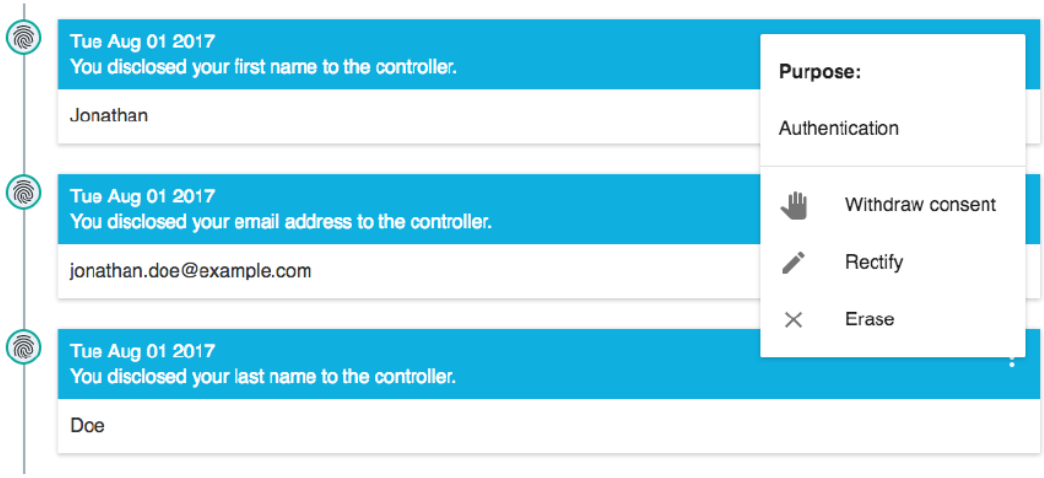

Figure 1.5: The SPECIAL Dashboard

The combination of legal and technical means to fulfill data protection constraints and make data self determination digestible again is taken to a new level. This can be best seen in chapter 2 where we take a new approach to consent interfaces that would not be possible without the SPECIAL system (section 7.2. It is worth noting that those interfaces can be multi-modal. In a smart city, if one is detected by a sensor on a streetlight, one has no way to interact directly with the streetlight. While being possible, it would be too complex or expensive to implement user controls into such a sensor. But it can be part of a SPECIAL system that allows an interface via web or some dedicated smart city application. Interfacing to the data subject can thus be done in many ways and not only at the data collection point.

The change with SPECIAL is that instead of profiling only the user, the SPECIAL system is adding policy data to data or to data categories. The benefit is a much better basis for intelligent new privacy enhancing technologies as the richer information allows for more sophisticated algorithms taking into account data protection rules. But such a system is also very beneficial to data controllers and the data industry at large. It allows in fact to mirror the legal bureaucracy into the system and create a full accountable system that can show its compliance upon pushing a button. All those are output mechanisms where parts of the data lake are selected and transformed to resolve some challenge.

The output to business partners will make contracts dealing with data much easier. Instead of a very complex document describing the system, the data and how they must be processed for each and every category, the system allows for a simple provision: The other part of the contract just has to promise to treat the data according to the metadata provided. A clause may determine the metadata the the other part must understand. But there also needs to be a default reaction promised in case the other part does not understand the metadata provided.

Finally, another output has to do with the investigative prerogatives of data protection authorities. They can investigate and visit a certain data controller. In this context they can ask for evidence that processing of personal data was done respecting the data subject's rights and having a legal ground for processing. So far, this is an evaluation of all the workflows and the code. Everything is looked at in general, which makes ad hoc exceptions difficult e.g. as they need some justification on paper etc. With SPECIAL, consent is recorded into the data lake. SPECIAL has also experimented with encryption and ledgers for better evidence. An evaluation by a data protection authority will become just a report of the compliance metadata stream where grounds of processing were stored back into the data lake at the time of processing.

5 Design Principles

As set out in Deliverable D1.8, SPECIAL has found and applied to following principles that govern the overall system design.

Principle 1) –Automated system rollout as much as possible. Using system deployment descriptions such as Terraform (system resources layer) and Docker Compose (services layer) the roll-out of an application becomes reproducible and reliable. Because the description is stored in a source control repository, changes over time and

variants can be maintained without the need of having them actively running. The consumption of system resources can then be dedicated to the active developments.

Principle 2) –Cloud enabled by design Our platform has to be hardware and Operating System neutral as much as possible. Using a service abstraction layer (i.e. Docker) addresses one part. Additionally the setup has to be decoupled from the local file system. Only then will the system be completely cloud enabled and runnable independently.

Principle 3) –Modular design, by preference following the micro-services pattern Micro-service design is the idea to create a system from the integration of a collection of services, each with a dedicated purpose. This approach enables scaling potential for the system: if one service is in high demand, adding new services of the same kind is a straightforward action. In addition, it allows to focus the development effort. The approach has proven results in the design of end-user facing software.

Principle 4) –Reuse best practice standards for well-known technical challenges As already mentioned in section 4, many privacy threats have industry supported mitigation strategies. Therefore, unless they are not sufficiently appropriate or adequate, it is our strategy to apply the best practices as much as possible.

Principle 5) –Payload data is preferable in the form of RDF, JSON-LD or JSON. Although the use-cases indicate that data from various sources with a multitude of formats are processed to create the desired value-adding services, it is our intent to keep the heterogeneity as much as possible under control by using preferable RDF, JSON-LD or JSON as payload data representation. Where-as RDF and JSON-LD are highly compatible with each other, JSON requires additional semantical lifting. This lifting can be defined by adding an LD context to the JSON payload. Thus, although not technically imposed that the data is exchangeable, these 3 data representation formats can form a uniform data landscape.

When a component does not comply with this preference, it may be required to create a dedicated payload translation layer for the component. To some extent, semantic lifting acts as such wrapping.

Principle 6) –The data-exchange channels are secure. The payload data has to be exchanged between the services. Most importantly is that the used data exchange channel is secured against penetration: HTTPS, secured database connectors (ODBC, JDBC), secure file access and a secure message bus (Kafka) are the preferred choices.

6 Conclusion

SPECIAL creates a system that ingests all kinds of information about personal data it knows about. It achieves this by semantifying all information to link it together in a semantic data lake that allows to extract graph structures that can be taken into account by data protection aware algorithms. Intelligent big data ingestion allows the

system to be policy aware and to react accordingly by making all processing provably compliant.

The SPECIAL system is also the basis for the creation of very innovative new interfaces that allow to preserve the paradigm of data self determination by showing the data subject the relevant information in context and thus allowing to make informed decisions about data collection, processing and sharing. As such, the SPECIAL system does not only further data protection and data self determination, it also allows to make better systems that create more trust, thus allowing people to share more data for the benefit of all. This in turn will further the development of businesses and further the Digital Single Market.

Chapter 2

Designing the Workflow

1 Design considerations

According to design principle No. 3, modular design should be applied, by preference following the micro-services pattern. Micro-service design is the idea to create a system from the integration of a collection of services, each with a dedicated purpose.

For the overall workflow design this means that SPECIAL is not a fixed platform with a mandatory number of elements being supposed to be present. To the contrary, SPECIAL was designed as a toolbox that allows to use certain tools for required data protection functionality. This means not all use cases have to implement the entire set of tools provided by SPECIAL. It is therefore important to match the functionality needed by a given use case to the set of tools available. It can be even important to further drill down and implement the tools adapted to the specific particular properties of a certain use case. This is due to the fact that SPECIAL created a very generic tool chain that can serve the an average eCommerce shop system as well as a very complex system like the verification of compliance policies within a system of financial operations.

As already laid out in chapter 1, SPECIAL is best understood if explained alongside the data life cycle. This includes data ingestion, data processing, data sharing and user interaction. One chapter of legal considerations is added as implementing a SPECIAL enabled system needs to think interdisciplinary from the technical to the legal to the legal tech side of a given use case.

In this report we exemplify the above with the use cases SPECIAL has worked on. Drawing parallels from the SPECIAL use cases is expected and encouraged. It explains how they did their raisin-picking to reach a higher state of compliance and data usage.

The location based services use cases were mainly interested in the recording of legal grounds for processing in the sense of the GDPR, including a new innovative way to obtain and manage consent. Here, SPECIAL plays the role of an enabler. It makes available the necessary systems, semantics and metadata to find or create legal grounds for new ways of processing personal data. It tries to overcome the opt-in vs opt-out dilemma that most businesses face.[25]. Detailed guidelines for the consent module within SPECIAL can be found in section 5 of D1.7, Policy, transparency and compliance guidelines V2[5].

For the *know your customer (KYC)* case, the challenge within the policy considerations was beyond this first attempt to enable data processing. The goal was to integrate

the SPECIAL components into the existing KYC system. Partner Refinitiv identified 19 permissions within the GDPR that allow for the transfer of personal data. Five of those permissions are based on adequacy decisions; six based on appropriate safeguards; and eight derogations. Using the formalism provided by the SPECIAL policy language[4], every permission could be expressed using the structure provided. It is thus a matter to first understand one's own workflow in order to identify the points where data protection and other policy metadata is needed to inform and semantify the workflow in order to check for compliance patterns. As already said above, the system integrated policy expressions from the KYC cases and the ones from GDPR requirements by translating between both worlds via classes, e.g. of permissions. SPECIAL allows also here the full power of integration of a Linked Data approach. Refinitiv was able to integrate the Open Digital Rights Language (ODRL)[10] and there, the Regulatory Compliance Profile into the system.

2 Data ingestion

There are two aspects to the SPECIAL system when it comes to data ingestion: Data reuse and new data acquisition of personal data. Both are closely related, especially in the use cases on location based services, as e.g. telecommunication companies have already location information that may be complemented by the GPS data from a handheld device. Data reuse on the service side has the advantage of limiting the battery drain on the device. To reuse, the application still has to communicate with the user to obtain consent for the use of location information. This has to be taken into account during workflow considerations.

2.1 Environmental information

The best way to start the evaluation of the ingestion or reuse of personal data is to start with the register of processing activities that have to be maintained following Art. 30 GDPR. This gives a good overview of the data available to a project. Those processes will also point to the environmental information available via protocols and surrounding policies. In fact, *environmental information* is used here to describe all information available to the system at data collection time. This can be data and metadata. It was quite a considerable amount of work to determine all the metadata required for the use cases. Which data points should be used, which environmental data should be ingested and what policy should be attached to it. The work on this list of metadata to consider was a decisive part. This was an iterative process, as later processing considerations made it important to go back to that step and reconsider the types, amount and instances of data and metadata to be ingested.

In the use cases around location based services, the challenge was to select those data points that were suitable for re-use and how to make them usable to the system. In this case, the normal data minimisation approach is recommended. An application, in order to function as planned, needs a certain number of data points. The data protection forces people to think about better data processing by forcing them to think about all data points and what their intended use and functionality in the system really is. If data is personal data or if the intention is to attach the data points to personal

data, GDPR applies. In this case, a legal ground for processing is needed. This legal ground can often be found in the relation between data controller and data subject. Throughout the project many reminders were needed that necessary data points do not need consent. Collection and processing of other data needed for some non-necessary functionality or data that is crucial for the intended business model will need the consent of the data subject. This is mainly the case for the creation of user profiles helping to filter the wealth of information available in a given application or on the internet. The distinction is crucial because the legal ground of processing will also determine the scope that is given.

The SPECIAL system forces people to already think about those workflows while being in the planning phase. Although the system is rather agile and flexible, the relevant metadata must be present in a digestible format. Depending on the system that should be SPECIAL enabled, there is a number of Linked data vocabularies that can be used. Those can be found at the W3C Technical Reports[1] section, at the Schema.org[2] repository or e.g. the European Commission repository for Core vocabularies. A similar approach could leverage the addition of provenance information into the data lake to derive data quality markers from it. But in order to do that, the provenance information has to be discovered and semantified, e.g. using the W3C Provenance Vocabulary[9]. In the case of the *know your customer (KYC)* case, the entire wealth of metadata and contextual information in the financial sector need to be taken into account before it can be augmented by SPECIAL components. As this makes big data even bigger, the limits are mainly to distinguish what environmental data needs to be preserved for later compliance checking.

For the use cases on location based services the difficulty was less the metadata, but how to deal with location information. What granularity is needed and how to cluster location information into the profile for the recommender engine and how to attach policy information to it. As the recommender engine itself was out of scope for the project, this wasn't resolved by the project itself. SPECIAL only provided the content of the policy and the mechanism to attach this policy to the relevant range of location points. But this had also repercussions on the UX challenges as it was not clear what object a given consent was reading on. It is understood that this couldn't be a single point on the map.

For the telecommunication service providers, most metadata was already present in their systems in various formats. This wasn't used directly, but similar data was generated for testing. As the industry partners all use bespoke systems, it was not an option to impose a Linked-data-only system. But in order to annotate data, some data had to be semantified and the system integrated those according to the finding in the deliverable D2.7[11] that has a detailed description on how to integrate bespoke or legacy systems into the SPECIAL processing. This does not only include the possibility to integrate data from line of business applications into the SPECIAL system. It works both ways as insights gained within the SPECIAL system can be written back into the line of business application. For simple systems with simple policies, it is even possible to only use SPECIAL as a mind model and do all the implementation within the framework of the legacy system or the line of business application.

[1] `https://www.w3.org/TR/`
[2] `https://www.schema.org/`

2.2 Policy information

The payload data, the data found in the environment and turned into metadata, or data having been categorised, can be augmented by policy expressions. Consent requests and sticky policies involve *data usage* policies, that are dealt with *usage policies (see section 6 of D1.7.)*. Compliance with such policies is meant to be checked automatically, exploiting the knowledge encoded in the transparency infrastructure. The formalisation of the GDPR has different requirements, since the constraints imposed by the data protection regulation are more difficult to assess automatically.

Initial guidelines to the formalisation of the GDPR were made. The semantics can be found in D2.6 that contains a formal representation for GDPR[3].It became clear that it was too complex for SPECIAL alone. SPECIAL was then instrumental in organising a Workshop on Data Privacy Controls and Vocabularies[7]. The community that came together in the Workshop defined the lack of privacy semantics as a prime obstacle to better policy expressions. SPECIAL helped to create the Data Privacy Vocabularies and Controls Community Group (DPVCG)[3]. The DPVCG created the Data Privacy Vocabulary[16]. This vocabulary is used in conjunction with the SPECIAL Policy Language[4] in order to provide the needed semantics for algorithms and compliance testing.

In order to build the policy for a use case, SPECIAL introduced an abstract core policy model (The Minimal Core Model or *MCM*), and derived a possible encoding with semantic web languages and preexisting policy languages. This helps to illustrate how policies are meant to be applied and queried in a SPECIAL enabled system. All aspects are clearly interrelated and place constraints on each other, and should be taken into account in the planning and policy discussion phase of a SPECIAL enabled system.

2.2.1 Usage policy model

From the above, as a first step, it is easy to derive simple data usage policies, whose main elements are summarised in Figure 2.1.

- "Data" describes the personal data collected from the data subject.
- "Processing" describes the operations that are performed on the personal data.
- "Purpose" specifies the objective of such processing.
- "Storage" specifies where data are stored and for how long.
- "Recipients" specifies who is going to receive the results of data processing and, as a special case, whom data are shared with.

The complexity of the usage policy model resides in the description of MCM's elements. The project to be implemented and its data life cycle needs scrutiny according to the descriptions given in more detail below.

[3] `https://www.w3.org/community/dpvcg/`

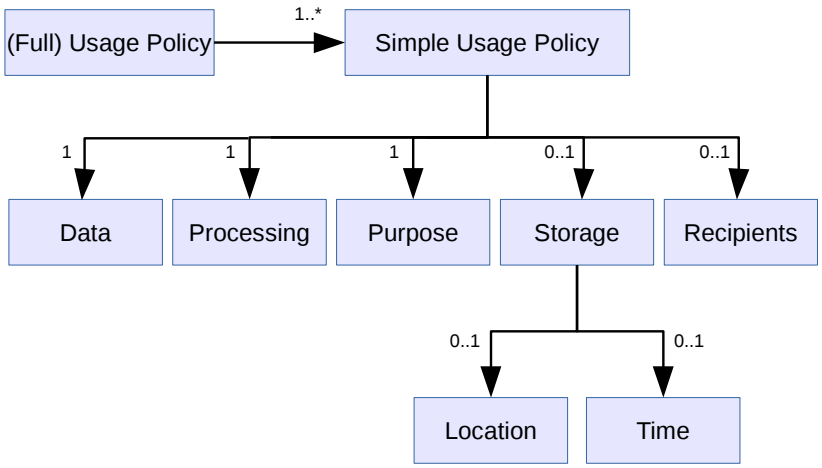

Figure 2.1: The minimum, core usage policy model (MCM)

The Data element: In order to describe which categories of data are collected, an ontology of personal data is needed. Most projects wanting to implement the SPECIAL system will need to express personal data. Developing such an ontology is an extremely difficult task, but luckily, DPVCG has made this a community effort and created the Data Privacy Vocabulary (*DPV*[16] that can always be taken as a starting point. Note that every possible piece of information that can be attributed to a specific individual is personal information, and as such falls under the scope of the GDPR (and possibly a future ePrivacy regulation).

In case Data Privacy Vocabulary (DPV) is not sufficient, the best approach is to leverage the extensibility and interoperability of semantic metadata, and to develop a complementary ontology of personal data covering the specific categories of personal data used in the project. This is the approach in the KYC use case, as DPV will not cover all the data categories and semantics needed. With semantic interoperability and the Linked data toolchain, the core can be extended with suitable profiles and/or integrated with further ontologies specialised for particular cases as needed.

From the SPECIAL use cases, only the one from the financial industry was complex and beyond DPV boundaries. But as Refinitiv has huge domain expertise in the area and already a Linked data toolchain, the integration was easily possible as described in D5.5[26]. The location based services reference scenarios use (a subset of) the information that can be found in IDs like passports, plus telephone numbers, physical and email addresses and thus integrated into the DPV. Moreover, the use cases of Proximus and Deutsche Telekom share common personally identifiable information (PII) that adheres to a small set of telecommunication standards. This facilitated the formalisation (semantification) of the data categories collected and processed in a wide range of applications

operating on telephone data. Similarly, it seems feasible to formalise the categories of data more frequently collected by social networks, that provide another wide range of applications with similar data usage modalities.

The Processing element: Data processing can be described with at least two approaches: (i) *algorithm oriented*, and (ii) *output oriented*. The former is particularly difficult and ineffective for several reasons. In many cases the same computational task can be carried out with several alternative algorithms, possibly quite different from each other, with complementary properties (e.g. time or memory consumption), but producing exactly the same output. Arguably, such differences are irrelevant in the policy context, since all alternative algorithms produce the same information and distribute it in the same way. Algorithmic descriptions are also intrinsically difficult to process: virtually all interesting properties of such descriptions are undecidable (e.g. algorithm equivalence and output properties). Furthermore, an algorithmic description of data processing is of little meaning to most data subjects; this makes algorithm-oriented descriptions unsuitable to the formulation of the usage policies enclosed in informed consent requests.

Due to the unnecessary complications introduced by algorithmic descriptions, we recommend an output-oriented approach to the description of data processing, that is, a categorisation of the data produced by data processing in terms of the information it conveys. For instance, data subjects are interested in knowing which information about themselves can still be found after data have been aggregated or analysed, and possibly the degree and kind of anonymisation of such information.

Therefore, data processing should be described through a suitable ontology. It may just describe the algorithm used. In Proximus' use case, for example, the result of data processing is an interest profile formulated in terms of a vocabulary of keywords extracted from well-defined sources (cf.D1.5[6]); in this case, the description of data processing could be simply collapsed to "*an algorithm that produces that type of profile*".

In the market, data taxonomies do not contain sufficient detail to model such aspects. Because the existence of common needs in important categories of applications encourages the structuring of the data ontology into a core taxonomy plus a set of profiles specific to application categories, further work is needed once there are sufficient corpora of such algorithms to start. The Data Protection Vocabulary Community Group is a place where such effort can be coordinated and where additional insight can be found.

The level of granularity of the output-oriented approach will be easier to handle for data controllers. Implementations can operate at the service level of the business logic, by introducing an abstract description of the effect of each relevant service, while services need not to be internally analysed to find the relevant semantics.

This may include semantics on the type and degree of anonymity of data-processing outputs. Concretely, by "type and degree of anonymity" notions such as *k-anonymity, l-diversity, ϵ-differential privacy* and the like, for specific parameters k, l, ϵ can be given as variables to the system, once they are part of a taxonomy. This may be the basis for negotiations with the data protection authority over sufficient safeguards for further processing outside the GDPR.

The Purpose element: Purpose is one of the most difficult aspects of data protection, especially for big data projects. Art. 5 (1) b GDPR states that data shall be: *collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes.*

One of the major difficulties of current systems is, that they do not record the purpose of processing into the data warehouse. Meanwhile, a SPECIAL enabled data lake can record the purpose. Later, an ontology can express what Art. 5 GDPR calls *compatible purposes*, including archiving purposes in the public interest, scientific or historical research purposes or statistical purposes.

To do so, the purpose element needs to describe formally why data are collected and/or processed. Not surprisingly, purpose descriptions are to be expressed through a corresponding ontology. Purpose descriptions are part of all of the usage policy languages developed so far, including P3P[24], ODRL[10] and especially the new Data Privacy Vocabulary[16]. None of those are exclusive. An implementation requires a raisin picking exercise over all those vocabularies, taxonomies and ontologies.

While personal data and data processing may vary widely across different domains, applications show much less variety in purposes. Objectives such as marketing, service optimisation and personalisation, scientific research, are pervasive across a variety of contexts. Accordingly, we expect the development of an ontology of purposes to be way less problematic than the ontology of data categories.

The Storage element: Cloud computing is a big part of the Digital Agenda. The storage element describes where data is stored and how long for. Accordingly, the Minimal Core Model attaches two corresponding subelements to the storage element: Location and Time. The level of granularity of both subelements needs not necessarily be fine-grained.

The GDPR is mainly concerned with two aspects related to storage location, namely: (i) whether data remains within the company boundaries or is distributed across different organisations (even if they are simply classified as “data processors”, or limit their activity to providing the storage service); (ii) whether data crosses national boundaries, since this may affect the applicable data protection regulations. Thus broad location classes, such as “within/without our company”, “our partner’s servers” may suffice for point (i), and again the vocabulary adopted by P3P may constitute a useful starting point for developing a location ontology. Nonetheless, in order to monitor and audit company processes, it may be helpful to refine such descriptions by keeping track of the hosts and files where data are stored (e.g. in the form of URIs). This information can be easily refined by adding the nation in which hosts reside, in order to address point (ii).

Guidelines for the duration of data retention can e.g. found in the DIN 66398 standard that creates a full data life cycle and erasure for commercial data, taking into account legal preservation obligations and laws about archiving. This means that the legal assessment of the duration of storage is complex while the technical implementation is less so. This is why SPECIAL did not foresee the need for complex time constraints. The GDPR, on the contrary, requires that storage is strictly bound to the service needs. This implies storage minimisation, hence the need to express *upper bounds* to storage duration, that may be expressed either in terms of the duration of the service that the

data have been collected for, or in absolute terms.

Guidelines for the duration of data retention can e.g. found in the DIN 66398 standard that creates a full data life cycle and erasure for commercial data, taking into account legal preservation obligations and laws about archiving.

Summarising, the storage time element should be able to express a single, possibly open interval, and temporal reasoning collapses to trivial interval membership and interval emptiness checks (for verifying, respectively, that a time point fits within the allowed storage period, and that the allowed storage interval has been correctly specified).

The Recipients element: In most cases where data is just processed internally, the recipient element will mostly serve to reassure the data subject and to help partition the data from shareable data. But in more sophisticated data value chains, the recipient considerations can get arbitrarily complex. One of the most complex issues here being current systems of real time bidding platforms for advertisement, where transitional permissions could be explored in a more scientifically oriented way forward.

The concept of data self determination, the notification duties in Art. 13 GDPR, but also the requirements for informed consent dictate the presence of this element. The GDPR does not clearly state to which level of detail this information has to be specified. Art. 13 (1) b) talks about *the recipients or categories of recipients*. Of course, categories can be created in a taxonomy or an ontology that may be used by the system. But there are diverging needs, such as the companies' desire to keep some of their business relations confidential, and the data subjects' right to trace the flow of their personal information. Depending on the use case, it is conceivable to adopt a coarse-grained categorisation such as *partners to which services are outsourced*, *business partners*, *unrelated third parties*, possibly *applying a same usage policy*. The Data Privacy Vocabulary has recipient elements, but those are not very detailed. But the P3P 1.1 Vocabulary[24] provides a core vocabulary at this level of detail. Should it be necessary to identify data recipients precisely, the ontology may be modelled around the existing standards that describe organizations and possibly their contact persons (e.g. X.509).

2.3 Line Of Business (LOB) Applications

The integration of legacy systems and LOB is a very important aspect of SPECIAL. It allows SPECIAL to integrate into *current* applications and workflows without requiring the replacement or putting risk on mission critical systems. There is a way to produce a tight coupling between SPECIAL and existing Line of Business applications in terms of both policy specification and enforcement. Firstly the data that will form part of e.g. a consent request and subsequently the usage policy needs to be based on the the type of personal data required by the company in terms of product or service provision, and contextual information relating to the purpose, processing and sharing. Secondly, companies need to ensure that personal data processing and sharing within the organisation and by its Information Technology (IT) systems complies with relevant usage policies.

2.3.1 Associating Policies with Data

In SPECIAL each consent policy will be given a unique URI[4]. This URI should be used to associate the policy with: the consent obtained from the data subject; the usage policy; relevant data processing and sharing events performed by the company; and possibly even related policies (e.g. if there is a need to maintain a history of policy updates).

Key considerations include how can we associate a URI with personal data stored in existing company systems, according to a variety of data models, possibly at different levels of granularity? How do we ensure that mappings between personal data items and policies are kept up to date? Additionally there is a need for flexibility in terms of policy retrieval. From a navigation perspective it should be possible to navigate from a policy to the data that it governs and also from the data to one or more policies that govern it. Other requirements include the ability to retrieve all policies based on contextual information, such as purpose, type of processing, data subject, to name but a few. Also, where more than one policy governs the data it is necessary to understand the interplay between such policies.

The most common way of integrating company systems is to implement the so called *semantic lifting* (See Section 2.4.1 of D3.1[23]). For use of the legacy data with SPECIAL, internal relative identifiers are translated into

2.4 Consent recording

SPECIAL has developed an innovative interface for the data subject called the *privacy dashboard*. The work on the interface was not only directed towards transparency of data processing, but it also has an interactive component to record consent. In the location based services use cases this was further refined to create an interface what SPECIAL named *dynamic consent*. The dynamic consent interface gives contextual information during the use of an application and asks the data subject for agreement or consent. This is the recorded into the SPECIAL data lake.

In SPECIAL, the data subjects themselves are a very important source of information. An implementation should plan for that by providing the necessary metadata to record the input from data subjects.

3 Securing the processing and creating evidence

SPECIAL is operating in a context called *legal tech* nowadays. The fact to use metadata to record permissions, obligations and rights in the context of data protection creates a system that also tries to generate legal consequences in the relation between a data controller and a data subject. And if such a system tries to create and manage those relations, the legal challenge is also a challenge about evidence.

SPECIAL provides two contributions here: First, SPECIAL allows to use Distributed Ledger Technology (DLT) like Hyperledger to secure the evidence over processing steps. It allows to record certain states of the system securely. Second, the metadata secured via DLT is then used in the compliance mechanism to provide a fully auditable and

[4]see Chapter 1 Section 2

provable check at a certain state, e.g. if a data protection authority audits the data controller. Such a system allows to provide the proof of consent for a certain type of operations at any moment in a machine readable form.

In the context of the SPECIAL project, information relating to data processing and sharing events could be stored in one or more distributed chains that are accessible via Application Programming Interfaces (APIs). These chains may include a hash of the data and a pointer to the actual data, which will be stored off chain in an encrypted format. Also, existing blockchain platforms and frameworks could potentially be used to perform compliance checking in a transparent manner. Alternatively, decentralised platforms such as Solid[5], which relies heavily on World Wide Web Consortium (W3C) standards, could be used to provide individuals with more control over their personal data and how it is used.

The goal is to combine transparency with security. The mechanisms are used to provide the desiderata for the transparency ledger (cf. D1.3 [18]). The principles are reminded here :

Completeness: All data processing and sharing events should be recorded in the ledger.

Confidentiality: Both data subjects and companies should only be able to see the transactions that involve their own data.

Correctness: The records stored in the ledger should accurately reflect the processing event.

Immutability: The log should be immutable such that it is not possible to go back and reinvent history.

Integrity: The log should be protected from accidental and/or malicious modification.

Interoperability: The infrastructure should be able to transcend company boundaries, in the sense that the data subject should be able to easily combine logs that they get from multiple companies.

Non-repudiation: When it comes to both data processing and sharing events it should not be possible to later deny that the event took place.

Rectification & Erasure: It should be possible to rectify errors in the stored personal data and/or delete data at the request of the data subject.

Traceability: In the case of processing it should be possible to know about any previous processing of the data. As such it should be possible to link events in a manner that supports traceability of processing.

The blockchain mechanism to secure the processing was implemented and tested in the framework of the financial use case. While it did work, performance needs to be improved. For a full overview of SPECIAL's research on Blockchain and distributed systems, see D2.8[2].

[5] `https://solid.mit.edu/`

4 Processing & Querying Data

SPECIAL concentrated on the value of metadata in order to meet the challenges that were subject to the research undertaken. This obfuscates lightly the goal of a SPECIAL system, namely to make data more exploitable while preserving humans against collateral disadvantages of such data exploitation. Meanwhile, such system can be used to not only do compliance checking, but also to create a more intelligent data lake.

On a very abstract level, the fact that data and metadata reside in a system that can be queried, that permission checking is a matter of checking metadata before checking data. This can be implemented in many ways, including a SPARQL endpoint. SPECIAL had to deal with a consent management system for the location based services and with a compliance checker for the financial service use case.

4.1 Location Based Services

The location based services investigated by SPECIAL combine a permission management with an application. This allows to create user profiles and to later use those profiles to filter information according to the preferences stored in the system. The system thus contains data and metadata. The permission events are stored in a log.

The purpose of this log is to provide data subjects and data controllers a way to manage their policies containing the consent values. These type of services are commonly referred to as CRUD services: they need to (C)reate, (R)ead, (U)pdated and (D)elete data entities. An architecture that is commonly used to implement CRUD services, augmented with an audit log, has been chosen:

- An API Layer which allows frontends and other clients to call its services. Data validation and authorization checks happen here as well (see Chapter 3 Section 3.1 of D3.4 [8]).
- A database layer which persists the data in a format which is optimized for use by the API Layer (Chapter 3 Section 3.2 of D3.4).
- Audit logs which record all transactions (Chapter 3 Section 3.3 of D3.4).

5 API Design

The consent management API allows the manipulation of 3 different entities:

1. Applications
2. Users
3. Policies

Each of these entities is manipulated in a similar way. D3.4 explains in detail how various endpoints are defined and gives examples by showing payloads. Depending on the use case, the needs of the API may vary. SPECIAL has defined those mainly for the location based services. Consequently, the API calls which allow for the retrieval of policy data are intended for use by UI clients which wish to render an individual users policies.

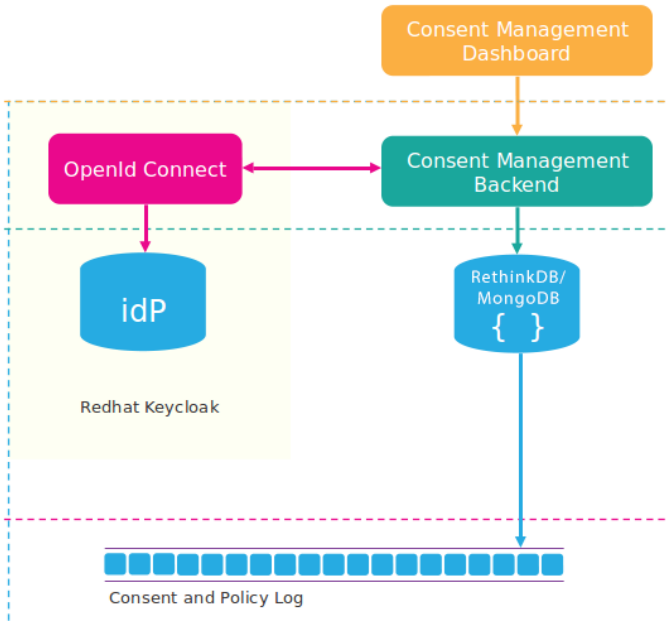

Figure 2.2: Consent Management in location based services

5.1 Compliance checker

The Know Your Customer (KYC) use case is a Service that requires much more intensive use of policy data. It is a compliance checker. Similar intensive use of policy data is expected e.g from an authorization server. Those services should preferably consume the policies from the full policy Kafka stream (see Section 3.3 of D3.4). This provides better decoupling, relaxes performance requirements on this service and provides consuming services with the option of reshaping the policy data to better fit their needs.

The purpose of the compliance checker is to validate that application logs are compliant with a users policy. These application logs are delivered in the format described in deliverable D2.3[12] using the policy language described in Deliverable D2.5[4].

Figure 2.3 shows an overview of the components that will be discussed in this section.

6 Data Flow

The compliance checker can be seen as a stream processor which takes in a stream of application logs and emits an augmented stream of logs. The system has the following data inputs:

- **Application Log Topic**: This is a normal Kafka topic that contains all application (processing event/request) logs which need to be checked for compliance.
- **Consent and Policy Topic**: This is a compacted Kafka topic which holds the complete policies for all data subjects.

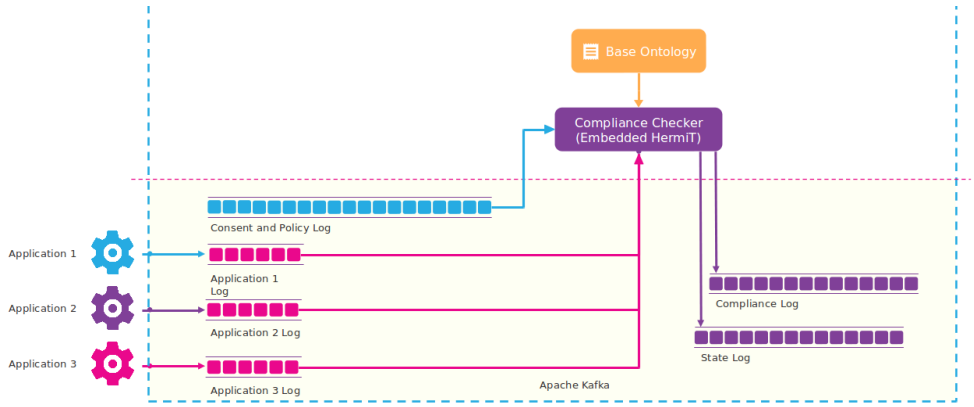

Figure 2.3: Compliance Checker

- **Base Ontology**: The base ontology are the vocabularies and class relationships which define the policy language as described in deliverable D2.5.

The system has the following outputs:

- **Compliance Topic**: This is a normal Kafka topic which contains the augmented application logs.
- **State Topic**: This is a compacted Kafka topic where the compliance checker can checkpoint the latest offset it has processed. This allows it to easily restore its state in case it needs to restart.

6.1 Policy Enforcement

In SPECIAL the sticky policy concept is used to tightly couple data and usage policies. When it comes to the state of the art, sticky policies are usually implemented by using cryptographic means to strongly associate policies with data. However, it is important to highlight that from a practical perspective it is not possible for said policies to be enforced automatically (i.e. it is an honors system whereby data controllers and processors can choose to either obey the policy or not).

Other open questions relate to using technical means to prove that usage policies are being adhered to. For example, if data subjects request that their data is deleted, how do we ensure that this data is in fact deleted and not simply made inactive. Another open research question relates to the inheritance of policies by derivative data. Considering the tight coupling between data and policies, data derivatives (e.g. in the form of aggregated and/or anonymised data) can not be covered by the same sticky policy.

6.2 Line of Business Applications

Irrespective of where the log resides, how much information goes into the log is dependent on what information is needed in order to automatically check compliance with both usage policies and relevant regulations. Given that event logging is a key component of many Line of Business systems, one option would be to re-purpose existing logs so

that they can be used to automatically verify compliance of existing systems. However, the suitability of existing logging mechanisms for this purpose requires further analysis. Alternatively it would be possible to have a dedicated log, however the attributes to be recorded would need to align with existing business processes. Ether way, the level of detail required to verify the compliance of existing business processes (that involve personal data) with respect to privacy preferences and legal obligations remains an open question. Likewise further analysis is required in order to determine where is the best place to hook into the existing company systems. Potential considerations here include the data tier, the service tier, or the business tier. Also, considering the complexity of existing business processes that are often only partially automated, a more in depth analysis is required in order to determine how much of the compliance checking can be automated. Key aspects here include the alignment of processing and sharing performed by the systems and what is specified in the consent.

7 Exposing data

The use of the word *expose* may be unusual. The normal business world talks about data sharing. But in a system-centric view, the data in the lake is exposed to either the data subject via the Privacy Dashboard or to other business partners. The latter cancan either query the data lake via some API or receive a data steram or a data package.

The Privacy Dashboard was developed within WP 4 of SPECIAL. The sharing of data beyond the data controller can be split into two sub-topics, namely the sharing of packaged data having sticky policy data attached to it or sharing via a sharing API.

7.1 The Privacy Dashboard

The Privacy Dashboard is a cornerstone in the SPECIAL system as it is an enabler for data self determination. By allowing not only to see what the SPECIAL system has about the data subject, but also by being interactive, it allows to make bargains to the *informedness* of a consent collecting mechanism. This directly affects the legal situation by giving a much bigger room for innovation, as lowering the consent barrier can be compensated by an ex-post control via the Privacy Dashboard. This was absolutely crucial for the location based services. It is less important for systems that have their center of interest in in-house compliance of data streams.

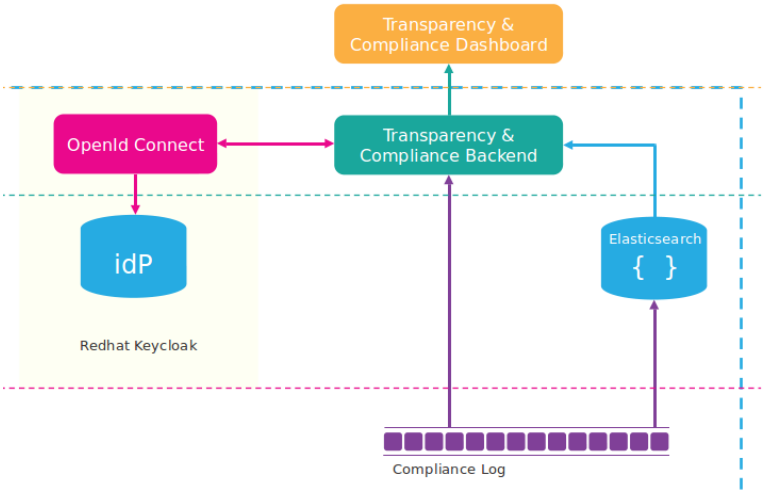

Figure 2.4: Transparency Dashboard

The Privacy Dashboard can only function well if it combines the richness of metadata in the system with an innovative frontend that does not overwhelm the data subject completely. This was a major challenge.

As can be seen in Figure 2.4, the proposal for the transparency backend as implemented within SPECIAL consists of the following components

- **Compliance Log**: This is the output of the compliance checker and serves as the reference for any visualisations
- **Elasticsearch**: Elasticsearch contains an indexed version of the compliance log and will provide faceted browsing, easy lookups and aggregations.
- **Transparency Backend**: The transparency backend is the sole entrypoint for any UI. The UI has to make innovative use of the semantics and data available. The backend will provide access control and enforce authorized access to the data in Elasticsearch or the compliance log.

Taking into account the capabilities of the backend, Work package 4 developed an innovative UI for the data subjects to get a glance at their state of data use, enabling self determination via the interactive parts. The initial structure of this UI is given in the mindmap shown in 2.5.

Additional challenges did appear within the use cases as the researchers had to either respect or to avoid the corporate design guidelines given by the use cases. SPECIAL created a neutral 4th use case in order to avoid further difficulties. This allowed researchers to experiment with the neutral use case and to transpose the solutions into the corporate constraints once there was agreement to go forward with a given solution. Given the regulatory context of SPECIAL and rather high anxiety over data protection, this was more work, but created a pragmatic solution.

As expected, people were trained from the currently used frontends in mostly US driven UIs to look for the less interruption possible. UI was seen as being too complex,

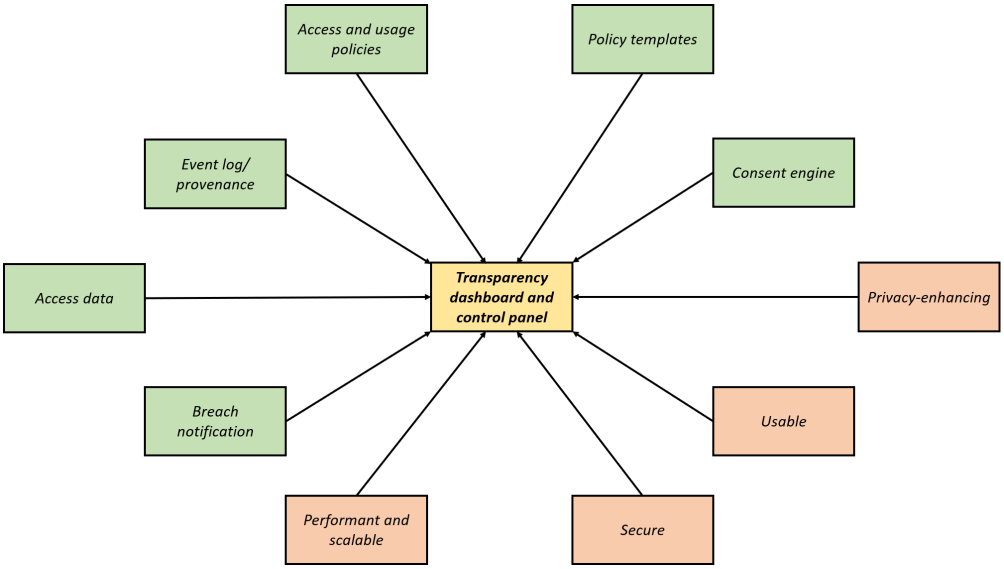

Figure 2.5: Dashboard - Mindmap

too burdensome, too rich. People did not know what to do with it. Only after they needed to explore what the system knows about them, they started to appreciate the capabilities given by the Dashboard. The more people had technical affinities, the more they liked to be in control[19].

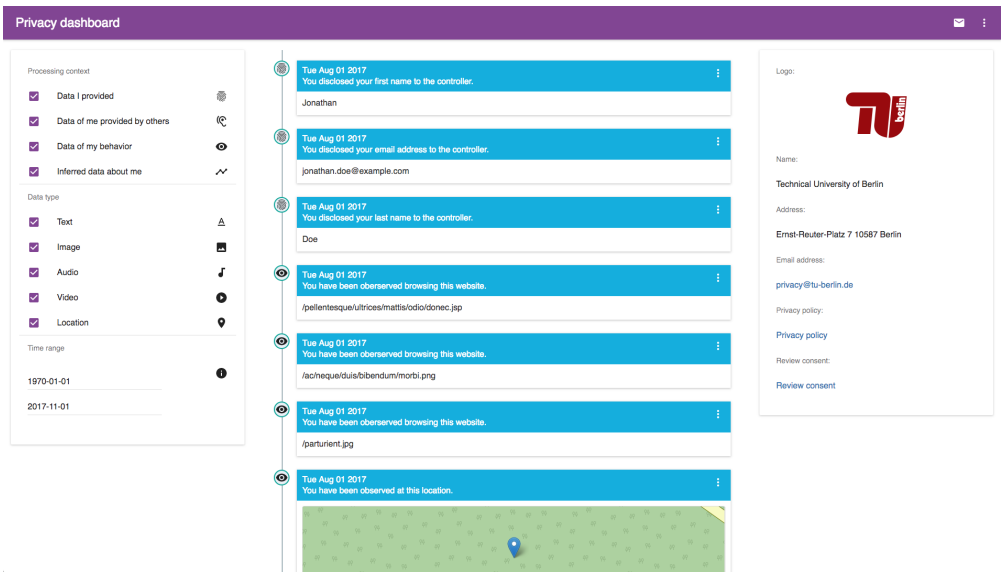

Figure 2.6: Dashboard - Interface

It was important to use metadata and a user centric approach to reduce the amount of data shown and expose the data subject to the most relevant data. One way of doing this was found to be an event driven log. But event driven logs are not the only way

of implementing the dashboard. A graph based approach that plotted the data subject concerning personal data and implications into a knowledge graph with nodes and edges remained misunderstood by most users and yielded very positive reactions from data protection experts.

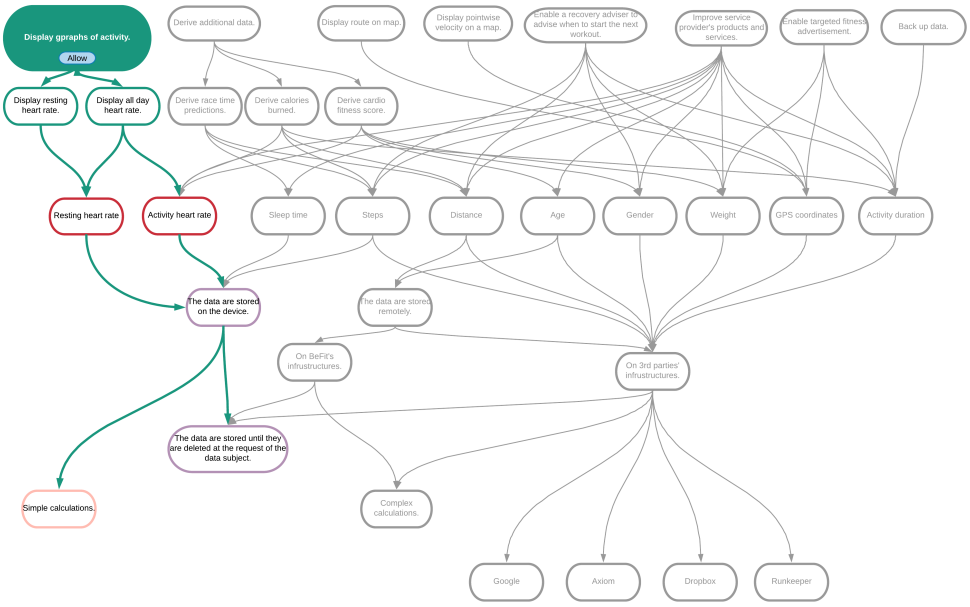

Figure 2.7: Dashboard - Graph

The graph as shown in 2.7 allows for a visual perception of very complex implications that privacy considerations can have. It visualizes not only the data, but also the relation between data points and between the data points and the data subject. By doing so, the dashboard allows a comprehension of data about the data subject that is unprecedented and will serve to re-establish the ability for self determination.

SPECIAL explored also ways to use this graph interface in order to provide a more complex consent interface for the data subject. The idea was that a data subject, by selecting a data item of interest, would see all the implications that would come when sharing that data item or category. This is a very very powerful tool, but it is avantime. People are used to the simple interfaces of the surveillance economy that are designed to avoid such deep understanding of the implications. The user testing had the corresponding results. It will be interesting to test this interface with users having a complex task, e.g. to avoid a certain processing while still agreeing to some other processing. In this case, the complex interface will provide the right level of information at the right time.

The dashboard mostly inspired the industry partners involved in the location based services. For Proximus there are even considerations to use the findings from SPECIAL to help the company wide preferences application to help customers understand their options.

7.2 Dynamic consent

As set out by the industry partners, the *dynamic consent* as a component was not yet ready for industry production. It was seen as too much cutting edge research, from a UX perspective, but also from a legal perspective. The concept was not well understood and it took a long time to overcome established reflexes about writing policies in an all encompassing document and try to get acceptance for that document. The early versions of consent still had a click-through policy that would have made later dynamic consent futile, legally.

It works so well because from a business management perspective, an all encompassing privacy policy is just yet another policy or compliance document. One document contains all provisions. This one document can be assessed by the normal company hierarchy. Changes can be controlled and are immediately visible to the managers. It just fits the management workflow. They have one package. Now the lawyers look over it and give their agreement. There is **one** decision over the policy. Now, they only need the data subject's agreement to their 22 page document. But in an opt-in scenario, user agreement does not happen. Because a normal contemporary privacy policy will have to describe the entire application in all its data collection and data use details before policy assertions can be made. This means a classic privacy policy will *always* be a huge document that nobody reads[13].

Dynamic consent is very hard for management. In fact, the entire policy is put into context with the data processing. Consent is not collected during the installation, but at the moment and in the context of the concrete data collection. This is nice for the data subjects as they understand the goal and scope of collection from the context it is happening in. There is no need to have a huge document describing a system in natural language and still remaining very ambiguous. If my public transport routing app wants my location data at the moment I look for going from current position to Vienna Praterstern, a data subject will easily understand why the application needs location data at this moment. If data subjects are curious, the SPECIAL system will allow them to also see, control and alter the information that may go into a profile. One way of doing this is to redirect them to the privacy dashboard.

For developers dynamic consent has also a lot of advantages. They do not need to document a certain state of their system in long natural language documents, because it is understood out of context. This spares a lot of work. Changes to the system may in some cases not even need adaption of the dynamic consent system as the context will be self explaining. It is also an advantage for complex systems, as the data protection questions will arise in context, not as some abstract task that needs a total overview of the system. This will make privacy by design and data protection systems so much easier.

But with dynamic consent, there is a system, and not a document. Given high fines, GDPR Angst and the tendency of managers to always look for more control, the main obstacle will not be technical or design barriers, but management barriers. How to present complex systems and partial data protection solutions like consent recording to management will have to be further explored. It is an interdisciplinary challenge as the technical changes have immediate repercussions into the management questions and presentation challenges.

7.3 Sticky policies

Sticky policies is the term for the approach to attach the policy to the data in a manner that ensures that the policy is tightly coupled to the data (which is especially important when data transcends company boundaries).

Those are created in the following way as can be seen in figure 1.1 and as it is explained in D1.8[14] :

- Data (both payload and the consent data) is harmonised by making the semantics of the data explicit.
- Data is augmented with consent approval and usage policies and other metadata.
- either ensure that the data is securely and efficiently accessible and provide an application programming interface (API), or
- package data and metadata together and make their link explicit (with URIs) before transferring the package across company borders.

All use cases experimented with sticky policies in order to provide correct data for the dashboard or the compliance checker. Research into it has not really materialized and there is only one proposal as to how to encode it. It is not a requirement for the Proximus use case but would nevertheless be seen within Proximus as an added value for other use cases. CERICT suggested to attach the policy to the data subject while the ERCIM concept attaches policies to payload data. It was also the question that in practice, it seems that any algorithm that has access to the data after understanding the policy can simply repackage the data without the policy. So how to secure the packaging and harmonisation between data and metadata remains a research question. Those can be solved in many ways that have to be matched against the use case and against each other.

Refinitiv assumed sticky policies when reporting their work on GDPR for transfers outside the EEA and also when exploring how to combine the SPECIAL system with smart contracts. They tested 4 options for compliance:

- Using inferencing (along the lines suggested by the policy language specification)
- Using closed-world rules (SHACL)
- Using a SPARQL query
- Using smart contracts in the context of the blockchain.

They concluded that all of the approaches worked, but that the performance profile of the Tenforce compliance engine was best for real time compliance checking. For smart contracts and blockchain, they noted that the compliance algorithm from first principles was painstaking and the performance profile was heavily dependent on the complexity of the policies checked. No further investment in this last option was envisaged. It was also noted that the system is a support within a workflow. Some of the decisions still need human intervention.

Deutsche Telekom was part of the W3C Do-Not-Track Working Group and was aware of the existence of sticky policies from the start. They consider sticky policies a charming feature and an attractive technical concept. But it only works on a conceptual level. The pilot uses sticky policies in the SPECIAL server. However, there is no end-to-end implementation of the concept available at this time. Deutsche Telekom used SPECIAL as a step towards the subsequent implementation in the productization phases.

The enforcement in such a system remains an open research question. SPECIAL made the assumption that if data is labeled correctly via metadata and policy information, the downstream data controller will also honor this. It is assumed that enforcement is done via out-of-band mechanisms like contractual obligations to honor the metadata system. The technical enforcement would need the adaption of this rather light attacking model and research new solutions.

Chapter 3

Legal considerations

With the interdisciplinary nature of SPECIAL, the challenge for the legal part of the project was to find solutions that fit the overall system design and architecture. This goes way beyond the classic legal scientific work that takes a given situation and examines its legal compliance. SPECIAL required not only understanding the legal environment of the use cases, it had to find new ways to satisfy legal requirements. Those new ways had to be checked for coherence and needed a plausible and convincing argumentation.

The legal challenge was twofold. SPECIAL explored new ways of expressing legally relevant facts in machine readable form. A privacy policy in PDF is understood by the lawyers. A privacy policy in RDF is understood by the engineers. In this context, it is a challenge to translate the legal expectations into something the machine can represent. Because there are as many legal ways to express something as there are technical ways to express certain needed semantics, the result is a many to many relation. SPECIAL solved this by making an informed choice.

A further challenge was the legal check whether the system designed does actually produce the evidence required and thus creates the desired legal effects. Only because there have been legal requirements in a PDF was not a guarantee that the final system would stand the final legal check against provisions like GDPR, but also against provisions and rules stemming from a contracted business relationship.

In SPECIAL the translation was done using experience from past research in PRIME and Primelife, but also from standardisation efforts like Do-Not-Track and P3P. There is no obvious way to translate the legal requirements into a semantified workflow. But by doing the translation, all use cases had in common that the slightest legal inconsistency or conflict between several laws came into spotlight and required a solution.

1 Personal Data

SPECIAL assumed that anonymisation, correctly done, would reduce the data quality and the usefulness of the information left so low, that no really interesting insights could be gained from it. There is an entire scientific field that deals with the question whether data is still personal data. GDPR defines it in Article 4 (1) in the following way:

> (1) *personal data* means any information relating to an identified or identifiable natural person ('data subject'); 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;

All SPECIAL use cases assumed personal data in that sense, although, for the project itself, the implementation and the testing, generated data was used in order to avoid liability issues.

Assume the applicability of GDPR, the following sections give an introduction. The considerations of the use cases exemplify the way a use case is analysed and how the SPECIAL tools can be used to escape from deadlocks and impasses. Although the SPECIAL system can do a lot, it is no panacea and we found one obstacle we were not able to overcome.

1.1 Special categories

While creating the workflow and evaluating the necessary safeguards to be put in place for legal processing, the special data categories of Art. 9 GDPR need a particular attention. The presence of such data categories alters the balance between the interests of the data controller and the interests of the data subject. As the data falling into those categories are considered *sensitive* and being of high risk, the processing safeguards have to be reinforced considerably. The requirements for consent become much higher as the presence of *explicit* consent is needed. Also other legal grounds for processing are narrowed. In general, it can be said that the processing of the special categories mentioned below requires a much stricter consideration of the necessity principle.

The *special* data categories in Art. 9 GDPR are:

1. Racial or ethnic origin
2. Political opinions
3. Religious or philosophical beliefs
4. Trade union membership
5. Genetic data
6. Biometric data for the purpose of uniquely identify a natural person
7. Information about health or sex life, and sexual orientation

SPECIAL use cases 1 & 3 use location data for the location based services. In the list of special categories of personal data, location data is not mentioned. Nonetheless, the processing of location data is sometimes seen as high risk processing since sensitive information related to the special data categories mentioned in Article 9 GDPR may eventually be derived indirectly. Therefore, in the context of location data, it deems advisable to treat location information like a special category of personal data. Particular care should be taken to deploy sufficient technical and organisational measures to achieve an adequate level of protection, including a prior data protection impact assessment[17].

For the second use case, a detailed description of data collected and processed is given in D1.2[22]. Depending on the parts of the *Know Your Customer (KYC)* system, that is using the SPECIAL system, there may even be processing of sensitive data categories. Given the high security standards in the financial industry, the focus of the research related to this use case remained in the area of legal grounds for processing.

2 Controller & Processor

Article 4 (7) GDPR provides a definition for a data controller, the entity legally responsible for the personal data collection and processing (bold highlights added):

> *controller* means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law;

Therefore, the entity making the decisions over which data is being collected and why is to be seen as the controller. Thereby, it is important to note that the GDPR explicitly allows joint controllership (alone or jointly with others) where under such circumstances, several entities can be responsible. The determination who is controller must be made taking into account the real circumstances of the individual case and the factual influence of the entity in question[15]. Consequently, not all recipients of personal data are controllers. Rather, in cases where another entity determines purposes and means of the processing, the recipient could be a processor. Article 4 (8) GDPR defines the term *processor* as well, stating:

> *processor* means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller;

Thereby, it is notable that the GDPR obliges both, the controller, as well as the processor with the protection of personal information. Nonetheless, when a processor agrees to process data on behalf of the controller the following is required:

- A precise allocation of responsibilities,
- managerial authority of the controller and
- based on this authority, the processor is bound to the instructions of the controller.

Any processing on behalf of the controller must be governed by a contract or legal act under Union or Member State law. The European Commission or a European supervisory authority may lay down standard contract causes to be used. What is new is that a contract can now also be in electronic form, not just in writing (Art. 28 (8) & (9) GDPR). Due to the broad territorial scope of the GDPR (Art. 3 (2) these rules also apply to cases where a controller or processor is located outside of the EU. If the processor wants to involve a sub-processor, the controller needs to agree first in written form (Art. 28 (2) & (4)). Then, the processor needs to oblige the sub-processor with the same duties corresponding to those imposed on him in his agreement with the controller. In case of an infringement of the GDPR, the data subject can turn to the controller and the processor(s) liable to demand compensation for material or non-material damage suffered (Art. 82 GDPR). This can under circumstances mean that the controller and

the processor can be liable jointly, whereas the data subject is free to decide to hold one of them responsible for the entire damage to receive effective compensation. In turn a controller or processor being held liable for the entire amount can claim back part of the compensation from the other responsible controller(s) and processor(s) (Art. 82 (5)).

3 The basic principles of data protection

In order to assess a use case, the evaluation of proposed solutions always needs to take the basic principles of data protection into account.

Article 5 (1) GDPR presents those basic principles to enable lawful personal data processing. It says personal data shall be [1] :

1. processed lawfully, fairly and in a transparent manner in relation to the data subject (lawfulness, fairness and **transparency**);
2. collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes (**purpose limitation**);
3. adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed (**data minimisation**);
4. accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay (**accuracy**);
5. kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) subject to implementation of the appropriate technical and organisational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject (**storage limitation**);
6. processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures (**integrity and confidentiality**)
7. The controller shall be responsible for, and be able to demonstrate compliance with, paragraph 1 (**accountability**).

In order to drill down into the detail of those principles, SPECIAL provided D1.2[22] containing a legal manual for GDPR.

[1] highlights in bold by the author

4 Legal grounds for Processing

GDPR follows Directive 95/46EC by establishing a general prohibition of processing of personal data. So instead of prohibiting certain actions, GDPR removes all freedom of processing and establishes several large classes of permissions. This influences the way things are modelled in the SPECIAL system. From a legal perspective it is important to look at Art. 6 GDPR:

1. Processing shall be lawful only if and to the extent that at least one of the following applies:
   (a) the data subject has given consent to the processing of his or her personal data for one or more specific purposes;
   (b) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;
   (c) processing is necessary for compliance with a legal obligation to which the controller is subject;
   (d) processing is necessary in order to protect the vital interests of the data subject or of another natural person;
   (e) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;
   (f) processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.

4.1 Consent

SPECIAL took the challenge to try reconcile potential big data innovations with the requirements of data protection. As we have seen above, a legal ground can be found in laws and contracts. But most of those laws don't yet take into account modern big data innovation or are too coarse grained to really help the SPECIAL use cases. This is the reason why SPECIAL concentrated efforts on getting to consent with the data subject in difficult systems.

4.1.1 Requirements for consent

The existence of valid consent must be demonstrable by the data controller (accountability principle). The legal preconditions for consent are laid down in Art. 4 (11) and 7 of the GDPR, requiring valid consent to be :

- freely given, specific, informed and unambiguous (for one or more specific purposes)
- possible to withdraw at any time
- a statement or clear affirmative action of data subject expressing agreement

Freely given, specific, informed and unambiguous (for one or more specific purposes) Art. 7 (4) GDPR prohibits bundling consent to other conditions. The performance of a service or contract can not be dependent on the collection of data that is *not necessary* for the performance of that contract or service. A lot of ink has been used to discuss about the fallacies of this prohibition of bundling consent to other things. Most Data Protection Authorities assume that globalized, generic consent for multiple vague purposes is not freely given.

While this is of high stakes in the surveillance economy that wants to extract a maximum of data, SPECIAL assumed a fair exchange between data controller and data subject. SPECIAL preferred to clearly distinguish between necessary data collection for the service and the data that is used to provide additional services to the data subject or data that is used as a monetization vector.

One of the major challenges in this context was to provide *a specific consent* to the data subject. As already mentioned in Section 7.2, the challenge to define specific purposes for an entire multi task application in a privacy policy leads either to people reading, being overwhelmed and not consenting or to people not reading and consenting to the unknown. Both options aren't in the interest of data protection or society.

SPECIAL provides two tools to help with the situation. First, the privacy dashboard (Section 7.1 allows ex-post control over data or data categories that have been collected without being necessary, thus easing the situation of those having consented to the unknown. Secondly, the concept of dynamic consent tries to partition the consent requests into understandable chunks that even anxious data subjects can consent to and that allow to build a larger permission over time. Both need a combination of legal and technical considerations to design a system producing legally valid permissions.

Withdrawal of consent Art. 7 (3) GDPR demands that consent can be withdrawn at any time, and the withdrawal must be as easy as giving consent. This is to prevent higher burdens for withdrawal, such as when giving consent would be with just one click online, but the withdrawal is required by the controller in written form or the like. The SPECIAL system finds support for its design in Art. 21 (5) GDPR. Art. 21 (5) states that automated procedures to enable the data subject to exercise the right to object are possible. The SPECIAL privacy dashboard uses this option to allow to not only withdraw, but also to alter the consent given.

But the data subject must receive information about the possibility of consent withdrawal prior to giving consent. SPECIAL achieves this especially in the location based services by informing about this option during the installation of the application as indicated in D4.5 (Chapter 5) [20].

Statement or clear affirmative action of data subject SPECIAL started with the aim to develop new, none-invasive forms of consent collection. This was inspired by the failure of privacy policies or cookie banners to really serve data self determination. But it was also inspired by the fact that it is very hard to get people to opt-in to data processing.

Talking about non-invasive interfaces, the challenge is to strike a balance between annoying the user and lack of information or possibilities of interaction. For SPECIAL it was thus important to understand what a clear affirmative action really meant.

The statement of consent does not necessarily need to come in written form. Recital 32 GDPR gives some examples, such as:

- ticking a box when visiting an internet website,
- choosing technical settings for information society services or
- another statement or conduct which clearly indicates in this context the data subject's acceptance of the proposed processing of his or her personal data

However, there are some examples what is not sufficient:

- silence,
- pre-ticked boxes or
- inactivity

Additionally, consent requests need to be clear, concise and not unnecessarily disruptive to the use of the service for which it is provided. It must be clearly distinguishable from the other matters. Art. 7 (2) GDPR makes clear that any consent given on the basis of a request not complying with these preconditions is not binding (i.e. invalid). Yet another occasion where, given the current complex globalised information systems, telling the data subject everything up front is making the data subject drinking from the fire hose. With the predictable result already mentioned, that either information is ignored, or it is overwhelming.

Exploring various consent interfaces as described in D4.5[20] and by inventing *dynamic consent* SPECIAL tried to break free of this dead end.

4.2 Dynamic consent

In Chapter 2, Section 7.2 it was already explained why giving a data subject all at once during the installation time is an approach that overwhelms the user and how *dynamic consent* can help in this situation by putting the consent request not at the beginning, but into the context of the operation it is meant to cover. With dynamic consent, a service or application provider does not have to explain the system to the user. Many things are just naturally explained by the context in which they appear. In such context, the assumption is that data subjects are much more inclined to give consent to things they do understand.

But there is no easy and explicit support for dynamic consent in the GDPR. To solve this and after a long and burdensome legal research, SPECIAL came up with a layered consent to give dynamic consent a legal justification. This legal justification was accepted by two very senior former Data Protection Commissioners as a good way forward while waiting for a new ePrivacy Regulation that could give legal value to preferences and consent requests that are contextualised.

In the be-fit use case, but also for the location based services, SPECIAL suggests to have an initial installation phase. During installation, the data subject receives most of the legally required information, except for actual data or data categories. Additionally, as such a system is totally new to a potential user, SPECIAL offers a sandbox in the

installation phase that explains and allows to play with the dynamic consent interface. And during this installation phase, we ask the data subject for consent to use the SPECIAL non-invasive dynamic consent mechanism. This means SPECIAL asks for consent for the future consent mechanism.

Legally, the biggest question was, whether this first request was *specific* enough. But as it is later filled with new information, SPECIAL was able to argue that the new consent mechanism has the intention to make even more specific consent requests in the course of action. Which is much better for data self determination.

The non-invasive interface consists of a banner that will allow the data subject to consent, to decline consent or to get more information (See D4.5[20]). But such a banner would be invasive like those current cookie-banners we see all over the Web. If there is no interaction, the banner goes away again.

The biggest question was now the result of such non-interaction. A first reflex assumed that the system would then report a *NO* to the SPECIAL system. This would even increase the rate of missed opt-ins. Over time, it would also annoy the data subject using the application by asking the ever same questions again and again. In this situation, the industry will further look into circumventing consent and will lobby for exceptions to data protection fitting a particular economically successful use case.

But SPECIAL could not just have said non-interaction means *YES* because of Art. 7 explained above. But given that the data subject would be annoyed by being obliged to always say *YES*, couldn't there be an automatic way of saying *YES*? As there are many different preferences possible, a system should not hardcode one option as the only one. In order to accommodate this fact, SPECIAL decided to ask for the default behaviour during the installation phase. Some people wanting to be more liberal with their data or already knowing a certain system would be able to configure the banner in a way that the frontend would send a *YES* by default, when the banner goes away.

This still wasn't good enough. Because one may make mistakes and one may miss to tap on the *NO* - button of the banner. What then? Luckily, in SPECIAL, there is a privacy dashboard that not only allows to look at one's own data, but also to verify and change data. A mistake is not final anymore. In the privacy dashboard, it is easily changed. This was sufficient to allow for the non-invasive behaviour of the banner. Of course, the defaults should not only be accessible for configuration at installation, but should be also part of the preferences of that application.

4.3 Other Legal grounds for Processing

During the design and elaboration of the SPECIAL use cases, we found out that there is a lack of understanding of law and contract as a legal ground for the processing of personal data. Privacy policies and consent requests are sometimes also far too large, because necessary data collection covered by contract or law are still mentioned.

Implementing a use case, the design phase should collect and prefer other legal grounds for processing. Those do not need a consent interface. But those reasons shall still be put in the metadata to prove compliance. And once such metadata is in the data lake recording the other legal ground for processing, of course such reasons can be exposed in the privacy dashboard.

5 Specific considerations on use cases

To asses and evaluate new use cases, one may benefit from the diligent legal analysis that has been done for the SPECIAL use cases. The work can be explored in full depth in SPECIAL Deliverable D1.6[21]. But the legal mentoring continued until the end of the project, underlining how interdisciplinary the work in SPECIAL was. The same counts for use cases wanting to implement the SPECIAL system or parts thereof.

In general, the implementation of the SPECIAL system forced people to remove white spaces and ambiguities. One can still hide a lot of ambiguity in a human readable privacy policy. There are unspoken rules that provide wriggle room for the processing of data. There are perpetuated misunderstandings what a certain rule would mean for data processing. When implementing SPECIAL, the ambiguity has no room anymore because the knowledge graph has difficulties recording and expressing uncertainty and leeway. The metadata brutally enforces a high degree of clarity concerning the data processing, the purposes and the rules and rights attached to that processing. Compromises found by committee suddenly show clearly why they can not be retained as solutions. All of that has significant repercussions into management hierarchies and project management. For one of the use cases and due to a conflict of two laws, we found a clear weakness in the currently implemented solution that was rather difficult to resolve. One should be prepared for such difficulties when starting to create a project of a SPECIAL system implementation.

6 Location based services

The generic location based service, the Proximus use case and the use case done by Deutsche Telekom all had location data and a profile at their center. Location data enriched the user profile and the user profile helped to filter and direct information. The generic use case allowed for a full view of the challenges for the entire data life cycle while the industry use cases were scoped from the perspective of the respective needs of the industry partners.

The generic use case was the basis for successful research and laid down the basic architecture. This was important in order to drive the agenda with the industry use cases. It allowed to determine which of the components could realistically be pushed in a telecommunication industry context.

D1.6 analysed the requirements for location based services and gave the following hints for the design, installation and runtime of a location based services :

1. Upon installation, the data subjects must be informed about the specific privacy configurations available to them in the application. For the website functionality, this should happen at first visit. A way to imagine such information is a wizard that helps on first run, or the option for first run expert configuration. See D4.5 for examples. The concrete messages then depend on the specific service that is offered on the basis of location data since they need to communicate the purpose of the processing to the user.

2. From Art. 21 (5) GDPR, we can also deduct that if consent is acquired over such a technical mechanism, the same mechanism, where possible, must also offer a way

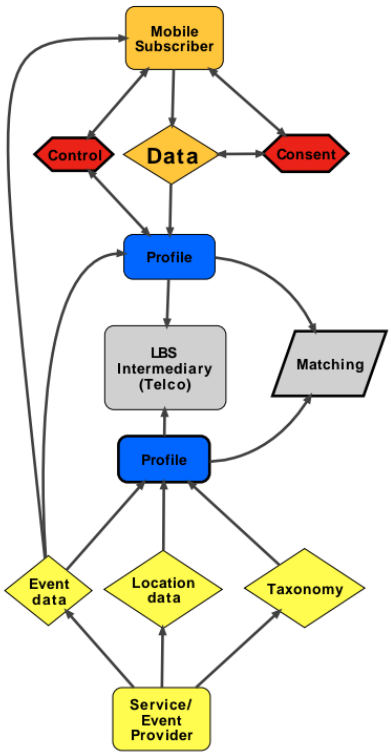

Figure 3.1: A Schema of Location Based Services

to withdraw consent. The *where possible* is important, because current operations may still be under way and can not be stopped immediately without disruption. Sometimes, withdrawal of consent may be in conflict with the need to invoice past uses of the service in the context of a certain invoicing period. There are many other potential situations where deletion only works in an asynchronous way and after some lapse of time. The goal must be to serve the user without creating undue disruption in the system.

3. The consent can evolve and extend over time, depending on contextual add-on requests. This will lead to a stateful consent system that may concern one or more objects, functionalities, services and relations. This means the application or website will have to offer a way to control those preferences, even after they have been set during the installation process or at first hit.

4. An agreement from the data subject regarding the modus operandi for the further gathering of consent is needed. This concerns the handling of the above mentioned extensions to the initial very basic consent needed to start the application. Concrete wordings need to be determined for each use case.

5. As soon as sensitive data in the sense of Art. 9 GDPR is concerned, the system

should never give automatic consent. Rather, it must prompt the user for explicit confirmation. But one can imagine a context or location-based system, like a medical system, where there is a constant collection of sensitive data. In this case, the challenging trade-off between click fatigue and the very specific requirements in Art. 9 GDPR have to be considered carefully to find a tailored solution. In this case the automatic functioning must have additional precautions and functions. The more sensitive the data, the more control functions are needed, during collection and later in the control interface.

The controller might obtain some information from the user's device that does not need the explicit consent of the data subject because other legal grounds allow for its collection and processing. This does not need to be communicated upfront towards the data subject when requesting consent for the usage of the location data for the service provision. Rather, a layered approach can be applied to make this information accessible in the user control interface (dashboard). This way, the data subject can drill down into all the data collection, whether based on consent or legal ground. As it is layered, this approach avoids exposing the data subject to information overload.

Example data categories which might not need to be mentioned due to other legal grounds applicable and which can be exposed in a second or third layer could be:

- Log data
- Collection done for security reasons
- Session cookies (an explanation in a sub-sub-layer will reassure the paranoid)

The installation itself was already described earlier. It must cover information duties according to Art. 13 GDPR. There was a light discussion on whether one could use a layered approach for the notification duties, thus only providing high level information. The Data protection board and the Art. 29 Working Party had issued a recommendation on the minimum set of information that was required. But those targeted Web sites with cookie banners without providing a full privacy dashboard. Still the long notification prevailed because it was easier to explain to the instances evaluation whether the installation phase fulfils all GDPR requirements.

The consent mechanism was able to reduce the amount of clique-fatigue when tested in the Proximus case. We realised that treating with live location data did not only require privacy considerations, but there was also an unease about security of potentially very sensitive location information. A crucial feature to have a pause button where the user would be able to stop enriching the profile was discussed, but not implemented or tested. But it rather looks like a crucial feature.

There were real design and technical problems in fencing the location data and how to tell the application when to be active, when to shut off and when to recommend something. This also affects the questions the system can ask to the user and e.g. the scope of the recorded consent. There is still room for improvement on how space is repartitioned in order to find the appropriate size of cells for a meaningful semantic entity that is still accepted by the data subject. Too big areas can have unwanted effects on cases were location data is accidentally recorded. Too fine grained location

data will burden the system with too much data. Finding the middle ground is again an interdisciplinary task.

The implementation of the Proximus client tried several components and tested them with users. A recommender system based on the location of the data subject allowed to explore technical and legal boundaries. Those can be seen in D5.5[26]. It helped the industry partner to understand the privacy concerns of the data subjects. Shortcuts on functionality and usability were made to keep the focus only on the privacy aspect and less on the commercial readiness or recommender accuracy. The consent mechanism worked, but the mobile platform provided difficulties concerning data deleting via the privacy dashboard. For future iterations one could imagine a multimodal interface that allows to consult the privacy dashboard rather via desktop/laptop than via mobile in case a data subject wants to rectify things. There are simple much less constraints imposed by the size of the device.

The use case of Deutsche Telekom involved two actors sharing data. So apart from having to take into account all the above, the question of interoperability comes into the focus. And this interoperability of data formats does not only concern the data, but also the metadata. Now what is the default behaviour if the other party does not understand a certain semantic provided by the metadata. A legal solution would be to require a minimum set of semantics that must be understood. A technical solution may involve a protocol that can provide feedback to the other party in case a semantic in metadata is not understood.

7 Tooling for compliance

The compliance use case had a very technical focus, especially on policy languages. In the highly regulated financial sector, the number of potential rules to be followed is near to infinite. The legal challenge for the use case was the open and unresolved legal conflict between the GDPR on the one side and the Anti Money Laundering Directive on the other side.

Attempts to solve the conflict of law by consent mechanisms or other constructions around processing permissions suffered from the number of involved data collection sources. Metadata helped in so far as some of the KYC criteria could be collected with a permission to complete a KYC profile thus helping future checks, making them easier and faster. Which is also in the interest of the data subject who is evaluated against Anti Money Laundering rules.

The industry partner already had experience with Linked data and knowledge representation. The internal translation from human readable guidelines into technical metadata and rules able to help with automatic compliance checking benefited a lot from that pre-existing knowledge. Contributions to the Data Protection Vocabulary Communnity Group included an ODRL[10] Regulatory Compliance Profile, where ODRL is used to map permissions and obligations from regulatory constraints into a policy language. The challenge was to overcome ODRL's origins as an language to express semantics for copyright licensing.

The SPECIAL system did not provide a key-turn solution for the need of the industry partner. But doing the exercise of applying the SPECIAL tooling to their own business processes gave new insights, tools and improved the situation for those involved in KYC.

It is interesting to see that all pilots, meaning even the KYC case, benefited from the fact that the clarity provided by translating the legal challenges into at technical system also provided new insights to create innovative and satisfactory new user experience. This is evidence for the fact that in the interdisciplinary cooperation around bureaucracy automation, the clarity forced by the technical implementation of social rules translates into a simplification of the user experience that fosters understanding of the situative expectations.

Bibliography

[1] S. Auer, S. Scerri, A. Versteden, E. Pauwels, A. Charalambidis, S. Konstantopoulos, J. Lehmann, H. Jabeen, I. Ermilov, G. Sejdiu, A. Ikonomopoulos, S. Andronopoulos, M. Vlachogiannis, C. Pappas, A. Davettas, I. A. Klampanos, E. Grigoropoulos, V. Karkaletsis, V. de Boer, R. Siebes, M. N. Mami, S. Albani, M. Lazzarini, P. Nunes, E. Angiuli, N. Pittaras, G. Giannakopoulos, G. Argyriou, G. Stamoulis, G. Papadakis, M. Koubarakis, P. Karampiperis, A.-C. N. Ngomo, and M.-E. Vidal. *The BigDataEurope Platform – Supporting the Variety Dimension of Big Data*, pages 41–59. Springer International Publishing, Cham, 2017. ISBN 978-3-319-60131-1. doi: 10.1007/978-3-319-60131-1_3. URL `https://doi.org/10.1007/978-3-319-60131-1_3`.

[2] P. Bonatti, S. Kirrane, J. D. Fernández, C. Galdi, L. Sauro, D. Dell'Erba, I. Petrova, and I. Siahaan. Deliverable D2.8: Transparency and compliance algorithms v2, Nov 2018.

[3] P. Bonatti, S. Kirrane, I. Petrova, L. Sauro, C. Kerschbaum, and E. Pirkova. Deliverable D2.6: Formal representation of the legislation v2, Dec. 2018.

[4] P. Bonatti, S. Kirrane, I. Petrova, L. Sauro, and E. Schlehahn. Deliverable D2.5: Policy language v2, 2018.

[5] P. Bonatti, S. Kirrane, and R. Wenning. Deliverable D1.7: Policy, transparency and compliance guidelines v2, 2018.

[6] P. A. Bonatti, F. De Meersman, S. Kirrane, M. Kurze, R. Wenning, B. Whittam-Smith, E. Schlehahn, J. Colbeck, R. Jacob, M. Piekarska, H. Zwingelberg, and L. Sauro. Deliverable D1.5: Use case scenarios v2, Feb. 2018.

[7] S. Decker and V. Peristeras, editors. *Data Privacy Controls and Vocabularies: A W3C Workshop on Privacy and Linked Data*, 2017. W3C. URL `https://www.w3.org/2018/vocabws/`.

[8] W. Dullaert, U. Milošević, J. Langens, A. S'Jongers, N. Szepes, V. Goossens, N. Rudavsky-Brody, W. Delabastita, J. D. Fernández, and S. Kirrane. Deliverable D3.4: Transparency and compliance release, Jan. 2019.

[9] P. Groth and L. Moreau. An overview of the prov family of documents. Technical report, Apr. 2013. URL `https://www.w3.org/TR/prov-overview`.

[10] R. Iannella and S. Villata. Odrl information model, Feb. 2017. URL https://www.w3.org/TR/2017/WD-odrl-model-20170223/.

[11] S. Kirrane, J. D. Fernández, A. Polleres, U. Milošević, and J. Langens. Transparency framework version 2, Nov. 2018.

[12] S. Kirrane, U. Milošević, J. D. Fernández, and A. Polleres. Deliverable D2.3: Transparency framework v1, Feb 2018.

[13] A. M. McDonald and L. F. Cranor. The cost of reading privacy policies. *ISJLP*, 4:543, 2008. URL http://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/isjlpsoc4&section=27.

[14] B. B. V. Nuffelen, U. Milošević, and W. Dullaert. Deliverable d1.8: Technical requirements v2, 2017.

[15] E. C. J. of (ECJ). Google vs spain. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62012CA0131.

[16] H. J. Pandit and A. Polleres. Data privacy vocabulary v0.1, July 2019. URL https://dpvcg.github.io/dpv.

[17] A. . W. Party. Guidelines on data protection impact assessment (dpia) and determining whether processing is "likely to result in a high risk" for the purposes of regulation 2016/679. *Opinions of the Article 29 WP*, 248, Apr 2017. URL http://ec.europa.eu/newsroom/document.cfm?doc_id=44137.

[18] R. W. Piero Bonatti, Sabrina Kirrane. Deliverable D1.3: Policy, transparency and compliance guidelines v1, 2017.

[19] P. Raschke, O. Drozd, U. Milošević, and S. Kirrane. Deliverable D4.4: Usability testing report v2, Mar 2016.

[20] P. Raschke, O. Drozd, and B. Bos. Deliverable D4.5: Transparency dashboard and control panel release final, Nov 2019.

[21] E. Schlehahn and R. Wenning. Deliverable D1.6: Legal requirements for a privacy-enhancing big data v2, Apr 2018.

[22] E. Schlehahn and H. Zwingelberg. Deliverable D1.2: Legal requirements for a privacy-enhancing big data v1, Oct 2017.

[23] B. van Nuffelen. Deliverable D3.1: Initial setup of policy-aware linked data architecture and engine, June 2017.

[24] W3C. The platform for privacy preferences 1.1 (P3P1.1) specification, 2006.

[25] R. Wenning and E. Schlehahn. Dynamic consent als weg aus der einwilligungskrise. In E. Schweighofer, F. Kummer, W. Hötzendorfer, and C. Sorge, editors, *Trends and Comminities of Legal informatics*, IRIS 2020. Österreichische Computergesellschaft, 2020.

[26] B. Whittamsmith, R. Jacob, and M. Kurze. Deliverable D5.5: Pilot implementations and testing plans v3, Nov. 2019.