Rigo Wenning and Sabrina Kirrane

Compliance Using Metadata

Book Section (Accepted for Publication)

Original Citation:

Wenning, Rigo and Kirrane, Sabrina ORCID: https://orcid.org/0000-0002-6955-7718

(2018)

Compliance Using Metadata.

In: *Semantic Applications. Methodology, Technology, Corporate Use.*

Springer, Berlin, Heidelberg.

pp. 31-45. ISBN 978-3-662-55432-6

This version is available at: https://epub.wu.ac.at/6497/
Available in ePub^WU: September 2018

Compliance using Metadata

Rigo Wenning, Sabrina Kirrane

3. Compliance using Metadata

Authors: Rigo Wenning, Sabrina Kirrane

Abstract (for Springer.com)

Everybody talks about the data economy. Data is collected stored, processed and re-used. In the EU, the GDPR creates a framework with conditions (e.g. consent) for the processing of personal data. But there are also other legal provisions containing requirements and conditions for the processing of data. Even today, most of those are hard-coded into workflows or database schemes, if at all. Data lakes are polluted with unusable data because nobody knows about usage rights or data quality. The approach presented here makes the data lake intelligent. It remembers usage limitations and promises made to the data subject or the contractual partner. Data can be used as risk can be assessed. Such a system easily reacts on new requirements. If processing is recorded back into the data lake, the recording of this information allows to prove compliance. This can be shown to authorities on demand as an audit trail. The concept is best exemplified by the SPECIAL project https://specialprivacy.eu (Scalable Policy-aware Linked Data Architecture For Privacy, Transparency and Compliance). SPECIAL has several use cases, but the basic framework is applicable beyond those cases.

Key statements:

1. Get conscious about the workflows and create a register of procedures that are affected by compliance requirements (e.g. Privacy requirements).
2. Model the policy constraints from the legal and corporate environment into Linked Data to create policy metadata.
3. Attach the relevant policy metadata to the data collected, thus creating a semantic data lake via Linked Data relations or Linked Data annotations
4. Query data and relevant metadata at the same time to only process data that has the right policy properties.
5. Write the fact of processing back into the semantic data lake and secure it with appropriate measures (e.g. Blockchain).

3.1. The Increased Need for Tools for Compliance

The digitisation of all aspects of our life results in systems becoming ever more complex. As they get more complex, humans have more and more difficulties in trying to understand what these system do. As billions of people live their lives online, they leave traces. Others have started to measure our environment in all kinds of ways. The massive amount of sensors now produces massive amounts of data. Moreover, because our society communicates in many new ways online, it creates new complex social models. The advent of open source software can be taken as an example. The open source ecosystem would not be possible without the Internet and the Web that allows complex governance structures to be built online[RAY]. Social networking, browsing habits and other interactions online are recorded. This leads to the creation of massive amounts of data. Data collection online and offline corresponds increasingly to the big data characteristics of velocity, variety and volume.

It is now tempting for certain actors to exploit the intransparency of those complex systems. This is done by harvesting and monetising data in opaque ways, or by just benefitting from protocol chatter. The internet as a whole has basic vulnerabilities in this respect. The most threatening example is certainly the pervasive monitoring of all internet traffic by the NSA and GCHQ [STRINT]. Additionally, the private sector is monitoring behaviour on the web, also known as “tracking”. A short term benefit is offered to a targeted individual and people do not realise the long term danger of the profile created. Various techniques are used to build profiles of people and sell them to the highest bidder. Entire platforms and toolchains are created and made available for free to be able to monitor what people do on a system.

Some call this the surveillance economy [SURV] as the prices for ads targeted to a profiled person generate much higher revenue than normal banner ads.

Malicious behavior in complex systems is not limited to eavesdropping on communications. The recent scandals on IT manipulations revealed that by manipulating one end of a complex system, there can be huge benefits on another end of such a system. An example of this is the Libor scandal [LIBOR] in the financial industry, where an index used for the calculation of interests was gamed to obtain certain results. The software in cars also had hidden functions that detect when the car is in a test cycle inside a lab and changes the engine's characteristics to comply with requirements said to be unachievable[1].

The combination of complex and intransparent systems with manipulations is undermining the trust people have in the correct functioning of those systems. This is especially true if sensationalist media with a hunger for audience and attention widely reports and exploits those topics. Verification by users is difficult.

As a consequence, people are more reluctant to use those complex systems. If they have an opportunity to avoid using the systems they will do so due to the lack of trust and confidence. This creates economic inefficiencies and hinders further progress of society by even more complex systems. The pace of innovation is seriously endangered if people start to mistrust the IT systems they use.

Consequently, governments all around the world create new regulations and demand compliance with those regulations. The aforementioned privacy abuses and tracking excesses have accelerated the reform of the European data protection law that lead to the General Data Protection Regulation (GDPR) [GDPR]. The scandals in the financial industry resulted in additional rules for reporting. However, lawmakers responsible for the regulations are often positivistic and underestimate the difficulties in implementing the regulation by transforming rules into code and organisational policies. Often, the compromise is to implement what is implementable and try to not get caught with the rest. The approach suggested here puts forward a different way. It suggests to use more technology: "*Social rules*" are translated into machine understandable metadata and can then steer and guide the behaviour of our complex systems using such metadata. "*Social rules*" in this sense include laws, usages but also user facing promises like privacy policies.

In order to demonstrate the compliance with regulations, especially in data protection and security, laws and implementation provisions very often recommend certification. The traditional way of doing certification is to engage some expensive consultant or auditor who examines the IT system and asserts that the system does what it says it does. The result is often an icon displayed on a website. This is very expensive and does not scale well. Additionally, those certification systems have a number of disadvantages. A slight change to the system can render a certification void[2]. Security certifications can even be harmful according to a study by KU Leuven [CSSe]. The privacy seals carry a dilemma that a service showing the seal also pays the seal provider. The seal provider has little interest in going against their customer. Not only has the manual certification disadvantages, but it is also inflexible. In order to address the trust issue, the SPECIAL project[3] develops a standardised transparent Web-based infrastructure that makes data sharing possible without destroying user confidence. This will benefit the overall economic impact and growth of the data value chain.

With the metadata approach, a flexible system is installed that can cope with policy changes and audits. Creating such a system requires some scrutiny over business processes, data collection, purposes and retention times. Implementing the approach thus serves the goals of renovating and optimising the business processes at the same time.

3.2. Making the data lake usable

[1] Also known as the Volkswagen emissions scandal, but many vendors are implicated.

[2] Except where the certification is meaningless or very imprecise.

[3] www.specialprivacy.eu

Since the Primelife project[4], we know about the issue with data reuse in Europe. Typically, data is collected at some point in time under certain conditions. Then this data is written into some data warehouse. Later, there is no knowledge anymore under which conditions the data was acquired and recorded into the data lake.

All those warehouses create a data lake full of valuable data. However, the data lake now contains a lot of muddy water; as there is no information about the purposes for which the data was collected, nor any information about the rights attached to the data. A potential new use of that valuable data faces the obstacle of legal uncertainty. This legal uncertainty creates a sufficient commercial and liability risk to deter commercial actors from realising the potential of the data lake. Regarding privacy, we can see that those services not following data protection rules can monetise their data and grow, but they will erode trust and usage of data driven IT services. The eroded trust will decrease the use of those services and generate less data, or it creates lower quality data because people lie about things. In our digitised world, networked services generate lots of data. Using this data is the most promising source for higher productivity and wealth. To avoid the erosion of trust and the decline of networked services, systems have to demonstrate how they follow the democratically established rules. But in the era of big data, the complex systems are too complex for pure human cognition. This means technology needs to help achieve demonstrable compliance that is understandable by humans.

Data protection often calls for data minimisation, meaning the collection and storage of less information. However, the suggestion here is to add even more information. This additional information is metadata containing policy and usage information that allows a system to keep the promises made to the data subject. It allows the data lake to be both usable and compliant. By having this information available, the risk and liability can be assessed. At the same time, it allows for a better automation of compliance processes required by law because it adds a layer of transparency and logging.

3.3. Concept and Architecture of a policy aware system

3.3.1. Data acquisition

The suggestion is to make the "*social rules*" available to the system machine readable as far as possible[5]. We refer to this as "*policy information*" in the following. This is best done using Linked Data (See Chapter 4). The policy information completes the "*environmental information*" available at data acquisition time. Environmental information here means all protocol headers and other information available to the IT system in question at data acquisition time. This information is normally spread over a variety of log files, but may also include other information sources like policy files and DNT headers e.g. Now the system has not only collected the actual personal data, but also metadata about the collected personal data. This could also be metadata about financial data or other data relevant for compliance with given policy information. The idea behind this extensive data acquisition is that, at collection time, we normally know perfectly well under which conditions data is acquired and, e.g., for how long it can be held. Today, this information is lost and forgotten once data is transferred to some data warehouse. However, the system suggested here will remember the constraints at collection time. Those constraints are stored as metadata and connected to the data collected. For various cases, there are existing ontologies to allow the conversion of environmental and contextual information into machine readable metadata. For other cases such ontologies or sometimes even taxonomies can be created for the specific requirements of a given business process.

[4] http://primelife.ercim.eu/

[5] See [PAW] and [PPL] were projects trying to implement some part of the idea

3.3.2. Connecting data and metadata

Now we have the payload data[6] that is subject to the intended processing, and we have metadata that tells us more about the data items processed, how long we can retain them, the purpose of the processing and other promises made to the user at acquisition time. Acquiring a full policy into an SQL database with every data item collected would be overkill. Instead, we need to link the metadata to the actual payload data, e.g. personal information like location data and mobile number.

For the system to work, the acquired payload data has to be transformed into Linked Data, a process commonly known as “semantic lifting”. Semantic lifting aims at adding “meaning” or extra meta (semantics) to existing structured/semistructured data following Linked Data principles and standard Semantic Web technologies [ODA]. A key feature of Linked Data is that it uses [IRI]s[7] to identify data uniquely. Once the payload data is identified uniquely, the metadata can point to it.

For example, let us imagine the acquisition of a mobile number for some service that has to be erased after 3 weeks. The IRI given to the mobile number at the semantic lifting could be http://wenning.org/ns/mtel/12345678[8] and the IRI given to the rule for the 3 weeks data retention could be http://www.w3.org/2002/01/p3prdfv1#retention, which equals 1814400 seconds according to the P3P vocabulary. The triple then indicates the phone number, the retention attribute and the retention time. As these are globally unique identifiers, this even works across enterprise boundaries.

Once the semantic lifting is done and once all payload data recorded is given a IRI, the policy information pointing to the IRI of the payload data record can be seen as an annotation of that record (fig.1).

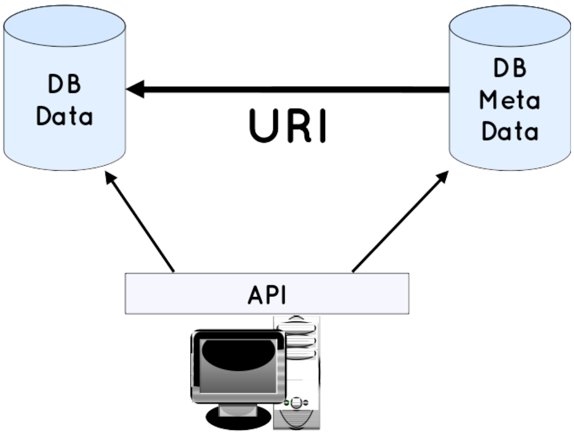

IMAGE: Chapter 3 Figure 1

Figure 1: Annotating data with metadata using RDF

While the W3C Annotation Recommendation [ANO] is about adding remarks on web pages, we annotate data records, but as a web page is a resource with a IRI, the principle is the same. The Annotation data model [AND](fig.2) consequently states: “An annotation is considered to be a set of connected resources, typically including a body and target, and conveys that the body is related to the target. The exact nature of this relationship changes

[6] Payload data means the actual data record, e.g. the name of a customer.

[7] IRI - Internationalized Resource Identifiers, the international version of URI according to RFC

[8] The IRI for the mobile number is a purely theoretical example, the retention time is from the P3P 1.0 Specification.

according to the intention of the annotation, but the body is most frequently somehow 'about' the target". This perspective results in a basic model with three parts, in both cases.

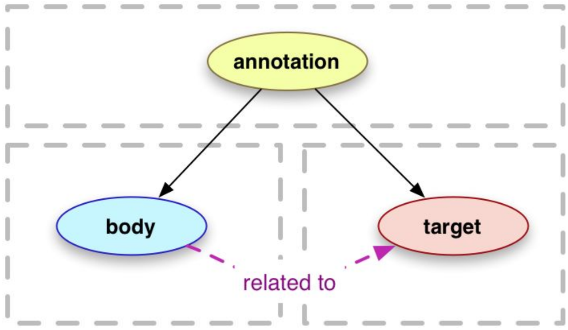

IMAGE: Chapter 3 Figure 2

Figure 2: The W3C Annotation data model
See https://www.w3.org/TR/annotation-model/
Copyright © 2017 W3C® (MIT, ERCIM, Keio, Beihang). W3C liability, trademark and document use rules apply.

During data collection, the environmental information (metadata) and the data receives a IRI each, which allows for linking them together in a triple. This is a form of semantic packaging. Through this semantic packaging, algorithms of the system can react based on that metadata. This rather simple idea leads to a significant amount of social and technical challenges that are not as simple as the basic idea. In the following, the concept is exemplified by a system that tries to achieve compliance with the [GDPR]. The SPECIAL H2020 research project explores ways to implement such compliance and make big data privacy aware.

The system will also work to implement the data retention framework of DIN 66398 that goes beyond privacy and also takes commercial archiving duties into account. It is also perfectly possible to implement reporting and diligence rules from the financial sector regulations. Of course this can be all done using a SQL database with hardwired semantics, but this would remain inflexible and become an island that would not be capable of connecting to the complex world surrounding it. This also means that we need Linked Data in order to allow for data value chains that can be adapted without reprogramming the entire system.

3.3.3. Applying constraints

As the system is using Linked Data, queries will be formed using SPARQL [SPA]. Once the system has data and metadata ingested, the compliance is a matter of applying the right query. It is now possible to make intelligent and policy-aware queries like: "find all data that can be further processed to provide a personalised service". It also allows, according to DIN 66398, to say: "list all data where retention time ends within the next week". In order to make such a query, of course, the system has to know about retention times. An important category of query will concern the constraint on data sharing by asking the system to return only data, for example, "that can be shared with business partner B1 for purpose P2".

3.3.4. Creating policy aware data value chains

Division of labour creates higher efficiency and adds value. In the digitised economy, division of labour also means sharing data. We know that the "sharing economy" created a lot of enthusiasm. At the same time people are more and more concerned about data sharing because the meaning of sharing remains unclear. Is data shared for any use? A solution may be to share data with the policy or constraint-information attached. This is not a new idea. It was most probably Michael Waidner[9] who coined the term "sticky policies" for this concept.

[9] IBM Zürich at the time, now Director of Fraunhofer SIT in Darmstadt

As mentioned, the use of IRI's allows for the preservation of the bundle of policy information and the payload data, even in a collaboration scenario. For a typical data protection scenario, the PrimeLife language uses the terms "data subject", "data controller" and "downstream data controller"(fig.3).

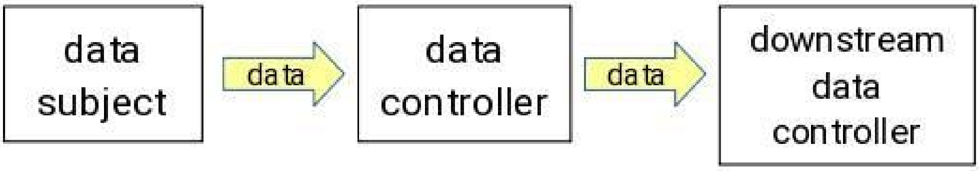

IMAGE: Chapter 3 Figure 3:

Figure 2: A data value chain from a privacy perspective

It is then a matter of choice and of the business model attached to the intended data flow, whether the metadata with policies, constraints and obligations is:

1. Packaged together with the data records in various ways.
2. Delivered in two independent files.
3. Made available via some API by the data controller to the downstream data controller.

The downstream data controller will then have to apply the same constraints as the data controller. Securing this relation is a matter of contractual provisions proscribing the adherence to the constraints received or open to cryptographically secured systems. Such cryptographically secured systems can be similar or even close to the rights management and rights labelling systems we know already today.

3.3.5. An automatic audit for compliance

If a certain personal data record is processed, this can be written as usual into a log file. Instead of using a log file, this fact can also be part of the metadata of the system. In this case the fact of collecting and processing that payload data is attached as an annotation to the payload data. The system can now reason over the meaning of terms like "purpose". With this in mind, a log of relevant information about processing, purpose of processing, disclosure and sharing can be held. Because it is not only a log file, but rather something that can be queried in sophisticated ways, the audit itself becomes a matter of a query into a certain stream within a business application.

It is now possible to ensure that the data controller actually did what they claim that they did. To ensure this, ensuring the security and the integrity of the log file is essential. This can be done by third parties, by some corporate rules or organisational measures. The modern way would be to use blockchain technology and write transactions and meaningful processing on that secured ledger. This is what the SPECIAL project's research is also focused on.

3.4. Finding and formalising the relevant metadata

A policy-aware system is only as useful as the policy information recorded as metadata. The primary obstacle to the realisation of the presented vision is the lack of reusable and machine-readable context information of sufficient quality linked to the actual data being shared. Additionally, policies are often defined only generically on paper. Likewise the consent to use data is often only collected on paper, containing an entire wealth of information and thus not capable of being specific enough to decide on real use limitations or data handling directives. It is important to provide an infrastructure that allows human-readable policies to be tied to their machine-readable counterpart. This will preserve context information and usage limitations and transport them with the data.

It is a huge task to model and formalise policy information and to allow for the semantic lifting described above. The complexity and richness of the semantic lifting depends on the variety of metadata added to the payload data, and in how far taxonomies and ontologies are already available for that use case. Here there is a strong interest in standardisation as it will allow for a wide reuse of the vocabularies established by those specifications. This may include industry codes of conduct, but also policy snippets that may be combined in new ways.

For privacy use cases, some progress has been made by the creation of the PrimeLife[10] [PPL] language extending the XACML language [XAC] to allow for the use of credentials and interoperable role based access control. For security assertions and access control data, the semantics of SAML [SAM] can be used. The SPECIAL project has a focus on using ODRL [ODR] to express usage constraints and obligations attached to the collection of data. A good way to add metadata about quality of data is to use the W3C Provenance framework [PROV]. For financial services the work has not been done yet, but looking into the semantics of the eXtensible Business Reporting Language XBRL [XBR] may help. Chapter 2 shows ways to approach this pragmatically.

Once the taxonomy or ontology of a given policy is identified and modeled, a receiving entity or downstream controller can receive the schema and thus adapt their system quickly to the requirements for cooperation with the data controller. A large variety of data sources and types are expected. The more that digitisation advances in our lives, the more context information will be available to the system. All this information will feed into and further increase the data lake. It is therefore of utmost importance to keep in mind the variety management described in Chapter 4. So far, most information is used for profiling and marketing. The justified fear is that to know your customers means to be able to manipulate them to their detriment. The proposal here is to use more data to give users more control.

3.5. Context-aware reaction as a decisive leap to usability

A system that collects policy data at collection time knows about the constraints, the process under way, the purpose of collection and many more aspects. The current state of the art in compliance and information is marked by two extreme ends. On the one end, there are complex and lengthy information sheets provided to people. Many people have received pages of information from their bank that could later be used to excuse foreclosures. McDonald *et al.* [MC1] found that the average privacy policy has about 2500 words. The results of their online study of 749 Internet users, lead the authors to conclude that people were not able to reliably understand privacy practices [MC2].

One of the reasons for this lack of understanding on the users behalf is an issue that already surfaced in privacy policies and also in a financial context. An all encompassing privacy policy or notification makes it necessary to cover an entire operation with all available branches and possibilities in one document. This leads to a legal document with a “one size fits all” mentality that is not geared towards the user, but entirely dedicated to the avoidance of liability. Lawyers are used to such long documents, but those documents are a disaster for usability. More and more voices declare the failure of privacy policies to generate trust while acknowledging that they are effective from a liability avoidance perspective.

The SPECIAL system allows for a radical change to this approach by allowing the building-up of an agreement between users and data controllers over time, step-by-step. This will serve user confidence and trust and, with the appropriate add ons, protect against liability. It will achieve this via the use of Linked Data, where a graph is created in which nodes are

[10] See footnote 4

connected to each other. In a stateful system a given point has information about the nodes surrounding it. The system can use this information about the surrounding nodes to create a context-aware user experience. Instead of the entire policy, the system now knows which parts of the policy, constraints or obligations currently apply. The user and commercial partner interfaces do not need to display all information, but can rather concentrate on relaying relevant information with respect to the current interaction. This is helped via categorisation during the modeling phase of the policy at collection time. The categories can be reused to help the interface achieve a layered information interface.

Applied to the General Data Protection Regulation [GDPR], the system shows the relevant and required contextual information. While a general policy is just an informational document, the SPECIAL system now allows for direct interaction. By implementing a feedback channel, an affirmative action of the data subject can be gathered from the interface, leading to an agreement the GDPR calls "consent". As the system is policy aware, it can store the consent back into the data lake and into a transparency log or ledger. The user may encounter such requests for consent in different situations. For every single situation, experience shows that the actual request is rather simple: "We will use your login credentials to display your profile identity to others in the forum for mutual help. This forum is not public". Over time, the user may be contacted to agree to some "research for trends within the online forum". Additionally it would be highly desirable to allow the data subject access to this information within a dashboard. There is even research on how to cumulatively add this policy information. Villata and Gandon [VIL] propose a mechanism to combine permissions from various licenses into an overall set of permissions. It is therefore possible to accumulate a variety of context dependent permissions into a known and machine readable set of constraints, permissions and obligations that will govern the relation either to the data subject or the downstream controller. Applied to financial or other contexts, the interactions will generate a cumulative set of machine-readable agreements that can be automatically implemented by the system itself with demonstrable compliance.

Being dynamic will enable policies that apply at different levels of granularity to be tied to different parts of the data with different levels of sensitivity. This facilitates removing the need to fix the purpose for data collection upfront, by allowing both monitoring and control by the data subjects: by using machines to help humans overcome their cognitive weaknesses in big data contexts.

3.6. Tooling for the compliance system

The system described above needs good tooling. If data and metadata are recorded, this creates such a massive flow of data that big data technology is needed. Most big data tools today are not well suited for Linked Data, but after 3 years of development the Big Data Europe project [BDE] has created a platform capable of dealing with Linked Big Data.

BDE is first of all a normal big data platform using docker containers to virtualise data processing units and docker swarm to orchestrate those into a workflow. It created ready to use dockers for most of the Big Data toolchain from the Apache foundation. BDE calls this the Big Data Integrator (BDI)

On top of the BDI, tools for semantic operations have been created. Not all of them are production ready, but further development is advancing rapidly. BDE ended in 2017 but development of the tools continues. As a user of this technology, the SPECIAL project will also further the development of the semantic toolchain and help an already well established community. In the following, these semantic tools of the BDI are explained.

3.6.1. Tools for the semantic data lake

Challenges in the suggested system are comparable to the Big Data challenges, namely volume, velocity, variety and veracity. Volume and velocity are largely solved by components such as HDFS, Spark and Flink [BCO]. However, in the BDE and SPECIAL use cases, variety is the biggest challenge[11]. A lot of different data types and non-matching terms in different datasets are found. As discussed previously, the best way is to tackle the problem of variety is head on using Semantic Web technologies.

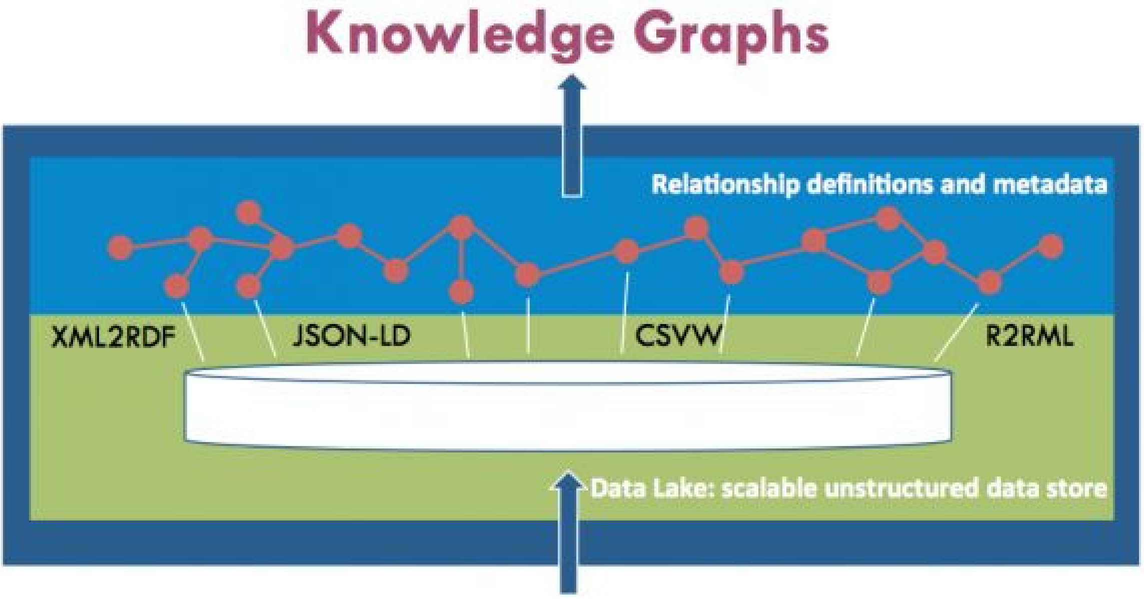
Knowledge Graphs
Relationship definitions and metadata
XML2RDF
JSON-LD
CSVW
R2RML
Data Lake: scalable unstructured data store

IMAGE: Chapter 3 Figure 4

Figure 4: The semantic data lake
Copyright: The Big Data Europe project https://www.big-data-europe.eu/semantics-2/ (accessed 20 Oct 2017)

3.6.2. Ontario or transforming ingestion?

BDE uses the "Ontology-based Architecture for Semantic Data Lakes" (Ontario) [AUER](fig.5). Data is stored in whatever format it arrives in, but it can be queried and analysed as if it were stored as RDF. Ontario has the option to accept SPARQL queries that are then re-written and run over one or more datasets in whatever the relevant query language may be. The results are combined before being returned as a single result set. SPECIAL also has the option of transforming ingestion of relevant data and metadata. In this case, the raw payload data is semantified by unique identifiers to make it addressable by annotations.

3.6.3. SANSA allows for semantic analysis

The SANSA stack(fig.5) [SAN] is a toolset that helps to streamline querying and reasoning over Linked Data. As we have seen previously, payload data and metadata annotations are stored in the system. Compliance is achieved through the filtering of data before the application of the intended processing. This means that the query/filter needs some level of sophistication to recognise the usable data sets. The SANSA Stack uses RDF data as its input and is able to perform analysis, e.g. querying, reasoning or applying machine learning over the Linked Data available in the platform.

[11] See Chapter 4

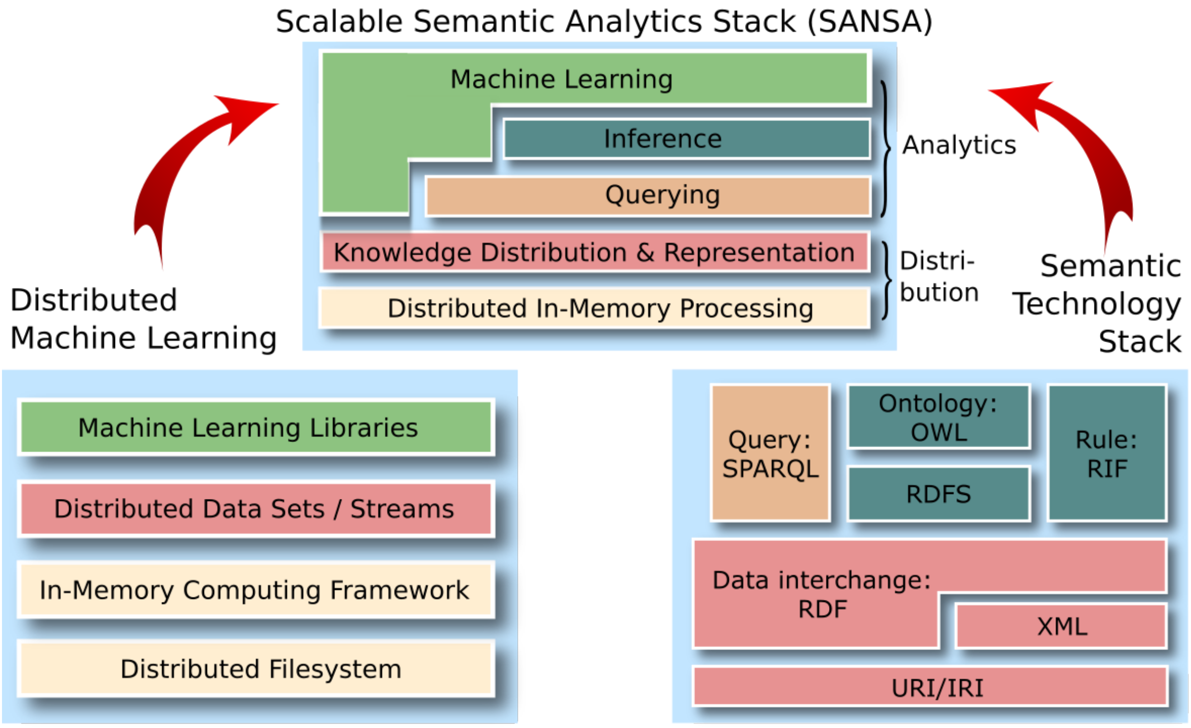
Scalable Semantic Analytics Stack (SANSA)
Machine Learning
Inference
Querying
Analytics
Knowledge Distribution & Representation
Distributed In-Memory Processing
Distri-bution
Distributed Machine Learning
Semantic Technology Stack
Machine Learning Libraries
Distributed Data Sets / Streams
In-Memory Computing Framework
Distributed Filesystem
Query: SPARQL
Ontology: OWL
RDFS
Rule: RIF
Data interchange: RDF
XML
URI/IRI

IMAGE: Chapter 3 Figure 5

Figure 5: The SANSA stack
Copyright: CC-BY Jens Lehmann http://sansa-stack.net (accessed 20 Oct 2017)

This allows exploration of not only the relation of payload data to metadata, but also the knowledge from the relations within the payload data or within the metadata. It helps construct the complex SPARQL queries needed to take account of permissions and constraints by providing algorithms that can be integrated into more complex and larger scale systems. Those will also typically need the parallelisation provided by BDE. It is still a challenge to parallelise SPARQL queries and reasoning. SANSA, although still under development, is well integrated into the BDE eco-system and provides docker-compose files as examples on github. This makes installation easy.

3.7. Recommendations

The advent of the GDPR will force companies to rethink their workflows. The time to think about adding a semantic layer for compliance is now. To do that:

1. Do the semantic lifting by giving IRIs to your payload data. The legacy systems can remain untouched as the IRI can point into it via some middleware.
2. Create the necessary taxonomies and ontologies according to Chapter 2 to allow for the appropriate semantics in the data annotations required for compliant data handling and checking.
3. Include respect for the metadata (annotations) provided into the contracts with business partners to insure respect for the data handling constraints.

3.8. Conclusion

The digitisation of our lives is progressing at high pace. The more aspects that turn digital, the more data we produce. The big data ecosystem depicts a situation where all the little streams from various ends form a big river of data: data that can be useful to fight diseases, but data that can also be used for manipulation. The possible options of dual use drive the call for more regulation. Data protection is only one field where the described system can be used for regulatory compliance. After the recent problems in the financial sector, new regulations about compliance and reporting were created. The SPECIAL system proposes the use of even more data to create a system of provable compliance. It also provides the basis for better integration of data subjects and users into the big data ecosystem by providing a dashboard and by organising a feedback channel. Additionally, it simplifies compliance management by private entities, and eases the task of compliance verification by data protection authorities. In essence, this brings trust to complex systems, thus enabling all big data stakeholders to benefit from data and data driven services.

3.9. Literature

[ANO] The W3C Web Annotation Working Group https://www.w3.org/annotation/ accessed 20 Oct 2017

[AND] Web Annotation Data Model, W3C Recommendation 23 February 2017, https://www.w3.org/TR/2017/REC-annotation-model-20170223/ accessed 20 Oct 2017

[AUER] Sören Auer et al. The BigDataEurope Platform – Supporting the Variety Dimension of Big Data, Web Engineering: 17th International Conference, ICWE 2017, Rome, Italy, June 5-8, 2017, Proceedings, 2017, 41–59

[BCO] Components supported by the Big Data Europe Platform https://www.big-data-europe.eu/bdi-components/ accessed 20 Oct 2017

[BDE] Big Data Europe https://www.big-data-europe.eu accessed 20 Oct 2017

[CSSe] Clubbing Seals: Exploring the Ecosystem of Third-party Security Seals, Tom Van Goethem , Frank Piessens , Wouter Joosen , Nick Nikiforakis, Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, Arizona, USA https://lirias.kuleuven.be/bitstream/123456789/471360/1/p918-vangoethem.pdf accessed 20 Oct 2017

[GDPR] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Official Journal of the European Union 59(L 119), May 2016, 1–88 ELI: http://data.europa.eu/eli/reg/2016/679/oj accessed 20 Oct 2017

[IRI] RFC3987 Internationalized Resource Identifiers https://tools.ietf.org/html/rfc3987

[LIBOR] https://en.wikipedia.org/wiki/Libor_scandal accessed 20 Oct 2017

[MC1] McDonald, Aleecia M, Cranor, Lorrie Faith: The cost of reading privacy policies, ISJLP 4, HeinOnline, 543, 2008 https://kb.osu.edu/dspace/bitstream/handle/1811/72839/ISJLP_V4N3_543.pdf accessed 20 Oct 2017

[MC2] McDonald, Aleecia M. and Reeder, Robert W. and Kelley, Patrick Gage and Cranor, Lorrie Faith, A Comparative Study of Online Privacy Policies and Formats. http://dblp.uni-trier.de/db/conf/pet/pets2009.html#McDonaldRKC09 in Privacy Enhancing Technologies, Volume 5672, Springer 2009 accessed 20 Oct 2017

[ODA] Tools for Semantic Lifting of Multiformat Budgetary Data, Deliverable D2.1 from Fighting Corruption with Fiscal Transparency, H2020 Project Number: 645833, http://openbudgets.eu/assets/deliverables/D2.1.pdf accessed 20 Oct 2017

[ODR] ODRL Vocabulary & Expression, W3C Working Draft 23 February 2017, https://www.w3.org/TR/vocab-odrl/ (accessed 20 Oct 2017) see also the Linked data profile https://www.w3.org/community/odrl/wiki/ODRL_Linked_Data_Profile accessed (20 Oct 2017) and the various notes linked from the WG page https://www.w3.org/2016/poe/wiki/Main_Page accessed 20 Oct 2017

[PAW] O. Seneviratne, L. Kagal, and T. Berners-Lee, "Policy-Aware Content Reuse on the Web," in ISWC 2009, 2009.

http://dig.csail.mit.edu/2009/Papers/ISWC/policy-aware-reuse/paper.pdf accessed 20 Oct 2017

[PPL] The PPL language, Primelife Deliverable D5.3.4 - Report on design and implementation http://primelife.ercim.eu/images/stories/deliverables/d5.3.4-report_on_design_and_implementation-public.pdf accessed 20 Oct 2017

[PROV] An Overview of the PROV Family of Documents, W3C Working Group Note 30 April 2013, http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/ accessed 20 Oct 2017

[RAY] Raymond, Eric S. (1999). The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary. O'Reilly Media. ISBN 1-56592-724-9.

[SAM] Security Assertion Markup Language (SAML) v2.0 (with further info) https://wiki.oasis-open.org/security/FrontPage accessed 20 Oct 2017

[SAN] SANSA - Scalable Semantic Analytics Stack, Open Source Algorithms for Distributed Data Processing for Large-scale RDF Knowledge Graphs, http://sansa-stack.net/ accessed 20 Oct 2017

[SPA] SPARQL Query Language for RDF, W3C Recommendation 21 March 2013, http://www.w3.org/TR/2013/REC-sparql11-query-20130321/ accessed 20 Oct 2017

[STRINT] A W3C/IAB workshop on Strengthening the Internet Against Pervasive Monitoring (STRINT), 28 February – 1 March 2014, London, https://www.w3.org/2014/strint/ accessed 20 Oct 2017

[SURV] Kenneth Lipartito, “The Economy of Surveillance,” *MPRA Paper*, vol. 21181, Mar. 2010. https://mpra.ub.uni-muenchen.de/21181/1/MPRA_paper_21181.pdf accessed 20 Oct 2017

[VIL] Villata, S., Gandon, F., Licenses compatibility and composition in the web of data, Proceedings of the Third International Conference on Consuming Linked Data-Volume 905, 2012, 124–135 https://hal.inria.fr/hal-01171125/document accessed 20 Oct 2017

[XAC] See eXtensible Access Control Markup Language (XACML), currently version 3, with various specifications https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml accessed 20 Oct 2017

[XBR] XBRL 2.1 https://specifications.xbrl.org/work-product-index-group-base-spec-base-spec.html accessed 20 Oct 2017