GDPR Transparency Requirements and Data Privacy Vocabularies

Eva Schlehahn, Rigo Wenning

GDPR transparency requirements and data privacy vocabularies

Eva Schlehahn[1] and Rigo Wenning[2]

[1] Unabhängiges Landeszentrum für Datenschutz (ULD, Independent Centre for Privacy Protection) Schleswig-Holstein, Kiel, Germany
[2] Word Wide Web Consortium/European Research Consortium for Informatics and Mathematics (W3C/ERCIM), Sophia Antipolis, France

Abstract. This tutorial introduced participants to the transparency requirements of the General Data Protection Regulation (GDPR).[35] Therein, it was explored together with the attendees whether technical specifications can be valuable to support transparency in favour of a data subject whose personal information is being processed. In the context of the discussions, past and present international efforts were examined that focus on data privacy vocabularies and taxonomies as basis work to enable effective enforcement of data handling policies. One example of a current undertaking in this area is the W3C Data Privacy Vocabularies and Controls Community Group (DPVCG) which aims at developing a taxonomy of privacy terms aligned to the GDPR, which encompasses personal data categories, processing purposes, events of disclosures, consent, and processing operations. During the tutorial session, the potential of such efforts was discussed among the participants, allowing for conclusions about the need to re-align and update past research in this area to the General Data Protection Regulation.

Keywords: General Data Protection regulation, EU law, transparency, data privacy vocabularies, technical specifications supporting GDPR compliance

1 Introduction

With the increasing digitization, the growing success of IoT devices on the market, and the incremental deployment of Big Data analysis of customer behaviour, it is apparent that ICT services and systems are by now widely used to link various types of information, recognize patterns and correlations, and to assess risks or chances on the basis of statistical insights. Data subjects whose personal data is being collected and processed are usually not aware of the scope and consequences of such assessments.

This leaves them exposed to the frequently opaque usage and commercialization of their personal information by data-driven companies. There is a significant lack of control by data subjects, since it is very difficult for individuals as end users of ICT services to obtain a clear picture how much data has been collected about them, from which sources, for which purposes, and with whom

it has been shared. This situation is compounded by deficiencies in terms of controller and processor controllability and accountability[3]. With the intention of changing the regime of opaqueness by data-driven businesses using long and for the layman customer incomprehensible privacy policies, the European legislators made an effort to give transparency and intelligible information of the data subject some increased weight with the GDPR. The purpose of this tutorial was to analyse the requirements of the GDPR in terms of transparency and information obligations of the controllers. The following sections reporting on the session content will also foray into the domains of ethics and technology to determine whether those can in addition to the legal demands provide some insight how transparency can be understood and realized.

Based on the three dimensions legal, ethics and technology, a more distinct insight can be gained what transparency actually means and what it needs to encompass to meet the threshold of GDPR compliance. In this context, limits and challenges to its realization are explored, taking into account the upcoming ePrivacy Regulation as well. Past and current approaches are explained on how privacy by design via technical specifications aimed to enhance data protection compliance. In the last section, conclusions are drawn that call for more work and research in this area.

2 GDPR Transparency Requirements

From European data protection law perspective, transparency is a core necessity to empower the data subject. This means knowledge and the means to hold controllers and processors of his or her personal data accountable. For instance, it has been relatively clearly stated in recital 43 of the GDPR by explicitly mentioning transparency as a tool to better *'balance out the power asymmetry between data subjects and organizations'*. In this context, the emphasis on the empowerment of the data subject is reinforced by the explicit requirement of transparent information, communication and modalities for the exercise of the rights of the data subject, Art. 12 (1) GDPR (bold highlights by the authors):

> *'1. The controller shall take appropriate measures to provide* ***any information*** *[. . .] relating to processing to the data subject in a* ***concise, transparent, intelligible and easily accessible form, using clear and plain language****, in particular for any information addressed specifically to a child. The information shall be provided in writing, or by other means, including, where appropriate, by electronic means. When requested by the data subject, the information may be provided orally, provided that the identity of the data subject is proven by other means.'*

[3] Cf. with regard not only to the GDPR, but also to the review of the ePrivacy Directive, see [12]:pages 3, 7, 10, and 11 as well as in [11], pages 4 f. The results of the public consultation and the Eurobarometer survey outcomes strongly indicate a lack of citizen's confidence of being able to control and protect own personal data online.

Beyond this obligation for the controller, the GDPR has a multitude of other sources also determining that the perspective of the data subject is the deciding factor whenever it seems doubtful whether transparent information was provided about a processing operation. This is a central difference to the domain of IT security, where the processing organisation, its business secrets and company assets are the paramount subjects of protection. Therefore, in the realm of personal data protection with its fundamental rights underpinning, the following questions present themselves whenever a personal data processing operation is intended:

- Which data shall be collected and processed, and to which extent?
- In which way shall the data be processed, using which means?
- For which purposes shall the data be processed, and by whom?
- Is a transfer to and/or storage at other parties/foreign countries foreseen?

A concise knowledge about the points above is a necessary precondition enabling data subjects to exercise their rights granted by the GDPR. Such rights include the right to be informed, to access, rectify, or erase one's own personal data. Consequently, it is essential to capture the complete life-cycle of personal data. This ranges from the moment of initial collection over all processing operations performed until the deletion of the information.

Yet, transparency is important not only for data subjects, but also for controllers, processors, and data protection supervisory authorities as well. For instance, data controllers usually desire to maintain internal knowledge and controllability of their own processing operations. This does not only benefit business efficiency needs, but is also important with regard to compliance efforts in order to properly guarantee the data subject‘s rights. Moreover, the aforementioned compliance efforts must be demonstrable by the controller (see Article 24 GDPR). By Articles 12-14 GDPR, the controller is obliged to fully comply with comprehensive transparency and information duties, which will in turn require the implementation of correlating technical and organizational measures. Besides the controllers, data processors have the obligation to assist the controller in compliance efforts while being bound to controller instructions and oversight. Furthermore, knowledge about the inner workings of a personal data processing operation is also crucial for data protection supervisory authorities to perform their supervision and audit duties.

In addition to the general transparency requirements in the GDPR, the conditions of valid consent play a significant role. According to Article 4 (11) in combination with Art. 7 GDPR, valid consent must be freely given, specific, informed and unambiguous. The statement of consent must be a clear affirmative action of the data subject, and given for one or more specific purposes. It is notable that the existence of valid consent must also be demonstrable by the controller of the processing operation. Consequently, transparency is a crucial element from many different perspectives. This includes fairness and lawfulness personal data processing. Below, a tabular overview is given. It shows the various articles and recitals of the General Data Protection regulation mentioning and requiring transparency:

Article	Title
5 (1) a.	Principles relation to processing of personal data
12	Transparent information, communication and modalities for the exercise of the rights of the data subject
13	Information to be provided where personal data are collected from the data subject
14	Information to be provided where personal data have not been obtained from the data subject
15	Right of access by the data subject
19	Notification obligation regarding rectification or erasure of personal data or restriction of processing
25	Data protection by design and default
30	Records of processing activities
32	Security of processing
33	Notification of a personal data breach to the supervisory authority
34	Communication of a personal data breach to the data subject
40	Codes of conduct
42	Certification
Transparency mentioned in Recitals:	
32, 39, 42, 58, 60, 61, 63, 74, 78, 84, 85, 86, 87, 90, 91, 100	

From an ethics perspective, transparency is a central requirement as well. Many ethical principles have evolved historically and are recognizable in values that are laid down e. g. in the European Convention on Human Rights[17], or the European Charter of Fundamental Rights[13]. According to the European Group of Ethics in Science and New Technologies, core values are e. g.:

- The dignity of the human being
- Freedom
- Respect for democracy, citizenship, participation and privacy
- Respect of autonomy and informed consent
- Justice
- Solidarity[21]

Already since 1985, ethical experts demand transparency in the context of ICT. Moor has introduced transparency as the crucial element to encounter the so-called '*invisibility factor*', which is inherent when information and communication technologies are being used. This '*invisibility*' has three dimensions:

- **Invisible abuse**, e. g. taking advantage by the use of ICT to adapt the program or to remove or alter confidential information.
- **Invisible programming values**, where a programmer (either consciously or even unconsciously) influences the workings of a software algorithm and embeds his own values or prejudices.

– **Invisible complex calculations**, which are '*beyond human comprehension*', the system as '*black box*' where no one can tell if the results generated are correct.[29]

The general purpose of transparency from an ethics perspective is making the underlying values of coded software (algorithms) recognizable. This aims not only at the processing itself, but also at the results generated by automated decision-making.

From a technical perspective within the ICT sector (esp. in the US domain), transparency was for quite a long time understood as the exact opposite, i. e. '*obfuscating*' all information about systems and processes – and not burden the user with it[20]. However, the GDPR's concept of transparency that wants to give users knowledge and control over the processing of their data and over the workings of the ICT systems, is increasingly recognized outside of Europe as well. Moreover, it gets increasingly recognized that transparency should aim not only at user interface (UI) aspects, but should encompass the whole ICT system including the system architecture and the data flows[23],[4],[18]. Typical high level examples how transparency can be supported by technical and organisational means are the verification of data sources (keeping track of data), the documentation of IT processes, logging of accesses & changes of the data stock, versioning of different prototypes/systems, documentation of testing, documentation of (related) contracts, or of consent (if applicable: given/refused/withdrawn), consent management possible from a mobile device, and the support of data subject's rights via technology, e.g. easy access to own personal data, possibilities of deletion or rectification of wrong, inaccurate or incomplete information[31][4]. From technical perspective, transparency generally aims at the predictability and auditability of the used IT, an aspect that is often also called provenance. This entails re-tracing and understanding past events, showing the technical and organisational setup, and avoiding or mitigating possible future issues. Usually, three different dimensions of provenance are being differentiated, namely the provenance of data, provenance of processes, and reasoning (or analytical) provenance. Provenance of data means that in all cases, the data flow is documented, while the documentation as well as the system itself can give insight about the source, type, quality and contextual allocation of the data, including the applicable data handling rules[6]. Examples how to realize data provenance with technical measures are sticky policies, differentiated data reliability scores, or automated deletion routines implemented. Provenance of processes means a proper documentation of the ICT components and analytic tools that are being used to process the data, which includes a documentation of the used analytical parameters as well to avoid such systems acting as kind of a black box producing non-retraceable results[32]. Finally, reasoning or analytical provenance is strongly related to the results of analytic processes. Here, transparency shows how analytics systems have been used to generate a certain output. In contrast

[4] With an exemplary list of transparency-enhancing technical and organizational measures referenced in the handbook of the Standard Data Protection Model recommended for use in Germany

to the process provenance, this aspect includes also the human or organizational factors around the ICT usage. Examples of measures to support reasoning provenance are the technical support of information and access rights in favour of data subjects, or the auditability of the processing operations (e.g. by human readable policies)[22]. All three perspectives – legal, ethical, and technical – have one thing in common: The requirement to involve stakeholders. This includes not only data subjects but also addresses controllers and processors. They all need to have relevant and sufficient information in order to understand the respective data processing operations. Only looking at GDPR-obligations for transparency might fall short ethically, if a holistic approach is the goal. In this case, it may be desirable to extend the requirements and understanding also to the risks involved and the decisions based on the results of the processing. Consequently, transparency could be understood as the property that all data processing – meaning all operations on data including the legal, technical, and organizational setting – and the correlating decisions based on the results can be understood and reconstructed at any time. Such an understanding of transparency entails a full capture of:

- the types of data involved,
- their source and quality/reliability
- processing purposes,
- circumstances of the processing,
- used systems and processing operations,
- the generated results,
- lawfulness and
- the related legal responsibilities (accountability)[28][5].

However, such kind of transparency is hard to formalize and measure so far. Fully comprehensive concepts are not yet state of the art. Current implementation approaches typically do not only concern the IT system and technical means alone. Rather, a comprehensive approach in consideration of legal, ethical, technical, and organizational/business expertise seems advisable to avoid discrepancies and to enable synergy effects. The fact that there is no 'universal' solution available should be recognized. Transparency solutions must therefore always be developed dependent on a careful assessment of context, individual case, and the processing purposes, means and foreseen execution. Only with such an earnest approach, the verifiability of data processing operations can be attempted. This goes beyond, but encompasses all transparency requirements of the GDPR in order to achieve coherently formulated functional requirements for automated processing systems.

The upcoming ePrivacy Regulation (ePR) contains interesting transparency requirements. As of now, the application scope includes electronic communicationsdata, meta- and content data. This extends the application scope of the current ePrivacy Directive that is still in force. Concerned with the new regulation

[5] Meis et al. constructed a set of requirements for a transparency focused ontology on the basis of the ISO/IEC 29100:2011standard, OECD principles, and the US fair information practices (FIPs), and which already entails some of these aspects.

will be all OTT[6], ECS[7] and software providers permitting electronic communications, including the retrieval and presentation of information on the Internet. This includes a lot of different apps such as IoT devices and many other things. In the commission proposal version, Art. 9 of the draft-ePR explicitly refers to the GDPR for consent requirements, which includes the correlating information and transparency obligations of the controller towards the data subject. In terms of transparency, relevant changes were recently hotly debated with regard to Art. 10 of the European Parliament version[34]. This article obliged hardware and software providers to ensure privacy by default, including the possibility of users to set their own privacy settings. However, in the later EU Council version, this article was deleted. Therefore, it is yet highly unclear when the Trilogue process of the ePrivacy Regulation will achieve a compromise result and how it will look like in the end.

3 Data Protection focus on technical specifications

In this section, some examples for basic approaches to GDPR-aligned technical specifications are given. Based on the transparency requirements explained above, the minimum core model for personal data processing policies should usually entail the data categories, processing operation and purpose, storage (retention time and storage location), and the recipients of the data to enable a coherent definition of a data usage policy. Such a core model is visualized below:

Going more into detail for the GDPR-aligned technical specifications, categories of personal data could for example entail differentiations like master record data, location and movement data, call records, communication metadata, and log file data. Moreover, special categories of personal in the sense of Art. 9 GDPR should be taken into account. This concerns personal data related to racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health, and data concerning a natural person's sex life or sexual orientation.

Beyond the categories of the data, technical specifications should support the documentation of processing purpose(s) and the correlating legal ground. If consent is determined as the applicable legal basis for processing, a link should be enclosed to the actual consent agreement itself, so the exact wording can be reviewed if needed. Furthermore, a versioning of consent agreements and the capture of the current status could be technically supported, e.g. by attaching pre-designed labels, such as given (if yes, specify whether explicit or implicit), pending / withheld, withdrawn, referring to the personal data of a minor, or referring to the personal data of a disabled person in need of specific accessibility provisions to manage consent. These are of course only initial ideas which could be further developed depending on context and need.

[6] OTT (Over The Top Services) are communication systems over data networks, e.g. skype

[7] Electronic Communication Services

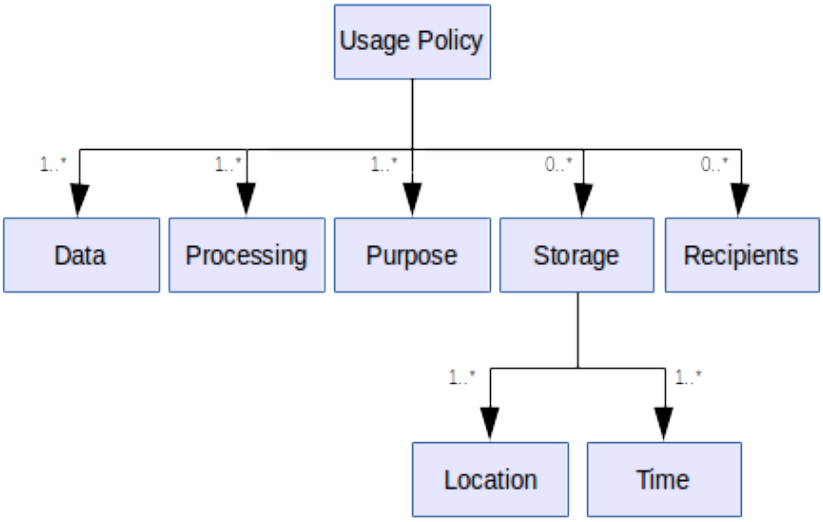

Fig. 1. The SPECIAL Core personal data processing policy model.

Technical specifications based on GDPR terms should enable a documentation of the involved data controller(s) and processor(s), as well as storage location and cross-border data transfers. For the latter, a specification of the involved foreign countries is also needed since this has significant impact on the legal assessment of the lawfulness of a processing operation. In this context, it is advisable to capture the location of the data centre where the processing and storage occurs, including the location of the controller establishment. For the latter, it is relevant to know whether the data transfer occurs only within the European Union, or to a third country with a legal basis for compliance acc. to Art. 44 et seq. GDPR (treating them as 'EULike'). Examples for such a legal ground for third country transfer could be an adequacy decision by the European Commission, appropriate safeguards, or existing Binding Corporate Rules (BCR). Where possible, a link should be provided that points towards any source documenting the respectively applicable legal document, e. g. to the Commission's adequacy decision or the BCR text. However, there might also be cases where data transfers to other third countries are foreseen for which any of the legal basis of Art 44 et seq. GDPR are not, or not anymore applicable. This could for instance, be the case if a court declares a BCR invalid, or the Commission changes an adequacy decision. To this end, it seems advisable to use country codes (e.g. according to TLD, ISO 3166), which allow for an easy later adaption of a policy in case of legal changes. Furthermore, in terms of data handling policies, it deems sensible to also incorporate labels that can express rules excluding data transfers to some jurisdictions (e.g. 'notUS', 'notUK').

Regarding the storage periods, specifications for certain deletion times and correlating actions required can be defined. Some rough examples are:

- delete-by_ or delete-x-date_month_after ¡event¿,
- no-retention (no storage beyond using once),
- stated purpose (storage only until purpose has been fulfilled),
- legal-requirement (storage period defined by a law requiring it),
- business practices (requires a deletion concept of controller),
- or indefinitely (e. g. for really anonymized data, and public archives).

Last, but not least, technical specifications should enable enforcing rules how to handle the data. For instance, defining the user or access activity that is being allowed or even determined for the personal data within an ICT system, such as read-only, write, rectify, disclose, deletion, anonymize, pseudonymize, or encrypt. Furthermore, notification obligations of the controller towards the data subject could be captured as well under certain preconditions, eventually with predefined action time, e.g. reminder of consent withdrawal right, or communication in case of a data breach.

4 Privacy enhancing technologies: The evolution of technical approaches for transparency and controls

In her PhD Thesis, Aleecia McDonald had found that most privacy policies are written beyond most adults reading comprehension level[26]. In another study, she showed that the time of reading all those natural language privacy policies would take an average person with average viewing habits around an average 201 hours of reading privacy policies per year[27]. For work, this would mean an average of 25 working days only to read privacy policies. Additionally, current privacy policies have a 'take it or leave it' approach. Data subjects can read the privacy policy, but mostly they cannot change any of the data collection. We see this change now with the advent of GDPR. But the interfaces are mostly very crude, sometimes culminating at an opt-in/out of more than hundred trackers. Under such circumstances, data self-determination can be exercised only for a few selected sites. The nominal transparency by privacy policies, carefully crafted by legally savvy people remains nominal. Aleecia McDonald had the merit to scientifically prove an assumption that was made very early on in the development of privacy enhancing technologies (PETs): Because the bandwidth of humans is very limited, the computer should help humans to better understand their situation.

4.1 From PICS to P3P

The quest for more transparency on the Web started very early, already in 1995. One of the very early and very successful adopters of web technologies was the porn industry. They were the first to make money with their content. At some

point, the porn content was spilling over to people who did not want to see it. This was also an issue of self-determination. And there was a technical challenge. People wanted to know before the actual content was downloaded and displayed to avoid bad surprises. This included filtering content for children. By labelling the content with a rating, a filter would be able to recognize the labels and react or block based on the rating. This was opposed to several governmental initiatives that wanted to install central filters on the Web to filter out so called 'illegal and harmful content'. The PICS[24] filtering would not have given such huge power to a single node on the Web. Freedom of information, which includes freedom to receive information, would be preserved. Not a central government institution would decide what people could see, but every individual or family or school could define a filter based on the ratings and their cultural preference. The plan to solve issues of self-determination with labelling came back in several iterations over the past two decades. And it became more sophisticated with each iteration. From a technology point of view, PICS was a very simple tool. There was no XML and no RDF yet in 1996. There was a very simple syntax that transported labels via an HTTP response header. The labels applied to the URI that the browser had requested. The labels themselves were not specified by W3C as the organisation did not feel competent in this area. But a real good labelling system was never found. The Internet Content Rating Association (ICRA) created the predominant labelling scheme. The issue with the system was that normal authors did not have an incentive to label their pages unless they were of a certain type. In 1999, all porn sides carried labels because they wanted to be found. Others did not want to label their pages because there was a (justified) fear that governmental nodes on the Web would then filter all content thus stifling freedom of speech. But most pages were not labelled meaning PICS remained of limited usefulness. Because of the criticism, the lack of incentives, and because the labelling was complex and coarse at the same time, PICS did not take off. The browsers finally removed support for the filtering and ICRA went out of business in 2010.

In 1997, when there was still a lot of enthusiasm about PICS, the idea came up to also use such a transparency scheme for privacy and data protection. In fact, people found out that a lot of http protocol chatter was stored in log files and used for profile building. People did not realise that such profiling was happening as the browser was totally opaque about it. The US Senate threatened the advertisement industry with legislative action. Industry feared that such legislation would damage their revenue model. It was better to improve the user experience than to continue to provide a picture perfect example on why legislation was needed. This created a discussion within W3C about technical remedies to the opaqueness of the HTTP protocol. Major vendors like Microsoft joined a W3C Working Group to make a tool that could replace legal provisions. A combination of researchers and industry started to create the Platform for Privacy Preferences (P3P)[14]. The idea was similar to the one in PICS: Servers would announce what data they collect and what they do with it. To express the data collection and usage, P3P created the P3P vocabulary. P3P came before

XML, so the file was in a similar, but not quite compliant data format with angle brackets. This allows a service to encode their practices in a machine readable file and store it on the server. For the client side, the Group created a vocabulary to express preferences and match those preferences against the P3P policy document found on the server for the resource requested via HTTP[37]. The idea was that APPEL[36] would enable privacy advocates to write preferences for consumers. But APPEL was finally taken off the standards track as it had technical and mathematical problems[1]. The development concentrated on the policy language and client side preferences. The preference exchange language was downgraded to a secondary goal.

4.2 The success and decline of P3P

Everyone in the technical community found the idea behind P3P compelling. The vocabulary was well respected and a lot of sites started to implement P3P on the server side. At the same time, browsers were complaining that it was too complex to implement P3P on the client side. In a last minute chaotic action, the P3P WG accepted a proposal from Microsoft to introduce so called 'compact policies'. The verbose and complex XML-like file expressing the P3P policy was replaced by a very coarse representation of abbreviated policy tokens transported by a HTTP header.

The power of P3P became shortly visible when Microsoft[8] announced that the new Internet Explorer (IE) 6.0 would only allow third parties to set a cookie if this third party would send P3P compact policy tokens about data they collect and about the associated purposes. Within a few month, P3P compact tokens where present in a majority of HTTP driven exchanges. But they were very different from what the Working Group had expected. Someone somewhere on a developer platform created a 'make-IE-6 –happy' string with P3P compact tokens in them. This would fool IE 6 to believe that the service would do the thing it announced and allow the third party cookie to be set. Within short time, a very large proportion of the tokens found were those 'make-IE-6-happy' tokens.Of course those were obvious lies using technical means. Critics of P3P had always said that P3P itself had no enforcement. And indeed, P3P just made declarations easier and machine readable to help humans assess the situation, remove the opacity. It was never meant to be an enforcement tool. It removed opacity and relied on the fact that the legal system would sanction lies and deceptive behaviour. And there were cases when this was applied successfully. Especially when people really tried to implement P3P and exposed in P3P what they were really doing. This way someone found out that the US Drug addiction online service for addicts had a tracking cookie. The cookie was subsequently retired.

The US Senate finally dropped the privacy legislation and the European data protection authorities preferred direct legal action within an orderly administrative procedure. There was no incentive for the industry anymore to invest into

[8] Microsoft had considerable market power on the Web in 2002.

privacy enhancing technologies. Paying for lobbyists was more cost effective and allowed proven technology to remain unchanged.

What remained from P3P was its vocabulary, especially the ¡STATEMENT¿ section. This section is still cited and influences the way people express policies in the area of data protection.

4.3 From the Web to the Enterprise world: Prime & PrimeLife

In Europe, research on the topic continued. From 2004-2008, the PRIME project[9] explored new ways. The assumption was that privacy policies can be turned into privacy rules. The system would then automatically enforce those rules. The principles of the project were the following:

1. Design starting from maximum privacy.
2. System usage governed by explicit privacy rules.
3. Privacy rules must be enforced, not just stated.
4. Trustworthy privacy enforcement.
5. Easy and intuitive abstractions of privacy for users.
6. An integrated approach to privacy.
7. Privacy integrated with applications[9].

The system had several use cases that were implemented using the technology created by the project. The project was very research oriented and decided for maximum privacy where ever possible. This included already anonymous credentials[8] and the machine-assisted management of pseudonyms. As it was very research driven, the assumption was that the client would issue preferences and the server would acknowledge receipt of those preferences or rules and act accordingly. But this did not work out in practice. In fact, most systems are designed for a limited variety of options. The user can chose between those options. If the user or data subject comes up with new preferences that are not yet implemented as a workflow on the server side, the system will simply fail. The PRIME project did use RDF[5] on both sides of the equation which cured some of the difficulties P3P had concerning the matching of preferences to policies. But it was too early because the use of RDF or Linked Data or graph data was not yet very widespread. The project decided deliberately against taking into account legacy enterprise databases and went for the pure research by requiring all systems to be RDF in order to work. The main argument was that integrating those legacy systems with their legacy interfaces would allow a malicious actor to circumvent the enforcement engine of the PRIME system. Given those practical constraints, the PRIME project produced good scientific results, but its practical relevance remained marginal. It was not really usable for real world systems unless someone would create a brand new system from scratch. PRIME had improved the understanding of the difficulties with transport anonymity and also advanced considerably the state of the art concerning the sticky policy paradigm.

[9] `https://cordis.europa.eu/project/rcn/71383/factsheet/en.`

Drawing conclusions from this initial experience with a privacy-enabled data management system, most of the partners of the PRIME project created the PrimeLife project[10] to now concentrate on the challenges for such a privacy-enabled system in relation to real world systems. When the PrimeLife proposal was created, the predominant data format was XML[3][10] and the predominant software architecture was Web Services[11]. PrimeLife concentrated on those technologies and reduced the dependency on RDF. The attacking model was changed. An assumption was made that a company wanted to do the right thing. In fact, rights management systems are very hard to enforce once the entire computing resources are in the hand of the attacker. Consequently, PrimeLife had a security boundary between the client side and the server side, and a focus on data usage control on the service side. This allowed efforts to push boundaries on both sides. PrimeLife continued integrating ways to achieve results and checks without consuming personally identifiable information while further developing a system to create data value chains across company borders. One special merit of PrimeLife was the invention of the term '*downstream data controller*'. This terminology allowed researchers to better repartition and assess the respective responsibilities between data controllers and responsibilities further down in the data value chain. It allowed clearing the fog in discussions where people partly did not realise that they were talking about the same thing. Or, that they believed they were talking about the same person, while meaning another one. The term '*downstream data controller*' greatly reduced the confusion. For PRIME, the sticky policy paradigm was rather easy to implement as RDF with its URIs[7][16][12] on every triple had a natural way of addressing a specific data packet or data record. In fact, this is like a sentence in English language. To make a policy sticky, it has to be attached to the data record it applies to. Within RDF data, every object has a URI. So it is sufficient to create a new triple with a policy that points to the data record it wants to apply to. The URI used is a world unique identifier. This means wherever the data travels, the relation to the policy remains intact. This allows data value chains across company borders to honour the policies expressed and even impose the respective rules to subsequent downstream data controllers. In PrimeLife, because web services were used, the stickiness of the policy made a rather complex system necessary. All was using standard data formats, namely XACML[30] and SAML[2] to create, manage, transport and execute policy statements along a given data value chain. While this was very pragmatic at the time, technology has moved on and different things have to be used today.

PrimeLife had a rather holistic view, taking into account the data life-cycle and providing new ways to manage access control in a way that was much closer to the needs for business, e.g. depending of the role of the person wanting to access certain information. PrimeLife had an entire work package concentrating on user interface issues. Data self-determination, if taken seriously, requires the

[10] `http://primelife.ercim.eu/`
[11] `https://www.w3.org/2002/ws/` accessed 2019-01-21
[12] RDF uses IRIs to identify objects

data subject to understand what is happening. This is not only an issue for transparency and the provision of information from the server/service side. It is also an issue of cognitive limitations by humans confronted with the huge wealth of information transitioning through today's communication systems. PrimeLife experimented with icons and logos and user interfaces and did usability testing in a lab environment[19]. One of the lasting outcomes was a logo: Users were tasked to find out about a privacy tool PrimeLife had developed. It was collecting data about who is collecting what to identify trackers and show them to the user[13]. Users had the task to find out who collects what about them. The challenge was, whether the users would find the tool and click on the right logo. All kinds of logos were tested. But only with an icon with footsteps on it, users found the tool easily. PrimeLife was also very successful in the scientific field testing out role based access control and a new type of group based social networking. It created a new policy language that was able to express all the metadata in standard languages and extended XACML to carry more metadata.

After PrimeLife, there have been further research and development efforts focused on providing better transparency for users of digital services. For example, the research work done in the A4Cloud project[14] built upon the PrimeLife Policy Language (PPL) and extended the identified requirements in order to create a proof of concept for an accountability policy language (A-PPL) for cloud computing contexts. This policy language addresses data handling rules corresponding to cloud provider's privacy policies, data access and usage control rules, retention period and storage location rules, logging, notification, reporting and auditability[33]. However, those efforts need further work since they have been made before the reform of the European data protection framework with its enhanced transparency requirements in favour of data subjects.

4.4 SPECIAL: From enterprise systems to big data and knowledge graphs

Enterprise systems have evolved a lot in the meantime. The European Commission has successfully implemented their public sector information strategy.[15] The aim is to provide public information that can be combined with private sector information to form new innovative services and products. The Big Data Europe project (BDE)[16] created an open source platform ready to use for everyone to provide an easy tool for processing analysis of information for the public sector in all seven societal challenges put forward by the Commission: Health, Food, Energy, Transport, Climate, Social sciences and Security. The project quickly found a first challenge. Most of the data found, personal or not, was of high

[13] The privacy dashboard was a Firefox extension that stored all data from the HTTP chatter into a local database and was able to show the tracking to the user. See `http://primelife.ercim.eu/results/opensource/76-dashboard`.

[14] `http://a4cloud.eu/`.

[15] See `https://www.w3.org/2013/share-psi/` for more information and pointers.

[16] `https://www.big-data-europe.eu`.

heterogeneity. Within the same societal challenge a wealth of databases in silos was found. The variety issue was solved by using RDF and Linked data to join data from a high variety of sources. To do so, BDE developed a method to semantify the data streams coming into the big data platform. They called it semantic lifting. With an all semantic data lake, we are back in a situation where the insights of PRIME can be used to make policies sticky, notably by just adding policy information to the knowledge graph created via the semantification and other transformations. For the analysis, the parallelization of the processing was not known so far to the normal inference engines. BDE started work on the SANSA[25] stack now further developed by University of Bonn. It allows accomplishing even more complex policy data processing and inference in an acceptable amount of time. Now this engine was again usable to come up with a privacy enhancing system for big data. The SPECIAL project[17] uses this system to produce a new tool for sticky policy and data usage control within a big data environment. Special, like the initial PRIME project, uses the Linked data properties to annotate data. Through semantification, all data has a URI and can thus be an object of a Linked data statement. For internal purposes and for performance, the Linked data can be transformed into something else to better fit the legacy systems. But the data must retain Linked data properties, especially being linked to a policy statement. If data from a system is given to commercial partners in a data value chain, of course with the consent of the data subject, the RDF Linked data platform plays the role of a transport format. After the technical challenges of the past projects, the SPECIAL project found rather social challenges in context of the technical issues. Of course, a deep understanding of the Linked data world is needed to design policy annotated workflows. While the technology stack has not yet arrived in many production systems, it is a rather mature area with a high potential for new use cases. We are now moving from the question of how processing is organised to the question about the semantics for the actual policy annotations. P3P has given some direction with its statement vocabulary. But this was really advertisement driven and is not sufficient for the very generic privacy transparency and data management tool that will be created by SPECIAL.

This is why W3C organised a Workshop on Privacy and Linked Data in April 2018[15]. The workshop assumed that most services today, lack the tools to be good citizens of the Web. Which is related, but not limited, to the work on permissions and on tracking protection. Because those permissions and tracking signals carry policy data, the systems have to react upon those signals. To react in a complex distributed system, the signals have to be understood by more than one implementer. The challenge is to identify the areas where such signals are needed for privacy or compliance and to make those signals interoperable. The Workshop concluded that more work on Data Privacy Vocabularies is needed. The Workshop participants decided to initiate a W3C Data Privacy Vocabularies and Controls Community Group (DPVCG).1 The DPVCG will develop

[17] Scalable and policy-aware linked data architecture for privacy, transparency and compliance (SPECIAL), `https://www.specialprivacy.eu`.

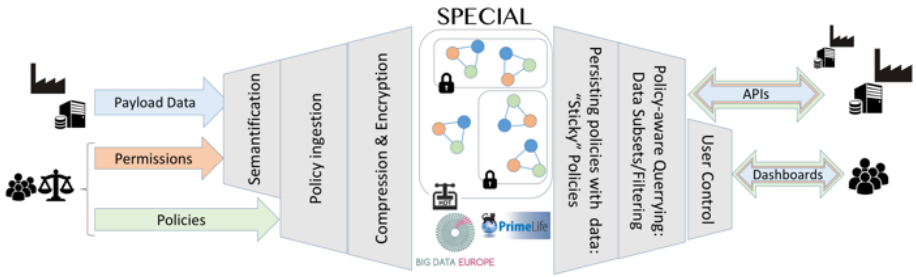

Fig. 2. A birds-eye view on the Special data flow

a taxonomy of privacy terms, which include in particular terms from the new European General Data Protection Regulation (GDPR), such as a taxonomy of personal data as well as a classification of purposes (i.e., purposes for data collection), and events of disclosures, consent, and processing such personal data. Everybody is welcome to join the effort.

5 Tutorial outcomes and conclusions

The main objective of the tutorial session was to introduce participants to the transparency requirements of the GDPR and to delve deeper into the possibilities of technical means supporting their realization. This was achieved by an introductory presentation, and discussions during and after this presentation that took place among all attendees. These discussions often evolved around the significance of earlier attempts of creating technical specifications and vocabularies for privacy terms, their usefulness for future endeavours in this field, and initial ideas what could be captured in such vocabularies and taxonomies to enable meaningful and enforceable data handling policies. The discussions showed that the work of the SPECIAL project as well as of the W3C Data Privacy Vocabularies and Controls Community Group are very important first steps into the right direction, while recommendations were made by workshop participants to stick very close to the GDPR and to avoid the pitfalls of former efforts in the area of standardisation. However, the discussions showed that especially in the context of big data applications, further research as well as real development work will be needed to inch closer to more GDPR-aligned processing taxonomies that can be adopted by businesses as well.

Acknowledgments. Supported by the European Union's Horizon 2020 research and innovation programme under grant 731601.

References

1. *W3C Workshop on the long term Future of P3P and Enterprise Privacy Languages*, 2003. W3C. URL https://www.w3.org/2003/p3p-ws/.
2. Security assertion markup language (saml) v2.0. Technical report, Mar. 2005. URL https://www.oasis-open.org/standards#samlv2.0;https://wiki.oasis-open.org/security/FrontPage#SAML_V2.0_Standard.
3. Extensible markup language (xml) 1.0 (5. edition). Technical report, Nov. 2008. URL http://www.w3.org/TR/2008/REC-xml-20081126/.
4. *Engineering Privacy by Design*, 2011.
5. Rdf 1.1 primer. Technical report, June 2014. URL http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/.
6. G. A. *Data Provenance*. Springer, 2009.
7. T. Berners-Lee, R. T. Fielding, and L. Masinter. Uniform resource identifier (URI): Generic syntax. Technical report, 2005. URL http://www.ietf.org/rfc/rfc3986.txt.
8. J. Camenisch and A. Lysyanskaya. Efficient non-transferable anonymous multi-show credential system with optional anonymity revocation. In B. Pfitzmann, editor, *Advances in Cryptology — EUROCRYPT 2001*, volume 2045, pages 93–118. Springer Verlag, 2001.
9. J. Camenisch, R. Leenes, and D. Sommer, editors. *PRIME – Privacy and Identity Management for Europe*, volume 6545 of *Lecture Notes in Computer Science*. Springer Berlin, 2011.
10. C. Collins. A brief history of xml. Mar. 2008. URL https://ccollins.wordpress.com/2008/03/03/a-brief-history-of-xml/.
11. E. Commission. Flash eurobarometer 443: e-privacy. Technical report, Dec. 2016. URL http://data.europa.eu/euodp/en/data/dataset/S2124_443_ENG.
12. E. Commission. Summary report on the public consultation on the evaluation and review of the eprivacy directive. Technical report, Aug. 2016. URL https://ec.europa.eu/digital-single-market/en/news/summary-report-public-consultation-evaluation-and-review-eprivacy-directive.
13. E. Council, E. Parliament, and E. Commission. *Charter of Fundamental Rights of the European Union.*, pages 389–403. Number 83 in Official Journal of the European Union C. European Union, 03 2010. URL http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2010:083:0389:0403:en:PDF.
14. L. F. Cranor. *Web Privacy with P3P*. O'Reilly & Associates, Inc., 09 2002. ISBN 0-596-00371-4.
15. 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/.
16. M. Duerst and M. Suignard. Internationalized resource identifiers (iris). Technical Report 3987, Jan. 2005. URL http://www.ietf.org/rfc/rfc3987.txt.

17. ECHR2010. Convention for the protection of human rights and fundamental freedoms as amended by protocol no. 11 and no. 14. http://conventions.coe.int/treaty/en/Treaties/Html/005.htm, jun 2010.
18. B. Goodman and S. Flaxman. Eu regulations on algorithmic decision-making and a "right to explanation. *AI Magazine*, 38(3), 2017.
19. L.-E. Holtz, K. Nocun, and M. Hansen. Towards displaying privacy information with icons. In S. Fischer-Hübner, P. Duquenoy, M. Hansen, R. Leenes, and G. Zhang, editors, *Privacy and Identity Management for Life*, pages 338–348. Springer, 2011.
20. F. Inchauste. The dirtiest word in ux: Complexity, July 2010. URL http://uxmag.com/articles/the-dirtiest-word-in-ux-complexity.
21. J. Kinderlerer, P. Dabrock, H. Haker, H. Nys, and M. Salvi. *Opinion 26 - Ethics of information and communication technologies.* Publications Office of the European Union, Feb. 2012. ISBN 978-92-79-22734-9. doi: 10.2796/13541. URL http://bookshop.europa.eu/en/ethics-of-information-and-communication-technologies-pbNJAJ12026/.
22. N. e. a. Kodagoda. Using machine learning to infer reasoning provenance from user interaction log data: Based on the data/frame theory of sensemaking. *JCEDM Special Issue*, (11):1, 2017.
23. B.-J. Koops. *On Decision Transparency, or How to Enhance Data Protection after the Computational Turn*, pages 196–220. 2013.
24. T. Krauskopf, J. Miller, P. Resnick, and W. Treese. Pics label distribution label syntax and communication protocols. Technical report, Oct. 1996. URL https://www.w3.org/TR/REC-PICS-labels-961031.
25. J. Lehmann, G. Sejdiu, L. Bühmann, P. Westphal, C. Stadler, I. Ermilov, S. Bin, N. Chakraborty, M. Saleem, A.-C. N. Ngonga, and H. Jabeen. Distributed semantic analytics using the sansa stack. In *Proceedings of 16th International Semantic Web Conference - Resources Track (ISWC'2017)*, pages 147–155. Springer, 2017. URL http://svn.aksw.org/papers/2017/ISWC_SANSA_SoftwareFramework/public.pdf.
26. A. M. McDonald. *Footprints Near the Surf: Individual Privacy Decisions in Online Contexts.* PhD thesis, 2010. URL https://kilthub.figshare.com/articles/Footprints_Near_the_Surf_Individual_Privacy_Decisions_in_Online_Contexts/6717041.
27. A. M. McDonald and L. F. Cranor. The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society*, 4(3):543–568, 2008. URL http://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/isjlpsoc4&section=27;https://kb.osu.edu/dspace/bitstream/handle/1811/72839/ISJLP_V4N3_543.pdf.
28. R. Meis, R. Wirtz, and M. Heisel. *A Taxonomy of Requirements for the Privacy Goal Transparency*, pages 195–209. TrustBus, 2015.
29. J. H. MOOR. What is computer ethics? *Metaphilosophy*, 16(4):266–275, Oct. 1985. ISSN 1467-9973. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9973.1985.tb00173.x;http://web.cs.ucdavis.edu/~rogaway/classes/188/spring06/papers/moor.html.

30. T. Moses. extensible access control markup language (xacml) v2.0. Technical report, 2005. URL http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-core-spec-os.pdf.
31. C. I. D. P. of the Authorities. The standard data protection model. Technical report, 2016. URL https://www.datenschutzzentrum.de/uploads/sdm/SDM-Methodology_V1.0.pdf.
32. H. Pandit, D. O'Sullivan, and D. Lewis. Queryable provenance metadata for gdpr compliance. *Procedia Computer Science*, (137):262–268, 2018.
33. J. G.-A. H.-J. L. Posegga, J. Herrera-Joancomartí, E. Lupu, J. Posegga, A. Aldini, F. Martinelli, and N. Suri, editors. *A-PPL: An Accountability Policy Language*, 2015. DPM, Springer. ISBN 978-3-319-17015-2. URL https://link.springer.com/book/10.1007/978-3-319-17016-9;https://doi.org/10.1007/978-3-319-17016-9.
34. B. Sippel and E. Parliament. Report on the proposal for a regulation of the european parliament and of the council concerning the respect for private life and the protection of personal data in electronic communications and repealing directive 2002/58/ec (regulation on privacy and electronic communications), oct 2017. URL http://www.europarl.europa.eu/sides/getDoc.do?type=REPORT&mode=XML&reference=A8-2017-0324&language=EN.
35. E. Union. 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), May 2016.
36. W3C. A P3P preference exchange language 1.0 (APPEL1.0), 2002.
37. W3C. The platform for privacy preferences 1.1 (P3P1.1) specification, 2006.