6

Technical Components

Sotiris Koussouris[1] and Yury Glikman[2]

[1]Suite5, Cyprus
[2]Fraunhofer, Germany
Email: sotiris@suite5.eu; yury.glikman@fokus.fraunhofer.de

Abstract

This chapter covers a selection of the technologies coming out of those projects listed in Chapter 2, which collectively contribute to a citizen being able to safely share their personal data.

The four sections reflect the different general issues being faced and have attempted to reflect the objectives of the European Commission when these projects were granted their funding.

They include:

- data owners and subjects controlling their own data;
- preserving data privacy and data quality simultaneously;
- information delivery on privacy metrics and data content and value;
- data platforms.

Additionally, reference is made to several initiatives which are designed to support this overall process in making it possible for a citizen to share their data securely and fully under their own control.

6.1 Introduction

As an indication of what is being produced by the projects listed in Chapter 2, some of the technologies that they are producing are highlighted, along with references to some of the other supporting initiatives, working within this field.

This chapter covers the technologies that provide the basis for enabling a citizen to be able to share their personal data securely and in a way that they themselves have control. These are a combination of established technologies and those emerging from the raft of projects funded through the Horizon 2020 Research Framework Programme, principally in the ICT-13-2018-2019 Call "Supporting the emergence of data markets and the data economy",[1] which was described in Chapter 2, in relation to why we have focussed upon certain projects. These projects were amongst other things, "Supporting the emergence of data markets and the data economy".[2]

The chapter should be treated more as a "catalogue" to act has a guide as to what the projects covered in Chapter 2, "have been up to", rather than as a learned treatise. It is divided into four sections, which reflect those technologies signposted within the above-mentioned, which, in the words of the call, are:

- "The personal data platforms shall ensure respect of prevailing legislation and allow data subjects and data owners to remain in control of their data and its subsequent use".
- "Solutions should preserve utility for data analysis and allow for the management of privacy/utility trade-offs, metadata privacy, including query privacy".
- "Solutions should also develop privacy metrics that are easy to understand for data subjects and contribute to the economic value of data by allowing privacy-preserving integration of independently developed data sources".
- "Industrial data platforms shall enable and facilitate trusted and secure sharing and trading of proprietary/commercial data assets with automated and robust controls on compliance (including automated contracting) of legal rights and fair remuneration of data owners".

6.2 Data Owners and Subjects Controlling their Own Data

Over the last years, witnessing the advent of various data-driven platforms that exploited in a very effective, sufficient, and profitable manner data belonging to users, a huge societal movement was constructed calling for

[1] https://cordis.europa.eu/programme/id/H2020_ICT-13-2018-2019

[2] "Supporting the emergence of data markets and the data economy. Programme H2020," *CORDIS European Commission*. https://cordis.europa.eu/programme/id/H2020_ICT-13-2018-2019 (accessed Jul. 25, 2022).

new methods and tools, both from the technology as well as the legislative level, for allowing data owners and subjects to regain control of their data and their subsequent use and bring them into the position of the decision-maker to resolute how, when, for which cause, and by whom data shall be used.

In Section 6.2, we explore novel technological approaches that contribute towards this direction.

6.2.1 User personas

User personas have their origins in the marketing world and are used as another way to preserve the privacy of individuals. A persona is a mechanism by which insight from aggregated datasets (groups of individuals) is generated and shared with those who are interested (data seekers in this case).

As an example, we can think of a data seeker who can ask for a Persona on the daily physical activity of people living in Paris. The technical workflow to generate such a persona is as follows:

1. The data that is irrelevant to the query is dropped from the dataset.
2. The persona generator clusters the data using a mean shift algorithm and adds a new column to the dataset and assigns a number showing which cluster each data point is a part of.
3. For each column, mean, mode, standard deviation, and variance are calculated and put into a dictionary.
4. Similarly, a coefficient matrix is produced and stored in this dictionary as well.
5. Then the same statistics are produced for any categorical columns in the dataset.
6. This process is iterated for each cluster in the dataset.

These insights are essentially demographics relating to the statistical properties of the aggregated data. The resulting persona of our example provides insights on the average daily activity of people living in Paris broken down by age and area, the main data points that determine higher and lower levels of activity and how closely these data points are related. This means a data sharer can still get some value from their data without actually letting a data seeker see his/her data, thereby reducing the risk of them being identified.[3]

[3] https://www.datavaults.eu/material/results-and-documentation/

6.2.2 Direct anonymous attestation (DAA)

Direct anonymous attestation (DAA) is a technology based on group signatures for enhancing privacy and trust amongst transacting stakeholders.

From a technological perspective, DAA is a cryptographic protocol that allows a trusted platform module (TPM) to serve as a trust anchor for a host platform it is embedded in. To do so, the TPM chip creates attestations about the state of the host system, e.g., certifying the boot sequence the host is running on. These attestations convince a remote verifier that the platform it is communicating with is running on top of trusted hardware and using the correct software. A main design goal of DAA is that attestations are made in a privacy-preserving manner. That is, the verifier can check that attestations originate from a certified hardware token, but it does not learn anything about the identity of the TPM. Another important feature of DAA is that it supports user-controlled linkability which is steered by a base name "bsn." If a platform uses a fresh or empty base name, the resulting attestations cannot be linked, whereas repeated use of the same base name makes the transactions linkable. A DAA can be seen as a special variant of group signatures with a central issuer controlling membership to the group of certified TPMs, and TPMs are able to sign anonymously on behalf of the group. Instead of the opening capabilities provided in group signatures, DAA controls privacy through the use of base names and user-controlled linkability.

The privacy requirements that are captured by DAA are the ones documented in the ETSI TS 102 941 standard[4]: anonymity (ability of a user to use a resource without disclosing its identity), pseudonymity (ability of a user to use a resource without disclosing its identity while being accountable for that action), unlinkability (ability of a user to make multiple uses of resources without others being able to link them together), and unobservability (ability of a user to use a resource without others being able to observe that the resource is being used). The project DataVaults[5] provides a good example of all the above cases where DAA is used for enabling data owners to both authenticate their platforms in a privacy-preserving manner and also share their data in an anonymous way by leveraging group-based pseudonyms.

[4] "ETSI ICT Standards." https://www.etsi.org/standards#Pre-defined%20Collections (accessed Jul. 25, 2022).

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

6.2.3 Access control policies

One of the key points when it comes to controlling data is to solve the fundamental issue of how citizens can become the actual owners of their data setting the conditions to be fulfilled, in a transparent and automatic way, for external data seekers. Currently, the perception of the citizens about being in charge of controlling with whom their data is shared is very low, making the decision to share not as easy as the modern smart cities would need. The GDPR, as a means to protect their data against abusers, is quite new and barely known by regular people. Moreover, there are not enough incentives (for example, monetisation) to promote data sharing. Nevertheless, more and more data is being generated with a real possibility of using them to improve lots of services.

One possible approach, as the one employed by the DataVaults project, builds on the premise to change this feeling of the citizens based on letting data owners decide what, how much, and in which manner they share their data, guarantee their privacy and security, and retrieve a fair share of the value their data generates. In such a platform, a personal data cataloguing system is presented to the data owner with its possible associated access policies, to allow the discovery of this kind of data and, at the same time, allow for control that only those who have been specified by the owner have access to them, under the terms previously established. This access control policies system opens the doors to new possibilities in the exploitation of personal data and even proposes a starting point to establish usage control policies to personal data. The KRAKEN project[6] works also on a similar direction offering as one of its main pillars the self-sovereign identity (SSI) solution which essentially is a user-centric access control to data, where the data owner controls their data. The KRAKEN approach is being piloted in two high impact domains, health and education, and contributes to data spaces by providing tools and solutions that can preserve the privacy and assure security, trustability, data integrity, and confidentiality, when data are shared between different stakeholders in a data space, and even between different data spaces.

From an architectural viewpoint, an access control service is composed by an access policies editor, envisioned as part of a data sharing configuration template, as a means for the individuals to define the conditions under which their data will be shared, and an access policies engine, the enabler of the

[6] "KRAKEN - broKeRage And marKEt platform for persoNal data." https://krakenh2020.eu/ (accessed Jul. 22, 2022).

access to the data in case the comparison between the attributes of the data seeker meets the conditions for sharing the data.

6.2.4 Data owners consent management

Consent and adherence to legislation is a very important aspect when it comes to data collection, and contrary to what many think, the technical implementation of such procedures is a very tough task, as it is very hard to offer a highly usable and enjoyable platform to users and, at the same time, tick all the boxes coming from the legal world.

The smashHit[7] project offers a platform that provides functionalities to different actors aiming to get consent creation, management, and observation functionalities. The smashHit platform for this reason includes components[8] such as the following:

- Semantic models of consent and legal rights – This component represents the used ontologies and resulting semantic data models used for the description of legal rights, consents, context and automatic contracting rules, and terms and conditions. The different ontologies are linked by a top-level ontology that describes very basic concepts to be extended by additional linked sub-ontologies.
- Context sensitivity support tool – This component adds context-sensitive features to the overall smashHit framework and enables a context-dependent behaviour of the system.
- Consent certification – This component includes functionalities regarding the life cycle of the consent certifications, i.e., the consent certification creation, consent management, and consent distribution among the parties.
- Data use traceability – This component will allow to find data leakages or misuses by using data fingerprinting or watermark techniques and provides data owners the power to manage their data contracts.
- Contract support tool – This component provides automatic contraction functionalities that enable automatic consent document generation and execution, as well as semantic consent representation and visualisation.

[7] https://www.smashhit.eu

[8] https://www.smashhit.eu/wp-content/uploads/2021/03/smashHit_D1.3_Public_Innovation_Concept_v100.pdf

- Consent tracing functionality will be provided, enabling the identification of broken consent chain and guaranteeing that the data exchanged are in agreement with the data consented by data owner.

6.3 Preserving Data Privacy and Data Quality Simultaneously

Though, in the recent years, many methods have surfaced regarding privacy preservation, many of them are a double edge sword when it comes to preserving the quality of the data, as the more privacy preservation measures are applied, the more difficult it becomes to process data and extract highly reliable and accurate analytics results. In this section, we focus on technologies that are already there when it comes to preserve the privacy of individuals and that do have an active research community that pushes forward innovations which aim to disengage data privacy from data quality.

6.3.1 Data anonymisation

Anonymisation is a core enabler for data sharing, especially when one works with personal, sensitive data.[9, 10] The anonymisation process renders data non-identifiable, such that the probability of re-identifying data sharers/individuals in the data is rendered sufficiently low. Identifiability can be viewed along a spectrum. As the data are increasingly transformed, the identifiability of the data is gradually reduced until it reaches a level that is below the applicable anonymisation threshold. At this point, the data are no longer identifiable. Statistical and machine learning mechanisms can be used to anonymise data so that the true values of personally identifiable information (name, age, gender, ethnicity, etc.) are effectively obfuscated and/or transformed (location data). Anonymisation can be incorporated into:

- a "privacy by design" approach to provide improved protection for data sharers;

[9] "Sharing Anonymized and Functionally Effective (SAFE) Data Standard for Safely Sharing Rich Clinical Trial Data." https://www.appliedclinicaltrialsonline.com/view/sharing-anonymized-and-functionally-effective-safe-data-standard-for-safely-sharing-rich-clinical-trial-data (accessed Jul. 25, 2022).

[10] "Guidance Note: Guidance on Anonymisation and Pseudonymisation," *An Coimisiún um Choisant Sonraí. Data Protection Commission*, 2019, Accessed: Jul. 25, 2022. [Online]. Available: https://www.dataprotection.ie/sites/default/files/uploads/2019-06/190614%20Anonymisation%20and%20Pseudonymisation.pdf

- a "data minimisation" strategy aimed at minimising the risks of a data breach.

Both approaches are valid and, in many cases, are combined. Nevertheless, the main challenge in existing anonymisation approaches remains how to manipulate the datasets so that they do not lose the essential information that they hold (from a data scientist perspective), but at the same time satisfy the requirements of the privacy guarantees they are offering to their end-users.

Despite datasets not containing any personally identifying information (PII), such as name, address, etc., individuals can be identified through their quasi-identifiers (QIs). QIs are the attributes whose combination can serve as a unique identifier for individuals. The safe-DEED[11] project offers a demonstrator that is able to perform de-anonymisability analysis of a dataset and check the above-mentioned factors.

6.3.2 Secure data analytic services

Another method to simultaneously support high data privacy guarantees whilst retaining analytics results of high value is the establishment of analytics services that are able to run in secure environments, such as isolated data spaces or on-premise or private hosted cloud infrastructures.

One of the recent infrastructures in this field is the secure analytic host service (SEAS)[12] created in the scope of the DataVaults project to set up and deploy a playground and visualisations host to execute and visualise the different machine learning experiments. The main idea of the SEAS is to allow the user to create a friendly environment which can run in a very user-friendly manner, creating a secure environment where analytics can be executed, as this environment will be ultimately chosen by the user; thus, no third parties are involved.

The SEAS component could be divided into two subcomponents. The first one is the "service analytic host", a component that manages the deployment and set-up of the playground and visualisation host. This is a component that is offered as a central service (in the case of the DataVaults project, this is hosted in the main DataVaults cloud platform). The user/data science/data seeker defines where he/she wants to deploy the "playground and visualisation host" (see the following subcomponent with regard to choosing

[11] https://www.safe-deed.eu

[12] https://www.datavaults.eu/datavaults-components-for-data-sharing-value-generation-and-intelligence-2-3/

the server, the ports, the datasets, etc.). Technologies such as sbt play[13] as the framework to create the graphical user interface, Java for the setup, and Docker[14] and Ansible[15] to deploy in an automated way the playground and visualisation host is used to enable this component.

The second subcomponent is the "playground and visualisation host". This allows the user to run, track, and visualise the machine learning experiments. This component could be deployed in another server or cloud service outside the operational environment of the "service analytic host", while it could be even deployed in-house, on the user's laptop using a ZIP file. MLflow[16] as a platform for the machine learning lifecycle is the technology used to track and run machine learning experiments. Apache Superset is an exploration and visualisation platform of data that allows the user to understand their data in a better way. Both MLflow and Apache Superset[17] will communicate using a PostgreSQL database.

6.3.3 Data management technologies

Data management in smart cities is a big and complex task. Data gathered and managed by the modern IT services in cities have all characteristics of big data: high velocity, high volume, high value, high variety, and high veracity. In addition to that, the data managed by cities' public administrations are very sensitive and, in many cases, are personal data of the citizens. The modernisation of the IT services in public administrations is a slow, continuous, and never-ending process. There are many good data management systems on the market, but their deployment in the existing complex infrastructures is always a huge challenge. The applications and underlying data management systems of several historic generations often are used in parallel and need to work together. Maintaining the interoperability between them remains a permanent challenge. The best precondition for it is always to follow recognised open standards and ensure the availability of APIs at all services in the public IT infrastructure.

One of the important topics in the context of smart cities is the integration and harmonisation of data coming from different sources. The data create a basis for optimisation of the existing processes and creation of new services, but they can be very heterogeneous. They may have different

[13] https://www.playframework.com/documentation/2.8.x/BuildOverview
[14] https://www.docker.com
[15] https://www.ansible.com
[16] https://mlflow.org
[17] https://superset.apache.org

structure and semantic, which may be not easily understandable by anyone apart from the creator of the service generating the particular data. This is why the documentation of data formats and APIs is very important. The data can be harmonised on the basis of a common data model, which would cover the needs of the planned informational system. In some cases, it is sufficient to harmonise data on the level of data description (metadata) without modifying data itself. Data market place is such an example. The data shared and exchanged through a market place may have different semantic and structure, but the description format of data has to be aligned to make the data discoverable and present the information about data to the data seeker. The best practice in designing data models for metadata is to create them using the semantic web technologies and, in particular, the standardised vocabularies.

Amongst the key technologies in data management is data cleaning, which, amongst others, includes operations such as data validation, data cleaning, and data verification. The aim of such technologies is to offer to all engaged parties the ability to identify all the incomplete, incorrect, inaccurate, or irrelevant parts of this data, and then replace, modify, or delete the dirty or coarse data in order to result in high-quality datasets, which have an improved added value for the data consumers.

One example of such a technology offering comes from the PolicyCLOUD[18] project which delivers a data cleaning component that incorporates algorithms and techniques for detecting and correcting (or removing) corrupt or inaccurate records from diverse types of collected data,[19] including also other intelligent features that have to do with the level of customisation in terms of rules for validation, cleansing, etc., tailored to the needs of each different party that exploits this component.

Another example is that offered by the PIMCity[20] project, which is also offering novel tools for data management.[21] The data knowledge extraction (DKE) component offers the means to extract knowledge from the raw data implementing machine learning and big data solutions. One of the biggest challenges here is the creation of value out of the raw data. When dealing with personal data, this must be coupled with privacy preserving approaches, so that only the necessary data are disclosed, and the data owner keeps the

[18] https://policycloud.eu

[19] "Data Cleaning - Policy Cloud." https://policycloud.eu/services/data-cleaning (accessed Jul. 25, 2022).

[20] PIMCity – Building the next generation personal data platforms." https://www.pimcity-h2020.eu/ (accessed Jul. 25, 2022)

[21] https://www.pimcity-h2020.eu/software/

control on them. The DKE consists of machine learning approaches to aggregate data, abstract models to predict future data (e.g., predict user's interests in recommendation systems), and fuse data coming from different sources to derive generic suggestions (e.g., to support decision by users, providing suggestions based on decisions taken by users with similar interests). Another component, the data portability and control (DPC) tool allows individual users to migrate their data to new platforms, in a privacy-preserving fashion. More specifically, it provides methods for extracting data from one PIMS (e.g., bank data through the TrueLayer API), process it by filtering out sensitive information or user-inputted data (e.g., remove login credentials or debit card numbers), and outport it into other modules, a new PIMS (e.g., EasyPIMS[22]) or an exported file in a common data interchange format, e.g., JSON.

PIMCity also offers tools for data provenance and aggregation. *Data provenance module* OpenAPI allows developers to insert watermarks of ownership in the datasets they share in the marketplace. In general, this component is used internally by the PDK and developers that are in need of controlling data ownership even after a dataset has left the platform. This is done by embedding difficult-to-remove watermarks into the datasets. When it comes to aggregation, the data aggregation (DA) tool enables data owners that hold a bulk of their users' data to aggregate and anonymise them. This allows sharing these data in a privacy-preserving way.

The Snap4City platform,[23] which is powering the REPLICATE[24] and is operative with services and data of several cities including Florence, Helsinki, Antwerp Valencia, Venezia, and Roma, includes also a set of tools/services[25] offering a secure and privacy respectful solution for data management, used in several scenarios and functionalities of the platform, working with "my personal data type" according to GDPR, extracting personal data from IoT devices of users, protecting the storage for personal data, etc.

6.3.4 Data models and interoperability

Data management is closely related to data models, as the latter are structures that are in a position to support interoperability in data management systems.

---

[22] https://www.easypims.com

[23] "Snap4City" https://www.snap4city.org (accessed Jul. 25, 2022).

[24] "REPLICATE's Final Results and Impacts – Replicate Project EU." https://replicate-project.eu/the-replicate-projects-final-results/ (accessed Jul. 25, 2022).

[25] "US11. Using tools/services of a secure and privacy respectfully solution - Snap4City." https://www.snap4city.org/drupal/node/166 (accessed Jul. 25, 2022).

When it comes to personal data, the DataVaults project has followed this approach as its data model is defined as a profile of the general data catalogue vocabulary (DCAT[26]). DCAT is a W3C recommendation providing an RDF vocabulary to facilitate interoperability between data catalogues in the web. The DataVaults data model reuses the main parts of the DCAT and extends it with the classes necessary to describe the personal data and domain-specific data of the DataVaults demonstrators. This makes it interoperable with high number of other data models based on the DCAT. The DataVaults data model describes a holistic personal data value chain addressing all the aspects of personal data management. This includes data protection, security, GDPR compliance, IPR management (compensation schemes, etc.), and representation of the main value flows in data marketplaces. It is based on existing semantic web standards that are very important for potential interoperability with other data models and for the model reuse.

A catalogue in DataVaults data model represents the top-level class assigned to an individual. The metadata of a catalogue as defined in the core DCAT vocabulary contains properties that allow description of profiles of individuals. There are two ontologies commonly used for describing an individual's master file data. They have many overlapping properties. In DataVaults, the focus has been laid on the friend of a friend (FOAF) ontology, which is complemented by the vCard ontology.[27] All FOAF-related fields are added via the dct:publisher property and encoded as a foaf:Person element. All vCard-related values can be found as a vcard:Individual attached via the dcat:contactPoint property.

The importance of having proper and interoperable data models is also pinpointed by the i3-MARKET project,[28] as the availability of common data models is a key enabler for establishing a scalable data economy. As such, i3-MARKET offers methods to access in a decentralised manner the semantic descriptions of the offered data assets in order to enable data discovery across today's silos. This enables federation among the individual data spaces and marketplaces, without the need of central control or coordination that has to be trusted by all parties. This is made a reality by using a secure semantic data model repository that enables data consumers to efficiently discover and access data assets (due to precise semantic queries) and integrate the data into their applications/services (based on a common

[26] https://dvcs.w3.org/hg/gld/raw-file/default/dcat/index.html

[27] https://www.w3.org/wiki/Good_Ontologies

[28] "i3-MARKETArchitectureandApproach-i3Market."https://www.i3-market.eu/i3-market-architecture/ (accessed Jul. 25, 2022).

understanding of the meaning of the data). In this way, independent data providers and consumers are enabled to exchange and use data in a meaningful way – without prior information exchange. We build on semantic data models defined and make them accessible via a public data model repository.

The PolicyCLOUD's interoperability component aims to enhance interoperability via the utilisation of linked data technologies, such as JSON-LD, and standards-based ontologies and vocabularies, coupled with the use of powerful tasks from the domain of natural language processing (NLP), in order to improve both semantic and syntactic interoperability of data and datasets. Through these coupled technologies and methods, the interoperability component seeks to provide a state-of-the-art approach to achieve interoperability in data-driven policy making domain. SemAI[29] entitles this hybrid approach and is a combination of commonly used semantic techniques coupled with the utilisation of NLP tasks and methods. The main goal of this hybrid mechanism is to design and implement a holistic semantic layer that will address data heterogeneity. SemAI introduces a multi-layer and hybrid mechanism for semantic interoperability across diverse policy-related datasets, which will facilitate semantic interoperability across related datasets both within a single domain and across different policy-making domains. This requirement relates to local–regional public administrations and business domain, but it also goes beyond the national borders as it also seeks to invoke a language-independent hybrid mechanism. To this end, this hybrid approach aims to enhance both semantic and syntactic interoperability of data based on the aggregation, correlation, and transformation of incoming data according to the defined schemas and models. The knowledge that is derived from these processes, shaped in a machine-readable way, can be used later from other tools for providing big data analytics, i.e., sentiment analysis, etc.

6.3.5 Digital twins for privacy preservation

The role of models in a digital twin (in addition to data) is to interpolate spatially where no direct measurement data is available; to integrate the presentation and analysis of complementary datasets; to cross-correlate data from different domains; to infer properties/attributes that are not directly measured;

[29] "SemAI: Enhanced Data Interoperability - Policy Cloud." https://policycloud.eu/services/semai-enhanced-data-interoperability (accessed Jul. 25, 2022).

to convert measurements and state info into KPIs; and finally to extrapolate (predict) business as usual (BAU) and what-if scenarios.[30]

As an example of digital twins for privacy preservation, one could have a look at the DUET project. The project has published the DUET-Cell architecture[31] as a plug-in interface to create a digital twin platform for urban regions where it becomes possible to add new (possibly third-party) data sources to an existing digital twin case, add simulation models to an existing case for comparison or to be used as input for another model adding visualisation clients, and finally extend the digital twin ontology to support more city domains or expand existing ones.

6.3.6 Cryptographic solutions for data privacy

Cryptography is amongst the key enablers to achieve data privacy and is one of the cornerstones when it comes to data protection. For this reason, all existing and emerging platforms do support cryptography out of the box, as one of the many features they employ to protect the data they hold. In the last years, many innovations in the area of cryptography have surfaced and such platforms do consider, or even integrate novel approaches which, at the same time, provide very high trust guarantees and, at the same time, allow various operations over the data, without jeopardising trust or privacy.

As an example of using such state-of-the-art technologies, we suggest KRAKEN[32] that creates a data-analytics-as-a-service component by using secure multi-party computation (SMPC) and functional encryption (FE) for performing computations on protected data. SMPC guarantees confidentiality preventing the data provider, the data processor (the service), and the consumer from learning about the plain data or the sensitive data. Only the consumer is able to learn the statistical analysis result. Additionally, the authenticity and integrity are assured by using zero knowledge proof (ZKP). Sharing data between producers and consumers involves ensuring a secure end-to-end data sharing. Also, ensure that only the consumer has access to the final data or result implies a fine-grained access control. Regarding

[30] "DUET D3.4 Smart City domains, models and interaction frameworks v2," *DUET Consortium*, 2021, Accessed: Jul. 25, 2022. [Online]. Available: https://www.digitalurbantwins.com/_files/ugd/68109f_d4c08a03f34e481695977b4fd2577b16.pdf

[31] "Open Technical Approach: T-Cell Architecture for DUET Digital Twins." https://www.digitalurbantwins.com/technical-approach (accessed Jul. 25, 2022).

[32] "Deliverables - Kraken." https://www.krakenh2020.eu/resources/work-package-deliverables (accessed Jul. 25, 2022).

the use of cryptographic methods on self-sovereign identity (SSI) solution, the extension of the SSI solution with ZKP allowing the citizen to disclose part of their attributes in a privacy-preserving way is being investigated in KRAKEN.

Besides ZKP, the use of digital signatures, proxy re-encryption, or attribute-based encryption (ABE) assures a secure end-to-end data sharing and authenticity and confidentiality of data analytics. This approach is being used in KRAKEN as well as in the DataVaults and in the ASSURED[33] project, where in the latter, ABE is handled by hardware TPM devices (of the stakeholders) that have embedded the list of attributes needed for the encryption and the decryption of the data, instead of the attributes being held by a centralised third-party authority.

6.3.7 Artificial intelligence threat reporting and response systems

As we are moving into a world of more and more connected infrastructures, it is essential to also come up with more intelligent ways to detect threats to privacy, and cybersecurity surely plays a very important role in this endeavour.

As an example, one could consider the IRIS[34] project, which addresses the challenge of protecting IoT and AI-driven ICT systems from cyberthreats and cyberattacks on their operation and privacy. The IRIS concept is proposed as a federated threat intelligence architecture that instates three core technological and human-centric components into the threat intelligence ecosystem. First, there is the collaborative threat intelligence that forms the nexus of the IRIS framework and core component of the architecture enhancing the capabilities of the existing MeliCERTes platform by introducing analytics orchestration, an open threat intelligence interface and an intuitive threat intelligence companion. All this supported by a data protection and accountability module. Second, there is the automated threat analytics that collects and supplies key threat and vulnerability assessment telemetry and responds to received intelligence, initiating autonomous response and self-recovery procedures. The third component is the cloud-based virtual cyber range, which delivers an immersive virtual environment for collaborative training exercises based on real-world environment platforms (and digital

[33] https://assured-project.eu

[34] "IRIS H2020 project." https://www.iris-h2020.eu/ (accessed Jul. 22, 2022).NOTE-Not to be confused with the IRIS project referred to in Chapter 2 which is https://irissmartcities.eu/

twin honeypots), providing representative adversarial IoT and AI threat intelligence scenarios and hands-on training.

6.4 Information Delivery on Privacy Metrics and Data Content and Value

Aside from the tools and methods used to allow engaged stakeholders to secure their data and build the necessary barriers to safeguard privacy and enable trust, there is also the need to explain to data owners how their data and, at the very end, their privacy is protected, what they can find out from their own data themselves, as well as how much value their data hold. In Section 6.4, we explore some of these technologies.

6.4.1 Privacy metrics and risk management and privacy metrics for personal data

Risk management contains both the notion of monitoring as well as evaluating the risks related mainly to privacy. The main root of security risks is the sharing of a dataset that contains private information, and a risk management system can offer two different ways to analyse this risk.

First, it is in the position to provide an overview of privacy risks and the privacy exposure to the platform administrator on the basis of the datasets that have been shared to the platform. This is based on a model used to serialise all information that can describe the datasets and then perform the calculation of relevant risks. The above privacy assessment is based on the analysis of asset chains in order to estimate the impact from the interconnections of the assets and datasets and to provide a single estimation for the overall impact of a specific asset/vulnerability combination. Risk management also provides the ability to identify and add new known vulnerabilities to enrich the risk knowledge base of the platform.

Second, it can provide to the end-user the calculated risk regarding the datasets that the end-user has added to the platform, and it also warns her/him of the risk exposure of a dataset that she/he is trying to share, in order to enable more informed decisions regarding the amount of personal information the user is comfortable sharing.

The DataVaults project is delivering such a risk management component where input is taken from both the datasets that an owner want to share as well as from the dataset that has been already shared by the same individual, to assess a unified privacy risk exposure metric.

The PIMCity personal information management systems (PIMS) development kit[35] (PDK) also similar components are provided, such as the following:

- Personal data safe (P-DS) is the means to store personal data in a controlled form. It implements a secure repository for the user's personal information like navigation history, contacts, preferences, personal information, etc.
- Personal privacy metrics (P-PM) represent the means to increase the user's awareness. This component collects, computes, and shares easy-to-understand metrics to allow users to know how a service stores and manages the data.
- Personal consent manager (P-CM) is the means to define all the user's preferences when dealing with personal data. It defines which data a service is allowed to collect, process, or can be shared with third parties.

6.4.2 Personal data analytics

Analytics at the side of an individual/citizen can offer to that person two main things in a platform that handles and manages personal data. On the one hand, they can allow her/him to have a quick view of the data she/he generates using cumulative and predefined graphs and charts. Doing that, a citizen is able to better understand the data she/he generates and can have some high-level insights of what other stakeholders might discover if she/he decides to share her/his data. Think, for example, that instead of seeing just bio signals, one could see the daily or hourly heart rate, or instead of having a long list of things she/he searched for in Google, to have her/his queries categorised based on identifying how much this person is, for example, querying for sport facilities, real-estate agents, and shopping.

On the other hand, the ability to export analytics allows the individual also to engage in data sharing activities by just sharing these analyses, which increase the privacy level of the data subjects, as instead of sending out raw data, already pre-processed data are provided.

The PIMCity personal information management systems (PIMS) development kit (PDK) includes a personal privacy preserving analytics (P-PPA)

[35] https://www.pimcity-h2020.eu/about-the-technology/pimcity-development-kit/

which allows extracting useful information from data while preserving users' privacy and leverages concepts like k-anonymity and differential privacy.

The DataVaults project showcases how all of the above can be delivered to an individual by having a lightweight analytics engine running at the side of the user, which will rely on predefined algorithms and data categories linked to the data sources that a citizen is willing to share with other stakeholders.

6.4.3 Data valuation

Understanding the real value that is included within a dataset is a crucial point for all engaged stakeholders in the value chain of the data economy.

The Safe-DEED project provides tools to facilitate the assessment of data value, thus incentivising data owners to make use of the cryptographic protocols to create value for their companies and their clients.[36] The data valuation applications, as part of the Safe-DEED demonstrator, describe the initial implementation of the data valuation component (DVC). The supported algorithms are selected regression, classification, and clustering algorithms (at ADAS level), and a rule-based algorithm for generating the economic value of the input data set (at S2VM level).

The PIMCity PIMS development kit (PDK) also includes data valuation tools. These tools, from the market perspective (DVTMP) module, leverage some of the most popular existing online advertising platforms to estimate the value of the audience, while from the user perspective (DVTUP), it provides estimated valuations of end-users' data for the bulk dataset they are selling through the marketplace.

6.5 Data Platforms

All of the above-mentioned technologies are being integrated into data platforms, either industrial ones or personal data platforms, which aim to bring different stakeholders together to make the notion of the data economy a reality, respecting both data owners, data consumers, and all other engaged stakeholders such as data brokers, etc.

In this subsection, some key technologies for the realisation of such platforms are presented as well as some examples of such marketplaces.

[36] "Safe-DEED Valuation tool." https://demo.safe-deed.eu/ (accessed Jul. 25, 2022).

6.5.1 Secure and trusted data communication channels

One of the main features that data platforms should include is that of secure data exchange which is a key factor in generating the trust between the transaction parties.

IDS is one of the most prominent initiatives for creating connectors for secure data sharing. The project DataPorts[37] which works on secure managing and sharing transportation and logistics data among trusted stakeholders and using these data to offer novel AI and cognitive tools to the seaport community, has based its architecture on the IDS concept. In more detail, the DataPorts platform is based on the concepts from the IDS reference architecture designed for secure data exchange and trusted data sharing. These concepts are represented by components in the DataPorts platform. The IDS reference architecture addresses such strategic requirements as enabling: trust, security and data sovereignty, ecosystem of data, standardised interoperability, data market, re-use of existing technologies, and contribution to standardisation.

To enable secure end-to-end data exchange, the i3-MARKET project makes use of secure and trusted APIs to allow data spaces and marketplace providers to obtain identities, to register data assets, to fetch their semantic descriptions, to create and sign smart contracts, to make payments, etc. This ensures complete openness, i.e., that any data space or marketplace provider can connect its local ecosystem with the global i3-MARKET data market ecosystem.

Secure communication is also established by peer-to-peer connections, as showcased by the InteropEHRate[38] project, where key health data is managed in "patients' hands", i.e., through smart EHRs (S-EHR) on mobile devices. In InteropEHRate, data is always transferred via highly secure channels including a direct device to device (D2D) communication. It is developing open interchange protocols supporting patient-centred exchange of health records between patients, healthcare actors, and researchers. It is guided by a future-proof prospective where most citizens will own mobile health repositories, called smart EHRs (S-EHRs), managing a wide range of their own personal health data on smart devices, regardless of whether the data have been produced by health professionals, any sensor, device, or the citizens themselves. As such, InteropEHRate gives citizens and patients a meaningful choice for the method their health data is being stored and exchanged, such that they can benefit from its usage everywhere it is needed and gain also more control making the process fully compliant with the GDPR. This choice

[37] https://www.dataports-project.eu/

[38] https://www.interopehrate.eu/about/

is made possible by developing an innovative health data storage on smart phones. Citizens can also initiate themselves standardised communications and transfer data, e.g., to hospital outpatient services by using secure remote connections via the internet, and, in addition, by building new connections without the usage of internet between the citizen's device and the healthcare provider systems using short-range device to device communication.

6.5.2 Immutable ledgers and smart contracts

Data exchange in novel marketplaces takes advantage of the recent innovations in the technological domain of blockchain, and, as a result, it is nowadays a commodity to store data sharing transactions in blockchain ledgers and employ smart contracts that can automate many of the necessary activities that are included in data exchange.

The i3-MARKET project makes use of immutable and auditable smart contracts for the trading of data assets across data space and marketplace boundaries. All stakeholders, namely data providers (for confirmation of the offer and its conditions, e.g., license, price, and SLAs), data consumers (for agreement of the contract conditions), and data owners (for consent to the data exchange) must sign these contracts. This developed solution can also be adopted by individual marketplaces for handling local contracts. Similar solutions are also followed by other projects, such as DataVaults where a double ledger approach is provided in order to keep data owners and data consumers apart, with the platform playing the role of the data broker, or as in BEYOND[39] project, where a similar ledger approach is used which also enables multi-party contracts.

KRAKEN[40] developed a decentralised solution such as SSI which provides a decentralised user-centric approach on personal data sharing; it uses verifiable credentials (VCs) with different levels of assurances for accessing online services. KRAKEN also provides a trusted environment comprising different trusted registries, in order to be ready for being aligned with EBSI[41]/ESSIF[42] and leveraging their capacities, supporting trustability in SSI solution. Also, KRAKEN provides a public DID infrastructure supporting access to different data ledger technologies (DLTs) and the management of public

[39] https://beyond-h2020.eu

[40] "Deliverables-Kraken."https://www.krakenh2020.eu/resources/work-package-deliverables (accessed Jul. 25, 2022).

[41] https://ec.europa.eu/digital-building-blocks/wikis/display/EBSI/Home

[42] https://www.eesc.europa.eu/sites/default/files/files/1._panel_-_daniel_du_seuil.pdf

DID, providing the associated services for creating/generating, publishing, and then using the public DIDs. The ledger uSelf component developed in the context of KRAKEN comprises two subcomponents. One is the SSI mobile app for managing the VC and key material, allowing the citizens to use their own smart devices for identifying themselves using the SSI solution. The mobile app provides a friendly graphical user interface designed to adopt and simplify the SSI workflows. The other is the ledger which is used as is used as a broker for integrating service providers (SPs), reducing the complexity of the protocols and processes associated with the SSI solution by providing a simplified interface. With the SSI app and the broker, the SP can use SSI key material for onboarding and login processes. A backup/synchronisation system is integrated with the SSI solution allowing the citizen to use several devices with the same credentials and recover their credentials in case the smart device is lost or stolen.

6.5.3 Crypto wallets

Crypto wallets are used to allow entities (be it individuals or organisations) to have a place where they can digitally retain information relevant to their crypto currency. As crypto currency is, of course virtual, these wallets are used to store cryptographic keys to the crypto currency that each user has. This means that the crypto currency is not actually stored in the wallet but leaves in the blockchain, and the role of the wallet is to hold securely the private cryptographic keys of everyone to be able to access their crypto currency. It is therefore of outmost importance that users do not lose access to their crypto wallets, as in such a case, there is no way to be able to regain control of the crypto currency the user has linked to that specific wallet.

Crypto wallets are highly important in online trading and can be used in operations such as data sharing. They offer an initial degree of privacy, as a user is identified by a hash; however, it is possible by exploring the transactions of a party in a blockchain network to reveal information about his identity. There are also approaches proposed, such as the one of the DataVaults project, which suggest using intermediate wallets owned by trusted third parties, to play the role of the broker between two or more entities, maximising the privacy guarantees and anonymity of the transaction parties, as they later will be actually transacting only with the broker who is the only one knowing their true identities of the different parties.

Transactions in the i3-MARKET project are based on a crypto currency/token as a means to provide a transparent, cost-efficient, and fast payment

solution for trading data assets among the participating data spaces and marketplaces. The crypto token is used to incentivise data providers to offer their data assets and thus accelerate the European data economy. The solution is designed in a way that the participating data spaces and marketplaces can also use the tokens as internal payment medium.

6.6 Other Supporting Initiatives

In addition to the outputs from the projects that have been referred to above, there are a variety of other initiatives offering supporting frameworks. Many fit into both camps, providing valuable technologies as well as providing a more general supporting function, such as RUGGEDISED and SmartEnCity.

6.6.1 EUHUBS4DATA

The federation aims at creating a solid ecosystem, bringing together relevant European initiatives around the data economy, fostering collaboration among those initiatives towards the objective of common European data spaces, attracting SMEs and start-ups to use and benefit from the federated services and data sources, and raising awareness in society about the benefits of data-driven innovation.

In fulfilling this aim, a catalogue of services was offered, datasets were made available, and courses on the offer have been provided.

The "catalogue" of EUHUBS4DATA[43] comprises:

- 172 services offered by the membership, ranging from "access and support to big data-AI stack environment" through to "very large speech dataset to build up automatic speech recognition and text to speech deep models".
- 160 datasets made available by the membership ranging from "air quality" through to "world bank open data".
- 80 courses offered ranging from "add-ons to courses: workshop on ML for energy, manufacturing, fintech, transportation, healthcare, etc." through to "strategies for data-based business models and hands-on development".

[43] "EUHubs4Data - European Federation of Data Driven Innovation Hubs." https://euhubs4data.eu/ (accessed Jul. 25, 2022).

6.6.2 MyData

The MyData Mission Statement[44] is: "We help people and organisations to benefit from personal data in a human-centric way. To create a fair, sustainable, and prosperous digital society for all".

"MyData Global aims to empower individuals by improving their right to self-determination regarding their personal data. The human-centric paradigm strives for a fair, sustainable, and prosperous digital society, where the sharing of personal data is based on trust and a balanced and fair relationship between individuals and organisations".

6.6.3 Solid Flanders

"Solid[45] is a specification that lets people store their data securely in decentralised data stores called Pods. Pods are like secure personal web servers for your data.

Any kind of information can be stored in a Solid Pod. You control access to the data in your Pod. You decide what data to share and with whom (be it individuals, organisations, and/or applications). Furthermore, you can revoke access at any time. To store and access data in your Pod, applications use standard, open, and interoperable data formats and protocols".

6.6.4 Big value data association (BDVA)

"The Big Data Value Association –BDVA, (from 2021, DAIRO - Data, AI and Robotics), is an industry-driven international not–for-profit organisation with more than 230 members all over Europe and a well-balanced composition of large, small, and medium-sized industries as well as research and user organizations. BDVA/DAIRO focuses on enabling the digital transformation of the economy and society through Data and Artificial Intelligence by advancing in areas such as big data and AI technologies and services, data platforms and data spaces, Industrial AI, data-driven value creation, standardisation, and skills. The mission of the BDVA is to develop the Innovation Ecosystem that will enable the data and AI-driven digital transformation in Europe delivering maximum economic and societal benefit, and, achieving and sustaining Europe's leadership on Big Data Value creation and Artificial Intelligence".[46]

[44] https://www.mydata.org/

[45] https://solidproject.org/about

[46] https://www.bdva.eu/about

6.7 Looking into the Future

As the current IT landscape is witnessing dramatic changes which are highly influencing and are also influenced by the changes in our societies, numerous changes are expected to happen in the area of data sharing, privacy, trust, and security. In order for someone to be able to remain at the forefronts of the technology, much effort has to be invested in operations such as technology watch and applied developments and testing of current solutions as those mentioned briefly in the previous subsections.

Regarding this chapter, it is mentioned that the results from all the projects referred in the previous subsections are and will be published on their respective web sites as they appear and will generally include all the supporting tools and manuals for the implementation of these results. The DataVaults "manual", the RUGGEDISED D6.6, the guidance within IRIS, etc., all add to the repository of knowledge for moving forward. In addition, a new platform has been set up by the Commission to also store relevant information.[47]

Chapter 7 will start to address some of the interoperability issues faced when augmenting existing platforms and approaches already taken by a smart city.

Further, Chapter 17 points to the technologies that will become available in the near future, from projects about to start, as given the time scale set by many cities of becoming climate neutral by 2030, there cannot be any unnecessary delay in making use of these emerging technologies if these goals are to be achieved. Advance awareness of what is coming next will shorten the process of adoption.

[47] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/horizon-results-platform