TRUSTS Whitepaper

on Data Governance in Collaborative Data Environments

Authors: Stefan Gindl, Michael Boch, Gianna Avgousti, Christos Skoufis, Alan Barnett, Matthew Keating, Nina Popanton

September 2022

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871481

Disclaimer

Copyright message

Glossary of terms and abbreviations

Abbreviation / Term	Description
GA	Grant Agreement
UC(s)	Use Case(s)
TRUSTS	Trusted Secure Data Sharing Space
AML	Anti-Money Laundering
FIs	Financial Institutions
AI	Artificial Intelligence
ML	Machine Learning
ERP	Enterprise resource planning
PII	personally identifiable information
QID	quasi-identifier
GDPR	General Data Protection Regulation
SSN	Social Security Number
KG	TRUSTS Knowledge Graph

1. Outline of this Whitepaper

Data governance is an emerging necessity in enterprise information management and should be a strategic initiative for all organizations. In this whitepaper, the TRUSTS Consortium will provide hands-on insights on their experiences on data governance mechanisms in one of the three use case fields: Anti-Money-Laundering.

2. Introduction

The **Trusted Secure Data Sharing Space** (TRUSTS) project develops a platform for trading data and data services in a trustworthy and reliable manner, which will enable a pan-European data economy in which privacy and security are at the forefront. This platform lets organizations share their data via a set of nodes. Each node is reserved for an individual organization. This guarantees that organizations remain in full control of their data assets. Therefore, data **governance** plays a major role in TRUSTS. It enables a homogenous view of the data as well as a standardized terminology for it. This is especially important for the TRUSTS platform since it has the interoperability of different existing data infrastructures, some of which are operated by the project partners or their customers, and some of which are operated by third parties, as its objectives. The architectural design of the TRUSTS platform requires the orchestration of different components, which in turn necessitates the exchange of information in an unambiguous and consistent manner.

For this reason, a knowledge graph was created called the **TRUSTS Knowledge Graph** (KG). It is the collection of metadata about different resources on the platform, and to the technical and organizational mechanisms for its maintenance, consumption, and governance. This metadata is expressed as a collection of statements about entities such as assets, participants, nodes, and themes. These statements, when read by components of the platform, empower the different functionalities of the platform.

3. Data Governance in Data Marketplaces

An important part of data governance, also of TRUSTS, is data privacy and security. The challenge is to get the best possible quality of data with the highest standards of privacy and security. TRUSTS aims to obtain as much information as possible with the highest level of security. In the EU, this is particularly important with regards to regulations such as the EU General Data Protection Regulation (GDPR)[1]. However, there are no clear statements regarding data anonymity in GDPR. Recital 26 states the following: "*The principles of data protection should therefore not apply to anonymous information, … data rendered anonymous in such a manner that the data subject is not or no longer identifiable*". It states that it does not apply to anonymized data, but the characteristics and requirements for data anonymity are not specified.

Persons can be easily identified via their name, address and SSN number, which is the so-called personally identifiable information (PII). But is the dataset really anonymized if you just remove every PII? For example, attributes such as gender, date of birth and ZIP code alone are usually not a threat, but in combination they can be used to infer individuals. This type of information is called a quasi-identifier (QID). So-called de-anonymization attacks try to infer individuals from anonymized data on the basis of QIDs. The more QIDs a dataset has, the higher the risk of potential de-anonymization. While it is still relatively easy to assess the risks of PII, it is not as straightforward with QIDs. For this

[1] GDPR: https://gdpr.eu/, accessed September 09, 2022.

reason, TRUSTS offers an application that provides a risk analysis and suitable anonymization methods. Here, the focus was no longer on data in classic tabular form, but also on complex data types such as location and textual.

4. Smart Contracts for Good Data Governance

In the context of blockchain-based data marketplaces, *smart contracts* facilitate core functionalities such as the creation, transfer, and updating of assets, ledger querying, as well as auxiliary functionality such as assigning reviews or ratings to assets, and these kinds of functionalities are facilitated by the TRUSTS platform's blockchain-based smart contract executor component.

Data governance policies surrounding smart contracts and blockchain technologies typically pertain to security patching, bug fixes, transaction legitimacy, resolving disputes, and operations on assets. Smart contracts and the blockchains on which they run, present the unique data governance challenges of immutability and decentralization. Immutability poses an issue if illegal content is put on the blockchain, as removal is not typically easy. The distributed ledger will propagate this illegal content to other users - who may violate the law by possessing it. This issue is yet to be addressed on many blockchain platforms and could benefit from rigidly defined and enforced governance.

Although consensus, dispute resolution and forking provide some control, more robust governance mechanisms for blockchain-based systems are still under development. On-chain, immutable smart contracts are problematic for security patches and bug fixes, and by extension governance of update version compliance. A solution to this is using on-chain proxies pointing to off-chain smart contracts, allowing them to be updated.

The **TRUSTS smart contract executor** component, similarly, supports updating already-deployed smart contracts. Smart contracts can be used as a tool to apply data governance by automatically applying and enforcing policies and rules, thus smart contracts would be part of both the governance system and the system being governed. A strong distinction between governance and operational smart contracts is needed here, but the TRUSTS platform has the capability to leverage smart contracts in this way to support data governance.

5. Data Governance in practice in TRUSTS

At the example of UC1 'Smart big data sharing and analytics for AML compliance'

As part of this project, the TRUSTS Consortium defined three business-oriented Use Cases (UCs) to demonstrate and realize the added value of the TRUSTS platform by showcasing the sharing, trading, (re) use of data and services, and the added value generated through innovative applications built on multiple open and proprietary data sources.

UC1 *"Smart big data sharing and analytics for Anti-Money Laundering (AML) compliance",* supports the main TRUSTS objective to create a European Digital Marketplace, providing dataset trading among business stakeholders. UC1 leverages the power of the TRUSTS Platform's concept for securely sharing data between organizations, applying smart big data analytics for AML compliance purposes as well as fairly trading the resulting data to the end-users such as Financial Institutions (FIs), internal or external auditors, fiduciaries, audit firms etc.

Acknowledging the significance of Artificial Intelligence (AI), Machine Learning (ML) and smart analytics for providing better efficiency in combating money laundering, the purpose of this UC was

to securely share closed-loop data to feed a next generation advanced AI/ML-based AML solution, benchmarked against the current state-of-the-art (i.e., traditional rule-based). Advanced AI and ML techniques were developed and integrated with the existing AML rule-based model that was used as an offering to support this UC, to enable finer grained resolution at the scale needed to detect money laundering activities, thus evolving it into a next generation AML data-driven model. These techniques include:

I. the design and development of adaptive AML risk scoring algorithm,

II. the transaction monitoring and anomaly detection based on ML models,

III. a combination of Federated Learning and Ensemble Modeling for utilizing data from different end-users.

These algorithms and models can detect malicious behavior through data analysis, work with metadata and provide deeper insights about existing and prospective customers.

Effective data governance is at the heart of managing the data used in operational systems. TRUSTS sits at the center of this data exchange between actors and facilitates the trading of the resulting data and data analytics services for a wide range of actors interested in progressing to the next step of AML compliance, to ensure proper use of data, data quality, and policy compliance.

The UC1 detailed overall plan and setup activities as well as an introduction to the UC1 components, actors, and assets deployed in TRUSTS emulated in this UC can be found in the public Deliverable 5.2.

6. Conclusion

This article shares insights in the research work of privacy-preserving technologies and smart contracts in TRUSTS, which can serve as examples for companies and other actors in the data community. Such approaches are also useful for them to implement and to help to maintain data governance while sharing their data. For instance, privacy-preserving technologies allow analyses of datasets while maintaining anonymity of the data, preventing personal information from being leaked to the public. TRUSTS has developed a two-stage process, combining a risk assessment for de-anonymization with a subsequent re-anonymization step. In other words, at first this process estimates how easy it is to identify individuals from an actually anonymous dataset. Subsequently, it applies a set of steps to further anonymize the data of these individuals. This helps companies reluctant to share data, which might accidentally compromise personal data of, for example, customers. Smart contracts, on the other hand, serve as a trustworthy technology to securely and robustly record transactions, without a single entity having monopoly over the records.

This article also elaborates on the application of TRUSTS research outputs in an anti-money-laundering use case. A combination of risk scoring algorithms, anomaly detection, and federated learning has been leveraged to detect malicious money laundering activities. Companies active in this highly sensible field can also benefit from these technologies.

Data governance is a crucial aspect for all organizations aiming to participate in a data-sharing economy and willing to sell and purchase datasets. A clear concept and strategy towards data governance is a fundamental pillar of data sharing and gives participating organizations the confidence to trade data for mutual benefit.