LEONARDO CYBER & SECURITY SOLUTIONS

PRIVACY QUERY PLATFORM

LEONARDO

In a world increasingly reliant on digital technologies, the ability to make strategic decisions for the public and business interests is driven by the availability of data and the value derived from processing it. The economic growth of countries, scientific and industrial research, and the security of nations are linked to the use of the huge amounts of sensitive data produced by an increasing number of connected devices, but also by people. For example, in the health sector, the study of confidential data can be very important for the public interest in analysing specific diseases and discovering new treatments.

However, the usage of these data must be supported by the adoption of measures to ensure the correct and secure use of the information. Indeed, when there's the need of processing personal data, information that allows the identification of a natural person or of his habits, his state of health or his lifestyle are subject to protection rules according to the European Union regulation on the processing of personal data and privacy, the GDPR. To make it possible, data must be deprived of some identifying components through the use of anonymization and minimisation techniques.

Anonymization operations modify the data irreversibly so that the identity of the person to whom the data relate cannot be revealed. In addition, data minimisation must be applied on the results of searches to make available only the data strictly necessary for the purposes of the analysis and research.

PRIVACY QUERY PLATFORM

Leonardo has developed the Privacy Query Platform to implement the necessary controls to allow the extraction and use of information containing personal data while ensuring privacy.
The Privacy Query Platform, based on innovative technologies such as Big Data and Artificial Intelligence is usable in Cloud or on premise and manages information. stored on heterogeneous sources. In particular, the Privacy Query Platform consists of:

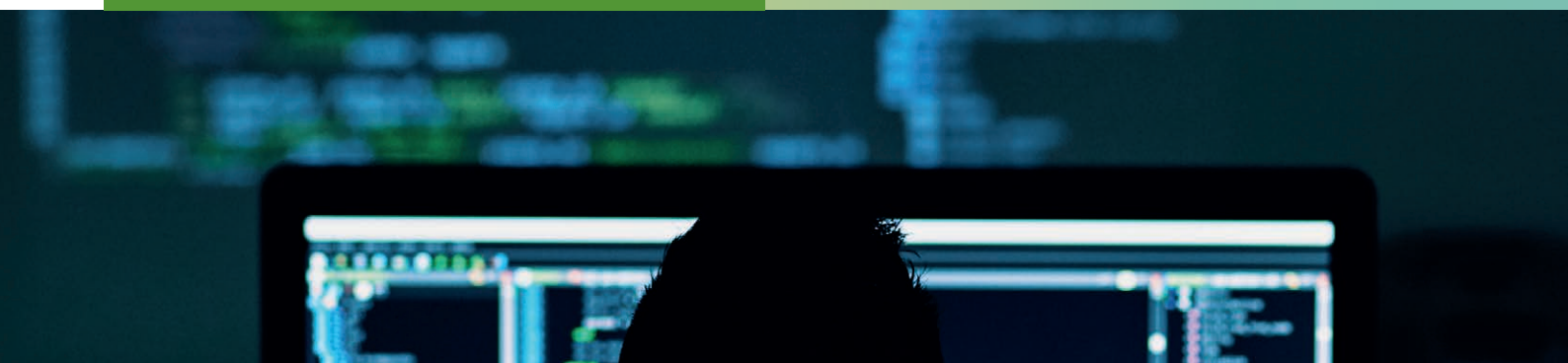
- a Data Protection Engine, which can be integrated with Business Intelligence tools, logically placed between the source containing the data to be anonymised and the users who need to use the information,
- a portal to access the system's functionalities, which allows, in an intuitive way, the use of data also through the guided generation of "queries" to search for the information of interest.

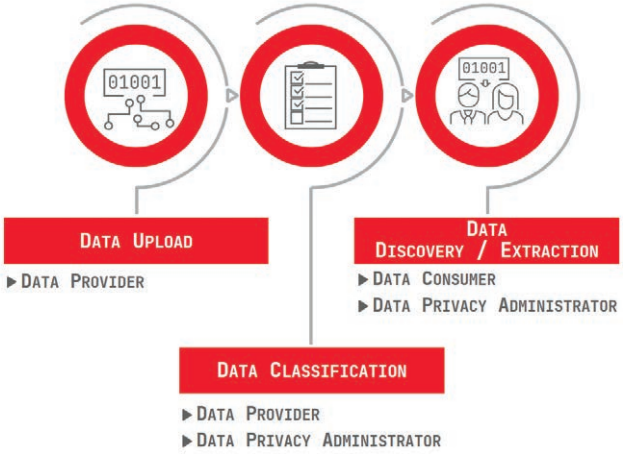
Privacy Query Platform allows both the definition of data research areas and the mapping and definition of the data to be exposed. This is done also through the use of metadata describing exposed data structure and classifying their contents.
Through the Privacy Query Platform functionalities, users, research bodies or organisations that need to use personal data, obtain access to the data in anonymised and minimised form and perform the extraction of all and only the necessary results according to the selected purpose and associated with the user's profile.

THE DATA PROTECTION PROCESS

The process for the complete management of data privacy, from uploading and classification operations to use by consumer, is divided into three phases – Data Upload, Data Classification and Data Discovery / Extraction described below -, while four roles are assigned to the users who interact with the platform:

- **system administrator:** responsible for managing and configuring the system in order to ensure data access
- **data provider:** is the owner and manager of the data and is responsible for cataloguing the different sets of data
- **data privacy administrator:** carries out the operations related to data sets privacy cataloguing, allowed research scopes configuration and enabling, user requested data views approval, also establishing information classification, and the audits
- **data consumers:** they can request access to data through customised views, perform searches and extract data from approved views and research scopes, and view details.





The three-step process guarantees the principles of minimisation and anonymization, thus ensuring the protection of privacy and adherence to data protection regulations:

- **Data Upload** for data sources and data set management, performed by the Data Provider, and necessary metadata definition.
- **Data Classification** for data categorisation through the use of three methods: “a priori” manual classification by the Data Provider, automatic classification through the use of Artificial Intelligence techniques and manual classification by the Data Privacy Administrator.
- **Data Discovery and Extraction** for data access by Data Consumers, who will be able to browse and extract only data belonging to their assigned scope of research. During this phase, the Data Privacy Administrator is responsible for approving new searches carried out by Data Consumers.

Privacy Query Platform uses a convolutional neural network to perform supervised learning in order to classify the images in the datasets to be anonymised. A model for Named Entity Recognition is used for the real-time classification and recognition of texts, which may contain sensitive information. This model allows the application of Artificial Intelligence techniques for text comprehension and the cataloguing of words into predefined classes.

THE DATA PROTECTION ENGINE

Privacy Query Platform is an essential solution to guarantee privacy protection when using data from systems that do not have their own anonymization or minimisation mechanisms. The solution also makes it possible to reinforce the native anonymization process of data sources that already apply these mechanisms by guaranteeing access only to information that is useful for statistical or research purposes.

AUTOMATIC PRIVACY CHECK

Automatic privacy checks, using Machine Learning techniques, performs automatic analysis of search results to ensure that the extracted data comply with the rules agreed with the privacy authority. The module analyses the information contained in a search result to identify the entity type represented by each individual field in the result. Several techniques are used:

- a priori classification of individual columns by the Data Provider
- classification based on the content of individual columns using a Natural Language Processing model based on Artificial Intelligence
- textual content extraction from unstructured data (e.g. images and PDFs)
- recognition and verification of confidential data within texts (e.g.: names, addresses, telephone numbers, age, gender)

GROUP DATA EXTRACTOR

The Group Data Extractor component, by means of the detailed analysis of the search results, ensures that the searches carried out contain results in aggregated form, with an aggregation level depending on the purpose of the search and consistent with the principle of minimisation. The results of the search will therefore be statistically relevant, compliant with privacy rules and deriving from a number of records greater than a given minimum (predefined for the for the search itself). The minimum number of data, which is essential to ensure respect for citizens' privacy, can be configured for different research fields. If they are not sufficient in number, the system does not allow the extraction of such data as it could be traced back to the natural persons to whom the information relates.

ANONYMIZATION

The Anonymization component uses the classification of columns contained in a search result to apply the most suitable anonymization according to the type of data:

- **partial anonymization:** part of the field value is replaced with non-significant characters (e.g. asterisks)
- **total anonymization:** the entire field value containing the sensitive data is replaced with non-significant characters
- **generalisation:** the field value is generalised through the use of a range of values (e.g. age 31-40 years, 41-50 years, ...)

Results’ anonymization is performed on data in transit without persistence of results, making the solution applicable to any type of source.

BENEFITS

- Privacy protection even from data sources that do not have their own anonymization or minimisation mechanisms.
- Configurability of search fields, usable data sources, minimum required aggregation level and anonymization rules.
- Automatic classification of data functionalities, to support users in privacy management, through the use of Artificial Intelligence techniques.
- Anonymization and minimisation of data in transit (without persistence), from heterogeneous sources, for compliance with the European Union regulation on personal data processing and privacy (GDPR).

2022 © Leonardo S.p.a. - MM09133 04-22

This publication is issued to provide outline information only and is supplied without liability for errors or omissions. No part of it may be reproduced or used unless authorised in writing. We reserve the right to modify or revise all or part of this document without notice.

For more information please email: cyberandsecurity@leonardo.com

LEONARDO