Synthetic Data Pilot with Banco de España

Findings and Recommendations

Written by:
Ivona Krchova
Dr. Paul Tiwald
Julio Zambrano
Andreas Ponikiewicz
Dr. Michael Platzer

September – 2022

MOSTLY AI

FISMA

EUROPEAN COMMISSION

Directorate-General for Financial Stability, Financial Services and Capital Markets
Directorate B – Horizontal policies

Unit B4 – Digital Finance

Contact: FISMA-B4@ec.europa.eu

European Commission
B-1049 Brussels

Synthetic Data Pilot with Banco de España

Findings and Recommendations

LEGAL NOTICE

Manuscript completed in September 2022

This document has been prepared for the European Commission however it reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

The European Commission is not liable for any consequence stemming from the reuse of this publication.

Luxembourg: Publications Office of the European Union, 2022

© European Union, 2022
Reuse is authorised provided the source is acknowledged. The reuse policy of European Commission documents is regulated by Decision 2011/833/EU (OJ L 330, 14.12.2011, p. 39).

For any use or reproduction of photos or other material that is not under the copyright of the European Union (*), permission must be sought directly from the copyright holders.
PDF ISBN 978-92-76-59303-4 doi: 10.2874/263371 Catalogue number EV-05-22-413-EN-N

EXECUTIVE SUMMARY

The European Commission, Banco de España and MOSTLY AI have successfully conducted a joint pilot on synthetic data during Summer 2022. MOSTLY AI provided a self-service synthetic data platform, that was installed and operated within the secure compute environment of Banco de España. Their staff then prepared and independently synthesized large-scale datasets successfully. The structure, the statistics and the patterns were very well retained, with a few noted exceptions, which are being addressed as part of this report. The valuable learnings from this pilot, and the corresponding recommendations by all stakeholders shall hopefully inform the continued establishment of a future project. We consider the initiative of utmost importance to foster cross-border, privacy-safe collaboration on Responsible AI within Europe.

1. INTRODUCTION

A thriving economy builds upon a healthy finance industry. Yet, the finance sector is also highly regulated, as the obtained information on individuals as well as organizations is highly sensitive. At the same time, this very same data provides new opportunities for smarter, data-driven, more-tailored services. From a privacy / confidentiality perspective it is important to protect data, but from an innovation perspective it creates a challenge when it comes to develop services & products, to collaboratively build fraud detection models, or to provide analytical insights for decision makers. Oftentimes, the collaboration across teams, across organizations, across borders is severely limited due to the inability to share granular-level data in a privacy-preserving manner.

1.1 MOSTLY AI

MOSTLY AI was founded in 2017 in Vienna, Austria, with a mission to empower all people with data to build a smarter and fairer future together for the benefits of everyone. At that time Europe was already preparing for the upcoming introduction of GDPR. While privacy and confidentiality is rightfully being recognized as a fundamental right that is to be protected, the broad definition of personal data was anticipated to severely hamper the development of an arising, yet crucially important data economy.

Inspired by advances in the field of synthetic image generation, and by building upon years of accumulated data science expertise both in academy as well as in the industry, MOSTLY AI successfully pioneered the field for AI-generated structured synthetic data. Their close collaboration with customers, in particular from the finance industry, allowed the company to develop their ground-breaking generative models further, and offer these as easy-to-setup, easy-to-use enterprise platform. By now the team has grown to a size of 49 employees across 22 nationalities, has received funding by leading European VC funds of more than €30m, and continues to lead this newly emerging category at a global scale.

1.2 Why Synthetic Data?

The dilemma between data protection and data utilization has always existed. And classic anonymization approaches, like data masking and data generalization, have been developed that allow a controlled trade-off between privacy and utility. E.g. the concept of k-anonymity provides an intuitive privacy guarantee, as it ensures that each released record is indistinguishable from k other records.

However, what has changed over the last decades is that the amount of data gathered for each individual has exploded in volume. And, as we need to consider each and every attribute that is associated with an individual as a quasi-identifier[1], concepts like k-anonymity are impossible to achieve in the era of big data. Any combination of seemingly innocuous data points allow to easily re-identify an individual even in large-scale datasets[2]. But, any classic data anonymization approach, that retains a 1:1 link to the original subjects, will hit ultimately a wall, due to an underlying mathematical law: the number of possible combinations grows exponentially with the number of data points per subject, and even for modestly sized databases quickly surpasses the number of atoms in the universe.

---

1 https://mostly.ai/blog/truly-anonymous-synthetic-data-legal-definitions-part-i/ (Platzer, 2020)
2 Simple Demographics Often Identify People Uniquely (Sweeney, 2000)

Thus, with a continuously growing number of data points collected, any individual in any database is increasingly distinctive from all others and therefore becomes susceptible to de-anonymization, no matter how much noise is being added[3]. A 2019 Nature paper[4] thus concluded:

> *"Even heavily sampled anonymized datasets are unlikely to satisfy the modern standards for anonymization set forth by GDPR and seriously challenge the technical and legal adequacy of the de-identification release-and-forget model."*

In addition, aside from basic linkage attacks, also AI-based model become available to carry out far more sophisticated re-identification attacks, that can succeed by merely relying on extracted patterns ("digital fingerprints"), as has been demonstrated recently[5].

In contrast to classic anonymization techniques, AI-generated synthetic data breaks the 1:1 relationship between the original and the released data subjects. This approach works by learning and distilling general patterns from an existing dataset to the sample based on these learned distributions completely new synthetic subjects, with synthetic data attributes. The result, if done right, is structurally identical, statistically representative yet truly anonymous data.

While the risk of re-identification is zeroed out, a provider of synthetic data still needs to control for any remaining risk of membership and/or attribute inference attacks. We, at MOSTLY AI, have from day 1 on engaged with 3rd party legal and technical experts to ensure the privacy compliance of the provided synthetic data[6]. In addition, every synthetization run is accompanied by an automated empirical privacy assessment, that establishes that also for the given dataset the generated data is not closer to the original data, than what can be expected based on a holdout data[7].

These advancements of AI-generated synthetic data were just recently confirmed by a JRC report on synthetic data for policy applications[8]. As part of their research, they successfully synthesized half a million of patient-level cancer records using MOSTLY AI and attested that synthetic data can become the key enabler of AI in business and policy applications within Europe.

---

[3] Unique in the shopping mall: On the reidentifiability of credit card metadata (De Montjoye et al, 2015)
[4] Estimating the success of re-identifications in incomplete datasets using generative models (Rocher et al, 2019)
[5] See https://mostly.ai/blog/synthetic-data-protects-from-ai-based-re-identification-attacks/ for an overview
[6] https://mostly.ai/blog/truly-anonymous-synthetic-data-legal-definitions-part-ii/ (Platzer, 2020)
[7] https://www.frontiersin.org/articles/10.3389/fdata.2021.679939/full (Platzer & Reutterer, 2021)
[8] https://publications.jrc.ec.europa.eu/repository/handle/JRC128595 (Hradec et al, 2022)

2. PILOT SETUP AND STEPS

2.1 Software Deployment

MOSTLY AI provided installation packages & instructions for version 2.2, their latest released version at the time of the pilot. The on-premises installation was then performed by the IT department of Banco de España. It is important to note, that MOSTLY AI (by design) neither had access to the system nor to the scoped data sets. For that reason, MOSTLY AI can also not share any analysis on top of the data, that goes beyond what is being shared by Banco de España.

2.2 Onboarding

The most focused part of the collaboration was between July 12 and August 6, the time from which our Synthetic Data Platform was fully operational until the deadline of the pilot. In this phase, the Data Science teams from Banco de España (the Banco de España Data Laboratory, BELab) and MOSTLY AI worked very closely together and had several touchpoints per week.

Most of the questions coming up were related to details of our AI-based synthesis approach as well as the privacy and accuracy metrics in the quality-assurance (QA) section of our platform. We spent a much smaller fraction of time on data pre-processing, i.e. discussing how to bring the data into a format that our platform can ingest. As soon as the latter question was clarified for the individual data sets that were part of the PoC, the BELab team was able to perform and trigger all data synthesis without further help. Questions related to operating our Synthetic Data Platform were hardly raised by the BELab team. They were enabled to become quite proficient users of the platform by the easy-to-operate user interface and the fact that, as soon as the input data is in the correct form little to no configuration of parameters is needed.

2.3 Data Preparation

A correct data format is the only requirement by our Synthetic Data Platform in order for the user to generate synthetic data. Apart from that, our synthesis approach and our platform are completely agnostic to the type of (tabular) data it is fed with. This opens our platform not only to any player and data sets in the financial market but also to any other domain such as insurance and the healthcare sector. If the EC plans to extend the synthetic data concept to other domains, the same technology can be used, and even creating synthetic, linked data sets across domains is possible.

Before ingestion of the original data into the MOSTLY AI synthetic data platform, two key requirements had to be met and jointly prepared by data scientists from BELab and MOSTLY AI:

1. Complying with **CSV format** requirements, as specified in the documentation. The provisioned data had to be appropriately formatted as column-separated, UTF-8 encoded data files, with missing values represented as empty strings, and dates represented in IS0-8601 format. Going forward, the need for this type of preparation will be greatly reduced, once data is either provisioned via a database or via properly typed Parquet files. This industry-wide file format is recommended to be used, as it's now also being fully supported from version 2.4 onwards.

2. Transforming the original multi-table star schema into a **two-table setup**. The original data used by Banco de España contains information about Spanish non-financial companies, their loan, and balance sheet data. All of that data is spread across several

tables. While the generative model is capable to retain any relationships between any number of attributes, within a table, and across tables, that information needs to be represented in a two-table setup. See figure 1. Thus, a handful of pre- and post-processing steps were required to shape the information into the required structure.

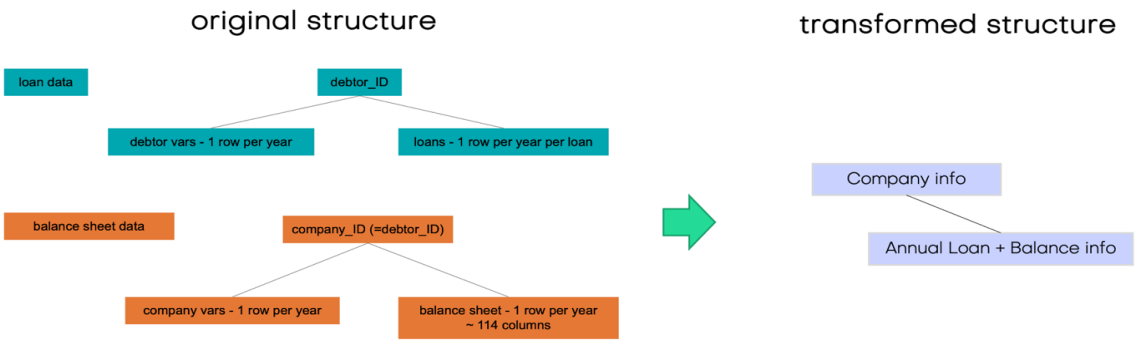

Figure 1 Schematic representation of the required data preparation.

2.4 Data Synthesis

Once the data has been accordingly prepared, the synthetization of the data itself is fully automated and easy to be performed end-to-end. See Figure 2. After the onboarding phase, the BELab team was able to perform and trigger all data synthesis without any further help. The synthetization process can take anywhere from half a minute to a full day. The runtime depends on the size and the complexity of the provided dataset, as well as on the underlying hardware.

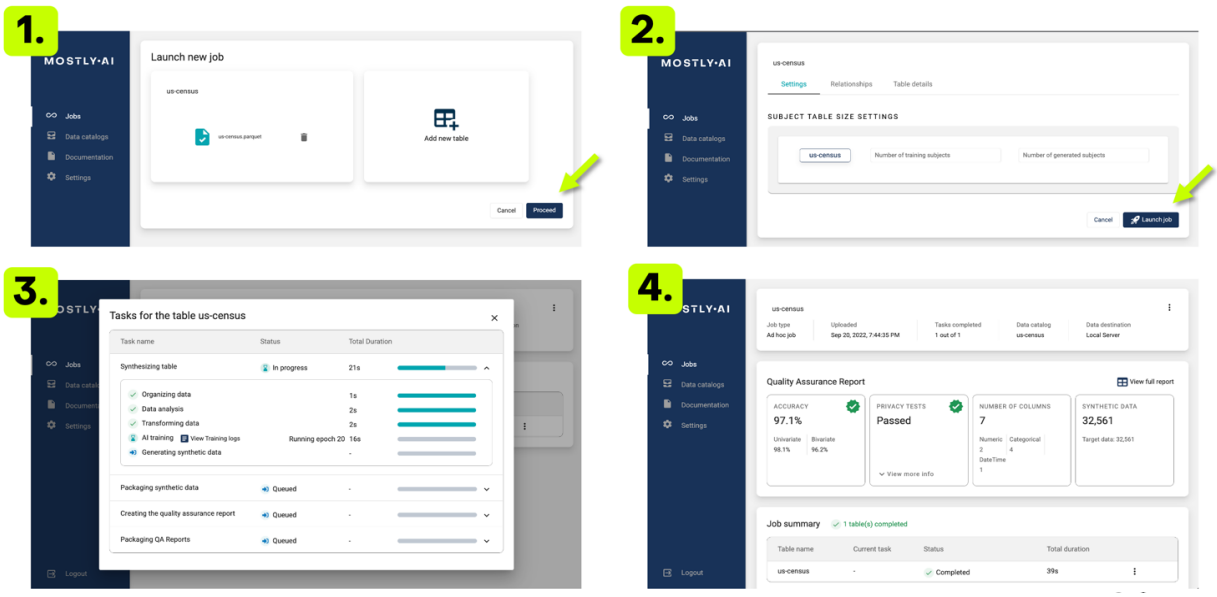

Figure 2 Web-based UI for synthesizing datasets end-to-end

2.5 Data Analysis

As MOSTLY AI (by design) neither had access to the system nor to the scoped data sets, all the analysis was done by the BELab team and, in the following, we report, summarize, and comment their quantitative results.

2.5.1 Univariate distributions, Bivariate distributions, and Correlations

The internal QA report of the MOSTLY AI Synthetic Data Platform shows very good results for univariate distributions, bivariate distributions, and feature-pair-wise correlations:

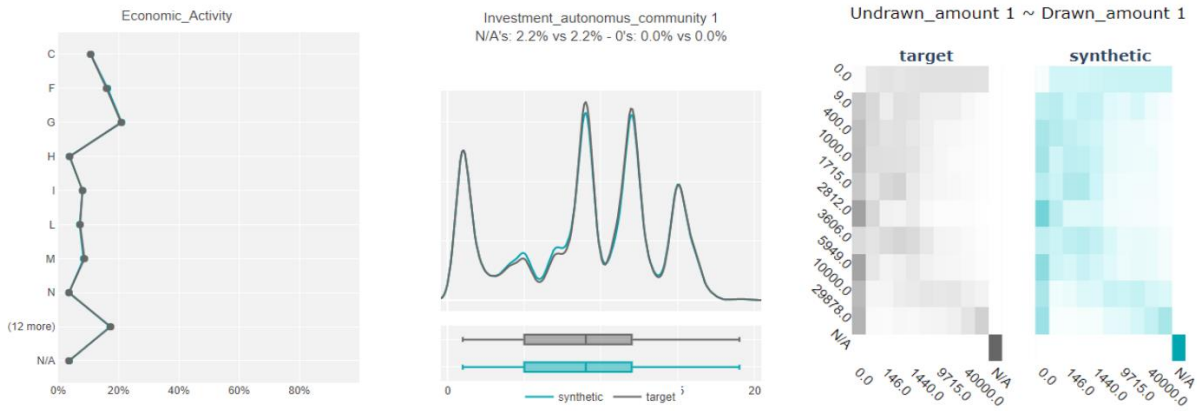

Figure 3 Comparison of original (grey) and synthetic (green) univariate and bivariate distributions of loan data within the MOSTLY AI Synthetic Data Platform

The BeLab team performed similar analyses outside of the MOSTLY AI Synthetic Platform that confirm the findings in the QA report.

2.5.2 Summary statistics of balance sheet data

The BeLab team performed further analyses on the balance sheet data, investigating summary statistics such as totals and, consequently, mean values of some of the attributes. For SMEs that make up the biggest part of the data set good agreement between synthetic and original data is reported. However, Banco de España observed discrepancies for the segment of large enterprises. It is known that the confidentiality of large enterprises, that dominate overall statistics, can only be guaranteed while adding significant noise to statistics. This is a general limitation for any (differential) private mechanism. However, this issue ONLY occurs for highly concentrated distributions, as for large enterprises, or as illustrated in Figure 3. For all other cases (in particular for SMEs, households, patients, etc.), also non-robust statistics, like the totals and the mean can be faithfully retained. This has also been confirmed by the observation that numerical values, which are represented as ratios, are indeed faithfully retained.

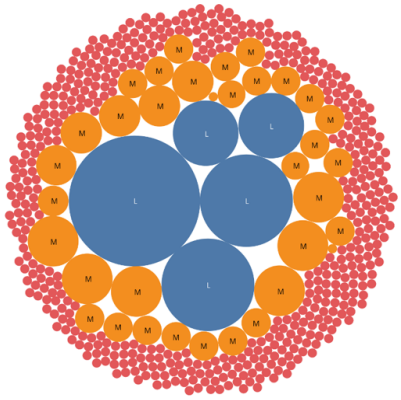

Figure 4 Illustration of the privacy challenge for highly concentrated distributions

2.5.3 Analysis of algebraic relationships

The data sets synthesized by the BeLab team contain many features that are linked via algebraic relationships. Banco de España reported some violations of these relationships within the synthetic data.

MOSTLY AI automatically detects and retains any possible statistical (=non-deterministic) relationships between any number of attributes. The generated synthetic data is only informed by the available data. Retaining dozens of deterministic, algebraic rules between numerical attributes (e.g. A+B=C) has been in the past 5 years of being in the field of synthetic data typically not been a use case for synthetic data.

We are at the forefront of research, and also plan to allow end users to incorporate pre-existing, known rules into synthetic data (see our recent paper[9]). An automated and guaranteed representation of an arbitrary number of algebraic relationships is currently not considered possible. Or if, then only by giving up on the representativeness of the generated data, which is commonly a key requirement towards synthetic data.

Nevertheless, we will direct our research efforts to investigate this area further, particularly how to reduce the severity of the misses, reported by Banco de España.

2.5.4 Further synthetic-data applications

The BeLab team and MOSTLY AI agreed that more time would have been desirable to explore the full potential of synthetic data and synthetic data use cases. A common use case for synthetic data is to provide truly anonymous, granular-level data to explore, train, and validate machine learning models. Particularly for credit records, MOSTLY AI has shown in the past, that the synthetic data exhibits the same utility as the original data for these types of use cases.

- In a published case study[10] on the detection of credit card fraud, MOSTLY AI demonstrated that synthetic data can even improve the accuracy of a downstream model compared to the original data, as it allows to increase the available training data.

---

[9] Rule-adhering synthetic data - the lingua franca of learning (Platzer, Krchova 2022)
[10] https://mostly.ai/blog/boost-machine-learning-accuracy-with-synthetic-data/

- In a joint study by MOSTLY AI and Nvidia[11] on validating machine learning models, it was demonstrated that synthetic data serves as a privacy-safe drop-in replacement for original data, and thus allows to safely conduct ML audits with 3rd parties.

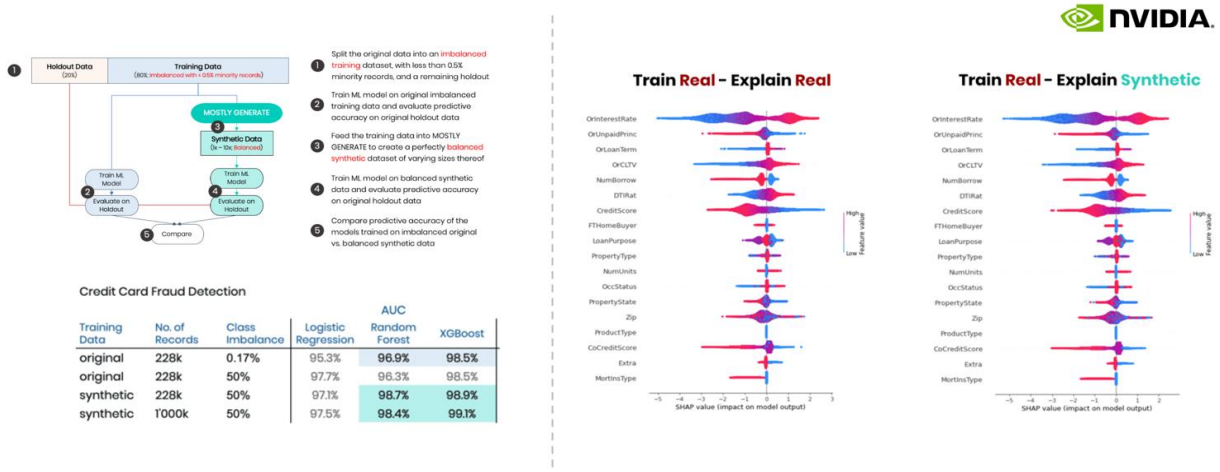

Figure 5 Machine learning use cases for synthetic data

3. FURTHER FINDINGS

- The guided installation & setup worked as expected

  The suggested alternative deployment option via Kubernetes or OpenShift is already part of our roadmap. This shall further drastically ease the setup and the operations of the MOSTLY AI platform.

- The self-service usage of the platform worked as expected

  We were happy to see that the high level of automation, and the ease of configuration allowed for a fast onboarding of users to the platform. The completely automated end-to-end process, which involves the analysis, the encoding, the training, and the generation of data, required little to no intervention.

- The automated quality assurance reports confirm a high match between all univariate and bivariate distributions

- The required pre- and post-processing of the original data structure for the use cases at hand did require close guidance from the vendor.

  However, these learnings can serve as blueprints with respect to the required two-table structure for a smooth roll-out to other data providers for future projects.

- Synthetic data is a novel, powerful approach, yet user acceptance also involves further education

  While classic anonymization approaches either remain vulnerable to re-identification (see the intro) or have to destroy a significant amount of information, synthetic data allows for a much-improved privacy/utility trade-off while remaining at a granular level. However, it needs to be acknowledged that also synthetic data underlies this

[11] https://developer.nvidia.com/blog/best-practices-explainable-ai-powered-by-synthetic-data/

fundamental privacy/utility trade-off, and a perfect privacy-preserving copy of a dataset is by definition impossible to achieve.

In addition, as there is typically no experience with assessing synthetic data, and no standardization has taken place yet in the industry, many questions do arise that relate to quality assurance. In version 2.4. of our product we have further improved our reporting significantly, and continue to support standardization efforts (among others, also as part of a IEEE workgroup)

4. RISKS AND RISK MITIGATION STRATEGIES

We conclude this report with the following list of recommendations:

- Continue plans for a cross-border unified data synthetisation project to foster innovation and collaboration. The need for a free flow of data, and thus information, to strengthen European sovereignty and competitiveness with respect to Responsible AI development is bigger than ever.

- Drive education on synthetic data, and motivate its need early on in the process. Transparent communication around benefits and limitations is essential to foster user acceptance.

- Focus on ML use cases, where synthetic data is shared to collaboratively and safely develop, train, and validate AI models.

- Focus on analytical use cases, where single entities do not dominate overall statistics. In particular, for the given use case, we strongly recommend to limit the analysis to small and medium-sized organizations.

- Leverage this pilot as a blueprint for the required data structures for other national entities, to smoothen the onboarding. A standardization of column names and attribute values can help but is not strictly required up-front.

- Take into account quantifiable accuracy & privacy targets. On the one hand, it is easy to be private, if utility is not considered. On the other hand, providing the most accurate data is trivial, if privacy is not respected. This needs to be taken into account, and frameworks do exist that help to measure the differences[12]. At the end of the day, it is the quality of the synthetic data, that will determine the success of the downstream use cases.

[12] Holdout-Based Empirical Assessment of Mixed-Type Synthetic Data

GETTING IN TOUCH WITH THE EU

In person

All over the European Union, there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact/meet-us_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:
- by Freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 2 299 96 96, or
- by email via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications from: https://publications.europa.eu/en/publications.

Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1952 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

Open data from the EU

The EU Open Data Portal (http://data.europa.eu/euodp/en) provides access to datasets from the EU. Data can be downloaded and reused for free, for both commercial and non-commercial purposes.

EV-05-22-413-EN-N

Publications Office
of the European Union

ISBN 978-92-76-59303-4
doi: 10.2874/263371