Annex C. Practical examples of data sandboxes

Spain

Banco de España participated in the synthetic data pilot led by the European Commission with the aim of supporting the creation of the future EU Digital Finance Platform's Data Hub. The pilot consisted of testing the procedure of generating synthetic data on the Central Balance Sheet (CBI) and Loans to Legal Entities (CIR) datasets available in the Banco de España's (BELab) data laboratory. The initial results obtained during the pilot were overall positive and encouraging for further exploration of the technology. However, some limitations were identified in the data such as suppression of outliers, imbalances among variables, and the impossibility of merging synthetic datasets.

The pilot highlighted that synthetising data should be considered as a cross-sectional project requiring collaboration from multiple departments within the supervisor and commitment from the software provider. Although BELab shortened the time needed to synthetise the data, the process was highly time and resource-consuming and required highly specialised staff.

Banco de España believes that synthetic data could be a valuable contribution to the sandboxes and cross-border testing, but the limitations detected in the data may hinder its use in certain cases. Therefore, market participants should be aware of these limitations. The results of the pilot should not be taken as the sole source of information when specifying the conditions of the tender or building the future data hub. Further analysis and more time are needed to obtain more reliable conclusions.

United Kingdom

The Financial Conduct Authority (FCA), together with the City of London Corporation, has successfully established a digital sandbox, with two pilot projects already underway offering access to synthetic data sets for testing and PoC development. The first pilot project ran from October 2020 to February 2021 and focused on improving SME finance, detecting and preventing fraud and supporting the financial resilience of vulnerable customers. The second pilot ran from October 2021 to March 2022 and focused on financial innovation linked to sustainability. In addition to synthetic data sets, the pilots offered participants access to a range of other development tools, such as APIs, programming environments, as well as access to expert mentors and observers.

The synthetic data was the most valuable feature cited by participants while simultaneously the one with greatest potential for improvement. Notably, referentially linked data sets and more granular data would enable more effective testing, and for products to be developed further. Overall, 84% of responders cited the pilot as having accelerated their product development. While it is difficult to ascertain or quantify this level of acceleration, analysis shows that the biggest factor was ready access to data in developing an early stage PoC. Several participants estimated they had accelerated their development by 4-6 months, with one going as high as 18-24 months, largely by negating the initial need to identify and work with an industry partner to get a PoC off the ground, or sourcing or generating data themselves.

Those who found that the pilot tools had not accelerated or improved their development generally noted that this related to the data being insufficient in some way for meeting their use case. This was as a result of one or a combination of the following:

Insufficient detail in the data, particularly a lack of relevant typologies or behaviours they required to model their solutions. For example, transactional spending patterns indicative of a customer experiencing fluctuating mental health (under the vulnerable consumers use cases) were not seeded into the synthetic data due to the complexity of doing so.

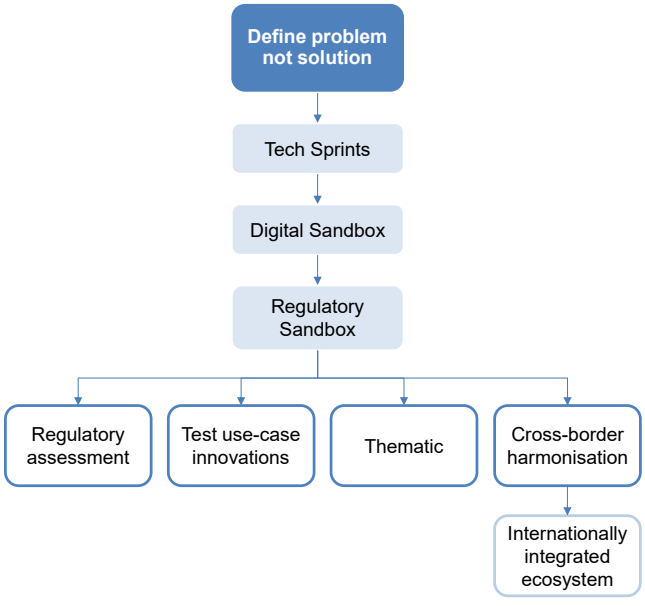
Required data sets were not available. For example, large volumes of unstructured data such as consumer complaints text, to train and validate natural language processing techniques.

Data sets not being referentially linked. For example, different data sets had been generated independently, so the behavioural patterns or characteristics of a synthetic individual 'John Smith', would not be consistent with 'John Smith' in a separate data set.

However, even with these limitations, participants noted the utility of the data sets for 'bootstrapping' product design. Even where the data could not be used to refine an algorithmic model, there was value in providing the data models, data structures and formats that were representative of what they would be working with in real production environments (FCA, 2021[1]).

It should also be noted that in the United Kingdom, the process of developing FinTech solutions has been optimised through years of experience and innovation in the sector. A systematic approach has been established, starting with the definition of a problem and letting FinTech companies explore possible solutions. The next step is a Tech Sprint, which involves defining the problem and providing datasets to develop proof of concept (PoC). The Digital Sandbox phase builds on the PoC, further testing the potential solutions and their viability. Finally, the Regulatory Sandbox assesses the regulatory requirements for the minimum viable product (MVP) developed during the previous stages. This step-by-step process provides a clear path for the development and implementation of innovative FinTech solutions, allowing for a seamless transition from concept to market-ready product (Figure A C.1).

Figure A C.1. Digital and regulatory sandboxes based on the systematic approach of the UK FCA



There are several countries that have implemented data sandboxes, including Australia, Singapore, the United Kingdom, and Canada. However, the only sandbox targeted at financial innovation that has offered participants access to financial data thus far has been the FCA, with the last two cohorts offering a data attribute.

Australia

The Australian Competition and Consumer Commission (ACCC) has established a consumer data right (CDR) Sandbox[1] to support the development and testing of CDR-related solutions by businesses. The Competition and Consumer Commission (ACCC) has created a “mock” registry, “mock” data holder, and “mock” data recipient CDR repository to help businesses develop and test CDR solutions within their own IT environment. This mock environment has been downloaded more than 20 000 times from GitHub and has received positive feedback from CDR participants and platform providers. The ACCC is now launching a hosted sandbox environment to build upon these mock tools, allowing businesses to test CDR solutions with other participants and to validate technical solutions early in the software development cycle. This sandbox has the potential to lower the barrier for businesses to join the CDR and increase the quality of solutions. The ACCC anticipates the sandbox will be broadly adopted by CDR participants and enable common platforms to be tested against other systems prior to widespread implementation.

References

FCA (2021), *Supporting innovation in financial services: the digital sandbox pilot*, https://www.fca.org.uk/publication/corporate/digital-sandbox-joint-report.pdf (accessed on 20 March 2023). [1]

Note

[1]Australia lists the type of data available for the sandbox through the following link: https://github.com/ConsumerDataRight/sandbox.