Regulatory Sandbox Final Report: Smart Data Foundry

A summary of Smart Data Foundry's participation in the ICO's Regulatory Sandbox

Date: September 2023

ico.

Information Commissioner's Office

Contents

1. Introduction

1.1 The Regulatory Sandbox ('the Sandbox') is a service the ICO provides to support organisations that are developing products or services which use personal data in innovative and safe ways, and will deliver a potential public benefit.

1.2 The Sandbox is a free, professional service that is available to organisations of all sizes who meet our entry criteria and specified areas of focus. We assess these criteria via our application process.

1.3 The Sandbox specifically sought projects operating within challenging areas of data protection. Sandbox participants have had the opportunity to engage with us, draw upon our expertise and receive our advice on mitigating risks and implementing data protection by design and default into their product or service. This helps ensure that appropriate protections and safeguards are in place prior to live processing.

1.4 Smart Data Foundry (previously known as 'Global Open Finance Centre of Excellence') is a wholly owned subsidiary of the University of Edinburgh. They are a data innovation organisation that serves the public, private and third sector. Its purpose is to improve people's lives and inspire financial innovation by safely unlocking the potential of financial data to enable research, innovation and skills development, supplying real data for research and synthetic data for innovation. In the Sandbox, Smart Data Foundry received advice to ensure that their research data facility complied with the UK GDPR principle of lawfulness, with particular consideration given to the research provisions and anonymisation.

1.5 Smart Data Foundry were accepted into the Sandbox in December 2021. The ICO determined that Smart Data Foundry's project aligned with several of the key areas of focus of the Sandbox at the time of application including data sharing in finance, protecting people with vulnerabilities and the use of privacy enhancing technologies.

1.6 The ICO and Smart Data Foundry agreed to work on the following objectives as part of Smart Data Foundry's bespoke Sandbox plan:

- **Objective one:** Consider the data protection implications of creating a repository of financial data to be processed for research purposes.

- **Objective two:** Consider the data protection implications of linking multiple datasets provided by multiple data controllers to support UK researchers across different research domains and use cases.
- **Objective three:** Consider the data protection implications of creating synthetic datasets from data held in the above repository to help support social and economic innovation.
- **Objective four:** Work with partners to test the utility and efficacy of the proposed documentation on an agreed project spanning the three previous objectives. Due to difficulties obtaining appropriate candidates for a practical test within the timeframe of the Sandbox project, we determined that this objective would be moved outside the scope of the Sandbox with the opportunity to use the exit report as a starting point for Smart Data Foundry's conversations with future data providers.

2. Product description

Smart Data Foundry's product is comprised of two parts; the first is the research facility, and the second is the innovation service which provides synthetic data for further research opportunities.

Research Data Facility

2.1 The purpose of the Research Facility is to provide a secure data storage and processing capability where separate, combined or linked datasets of pseudonymised and anonymised granular financial data – termed in Smart Data Foundry as 'research ready data' are assembled in appropriate technical data infrastructure. Datasets in the research database can be used for statistical and research purposes and support UK researchers across many different research domains and use cases. Smart Data Foundry intends to use the research facility to assist the linkage of multiple datasets that have been input into the research facility by different data providers. The data providers are mostly financial institutions such as banks and credit agencies; currently Smart Data Foundry has onboarded Equifax, FreeAgent and Moneyhub as data providers for the purpose of the research database, while having purpose-specific research agreements with NatWest Group and SAGE. An agreement with NEST Insights was also in place for a project that is now complete.

2.2 The development of the research facility involves the following steps; firstly, the scope and objectives of a project is identified either by Smart Data Foundry's in-house research data science team or a collaboration with an academic researcher, which highlights the need for a specific dataset. After assessing whether the data is already present within Smart Data Foundry's research facility due to a pre-existing agreement, the data will be made available to researchers for specific research projects that must meet the criteria of Smart Data Foundry's specified missions (see 2.8 for Smart Data Foundry's organisational missions). If the project requires a dataset that is not currently available, Smart Data Foundry liaises with a data provider that can provide the necessary financial data. They will carry out appropriate due diligence prior to signing a data sharing agreement for the ingress and continued processing of the shared data for research purposes. The data will be securely transferred into Smart Data Foundry's technical environment (provided by the Edinburgh International Data Facility) where Smart Data Foundry will undertake data quality checks to make sure it matches the data specified for the research purpose and that no additional personal data has been added beyond what is required. Smart Data Foundry will also carry out identifiability assessments to determine if further pseudonymisation or anonymisation techniques are needed (see s.3.28-31). When results from the specified research missions are ready to share with the wider public, the results are checked by the Smart Data Foundry information governance team to ensure the exported data is sufficiently aggregated to be anonymous.

2.3 The personal data involved in this processing may include:

- Pseudonymised identifiers as a subset of financial data;

- Indicators of financial activity such as specific transactions and vendors, including categorisation of income and expenditure (such as housing, tax, salaried income, benefit income);

- Personal characteristics such as age, gender and location; banded so as to prevent re-identification;

- Indicators of vulnerability such as disability or benefit provision may be processed, where a research mission requires this; and
- Special category data such as health or ethnicity may be utilised, where a research mission requires this.

Innovation Service (Generating synthetic data)

2.4 The second service that was included in Smart Data Foundry's participation was the generation of synthetic data for innovation. The primary purpose of this service is to create high-quality innovation opportunities from data held in the research database (see 2.1 – 2.3), including exploring the creation of anonymous models and synthetic datasets to help support social and economic innovation.

2.5 The undertaking of the service is to explore creating and generating synthetic doubles or 'twins' of datasets provided to Smart Data Foundry. These datasets would primarily be developed from pseudonymised financial datasets, as well as developed from reference data and metadata gathered from previous data analysis. Each dataset will be subject to an anonymisation assessment as well as privacy enhancing techniques and software developed by academic and technological experts.

2.6 There are broadly two approaches to the creation of these synthetic datasets:

- The use of simulation – known as 'agent-based modelling' - where data is generated from approximations and predictions of behaviour using characteristics given to a computer generated population to understand how they would interact. This processing does not use personal data beyond some aggregate information generated from real data to test and improve parameters. This is the synthetic data approach that Smart Data Foundry is already using.
- Using 'learning-based' synthetic data generation to create synthetic doubles of existing datasets utilising differential privacy and modern learning-based approaches which look to (1) learn all the meaningful patterns in data, and (2) use this learnt knowledge of patterns in the original data to generate new data that exhibits similar patterns, without recreating any of the

input data. It is this second form of synthetic data generation that was the focus of Smart Data Foundry’s ICO Sandbox participation as it relates to processing existing datasets which include personal data. This could be an approach that Smart Data Foundry implements in the future.

Objective and Benefits

2.7 Smart Data Foundry’s objective for making data available for research is to solve societal, economic and environmental problems that are broadly defined as ‘in the public interest.’ Smart Data Foundry outlined the purpose of the above processing activities as the following organisational missions:

- **Stop the Squeeze:** To help UK households and small businesses, particularly people with vulnerabilities, withstand the rising cost of living post-COVID and Brexit. By providing data on financial wellbeing and resilience, Smart Data Foundry will help the government and industry take real actions to assist people on the poverty line to survive and thrive.

- **Countering Climate Change:** Understanding how satellite data can be better interpreted to reveal patterns of activity on the planet’s surface that indicate positive or negative climate change impact.

- **Igniting Innovation:** Smart Data Foundry’s synthetic data engine is designed to accelerate innovation using high utility synthetic data to unlock progress and foster collaboration. This approach allows innovators to use synthetic datasets to prove ideas, build and release new products with confidence, safety and speed.

3. Key data protection considerations

3.1 During its Sandbox participation Smart Data Foundry and the ICO considered a number of key data protection themes in relation to the research facility and innovation products. The following sections represent the key data protection considerations that were considered in Smart Data Foundry’s participation and the work undertaken to address them.

Purpose compatibility (Research)

3.2 A key challenge for Smart Data Foundry was how to comply with the UK GDPR 'purpose limitation' principle when carrying out further processing on the data utilised in specific research projects, for example using these for additional secondary research.

3.3 Specifically, the challenge was applying the purpose limitation principle in a research context. Article 5(1)(b) UK GDPR states that processing data for research-related purposes is compatible with the original purpose. This means that an organisation will not need to determine a new lawful basis unless the original lawful basis is consent, in which case the specific research activity needs to be identified at the point the personal data is collected.

3.4 The ICO highlighted the key elements from the draft research guidance (now published) that would help Smart Data Foundry apply the purpose limitation principle in practice. Smart Data Foundry will be conducting research and maintaining the research facility using data collected from another organisation, they are not repurposing data they have collected directly from the individual and will therefore need to identify their own lawful basis to process the data and cannot rely on the assumption of compatibility with the original organisation's purposes.

3.5 After determining that Smart Data Foundry would need a lawful basis for processing, and could not rely on purpose compatibility, the next issue to address in the Sandbox was what lawful basis Smart Data Foundry could seek to utilise for their research-based processing activities instead. As per the ICO's research provisions guidance, the lawful basis for research-related processing activities is either Article 6(1)(e) 'public task' or Article 6(1)(f) 'legitimate interests'. Given that Smart Data Foundry is not a public authority and the task did not have a clear basis in law it was determined that legitimate interest would be the most appropriate lawful basis for processing. Smart Data Foundry had already produced a legitimate interests assessment ('LIA'). Our guidance on legitimate interest assessments can be found here.

3.6 The next task was to understand what this would mean for the data providers working with Smart Data Foundry. It was established that the financial organisations acting as data providers (the original controllers of the personal data) will need to conduct a compatibility test for the transfer of the data to Smart Data Foundry for research purposes. This can be done as part of a Data Protection Impact Assessment to assess compatibility of the new research purpose against the original collection purpose. Since the purpose limitation principle says you can reuse existing personal data for research-related purposes as long as you have appropriate safeguards in place, the assessment is likely to be compatible. Alternatively, the providers should declare the data sharing as a specific purpose at the point of collecting the data from the individual if they are relying on consent as a lawful basis.

Special category data (Research)

3.7 Smart Data Foundry raised that some of the datasets they will be attempting to link, process and store may include fields of special category data (in particular when assessing the impact of health conditions on individuals' finances and researching algorithmic bias). The ICO advised Smart Data Foundry how to document that they were using the appropriate Article 9 condition for processing (in particular around the application of the research condition and Smart Data Foundry's alignment to the ICO's research guidance). Smart Data Foundry identified Article 9(2)(j) as their condition of processing. This article provides a condition for processing special category data if it is necessary for research or statistical purposes.

3.8 During our review of Smart Data Foundry's DPIA we identified that the requirements for processing special category data need further consideration. Whilst some effort had been made to address the requirements for relying on this condition found in Schedule 1 Paragraph 4 of the DPA 2018 there was not enough detail to address the individual aspects of meeting the condition. We asked Smart Data Foundry to provide further information about how they proposed to meet each of the requirements such as:

- Measuring public interest not just in congruence with the research organisation's interests but to the wider public or society as a whole.
- Demonstrating that the research is scientific in nature.
- That technological and organisational measures ensure data minimisation as well as safeguard individual rights.

- Evidencing why anonymous data cannot be used and that they have considered whether they could use pseudonymisation as a safeguard to make it more difficult to link the personal data back to specific individuals.

3.9 In response, Smart Data Foundry provided further explanation as to how they would meet the conditions for the Article 9(2)(j) condition for processing. They agreed that where a research mission requires that they process special category data and it is necessary for scientific research or statistical purposes, they will carry out a legitimate interest assessment to demonstrate the necessity of processing that special category data. In addition, they will make sure that the appropriate safeguards are in place to uphold individual rights, including identifying and establishing that a project is not likely to cause substantial damage or distress to an individual and that special category data will not be used to take measures or decisions in relation to particular individuals. Smart Data Foundry will work with their governance structure to ensure that the public interest is congruent with wider public societal interest and demonstrate that the research is scientific in nature by relying on the expertise of academic staff. To ensure that all of these elements of meeting the condition for processing special category data under Article 9(2)(j) have been met they will be considered in the data protection impact assessment of the specific research mission.

Retention

3.10 A key data protection challenge to the Sandbox engagement, was helping Smart Data Foundry to understand whether it is possible to produce an 'enduring' database for active and future research which complies with the UK GDPR's retention (storage limitation) principle.

3.11 The storage limitation principle means data should be deleted when it is no longer needed. We discussed with Smart Data Foundry that to comply with this principle when they are carrying out a limited specified research activity, they will need to destroy the data at the end of the defined period and when the specific purposes and needs of the research activity are complete.

3.12 Although the general rule is that you cannot hold personal data indefinitely just in case it might be useful in future, Article 5(1)(e) provides an exception to the principle of storage limitation for research related processing. This means that personal data can be retained indefinitely if you are processing it for one of the research related purposes. Due to the exception for research processing in the storage limitation principle, as a data controller Smart Data Foundry would be able to process the data obtained from third parties indefinitely in order to maintain the research database. This data could not be used for any purpose other than research. The database would require appropriate organisational controls in place to prevent the misuse of the research data for reasons not covered by the research provisions in UK GDPR Article 5(1)(e). They will also apply technical measures to keep the data aggregated. In addition, they will carry out anonymisation assessments to reduce the risk of identifiability of the personal data or determine that it has been effectively anonymised. These anonymisation assessments are discussed further below.

3.13 The ICO advised Smart Data Foundry that a practical application of the retention principle means that they would need to have a policy in place to set the requirements that demonstrate when data needs to be kept indefinitely for research and when that data is no longer needed. To ensure that this principle can be upheld while also fulfilling their functions in research, Smart Data Foundry will assess the following criteria when deciding whether to retain a dataset for future research purposes:

- Is this dataset currently utilised to deliver an ongoing service or product?
- Have all elements of the dataset, i.e. individual fields and length of time, been utilised as part of a wider research bid or innovation project within a six-month period?

3.14 To enable these checks to function, they will conduct a quarterly review of datasets against these criteria, as well as ensure all research bids and innovation projects are congruent with their organisational missions.

Applying the research exemption provisions in the GDPR

3.15 The ICO advised Smart Data Foundry on how to apply the research exemption when considering individual rights. This work supported the identification of the correct research exemptions under the UK GDPR, and Smart Data Foundry updated their documentation to ensure the correct provisions were associated with each individual right.

3.16 Firstly, the ICO helped Smart Data Foundry to apply the research provisions on transparency to Smart Data Foundry's processing. Smart Data Foundry was aware of the exemption in Article 14(5)(b) which provides an exception to individuals' right to be informed when Smart Data Foundry receives their personal data from a third party if providing the transparency information would require disproportionate effort or seriously impair the purpose of processing; read our guidance on this research exemption. The ICO advised Smart Data Foundry that when determining that the provision of transparency information to individuals would be infeasible under the research exemption this should not be done in a blanket fashion. It should be subject to a written assessment setting out why the provision of transparency information would be infeasible. In response, Smart Data Foundry have included in their DPIA that where the information processed is proven to be identifiable they will make assessments in each individual research mission to determine if the volume of individuals makes sending transparency information reasonable rather than a blanket application of Article 14(5)(b).

3.17 The ICO also advised Smart Data Foundry that the DPIA should include an assessment of whether the processing would be within the reasonable expectations of the individuals impacted by potential harms. Following this Smart Data Foundry identified that, while the primary source of transparency information would be the data provider, it would be useful to have their own privacy notice for individuals who are aware of the further processing.

3.18 The research exemption for facilitating rights to access, rectification and objection was considered next. The research exemption in Schedule 2, Paragraph 27 DPA 2018 will only apply:

- to the extent that providing access to the data would prevent or seriously impair the achievement of the purposes for processing;
- if the processing is subject to appropriate safeguards for individuals' rights and freedoms;

- if the processing is not likely to cause substantial damage or substantial distress to an individual;
- if the processing is not used for measures or decisions about particular individuals, except in the case of approved medical research; and
- if research results or any resulting statistics are made available in a way that does not identify individuals.

3.19 The ICO asked that Smart Data Foundry amends the DPIA to reflect that the right to erasure has a built in research exemption in Article 17(3)(d) which states that if you are processing data for research related purposes, the right to erasure does not apply as far as giving effect to the right is likely to render impossible or seriously impair the achievement of your research objectives.

3.20 Smart Data Foundry were advised in their DPIA review that individual rights should only be restricted if the specific exemption for that individual right applies. Even though most processing will likely fall in scope of the research exemption, the DPIA must not suggest that the exemptions will be applied in a blanket fashion instead of considering the application of the exemptions on a case-by-case basis.

3.21 Smart Data Foundry should also make sure that they are meeting Article 89 UK GDPR conditions which require having appropriate safeguards such as data minimisation, pseudonymisation and anonymisation techniques in place where the processing is carried out for research purposes. Smart Data Foundry is going to document this requirement within the specific DPIAs completed for each specified research activity as part of an identifiability assessment for the data processed.

Controllership

3.22 Smart Data Foundry asked the ICO to advise them on their controller designation in the context of processing for the Research Facility (i.e. whether Smart Data Foundry is an independent controller or a joint controller) and if this designation would be subject to change when linking different datasets. An example would be linking financial data obtained from one company with a dataset from an energy company to analyse cost of living being impacted by energy prices.

3.23 Where Smart Data Foundry has been commissioned to carry out research on behalf of an organisation that is already a controller for that data, the ICO advised that Smart Data Foundry's role in determining the means of processing – such as the information that is collected and the manner in which the research is carried out – means that it is likely to be a joint controller. The commissioning organisation would retain control of the data because it commissions the research and determines the purpose the data will be used for, but the degree of agency that Smart Data Foundry has in the research design leads them to be a joint controller. Smart Data Foundry confirmed that they have a large role in research design so they would likely be a joint controller, read more in our guidance on how to determine whether you are a controller.

3.24 In contrast, Smart Data Foundry would be acting as an independent controller in regard to creating and maintaining the research data facility. This research data facility holds datasets from different third party data providers that have been linked for research purposes.

Data Linking (Research)

3.25 Smart Data Foundry will use the research facility to evaluate multiple datasets for deeper insights and to create linked datasets where the research purpose requires this. A key aim of Smart Data Foundry's participation was to consider the data protection risks and mitigations when linking datasets provided by multiple data controllers.

3.26 The ICO noted that when Smart Data Foundry is linking datasets within the research facility, the invasiveness of the processing and the risk of identifying an individual are likely to increase. Smart Data Foundry were advised that both the LIA and the DPIA should be reviewed when new datasets are added or linked to the environment to ensure that the data processing remains proportionate and necessary, and that any additional risks to individuals are effectively mitigated. This includes ensuring the technical environments are suitably secure.

3.27 Using the information provided by the ICO, Smart Data Foundry took the decision that they would only link the datasets in the research facility if they had a specific purpose that strongly aligned with their organisational missions and where they have sufficiently pseudonymised (anonymised where possible) the data in line with the identifiability assessments (see ss3.28-31 below).

Identifiability

3.28 Any processing for research should consider the possible role of pseudonymisation and anonymisation as appropriate safeguards to the processing as found in UK GDPR Article 89(1). Smart Data Foundry will need to carry out identifiability assessments to ensure the data can meet the bar of 'effectively anonymised' where possible. If this bar is not met, the data should be treated as personal data and have accompanying provisions around transparency and individual rights as appropriate.

3.29 The identifiability assessment is a qualitative anonymisation tool to help identify the risk of re-identifiability. The first stage of the assessments will focus on the three key indicators of identifiability from chapter 2 of the ICO's draft anonymisation guidance: singling out, linkability, and inferences. The ICO advised Smart Data Foundry on how to develop their assessment of these three factors following a review of some existing examples of Smart Data Foundry's identifiability assessments.

- **Singling out:** Singling out means that you are able to tell one individual from another individual in a dataset. For example, if you can isolate some or all records about an individual in the data you process, then that individual is singled out. The ICO advised Smart Data Foundry that singling out does not have to identify an individual by name. If a person's record can be individuated from others in a dataset it is still singling out. Smart Data Foundry would need to make sure that their assessments considered the richness of the data and how potentially identifying different categories are.

- **Linkability:** Linkability is the concept of combining multiple records about the same individual or group of individuals together. Individual datasets may seem non-identifying in isolation but can lead to the identification of an individual if combined.

- **Inferences:** An inference refers to the potential to infer, guess or predict details about someone. In other words, using information from various sources to deduce something about an individual (eg based on the qualities of others who appear similar). Smart Data Foundry were recommended to pay particular attention to how they deal with outliers. For example, high earners or individuals with large debts could be outliers in a dataset and there could be a risk that such an outlier could be connected by the media to a high profile bankruptcy or wealth.

3.30 Secondly, the assessments would look at specific technical and organisational measures to determine disclosure risk. These include access control, the motivated intruder test (see "What is the "motivated intruder" test?" in our draft anonymisation guidance) and the form that data will take when it is disclosed.

3.31 Finally, the assessments will have a mitigation stage for the risks identified in the first two stages. Smart Data Foundry will seek to apply anonymising or pseudonymising techniques as mitigations. These include techniques such as extreme value filtering, reducing information in a field (i.e. partial postcodes), aggregation, differential privacy (Laplacian noise) and random category generators.

3.32 When the anonymisation assessment is complete they will be able to make a decision as to whether the data has been effectively anonymised or needs to be classified as personal data. Smart Data Foundry's data classification reflects that confidence in the identifiability of the data will range from pseudonymised data with a risk of identifiability to data that has been effectively anonymised. The weighting of the assessment is classified according to the risk of disclosure and the subsequent harm of disclosure with the effectiveness of the mitigating anonymisation method and remaining risk of identifiability. The ICO recommended that this conclusion will need to consider whether the data is identifiable both in the hands of Smart Data Foundry and also in the hands of the researcher seeking to use the outputs of the research database. As the researcher will also need to assess the identifiability of the data it may be appropriate, and indeed more practical, for both organisations to undertake this assessment jointly.

3.33 In the anonymisation assessment, Smart Data Foundry considered that the combination of controls and anonymisation techniques reduced the risk of re-identification in the hands of the researcher within the research data facility such that it was considered sufficiently remote (and therefore effectively anonymised). They determined that the researchers would have no access or way to link additional information and no means reasonably likely to be used to obtain it for a specific dataset. They reached this conclusion on an assessment of a specific banking dataset and are aware that if other datasets are available and more identifiable, then this assertion may no longer be true for those datasets. Smart Data Foundry would also need to consider the risk of linkability between the available datasets.

Synthetic data: anonymisation (Innovation)

3.34 Using the ICO's draft guidance on ensuring anonymisation is effective, the ICO considered whether the methods utilised by Smart Data Foundry to synthesise datasets were sufficient to meet the bar of anonymous data.

3.35 Smart Data Foundry were advised that they should demonstrate that they have evaluated the potential harms that could result if synthetic data is shown to enable identification and also the mitigations that they have implemented. This advice was given following a review of their LIA.

3.36 The LIA indicated that use of differential privacy was being considered which could be used as an example of the extent to which an organisation is seeking to mitigate against risks of identifiability in synthetic data. However, the ICO advised that relevant information should also be included around the risks related to identifiability when using differential privacy. For example, managing the balance to determine a privacy budget that does not impact the utility of the data whilst ensuring the risk of re-identification is sufficiently remote. It was considered that the DPIA would benefit from addressing matters such as the privacy budget that would be set and whether differential privacy is being applied globally (to an aggregated dataset), or locally (to each individual data record).

3.37 In response, Smart Data Foundry amended the LIA and created a supplementary policy document that shows how they will handle the risk of disclosure when making synthetic doubles. The document provides comprehensive detail of how Smart Data Foundry uses learning-based synthetic data generation to create synthetic doubles of existing datasets. For example, it includes the following steps:

- Technical measures to reduce the identifiability of the data before starting. This includes removing direct personal identifiers such as names and addresses, and replacing any unique identifiers. Data providers would be asked to do this before supplying the data to Smart Data Foundry.

- Consideration of applying differential privacy to the input data before supplying it to the machine-learning system. They would have to choose appropriate parameter values used in differential privacy, and that might require iteration.

- Their identifiability assessment would use the three indicators of identifiability from the ICO's draft anonymisation guidance (singling out, linkability and inferences). This was advice given by the ICO when reviewing Smart Data Foundry's LIA.

3.38 The document also offered many proposed metrics for measuring the risk of re-identification (disclosure). Smart Data Foundry will select a suitable subset of these metrics and compute them to test if they reveal a risk of disclosure. These tests would require access to both the synthetic and the raw (mostly pseudonymised) data in the research facility database. They would test whether any whole records from the input data are present in the output data but would also utilise other metrics that measure less observable indicators that there is a risk of identifiability in the synthetic data. The document details how they would graph the original data and the synthetic double to look visually for possible problems that might be visible to the eye, but hard for a computational metric to pick up. Finally, they will perform other exploratory analysis to try to find flaws in the synthesis with respect to disclosure risk, including looking at raw values as opposed to only graphing them.

The document makes clear that the level of risk that Smart Data Foundry would accept for any dataset would depend on the recipient and their intended use.

Synthetic data: lawful basis (Innovation)

3.39 Smart Data Foundry and the ICO looked at the application of the purpose limitation principle and lawfulness when creating synthetic datasets from personal data held in the research facility.

3.40 Firstly, it was considered whether Smart Data Foundry's further processing on data stored in the research database in order to generate synthetic data doubles aligns with the purpose limitation principle of the UK GDPR. Anonymisation for the purpose of further research would be a compatible purpose unless there is a legitimate expectation by the individual that their data is to be kept in identifiable form or they had been informed that it would be kept in that form. However, we determined that as Smart Data Foundry is not the originating controller at the point of data collection, since it is obtained from third parties, then the purpose limitation principle does not apply. The processing of personal data into synthetic data is a separate processing activity that would require a lawful basis. The anonymisation processing activity to generate synthetic doubles is not part of Smart Data Foundry's original purpose. Find more information about purpose compatibility and anonymisation in the section "Anonymisation as part of your processing activities" in chapter 4 of our draft anonymisation guidance on accountability and governance.

3.41 Creating synthetic doubles as an anonymising process is a separate processing activity and any personal data processed under the activity should be using its own lawful basis. Smart Data Foundry needed to make clear that the two processing activities (the research database and synthetic doubles) are using separate lawful bases with separate legitimate interest assessments, as initially this was not clear when the ICO carried out its review of Smart Data Foundry's legitimate interest assessment for the generation of synthetic data. For further guidance on anonymisation as further processing, read the section "If we anonymise personal data, does this count as processing?" in the first chapter of our draft anonymisation guidance.

3.42 Following the identification that the processing would use legitimate interest as a lawful basis, the ICO Sandbox reviewed Smart Data Foundry's legitimate interest assessment for the process of creating synthetic datasets. In particular, Smart Data Foundry wanted guidance on demonstrating the controller's interests when the developments from the synthetic datasets may yet be unproven with regards to significant public benefit. The ICO's response was to highlight how benefits of processing will be embedded into the project approval process. The process will evidence that the specific organisational objectives are considered every time and therefore contribute to the consideration of the benefits of the processing when determining the weighting of the legitimate interest. In response, Smart Data Foundry have clarified that datasets will undergo a check against the organisational objectives and also included a summary of the Strength in Places objectives of UK Research and Innovation (a non-departmental public that directs research and innovation funding) which will demonstrate the expected benefits of the processing for their stakeholders.

4. Ending statement

4.1 Smart Data Foundry's participation has helped the ICO to understand the data protection challenges that research organisations face in understanding and applying the UK GDPR research provisions and determining if their anonymisation and pseudonymisation methods are effective. It has highlighted the special considerations that need to be given when combining datasets for research purposes and the importance of considering the role of identifiability assessments and the lawfulness of processing at the early stages of building a research database in order to embed data protection by design and default.

4.2 By engaging with the Sandbox, Smart Data Foundry has developed several important privacy artefacts under the guidance of the ICO, including a wider DPIA for the research facility, Legitimate Interest Assessments for the research facility and synthetic doubles propositions, as well as a qualitative Anonymisation Assessment toolkit. These are vital tools to ensure Smart Data Foundry are conducting research and innovation in the public interest with appropriate controls when handling data that is complex, important and potentially personally identifiable. The key areas of focus going forward for continuous improvement are: to ensure Smart Data Foundry has appropriate governance in measuring and validating that research projects align with its missions, that necessity and minimisation are appropriately balanced with each new research proposition and dataset, and working with experts to further develop quantitative anonymisation tools that complement the qualitative tools currently in place. This will assist continually monitoring and assessing re-identification risk.