Project: Smart Data Foundry – National Data Utility

Document: Mission-driven research – Data Protection Impact Assessment

Author: Adarsh Peruvamba

Date: 08/07/2022

Version: DRAFT

Assessment Details:

Name	Mission-Driven Research - DPIA
Respondent	Adarsh Peruvamba, Information Governance Manager, Smart Data Foundry
Approver	
Approver date and comments:	
Next review point:	

Assessment Questions

DPIA Question / extract from UoE DPIA	**Response**
Activity Name	Project for Smart Data Foundry ("we", "us") to provide a secure data environment where real consumer data from UK financial institutions can be safely shared with organisations working on big societal economic and environmental problems and conduct data-driven research
Activity outline *Explain broadly the scope of the activity, particularly what the activity aims to achieve (e.g. the benefits to the University/GOCOE, or to data subjects etc.) and what type of data processing it involves. Explain why the activity is necessary to achieve these aims.*	**Objective and Benefits** The objective for this research is to solve societal, economic and environmental problems, broadly defined as 'in the public interest'. We have outlined our objectives as Missions, and these include: **Stop the Squeeze:** We want to help UK households, particularly those classed as vulnerable, withstand the rising cost of living post-COVID and Brexit. By providing data on financial wellbeing and resilience, we will help the government and industry take real actions to help people on the poverty line survive and thrive. **Countering Climate Change:** We are working with academics who understand how satellite data can be better interpreted to reveal patterns of activity on the planet's surface that indicate positive or negative climate change activity, e.g., new trees being planted, methane emissions being reduced, CO2 sinks being deployed. We are also producing Enriching sandbox environments supporting innovation in new ESG-related FinTechs by creating synthetic data sources for those businesses to test their business models against. We recently worked with the FCA on their TechSprint, designed to

Commented [SS1]: As discussed this area will be particulary important to expand upon significantly

	support the development of new businesses that could create better ESG data. **Open Finance for All:** We see the future of Open Banking as the blueprint for Open Finance and Smart Data, available to all parts of society, especially groups that are traditionally excluded, designed to protect against crime, resilient to failure and future shocks to the system. Smart Data Foundry is ideally placed to drive collaboration with Government, Regulators, FinTech's and the wider Financial Services community. We will make sure that Open Banking works for everyone and paves the way for a better future. **Strong Small Business:** Smart Data Foundry recently partnered with Sage Group, FreeAgent and Equifax to work with their data. We are talking with three of the UK's major high street banks to contribute banking data to create a unique view of what's really going on. We'll share these insights regularly with Government and others to make sure they're making decisions on how to support this sector based on facts. These insights will mean better lending options for small businesses, better cash flow, and better productivity. All of which help accelerate the recovery of the wider UK economy. The criteria for project approval within Smart Data Foundry involves: • Strategic: Is this project a strategic enabler? How does it align with our missions as outlined above? • Reputational: Does this enhance Smart Data Foundry's external reputation? Does it create / strengthen relationships? • Commercial: Is this a revenue generating project? • Operational Effectiveness: Does this improve the efficiency of Smart Data Foundry? For research, the alignment with strategic missions as stated in purpose section will be a key milestone to approving a course of action. **Scope** This DPIA focuses specifically on the mission-driven research element of Smart Data Foundry's activities. This involves the following stages: - Agreement with data provider for the ingress and continued processing as a controller of data for research - Ingressing the data for this purpose within EPCC technical environments, including data quality and confidentiality checks - Making data available for specific research projects that meet the missions specified above, with clear written criteria as to how these missions match that criteria - Curating the data, ensuring that combinations of the datasets are tested to establish risk of reidentifiability and ensure the technical environments are suitably secure - Conducting checks on data that is egressed to ensure this is aggregated to be **information** rather than data

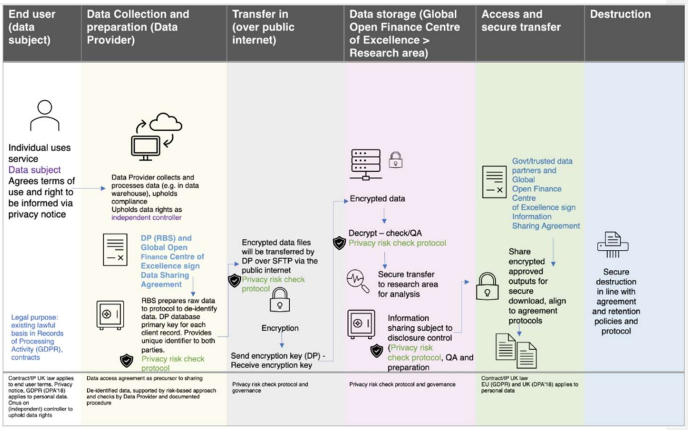
	This DPIA does not include the 'research processing' of the data within its scope. These will be separate DPIA assessments per research use case or mission. **Data Providers** This list of data providers will be updated as more data providers are brought on board. In terms of data being able to be used and re-used for mission driven research, these are the data providers where this form of processing has been agreed: - Equifax - MoneyHub Other data providers currently only have agreements in place for specific missions. These include Natwest (Covid Dashboard project, Later Life research), SAGE (Slow and Late Payments research) and FreeAgent (SME resiliency dashboard). **Nature of the data** Detailed information on this will be included as part of the Data Inventory, which is currently under development. **Necessity** Data-driven processing is necessary to achieve the missions and purpose related above. Without the processing of data at this scale, it will not be possible to perform effective data-driven research across the missions. It is also important to have these datasets utilised across different missions to ensure re-usability and the ability to address new problems as society's needs and issues evolve. Without access to a wide range of datasets for the use of data-driven and mission-led research, the research outcomes would be incomplete with either data that is not of sufficient volume to be statistically valid, or data that does not include the requisite fields to measure indicators such as financial health. The NDU also utilise pseudonymisation techniques as default, ensuring that unless strictly necessary for a noted purpose, data will be pseudonymised with SDF not having access to the key. This will help ensure data is effectively anonymised where possible. These controls will ensure that the data is being processed in the least intrusive way possible. As a counterfactual, it will not be possible to achieve mission outcomes without the requisite data. For example, to carry out the poverty premium research, without data relating to personal financial health and utility spend, it will not be possible to calculate an appropriate metric. That said, the date of birth of an

	individual will not be required for this either, and as such will not be collected or utilised. **Outputs:** All research processing outputs where the processing has involved the use of data that has any risk of reidentification will be aggregated to a minimum checksum of 10, ensuring outputs can be classed as 'information' rather than data. This applies for all data that is classified as Tier 2 and upwards. For broad reference, see Appendix A of data classification labels and descriptions. This will be further expanded for each project-specific DPIA. **Dissemination** This DPIA does not focus on the specifics of how the data results will be disseminated as this will vary from project to project. However, some broad principles are: - Smart Data Foundry will be transparent on how the data was obtained and utilised within the results of every publication - Smart Data Foundry will name the data provider partners involved in each production of result - Smart Data Foundry will only make available aggregated numbers – as stated above. Where possible, Smart Data Foundry will endeavour to share these results with the wider public if the dissemination allows this to be possible.

Describe the data processing being proposed to achieve the purpose? *To do so, you should describe the collection, use and deletion of personal data here and it may also be useful to refer to a flow diagram*	The path to establishing the NDU and achieving the purpose of providing secure data storage and data processing for statistical and research purposes consists of three phases : • Negotiate a series of bilateral agreements with Data Partners in which the NDU produce insights from their data for clearly defined purposes • Evolve agreements with Data Partners to allow the NDU to use datasets for multiple purposes, evaluate multiple datasets for deeper insights, and to create linked datasets • Run mission-led projects involving one or more datasets to analyse and produce insights. An example of this would be the 'Poverty Premium', which would involve utilising financial data combined with utility data to help measure the level of 'premium' that less financially able customers may have to pay on essential services. The processing will require data suitable for each research purpose. As a result, this may include personal or sensitive personal data where the research outcome requires it. However, the data requests will ensure only data that is required to assist the strategic missions stated above will be collected. An example of a project lifecycle of processing would be as follows: - agree scope of the project and datasets required to solve the problem identified - if requiring a dataset that is not currently available, liaise with data provider and conduct appropriate due diligence prior to signing a DSA - including a project-specific LIA (if utilising standard system and not requiring a full DPIA) as well as anonymisation assessment to determine level of risk and mitigation required. data ingress by Information Governance team including ensuring the data matches what is expected and that there is no additional personal data than what is expected. This will include further deidentification if required by anonymisation assessment (and deemed not to affect the outputs of the project mission) - data available to data scientists within controlled EIDF (Edinburgh International Data Facility) 'Safe Haven' environment - when results are ready to share wider, the results are checked by the information governance team to ensure they are aggregated to a minimum checksum of 10 per subcategory, ensuring the exported data is 'statistical'.

Commented [PA2]: Would include further diagrams on data collection etc within DPIA

Below is an illustrative example of the likely data flow of research projects involving the National Data Utility when involving a third party data provider.



We will adopt different levels of scrutiny on anonymisation and reasoning around the circumstances required to hold personal data dependent on the data classification we assign to the data when originally ingressed. This will also be reviewed when that data may be combined with another dataset.

Classification	Description	Anonymisation approach	Circumstances required
Tier 4	Personal data where disclosure poses a substantial threat to personal safety, security or health	Look to reduce level of granularity in any other specific linkable characteristics – such as Age or Location – to move data towards lower Tier of security	Please see section on using 'special category data' below
Tier 3	Non-pseudonymised personal data, or pseudonymised/synthetic data where confidence in quality of	Look to reduce level of granularity in any other specific linkable characteristics – such as Age or Location – to move data towards lower Tier of security	Initial Data Sharing Agreement and DPIA from data provider, as well as original specific research conditions congruent with missions of the organisation. Would be stored in appropriate

Formatted: Font: (Default) Poppins Light, 9 pt

Formatted Table

	deidentification is weak		security zone as detailed in technical environment security section below.
Tier 2	Pseudonymised or synthetic personal data where confidence in deidentification is strong	The default state of which the data would ideally be stored for future projects	Initial Data Sharing Agreement and DPIA from data provider, as well as original specific research conditions congruent with missions of the organisation. Would be stored in appropriate security zone as detailed in technical environment security section below.
Tier 1	Pseudonymised or synthetic personal data where confidence in deidentification is absolute	Derived data from Tiers above to be made more widely available for research and innovation	No specific circumstances required for this level of data as it is no longer in scope of GDPR

Anonymisation Assessment

In collaboration with the ICO and based on their anonymisation guidance published in 2022 – which is undergoing a feedback process – we have developed a qualitative anonymisation tool to help identify the risk of re-identifiability. This involves three stages for each dataset – both linked proposals and singular datasets:

- Isolating what elements of the dataset are identifiable, looking at:
  - Singling out: what data points allow you to differentiate an individual from another where you can isolate them in a dataset from other individuals in a discernible way
  - Likability: what data points are here which are likely to be found in other datasets and enable links – e.g. Date of Birth/IP address
  - Inferences: Are there details about individuals that can be inferred from making correlations based on information in the dataset
- Processing for the purpose:
  - information about the data environment and whether there's technical or organisational measures to control access to the data and reduce identifiability risk

	○ Reasonable means available to a motivated intruder, including possible motivation, likely cost and time required, and any legal gateways or likelihood of use ○ Data disclosure and release: What forms of the data are likely to be disclosed, to whom, and in what form and stage - Assessment and mitigation ○ Listing all identifiable elements from first stage ○ Listing all active risks from second stage ○ Listing possible active mitigations against each of these risks in terms of **anonymisation techniques (expanded on in Appendix A).** The anonymisation techniques involved would broadly involve - Decision ○ Labeling if the data is adequately anonymised as a result of the controls ○ If not, judging whether we would consider it to be Tier 3 or Tier 4 in the above classification schema
List of Stakeholders - Activity or Proposal Stakeholders	Our collaborators in general can be summarised to be: - UK Research and Innovation: Smart Data Foundry was created from a significant grant bestowed by the research and innovation department in the government. They are a significant stakeholder that Smart Data Foundry will report findings to. - Fin Techs and Startups: Stakeholders and interest parties in the goal of opening finance data and making research data available for further access and innovation - Academics and Researchers: Utilising the data to work on issues aligned with the missions described above, as well as enhance their own reputations and develop curriculum for higher education and further research - Corporates and Businesses: Alignment on the mission of resilient small businesses, as well as interested in helping develop Smart Data Foundry's work to meet their own CSR causes and enhance their reputations

	- Government bodies, policy groups and regulators: Interested in the development from the research and collaborations to develop and influence policy - The wider stakeholders are the data subjects themselves – specifically data subjects who are economically active and have a financial data footprint, sampled from the datasets of data providers that Smart Data Foundry will have access to. The broad missions detailed above , if research is successful, should lead to outcomes such as SME resilience, income and expenditure fluctuations during crisis, and poverty premium facing lower-income households when accessing essential goods and services.
List of data subjects affected by the activity	Users of financial products such as banking, credit and pensions – this will include business banking account holders and personal banking account holders, including sole traders, public and private pension holders as examples. The scope of the project is primarily focussed on financial data although there may be examples where this data is linked with supplementary reference data around deprivation, geography and other comparative factors. As a default, we are not purposely processing personal data; rather there may be identifiers in the data set, which we will implement privacy risk checks and protocols for.
List the personal data you are going to process If involving special category data, please state the supplementary conditions met	- Pseudonymised Identifiers as Personal Information 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 - **Where research mission requires this,** indicators of vulnerability such as disability or benefit provision may be processed - **Where research mission requires this,** special category data such as health or ethnicity may be utilised. This is specifically the case under Article 9(2)(j) in that, it will only be processed if it is deemed **necessary** for the scientific research or statistical purposes. To do this, Smart Data Foundry will have additional controls for the stipulations required: o Necessity of purpose: A separate LIA will be prepared that challenges necessity and minimisation appropriately. This will particularly challenge the reasons why the data cannot be further anonymised or pseudonymised to the point of not being able to link to specific individuals o Appropriate safeguards for individuals rights – including upholding transparency to ensure data subjects can uphold their rights – will be observed. o The project **cannot** be likely to cause substantial damage or distress to an individual

	○ **Not used to measures or decisions** about particular individuals ○ We will work with the governance board of Smart Data Foundry – which includes stakeholders across the finance sector and third sector – to ensure that public interest is congruent with wider public societal interest, as well as demonstrating why the research is scientific in nature (relying on the expertise of university member of staff on board for this specific measure) ○ Only once all of these purposes are covered in **a separate case-specific DPIA** will processing begin on a research project involving sensitive personal data. In the cases of using personal data, the research will not be utilised to make decisions about particular individuals or cause any substantial damage or distress to an individual Data will undergo assessments on anonymization to ensure the data can meet the bar of 'effectively anonymised' where possible. If this bar is not met, the data will be treated as personal data and have accompanying provisions around transparency and data subject rights as appropriate.
How are individuals being made aware of how their personal data will be used? (right to be informed)	Where possible, Smart Data Foundry works with data partners to ensure their privacy notice contains appropriate information with regards to transferring data to third parties to enable research in the public interest. However, it is also reasonable to expect that since this is a secondary purpose to the primary 'expected' purpose in utilising a financial service, this is less likely to be read in detail and acknowledged. Smart Data Foundry has its own privacy notice that sets out its processing of data. This is unlikely to be viewed directly by a data subject unless directed to by the data provider. However, if there are concerns for the data subject with regards to their rights, it would be realistic to assume the Smart Data Foundry privacy notice is an area the data subject may search for as long as they are aware of the organisation.. Article 14(5)(b) clarifies that Smart Data Foundry does not need to send transparency information to data subjects where personal data is provided to us if it would be unreasonable or impossible to do so. As the data would predominantly be unidentifiable and pseudonymised without Smart Data Foundry having access to keys, with re-identification not permitted by staff, directly contacting data subjects would not be possible. In the event of processing data that is identifiable, Smart Data Foundry will assess if the volume of data involved makes directly contacting data subjects reasonable; but this is likely to be a remote possibility when considering volume and level of intrusion.

Does the activity involve the use of existing personal data for new purposes?	Yes - the project involves utilising the current data set provided by data providers with the addition of reference data elements. While companies are allowed to re-use data they control for research purposes, transferring that data for use within Smart Data Foundry is a separate processing activity that needs accounted for. As a result, each new use of data will undergo a Legitimate Interest Assessment to ensure this use has a valid legal base, which will be shared with the data providers involved to help form the basis of their legal basis for processing.
If processing personal data, what Lawful Basis is this data going to be processed under?	The basis for processing this data is Legitimate Interest. Please see the broader Legitimate Interest Assessment – Research for more detail. The summary of the assessment is as follows: "Due to the research focusing on the public interest, legitimate interest applies for this processing when weighing the benefits against the possible impacts on data subjects' rights and freedoms. " There are a list of controls required to ensure transparency, reidentification risk and data minimisation are to the requisite level to meet the balancing test. These have been included within the risks in the final section. Each specific research mission will undergo a separate legitimate interest assessment, with a full DPIA being completed if the circumstances around the processing conditions or data sensitivity are materially changed from what is covered within this DPIA. Full information on how we assess necessity and adequacy as part of this is below in the data collection section.
Can you confirm that data collection procedures are adequate, relevant and not excessive, i.e. that you are not collecting more information than necessary?	Proportionality of the processing will be directly related to the research outcomes. Personal data or sensitive personal data will not be utilised by data scientists and researchers if this is not required by the research question. The NDU also utilise pseudonymisation techniques as default, ensuring that unless strictly necessary for a noted purpose, data will be pseudonymised with SDF not having access to the key. This will help ensure data is effectively anonymised where possible. Each wider research mission is tested on necessity and adequacy utilising our Legitimate Interest Assessment template. This involves: - Accurately summarizing the purpose, including summarizing the benefits, stakeholders and ethical implications - Establishing necessity – this will look at the data processing in detail, articulate if this is **reasonable, effective,** and if there is a **less intrusive** way of achieving the same purpose. To do this robustly, we will interview the data scientists involved against each field of data involved in processing.

Commented [PA3]: Check that this risk has been added.

	- We will then balance the **reasonable expectations** of the individual with the **likely impact or harm** from a privacy subject rights perspective or material perspective – based on our assumptions and personal experiences of being data subjects for now, as there is no wider stakeholder opinion gathering planned as of yet. This will include what information about the processing they are likely to encounter and how clear the purpose is, as well as if there is a tangible benefit to the services used by the individual, whether they can opt in and out, and if the interests of the individual align with the organisation. - If we are satisfied with the balance, then we will approve the adequacy and necessity of the processing as part of our legal basis of legitimate interest All new projects will be assessed by company stakeholders in Business Development and Information Governance to see if it falls under the specific mission. If the project does not, it will likely not be pursued – however, if there is an exceptional circumstance, a separate LIA will be prepared. Projects will ensure that the data collected and utilised is the minimum volume of data required to ensure statistically valid results, as well as ensuring the fields processed are limited. The datasets processed will be directly related to the missions and subprojects within missions – for example, the poverty premium project will utilise data involving financial health, as well as data involving utilities. These datasets are required to effectively measure the research outcomes.
How will the personal data be checked for accuracy?	We have a Quality Assurance (QA) checking process for data quality and to check that we are not receiving inaccurate data or personal data, and to check specifically for identifiers. We check for accuracy and risk at every stage gate, so data in transit and at rest is checked.
How long will the personal data be retained for?	The majority of data agreements currently in place require us to delete this data once the project is completed. For more open ended agreements in the future, as this processing focuses purely on research related processing, Article 5(1)(e) provides an exception to the principle of storage limitation – and we believe our controls are sufficient to ensure the processing is only for research-related processing or further anonymisation at which point it will no longer be personal data.
What technical and organisational security measures will be in place to prevent any unauthorised or	The data will be hosted in a secure virtual machine environment at the Edinburgh International Data Facility (EIDF). Here is a summary of the policies and procedures in place for the EIDF facility. These are reviewed regularly with EPCC – the third party that runs the environment – as part of the service management monthly meetings, and

unlawful processing of the personal data?	recorded in a Security Services Protocol (SSP) document that is kept up to date. SDF and EPCC have an IT Services Agreement which details their relationship as controller and processor. As a part of this, responsibilities with regards to security as summarised in the SSP document stated above. **Confidentiality:** Only authorised project researchers will be allowed to access the virtual machine and the data, with role based access documented and logged by Information Governance and the service provider. All users require access to university VPN and then MFA access for the VM - this is enforced by default. VPN is university managed and is a prerequisite for entry to Data Safe Haven unless on a specific location-bound login from on premise ethernet. MFA – multi-factor authentication – involves two sets of logins as well as a security code which is text-messaged or app-generated – requiring multi-device login as well as multi-password. Data is protected in transit by AES-256 level encryption, utilising a web-based portal Serv-U or SFTP over a secured network channel to enable transfer. The database is not encrypted at rest - however the disk storage is fragmented to ensure security in this aspect. All data entering the facility will transfer through an ingress quarantine area where it will be checked and further de-identified where necessary to mitigate confidentiality and re-identification risks. Similarly, all outputs from the project will transfer through an egress quarantine area and be subject to disclosure control to ensure de-identification. Staff use a variety of devices and there is no embedded method of ensuring security/antivirus usage, encryption or USB port risk. EPCC - the third party providing the EIDF environment - does have systems staff who have root access to all areas. No other users have this access. In terms of root access, administrative access to the Safe Haven is controlled by a training and vetting process. Each Safe Haven administrator must complete training including - UoE Information Security and GDPR Training, MRC training on Research Data, Confidentiality and GDRP, Disclosure, and BPSS. Administrative access to the Safe Haven is logged either at the virtual desktop level or the Administrative VPN access service so we have a log of which administrators accessed the Safe Haven and when Administrative access to the Safe Haven is reviewed as part of the Staff Development Process which includes the leavers and joiners process and independent at least once a year as part of the SHS EPCC Administrator

	Review Process. No external third parties have credentials to access any parts of the Safe Haven Service other than access to built-in diagnostic support services **Integrity:** No network access to the internet with data only entering via ServU or SFTP within a secluded zone. This mitigates malware/AV checks. There are also scheduled maintenance windows to address software or hardware issues. The organisation is accredited to ISO27001 standard. Technical restrictions preventing any information or data being copied on or off systems. Data cannot be exported unless by the information governance team utilising ServU, which has an audit trail. Original copies of data are segregated and checked before copies being made available to other users. As a result, golden source of data is maintained. **Availability:** All backups are initiated from the scheduled processes on the backup server and operate in a pull fashion where the server initiates contact with the backup client installed on the VM. Files from specified data directories will then be copied via SSH (thus encrypted in transit) to the backup server. The backup server is also located in the Safe Haven and uses a different physical disk to that used by the Smart Data Foundry system. The backups for the Safe Haven are currently located on site on alternative disks i.e., the backup is not being stored to the same SAN unit that is running the live version /safe_data The backup solution makes use of deduplication as a way of tracking changes and only backing up data blocks that have changed since the last backup, in essence tracking incremental changes. On first backup the entire data directory will be backed up and then subsequent changes on the data directory will be tracked and added to the backup repository. Files will be retained in the backup for seven calendar days.
Will you be transferring personal data to a country outside of the European Union or the European Economic Area (EEA)?	No, all data is securely stored on premises in Edinburgh
If the data will be anonymised, is it likely that a 'motivated intruder' will be	Yes. Note we have included for this risk in our organisational and technical measures and governance design to protect data.

interested in attempting re-identification by linking the data with other information available to them?	We have augmented controls, to include daily log files which are monitored for intruder risk / unusual activity.
Data subject rights: If a subject access request (SAR) is received for personal data included in the activity, how easy is to comply? Is the data easily accessible elsewhere? Are you able to comply with requests for erasure or restriction of processing? Can you apply an exemption?	In the case of data we receive that is pre-anonymised by data providers to a level deemed **effectively anonymised:** we are not able to uphold data subject rights as we are not processing sufficiently identifiable information to feasibly do so. However, we will ensure that we will be transparent with the data subject about our data providers, and ensure they have a method of raising their subject access rights with the data providers directly, if applicable. Note that per Article 89(1) through our processes, the data will no longer permit the identification of data subjects and note related derogations of data subject rights. In the case of data we receive that could be defined as personal data in any way, we have the following approach in place for each right listed below We will publish – as part of our commitment to transparency – the method to enact each of these rights on our external facing privacy notice in the event of handling data defined as personal: **Right to be informed** As part of the legitimate interest assessment for each research mission, we will assess the benefits for the data subject and congruency of purpose with the possible impacts on the data subjects including infringement of their data subject rights. As a part of this, we will assess the volume of data involved compared to the possible infringement, and assess if the burden to inform would be unreasonable – either logistically or in terms of expense. Examples would include – mailing every data subject if the research does not directly impact them, or expending resource into identifying a data subject if that is not immediately possible. We do not directly collect personal data as part of this processing. As a result, our privacy notice will list our data providers to help subjects understand if their data may be included in research processing. As stated in separate sections, we will also liaise with data providers to ensure their privacy notices include our processing activity. **Right To Access** Individuals – if providing information that enables us to isolate and identify them in records – will be provided confirmation of what personal data is held

	in relation to them, and the ability to provide them a copy of their personal data. The timescale for providing this will be one month by default. **Right To Rectification and restrict processing** Individuals are also given the right to request that any personal data is rectfied if inaccurate and to have the use, by the controller, of their personal data restricted for a particular purpose(s) in certain circumstances. **Right To Object and Erasure** Since the legal basis for processing is based on SDF's legitimate interests, individuals have the right to object to this processing. If – as part of the objection or right to access – they decide to request the erasure of this data (and that this request is successful against the criteria for erasure) – Smart Data Foundry have processes in place to process this erasure. **By default**, in our agreements with data providers, we also integrate a process where 'Bad IDs' (i.e. IDs for deletion) are flagged as part of data refreshes. These IDs are removed from all records (but not from historical summarised derived reports – i.e. where this data is now 'information' instead). Smart Data Foundry reserves the right to apply an exemption to complying with a subject access right if and only if providing that data would prevent or impair the purpose. With the broad missions of Smart Data Foundry, this is not likely to be the case for most research projects. The criteria for the exemption to apply – including the purpose impairment mentioned above – would include that the research was not likely to cause substantial damage or distress, not be making measures or decisions about particular decisions, and the results were not being made available in a way that identifies a subject.
Are provisions in place in case a data protection breach occurs as part of the activity?	Smart Data Foundry adhere to the University of Edinburgh's data protection policies and there are documented incident response processes. A response plan for specific for a high-risk personal data breach is not required for the nature of this data.
Will any other organisations outside the University have access to the personal data?	Some parties will have access to outputs of the data – however these will be aggregated to a minimum checksum of 10, thus classifying this data as 'information' rather than personal data.
Will this involve many individuals	Yes –it is anticipated the volume will be often be in excess of 100,000 for the missions in scope
Will there be changes to data quality	No

assurance or processes and standards?	

Risk	Risk Description (detail in Register)	Impact	Likelihood	Mitigation	Residual Impact	Residual Likelihood
Data is shared inappropriately	Possibility that personal data is shared inappropriately. The data is confidential and commercially sensitive. As a result, the impact of the data being inappropriately shared is high.	High	Medium	Data will be held in a secure environment with access allowed only to authorised project personnel. Outputs will be checked by the Research Coordinator before being passed on to Government. We will use a minimum cell size of 10, as described earlier, meaning that no result shown will be derived from fewer than 10 businesses. Additionally, deidenfication methods such as grouping and rounding location data as well as banding data where possible means that the confidentiality risk impact of the data is further reduced. We are also protected by contract against approved researchers sharing data inappropriately.	Low	Low
Data used for different purpose	Personal data may be used for a new and different purpose without the knowledge of the data subjects, perhaps due to a change in the context in which the data is used. While our data providers broadly discuss the possibility of sharing data for research in the public	Medium	Medium	Reutilising data for the purpose of research is allowed as long as it is for scientific or historical research purposes - additionally, we will complete an LIA and share with data provider if required for their purpose. Furthermore, de-identification measures as well as the contractual obligation not to identify helps protect data subjects' confidentiality.	Low	Low

	interest and display this within their privacy notices, it is likely that data subjects may not realise that the purposes collected includes this. It is likely the sole purpose - utilisation of the financial function - is the only purpose considered.					
Data Linkage causing reidentification	Collecting information, matching and linking identifiers or whole datasets might mean that data are no longer anonymous if anonymity is what people were led to expect. If that were to happen, the data would have to be treated as personal data , with all the controls and rights expected to be upheld from that.	High	Medium	If datasets will be linked, there will be a process assessing likelihood of identifiability as well as testing it against the 'motivated intruder' test i.e. considering if further access to public datasets would make an individual identifiable. If linkage is pursued, the linking of the datasets will be undertaken by a third party, ensuring Smart Data Foundry do not have access to pseudonymisation keys. We will utilise differential privacy tools to ensure exceptional records do not stand out, utilising the ICO's latest anonymisation guidelines.	Low	Low

				We also have contractual stipulations to cease processing if any data is identifiable. Lastly, all outputs will be aggregated to a minimum checksum of 10, ensuring that there is no impact from broader access to research outcomes from a confidentiality point of view. When linking data, we will consider it through the 'anonymisation assessment' and apply mitigations at that stage.		
Duplication of data	Excess information collection or information not properly managed can lead to creation of duplicate records.	Low	Medium	We will integrate data management controls to ensure if multiple datasets are received, duplicates are deleted. Data scientists will also be reminded to check for duplicates when processing multiple feeds of data	Low	Low
Public distrust	Public distrust regarding use of data is a significant risk for all projects involving the use of 'personal data' in any form - identifiable or pseudonymised or even significantly deidentified to	High	Medium	In addition to the de-identification controls, we will publish our method of de-identification and the associated risks. We will work with data providers to ensure the transparency within their	Medium	Low

	the point of anonymised - because of the lack of transparency within the sectors the organisation is looking to work in.			privacy notice is appropriate for the data looking to be shared. We will ensure strict access controls to make sure only appropriate personnel have view of the data prior to further deidentification (utilising ingress checks) and then for project work prior to outputs. All outputs will be aggregated to a miinimum cell size of 10. In all endeavours, we will attempt to be transparent about the data and our methodology.		
Risk of motivated intruder	Despite proper security, due to the commercially sensitive nature of the data, there is a possibility of there being a motivated intruder looking to 'hack' the system. This is further enhanced as the organisation grows in reputation and has access to more granular and varied data.	High	Low	Controls on the security environment include strict 2FA access controls, prior checking on backgrounds prior to access, as well as no access to the internet. While network access is available to update oS and software packages, these will be narrow in scope and will not materially increase this risk. Data will also be encrypted in transit. All datasets are stored in separate groups, further ensuring it is more difficult to access. Generally, the data we receive from third parties is pre-aggregated and de-identified to a level where risk of entry and data loss is greatly reduced	Low	Low

				from a confidentiality perspective. We should explore file management within the safe haven area further. PEN test will be completed to help assure the likelihood and risk of intrusion in this regard.		
Malware/Antivirus protection	Lack of antivirus and malware protections on devices, as both university issued devices and BYOD don't have a regulated anti virus software beyond stock install. Further work is required to ensure the devices used by staff has regular security/antivirus updates, are encrypted appropriately and have disabled USB ports. However, as access will occur via VM client and that staff cannot ingress data or software manually, this risk is not likely.	Medium	Low	Organisation will look to institute a Smart Data Foundry specific Acceptable Use and BYOD policy to mitigate against this in future. However, as access to Safe Haven is via a VM, this is not an urgent issue when it comes to data protection as much as the protection of the data to be output once on local machines from a commercial sensitivity standpoint	Low	Low
EPCC Root Access	EPCC - the third party providing the EIDF environment - does have root access to all areas	Medium	Medium	There is a contract in place to mitigate against this, and that this is standard practice as data processors.	Low	Low

				Extra information on root access controls		
New software in VM - corruption	Risk of data injections or corruption from additional data manipulation software being available within EIDF Safe Haven environment	High	Low	Currently, the production of the new VM will end with a test of the VM populated with the new means we have developed, and which passes the agreed User Acceptance Tests. The VMs will only be commissioned once deemed safe. To help deem it safe, the list of software will be peer reviewed by data scientists and information governance. Any additional software will have to be tested in an environment with no client data within it. Once that testing/quarantining is complete and it is ascertained the software works as expected Check market for code-checking from a destructive sense to ensure more automatised checks are possible in the future	Low	Low
New software in VM - integrity	Risk of data integrity issues from additional data manipulation software being available within EIDF Safe Haven environment	Medium	Low	We have clear delineation between golden sources of data managed by information governance, and the data released for use by the wider team. As a result, if there are issues from software	Low	Low

				behaving in an unpredictable manner, this will not be irreversible. We will plan to have two repositories of original data along with the codes that assist the cleaning of the data We will integrate data management controls on outputs to ensure these are validated and sense-checked for data corruption.		

Appendix 1 – Deidentification guidelines

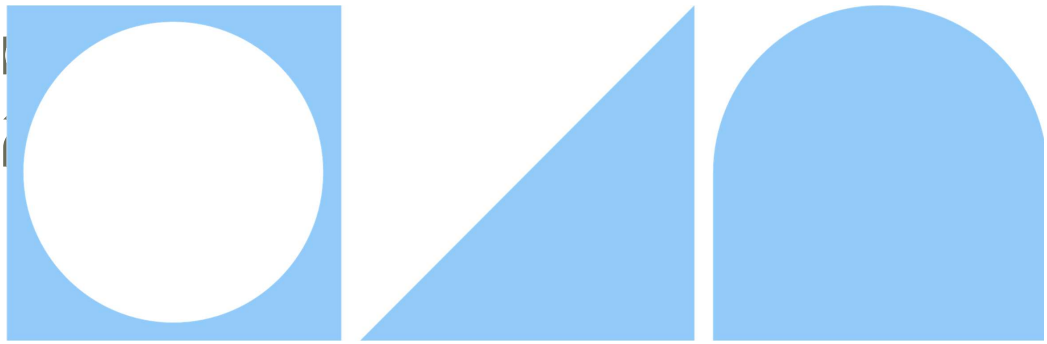
Anonymisation techniques and de-identification guidelines

In general, our approach to de-identification has been adapted from the ICO's guidelines on anonymisation and case studies of the same. Some of the broad techniques can be summarised in appendix 1:

Method	Description	Type of fields	Change in risk profile
Pseudonymisation of records/ID fields	More information on recommended pseudonymisation approaches here – essentially removing ID indicators and utilising keys/hashes/salted hashes instead	ID Fields	This is generally adopted by default – it is a significant control in ensuring personal fields such as name fields or ID fields are not necessary

Removal of field	Removing columns that have personal data completely from the dataset	Any field	While this approach can only be utilised if the dataset would still have utility for the function of the project, this is the simplest and most effective way of eliminating the risk profile of that particular data characteristic
Extreme Value filtering	If the analysis is focused on the median or mean rather than extremes, it can be prudent to filter out extreme values (example being 'high net worth')	Primarily 'Value' fields (Balances, Credits/Debits)	As probability of reidentification is higher for these individuals, this may be a method utilised. However, care must be observed to ensure this filtering does not materially affect the project outcomes.
Reducing resolution	Reducing information in a column by taking only a partial section of it – such as part of a post code instead of a full post code – to effectively create larger 'pools' of data	SIC Codes, Post Codes	Reducing the resolution and reducing specificity in category fields in this manner would group more records together and reduce the chance of re-identification
Rounding	Dropping the number of significant figures to	Primarily 'Value' fields (Balances, Credits/Debits)	The effect is similar to 'reducing resolution' in effect – by creating less specificity in the values, you reduce the specificity in the record and reduce chance of re-identification
Aggregation/Banding to form categories	Creating groups/category fields instead of a specific value field	Age, Free Text columns (e.g., transaction recipients grouped)	Creating a category for these identifiers rather than providing the specific identifier helps reduce the specificity of the data and reduce chance of re-identification
Laplacian noise	Utilisation of differential privacy to create new set of *responses* (out-put of operations applied to the in-put dataset) enjoying epsilon-differential privacy.	Primarily applied on subsets of the in-put dataset, typically on the Real 'Value' columns	The created *response* would have an added noise, for the case of queries on the in-put dataset the noise can have the form of a scaled symmetric exponential distribution with epsilon dependent standard deviation.
Random category generators for de-identification	Using random category generators, we can select or simulate different values for the desired category. An example of such generators is given by the python package Faker	Categorical 'Value' fields, such as Country, Name (linked to Country or Religion), Addresses (linked to the Country), Sex (linked	The generated value for the selected category, can be adjusted to the desired risk level. The level of risk might depend on the combined use of the generator and the way of synthesis. In the case of Faker together with the Synthetic Data Vault synthesiser, it'll generate new values for such categories using a learning-based algorithm.

		to Name or Country or Religion or Race)	



Legitimate Interest Assessment – National Data Utility - Innovation

Purpose

Assess whether there is a legitimate interest behind the processing

What is the purpose of the activity?	The purpose of the Smart Data Foundry is "To open finance for good." We unlock the power of financial data and smart data. For research, to drive data-driven innovation and to grow specialist talent so more can thrive. To help achieve that purpose, one activity we undertake is to **create and provide synthetic data** within secure data environments where real consumer data from UK financial institutions can be safely shared with organisations working on developing their own products and services. There are broadly two approaches to the creation of these synthetic datasets: - he use of simulation – known as 'agent-based modelling' - where data is generated from approximations and predictions of behaviour. This broadly involves setting parameters based on known outputs, expected ranges, using known expectations on format for output ,and generating data based on these variables. There is no personal data whatsoever utilised in the creation of this kind of output, resulting in a low risk dataset that can be utilised by companies for testing, scenario modelling and other innovation. - Using 'learning-based' synthetic data generation to create synthetic doubles of existing datasets, utilising initially differential privacy and then modern learning-based approaches which look to (1) learn all the

Commented [AP1]: This section re-written to address scope but also simply because this activity's proposal has changed significantly since this point.

CAPTIONS

Sub heading
Body caption copy

Sub heading
Body caption copy

SMART DATA FOUNDRY

	meaningful patterns in data (2) use this learnt knowledge of patterns in original data to generate new data that exhibits similar patterns without recreating any of the input data It is the latter proposal which requires thinking with regards to potentially processing existing datasets which include personally identifiable data, and judge the legitimate interests in doing so by weighing the benefits and purpose against potential infringements to privacy rights of individuals. As part of mission-led research Smart Data Foundry conducts, there will be datasets containing personally identifiable data held within secure managed storage – there is a separate Legitimate Interest looking at the research use case. The process activity under scope for this LIA is: - Generating synthetic doubles of datasets Smart Data Foundry are controllers of and have a valid legal basis to utilise for mission-driven research to assist in **effectively anonymising the data** for further use in research and innovation within an appropriately secure and risk-controlled environment. - Instituting appropriate privacy and disclosure controls to ascertain the reidentification risk of these datasets generated from research data - Providing vetted access to these "high utility and low re-identifiability" datasets, ensuring use cases are congruent with the SIPF objectives of enabling innovative approaches to improve financial services (more detail below). The intention for this is for company to be able to test and develop their product on high quality data that will not carry the same privacy and compliance risks due to the data no longer carrying the risks of re-identifiability and associated data protection risks. Ultimately, the purpose of this data processing in generating this synthesised data is to help technology startups and research organisations accelerate research methods and testing to ensure these entities can validate their products and prove their ideas on quality data which has a lower data privacy risk. As a result, this would help organisations on their journeys in becoming viable and profitable businesses. This is in alignment with our core objectives enabled by Strength In Places Funding from UKRI, which includes the objective of "Catalyse innovative approaches to improve financial services, increase fintech success and deliver benefits to the regional economy". Research and innovation purposes as per provisional guidance from the ICO defines some of the indicative criteria for research.

	From this indicative list, the following apply to the purpose of this activity: • Formulating hypotheses, isolating variables, designing experiments • Objective observation, measurement of data • Findings do not lead directly to decisions about individual subjects To help ensure the synthetic data generation is appropriate for scope when considering organisational objectives and the research and innovation purposes, each dataset that is considered for synthetic double generation will undergo a legitimate interest balancing test to ascertain the utility and suitability of the dataset against the organisation objective for "innovative approaches to improve financial services", as well as ensuring the dataset is appropriate for and congruent with the three indicative research purposes of "formulating hypotheses….", "objective observation….", and ensuring "findings do not lead directly to decisions about individual subjects".
What will be the benefits be? (include detail on if it meets a specific organisational objective)	The organisational objectives that the generation of synthetic data aids are as follows: **Open Finance for All:** We see the future of Open Banking as the blueprint for Open Finance and Smart Data, available to all parts of society, especially groups that are traditionally excluded, designed to protect against crime, resilient to failure and future shocks to the system. Smart Data Foundry is ideally placed to drive collaboration with Government, Regulators, FinTech's and the wider Financial Services community. We will make sure that Open Banking works for everyone and paves the way for a better future. This data processing purpose may **indirectly** help assist the other organisational objectives – **Stop The Squeeze** on cost of living, **Countering Climate Change** and supporting **Strong Small Business**, but these would be indirect outputs from the development of a more open, vibrant financial technology market with more resilient operationalised research. Furthermore, the criteria for project approval within Smart Data Foundry internally involves: • Strategic: Is this project a strategic enabler? How does it align with our missions as outlined above? • Reputational: Does this enhance Smart Data Foundry's external reputation? Does it create / strengthen relationships? • Commercial: Is this a revenue generating project? • Operational Effectiveness: Does this improve the efficiency of Smart Data Foundry? For this specific project with regards to synthetic double generation, we will institute an additional control to ensure the 'doubling' of this dataset is congruent with our organisational

	aims as approved by the SIPF bid from UKRI. Our 'Strength in Places' application specifically focused on the below: 1. Catalyse financial data collaboration in a trusted environment to enable research, innovation and social inclusion, thus accelerating industry adoption of Open Finance at scale in Central Scotland 2. Decrease time-to-market and development costs, thereby improving start-up survival, attracting inward investment, securing jobs and creating export opportunities After developing our offering further and conducting market analysis of the finance sector, we have focused on synthetic data provision as a primary enabled for fintechs and organisations focused on innovation, and this aspect has been agreed with the Strength In Places fund officer as our primary contribution to enabling financial innovation. With the generation of these data sets for use, the hope is that the use of the synthetic data outputs can affect real societal change within government and industry through the development of innovative and successful technology. With the provision of high utility but well synthesized data, Smart Data Foundry will have assets to help generate meaningful outcomes and benefits within the technology sector in south-east Scotland; a key objective from the Strength In Places Funding that Smart Data Foundry reports against. For furthering innovation, the alignment with strategic missions as stated in purpose section will be a key milestone to approving a course of action and ensure the project benefits are clear and tangible. There is also an **indirect benefit** to data subjects whose data is being utilised for the generation of synthetic data. In encouraging organisation to do their testing on synthetic data which has **no personally identifiable characteristics,** the usage of more personal data that the organisations may be using without adequate controls is being discouraged. There is also a revenue generation benefit for Smart Data Foundry; the organisation requires to be self-sustaining, and the generation of these synthetic data outputs and provision of a technical environment for utilising these datasets can help generate revenue to fund further research and innovation within the organisation. This overlaps well with the examples of public benefit suggested by the provisional ICO research guidance, which includes 'improved financial or economic outcomes', 'advancement of academic knowledge', and 'the provision of more efficient or more effective products and services for the public'.

Who are the stakeholders? (Including segments of data subjects)	Our collaborators in general can be summarised to be: - UK Research and Innovation: Smart Data Foundry was created from a significant grant bestowed by the research and innovation department in the government. They are a significant stakeholder that Smart Data Foundry will report findings to, such as alignment with the objectives described in the Strength In Places research funding application. - Fin Techs and Startups: Stakeholders and interest parties in the goal of testing and innovating new products, who require different forms of data to scenario-test and develop these products on - Academics and Researchers: Utilising the data to test code and analysis work in a safe environment with low risk data prior to proposing utilising this analysis on real data, to test hypotheses and methods. The wider stakeholders are the data subjects themselves – specifically data subjects who are economically active and have a financial data footprint, sampled from the datasets of data providers that Smart Data Foundry will have access to. , The broad missions detailed above , if innovation is successful, should lead to outcomes such as a more vibrant financial technology market with high quality jobs within the sector, more effective research and innovation with a lower risk of re-identifiability than if using 'real data', and generally a more active commercial marketplace for financial technology along with the societal benefits that brings.
Any ethical implications? *Ethical framework within Smart Data Foundry broadly looks at Transparency, Accountability and Fairness based on UK Government framework*	Upon formation, Smart Data Foundry developed an ethics checklist which was utilised in creating the initial business case. Since then, we have utilised the UK Government ethics framework of Transparency, Accountability and Fairness to develop a rating. This is some text summarising the National Data Utility's contribution to these three principles: Transparency: This processing activity is one of the core purposes of Smart Data Foundry and will be widely advertised as such; as a result, there will be significant communication from the organisation centred on the availability and usage of these innovation datasets. There can be further work done to ensure the wider data flow - I.e. the data partners we work with, the type of data collected and in what form, and the outputs – are clearly communicated on to ensure transparency is at its fullest. Broadly there are no significant transparency quandaries with this use of data. Accountability: Each approved innovation project requires approval from the project board, which includes representatives from across the organisation and senior leadership. Each project will include a steering group ensuring stakeholders listed above – such as the data providers and relevant representative bodies – are represented.

	Fairness: There are no direct decisions – automated or manual – being taken as a result of the research projects. The outputs may be used to influence or campaign for policy changes which may have indirect influences on decisions made. Where possible, given further opportunities to develop synthetic double datasets for specific use cases, we will attempt to produce data that accounts for biases within financial data such as within gender and race, but this requires further exploration. This is also further helped by the fact that organisations will have access to **synthetic data** for this innovation where the risks of reidentification are remote or non-existent. We are working on integrating these principles to be tested against missions and projects we undertake with data available for research.
Any regulatory implications other than GDPR?	Depending on the sector of the research, there will be regulations around access to certain types of data. This will be considered on a case by case basis dependent on the dataset.

Commented [AP2]: the example given was to generate data that more accurately reflects joint accounts not just predominantly being in the name of a male partner in cishet relationships but we can develop this further if essential - it will be an aspect we further develop as we test synthetic doubling and its use cases. I will reflect that.

Necessity

Assess whether the processing is necessary for your purpose

What is the data processing being proposed to achieve your purpose?	As previously stated, there are two stages to the data processing; generation synthetic data and enabling the use of this. These are further detailed below: **Generation of Synthetic Anonymous Data:** The path to generating synthetic data consists of the following phases : • Negotiate a series of data sharing agreements with Data Partners in which Smart Data Foundry achieve appropriate usage rights as controllers of data to generate synthetic doubles of data, for further use by other third parties. This data may originally be **personal data** for which Smart Data Foundry will have to ensure an appropriate legal basis for processing, with the aim of ensuring the data is fully anonymised following the synthesis of the data. This anonymisation will help safeguard privacy of individuals for future use by other parties. • To do this, we will import the data into our secure Safe Haven data environment which is purpose built to securely hold personal data, as detailed within the DPIA for further information. • Prior to conducting any learning-based synthesis, we will focus on (or work with data providers to) deidentifying the data to ensure re-identification risk is already mitigated as much as possible while considering utility. This will include using common

SMART DATA FOUNDRY

techniques such as banding categories such as age, using larger categories for location such as partial post code, and using aggregates or lower significant figures for specific amounts where this does not contaminate innovation or research use. This prior reidentification – which will include differential privacy techniques to ensure outlier data figures or data points that stand out from samples are sufficiently 'noised' and hidden – will ensure that reidentification risk is mitigated even in the case of accidental breach or misuse by subsequent data partners.

- Following generation of the synthetic data, there will be a series of tests to ensure re-identification risk has been eliminated to the degree possible to test. This will include using privacy metrics that measure disclosure risk – such as ensure no duplication of rows against original data, ensuring no rows that are measurably close to the original data. These metrics will be derived from generated academic research from MIT as well as open source metrics proposed by two synthetic data generators in SynthPop and Diveplane. On top of these metrics – and starting from deidentified data rather than personal, and also applying differential privacy prior to synthesis – we would also simply do tests to eyeball the data and try to judge whether specific records feel "too real" compared to real data. Further information is available in the file "Synthetic Doubles and Disclosure Risk" which we will keep current and work on further detail.

- This environment will also contain the necessary code packages and software to adequately synthesize new data which would be **effectively anonymised** with a remote change of reidentification. Once effectively anonymised, this dataset can be moved into a different and more flexible technical environment where it can used and re-used by other organisations as required, provided we are confident this data is now fully anonymised.

Enabling the use of synthetic data

Once this data is no longer personal data, this data is not within the scope of the GDPR and related data protection legislation. However, we will still ensure some controls are in place to ensure the security and safety of these data assets. This pathway will consist of the following:

- Organisations will have to register for access to a secure portal to access these synthetic doubles, including clearly stating their use purpose to ensure

	they are congruent with our company objectives and missions. • Data will be shared and updated with future versions. Different tiers of data will have differing controls on export – including Information Governance checks (on minimum field quantities) on data that is exported to ensure any possible re-identification issues are further mitigated.
Is this processing a **reasonable** method to carry out the task?	With the missions outlined above, and the necessity for synthetic data to help improve and accelerate technology development for research or innovation businesses, it is reasonable that data processing on a large scale basis is required to produce high quality synthetic data that is no longer personal data and complete anonymised. Since the data will be deidentified prior to the synthetic doubling processing, the risk of material impact as a result of accidental breach or misuse is further mitigated. Protecting against reidentification in the first instance will help ensure this processing is reasonable in carrying out the task. Since the output is data that will now be anonymised to the level where there is not a probable effect on a data subject from the disclosure of the data, the processing feels reasonable to ensure the data subject is protected.
Is this processing an **effective** method to carry out the task?	To generate synthetic data that enables the benefits listed above – which include effectively anonymising datasets to protect data subjects - you must have controllership of the data and process this data. In ensuring Smart Data Foundry will employ and utilise experts in the field both in personnel and software, we can ensure this processing is an effective method to carry out the task. As validation of the team's approach to generating synthetic data being an effective one, the data science team – in collaboration with the GEMINAI synthetic data system – developed a "Ten Steps To Synthesis" process which won an award from the High-Level Group for Modernisation of Official Statistics (HLG-MOS) - a report of which can be found here Generating Synthetic Data (smartdatafoundry.com)
Is the processing proportionate or can you achieve the same purpose in a **less intrusive** way?	It is necessary and proportional to generate synthetic, anonymous datasets to help enable technology and research growth for small companies who may not have access to the requisite data and technology to build their products. Since the output and result of this processing is to further protect the privacy of data subjects in enabling more protection against reidentification, it can be argued that this is the least intrusive method in doing so as it is being done within: - a highly secure environment with no external access

	- With data which will have a previous legitimate interest in holding that is additional to creating these synthesized versions The security controls of the technical environment include maintaining a segregated area ensuring that all original data collected will not be directly accessible to data scientists prior to Information Governance checks on data minimalisation, data quality and assigning an appropriate information handling level. This ensures that only the data required for achieving the task will be made available for further processing and synthesizing. As a counterfactual, it will not be possible to achieve mission outcomes without the requisite data – the original data is required to generate the synthetic data. The processing does help ensure the **least intrusion** possible for future use cases once synthesis is complete.
Will this processing help achieve the purpose? *Include a counterfactual if helpful to highlight problems faced if processing was not conducted*	Yes – this processing is necessary to achieve the missions and purpose related above. Without the processing of data to generate synthetic data at this scale, it will not be possible to provide data to improve technology products of small research and technology businesses. Without synthesizing the data, you also cannot avert the usage of less depersonalised/more risky data being utilised by these businesses, which would carry higher data privacy repercussions and risks.

Balancing Test

Consider the impact of the processing on the data subjects' interests, rights and freedoms

Is it special category, criminal offence, children's or sensitive data?	The data techniques are currently too new and have not undergone enough uses **in anger and in production** to risk the possibility of accidental exposure when involving special category data. This will also be difficult to implement when completing the deidentification stage on datasets which have low density special category information in specific crosstabulations – for example race categories within low density population areas. If utilising this form of data, a separate DPIA will have to be completed which looks at all requirements of using special category data for research.
Would the individual expect the processing activity to take place?	It must be recognised broadly that within the sectors Smart Data Foundry operates – which is predominantly the finance sector – there is not a wealth of data being made available to share for innovation. As a result, it is reasonable to expect that the individual is not expecting the processing activity to take place.

	However, while not 'expected', it can be reasonably forseeable and also within the data subjects' interest that this data is anonymised further to prevent the possibility of reidentification, and that innovation tests and being conducted on this synthesised data instead of their personal data. With the aims and missions of Smart Data Foundry being advertised, it is also reasonably forseeable that such a research institution has partnerships with other organisations that provide datasets for mission-driven research and innovation. It is also likely they would expect stringent security controls and the data to be effectively anonymised wherever possible to ensure the least amount of intrusion, which is the approach being adopted by Smart Data Foundry.
What is the communication with the data subject? (include detail on 'how' and 'when')	Where possible, Smart Data Foundry works with data partners to ensure their privacy notice contains appropriate information with regards to transferring data to third parties to enable research in the public interest for the original transfer of data. However, it is also reasonable to expect that since this is a secondary purpose to the primary 'expected' purpose in utilising a financial service, this is less likely to be read in detail and acknowledged. Smart Data Foundry has its own privacy notice that sets out its processing of data. This is unlikely to be viewed directly by a data subject unless directed to by the data provider. However, if there are concerns for the data subject with regards to their rights, it would be realistic to assume the Smart Data Foundry privacy notice is an area the data subject may search for as long as they are aware of the organisation.. Article 14(5)(b) clarifies that Smart Data Foundry does not need to send transparency information to data subjects where personal data is provided to us if it would be unreasonable or impossible to do so. As the processing would render the data impossible to re-identify, sending transparency information does not apply for this.
Is the purpose clear to the data subject?	From the privacy notice and general purpose/mission of Smart Data Foundry, the purpose of the data usage is clear. The visibility and expectation of the processing is less so.
Does the processing add benefit or value to a service that the individual uses?	**Indirectly –** there are benefits to technology and research organisations being more effective in starting up with access to this data, including higher quality jobs and better services. Alternatively, in creating anonymous synthesised data, there is an absence of privacy impact on the individual as a result of this processing, which is a benefit and value.
What are the possible impacts on the individual's rights? How likely and severe? How could you mitigate?	It is reasonably likely that the individual may not be informed about the processing as the volume of personal data may prove informing to be a disproportionate effort. However, as stated above, it is reasonable forseeable that data is utilised for this form of processing – for research in the public interest.

	The creation of synthetic data for further use as a result of the processing effective negates possible negative impacts on an individuals rights – thus there are no further impacts to mitigate as a result. As this is new technology, we must account for the remote possibility of the data not being sufficiently anonymised due to either user error or technological accident. In case of this, the deidentification of the data prior to any synthetic doubling will be an important step to ensure it is difficult to re-identify a particular individual even in the case of the data being accidentally made available without synthesis.
Is the processing likely to result in unwarranted harm or distress to the Individual?	No – there are significant information security safeguards on the confidentiality, integrity and availability of the originally received data. Furthermore, from the creation of synthetic data, there is unlikely to be a harm or distress to an individual as individuals will not be identifiable. When accounting for the possibility of deidentified and not sufficiently anonymised data being leaked or made available wider than the individual expects, it is reasonable to predict that private data on individuals financials being made available to a person or organisation could cause distress to the individual as a breach of privacy. However, deidentification controls even prior to synthesis should ensure there is no specific detail available within the datasets that can be exploited to make a specific material decision affecting an individual that causes material harm. There will not be any original data containing specific identification that can be exploited for harm such as ID fraud.
Can individuals opt in/out?	The option to directly opt in or out is not currently available to data subjects due to the legal basis utilised by data providers and Smart Data Foundry. Individuals under the legal basis of legitimate interest do have the right for the data subject to object to processing and be removed from the dataset in that way – however, this will not be enacted in aggregated information that is already published as per the research exemptions within the UK Data Protection Act. To enable this right to object, individuals will need to know the third parties Smart Data Foundry works with, as we will not hold the direct identifiers for these individuals. **Right To Object and Erasure** Since the legal basis for processing is based on SDF's legitimate interests, individuals have the right to object to this processing. If – as part of the objection or right to access – they decide to request the erasure of this data (and that this request is successful against the criteria for erasure) – Smart Data Foundry have processes in place to process this erasure.

Are the legitimate interests of the individual aligned with those of the organisation or third party?	The research and technology development outcomes broadly would align with individuals – however there is no empirical research or evidence to substantiate this.

Decision

Does legitimate interest apply?

Does legitimate interest apply to this processing? *Include additional detail to supplement the reasoning of the balancing test decision made in above question*	Due to the original research focusing on the public interest and then this processing ultimately resulting in creating data that does not require a legal basis, legitimate interest applies for this processing when weighing the benefits against the possible impacts on data subjects' rights and freedoms. This is predominantly because the possible impacts on the data subjects' rights and freedoms and negligible due to the nature of the processing being to achieve effectively anonymised data in the form of synthesized data. However, to mitigate against any possible the risks of the data subject not being informed of this processing or not expecting the processing, Smart Data Foundry will endeavour to: - Make public-facing announcements and discussions on research projects and outcomes, including naming data providers it works with for the original research projects. This will include utilising the press when the product is used for high profile or potentially innovative advancements. - Detail the broad approach to synthesis and data provision within a section of the Smart Data Foundry website; since privacy is a benefit of this processing, transparency around how privacy and governance is done on this process should be part of it. This will include any possible outputs from ICO discussions on the topic. - Constantly innovate and utilise all possible controls in ensuring the anonymity of the data post-synthesis and deidentification pre-synthesis, including accounting for human error. - Ensure all original research proposals for data collection are tested against the criteria of meeting 'research in the public interest' as well as Smart Data Foundry's missions
Completed by *(Name/Title)*: Date:	Adarsh Peruvamba – Data Manager 26/05/2022

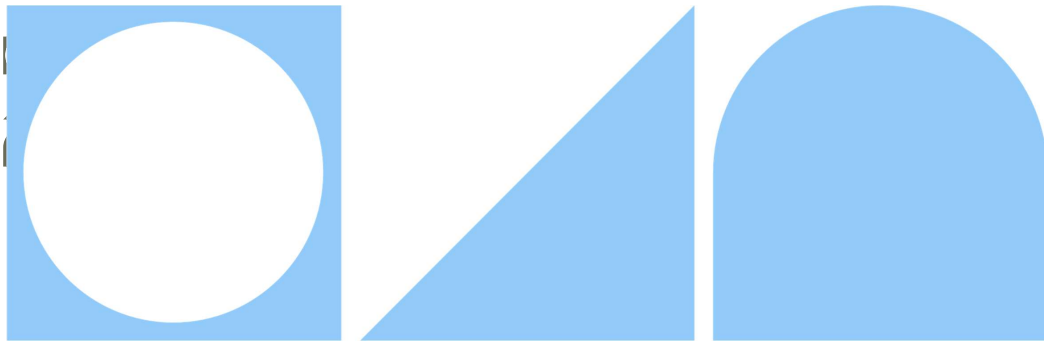
SMART DATA FOUNDRY

SMART
DATA
FOUNDRY

CAPTIONS

Sub heading
Body caption copy

Sub heading
Body caption copy



Legitimate Interest Assessment – National Data Utility

Purpose

Assess whether there is a legitimate interest behind the processing

What is the purpose of the activity?	The purpose of the Smart Data Foundry is "To open finance for good." One of the focus areas of the Smart Data Foundry is to use financial data (eventually including other smart data) for solving big societal, economic and environmental problems. Within the Smart Data Foundry, the purpose of the NDU is to provide a secure data storage and processing capability where separate, combined or linked datasets of deidentified granular financial data are assembled in a safe-haven archive. Datasets can be used either - For statistical and research purposes in the NDU - To create high quality synthetic data for testing, research and innovation in the Innovation Environment Research and innovation purposes as per provisional guidance from the ICO defines some of the indicative criteria for research. From this indicative list, the following apply to the purpose of this activity: • Formulating hypotheses, isolating variables, designing experiments • Objective observation, measurement of data • Publication of findings • Supporting diverse and inclusive research

Commented [SS1]: This is ambiguous do you mean for good as in forever or for good as in for the moral good? Or both?

Commented [SS2]: It may be helpful to out line phases of data inclusion and how you will consider relevance to the the data as a whole

Commented [SS3]: Will all data in the ultility be deidentified and granular? It may be helpful to define the term deidentified as in is not a legal term and may beused differently but different organisations

Commented [SS4]: This is what you are doing but it would be helpful to understand your purpose in terms of outcomes, eg better decision making, government policy, better societal economic outcomes?

Commented [SS5]: I think this could be tied to my comment about being clearer about outcomes

CAPTIONS

Sub heading
Body caption copy

Sub heading
Body caption copy

	• Findings do not lead directly to decisions about individual subjects
What will be the benefits be? (include detail on if it meets a specific organisational objective)	The objective for this research is to solve societal, economic and environmental problems, broadly defined as 'in the public interest'. The criteria for project approval within Smart Data Foundry involves: • Strategic: Is this project a strategic enabler? How does it align with our mission? • Reputational: Does this enhance Smart Data Foundry's external reputation? Does it create / strengthen relationships? • Commercial: Is this a revenue generating project? • Operational Effectiveness: Does this improve the efficiency of Smart Data Foundry? For research, the alignment with strategic missions as stated in purpose section will be a key milestone to approving a course of action. This overlaps well with the examples of public benefit suggested by the provisional ICO research guidance, which includes 'improved health and wellbeing outcomes', 'improved financial or economic outcomes', 'advancement of academic knowledge', and 'the provision of more efficient or more effective products and services for the public'.
Who are the stakeholders? (including segments of data subjects)	Projects are funded by: - Research funding councils - Collaborations with financial institutions - Public sector funding Project boards often include representatives from: - Representative bodies for stakeholders (such as Small Business Commissioner) - Financial institution representatives The wider stakeholders are the data subjects themselves, as outcomes focus on solving 'big problems' such as SME resilience, income and expenditure fluctuations during crisis, and poverty premium facing lower-income households when accessing essential goods and services.
Any ethical implications? *Ethical framework within Smart Data Foundry broadly looks at Transparency, Accountability*	Transparency: This processing activity is the core purpose of Smart Data Foundry; as a result, communication from the organisation is mainly centred on the research process and outputs. There can be further work done to ensure the wider data flow - I.e. the data partners we work with, the type of data collected and in what form, and the outputs – are clearly

Commented [SS6]: As discussed this area will be particulary important to expand upon significantly

Commented [SS7]: Can you define your strategic mission in the document?

Commented [SS8]: Can we be clear on who the data subjects are? I assume it will be a broad section of the econnmically active populace with no other discriminating factors, but it would be good to clarify

Commented [SS9]: Have you conducted an ethics assessment and was the outcome positive?

and Frairness based on UK Government framework	communicated on to ensure transparency is at its fullest. Broadly there are no significant transparency quandaries with this use of data. Accountability: Each approved research project requires approval from the project board, which includes representatives from across the organisation and senior leadership. Each project will include a steering group ensuring stakeholders listed above – such as the data providers and relevant representative bodies – are represented. Fairness: There are no direct decisions – automated or manual – being taken as a result of the research projects. The outputs may be used to influence or campaign for policy changes which may have indirect influences on decisions made. Where possible, data will be processed in a manner that accounts for biases within datasets.
Any regulatory implications other than GDPR?	Depending on the sector of the research, there will be regulations around access to certain types of data. This will be considered on a case by case basis dependent on the dataset.

Necessity

Assess whether the processing is necessary for your purpose

Will this processing help you achieve your purpose?	The path to establishing the NDU and achieving the purpose of providing secure data storage and data processing for statistical and research purposes consists of 2 phases : • Negotiate a series of bilateral agreements with Data Partners in which the NDU produce insights from their data for clearly defined purposes • Evolve agreements with Data Partners to allow the NDU to use datasets for multiple purposes, evaluate multiple datasets for deeper insights, and to create linked datasets The processing will require data suitable for each research purpose. As a result, this may include personal or sensitive personal data where the research outcome requires it.
Is the processing proportionate or can you achieve the same purpose in a different way?	Proportionality of the processing will be directly related to the research outcomes. Personal data or sensitive personal data will not be utilised by data scientists and researchers if this is not required by the research question. The NDU will maintain a segregated area ensuring that all original data collected will not be directly accessible to data scientists prior to The NDU will utilise pseudonymisation techniques as a

Commented [SS10]: The necessity test can be quite complex but should in its simplest terms consider if the this is the least invasive reasonable effective method to carry out a legitimate task. It will be helpful to highlight other ways that have also be considered to carry out the task and reasons why they have not be chosen

Commented [SS11]: It may be helpful to clarify that this processing will help to achieve the purpose? It is also helpful to construct a counterfactual assessment in highlighting the problems that could be faced if the processing wasn't conducted

Commented [SS12]: tbc

SMART DATA FOUNDRY

Balancing Test

Consider the impact of the processing on the data subjects' interests, rights and freedoms

Is it special category, criminal offence, children's or sensitive data?	Dependent on the use case, there may be special category or sensitive data involved. In these cases, we will require to ensure that this data is necessary for the purpose - I.e. it is directly related to the outcome and benefit of the research – and that appropriate safeguards around the privacy and freedoms of the individual are observed. In these cases, the research will not be utilised to make decisions about particular individuals or cause any substantial damage or distress to an individual
Would the individual expect the processing activity to take place?	It must be recognised broadly that within the sectors the National Data Utility operates – which is predominantly the finance sector – there is not a wealth of research data being made available to tackle problems facing the public. As a result, it is reasonable to expect that the individual is not expecting the processing activity to take place.
What is the communication with the data subject? (include detail on 'how' and 'when')	Where possible, Smart Data Foundry works with data partners to ensure their privacy notice contains appropriate information with regards to transferring data to third parties to enable research in the public interest. However, it is also reasonable to expect that since this is a secondary purpose to the primary 'expected' purpose in utilising a financial service, this is less likely to be read in detail and acknowledged. Smart Data Foundry has its own privacy notice that sets out its processing of data. This is unlikely to be viewed directly by a data subject unless directed to by the data provider.
Is the purpose clear to the data subject?	From the privacy notice and general purpose/mission of Smart Data Foundry, the purpose of the data usage is clear. The visibility and expectation of the processing is less so.
Does the processing add benefit or value to a service that the individual uses?	Yes – the objectives of the processing are to solve societal, economic and environmental problems. While the research will broadly be focused on breadth and solving wider problems, the publishing and utilisation of results should help the advancement of knowledge and eventually provision of more effective services to improve health of society.
What are the possible impacts on the individual's rights? How likely and severe? How could you mitigate?	It is reasonably likely that the individual may not be informed about the processing as the volume of personal data may prove informing to be a disproportionate effort. In the cases of holding pseudonymised data, it may not be possible without the data subject divulging further personal information to enact the right to object. Mitigations would include making it clear in the privacy notice as to the data that is held and in what form, and in partnership with which third parties. Right to object and erasure can still be enacted provided the data providers assist in doing so.

Commented [SS13]: Even if not expected it could be seen as foreseeable but there may be further expectations of the data subject such as not specific bits of their data

Commented [SS14]: Maybe helpful to directly reference the legislation in terms of transparency notices to data subject where data has been obtained from a third part

Commented [SS15]: You may wish to consider the wider rights of the data subjects outside of the specific DP rights. This processing may not impede those rights and if so it would be good to document here

Is the processing likely to result in unwarranted harm or distress to the Individual?	No – there are significant information security safeguards on the confidientiality, integrity and availability of the data. The outcome of the data being processed is for research in the public interest, with no decisions being made against individuals.
Can individuals opt in/out?	The option to directly opt in or out is not currently available to data subjects due to the legal basis utilised by data providers and Smart Data Foundaries
Are the legitimate interests of the individual aligned with those of the organisation or third party?	The research outcomes broadly would align with individuals – however there is no empirical research or evidence to substantiate this.

Commented [SS16]: Maybe helpful to consider the risk of harm from a breach given the deidenifed nature of the data

Commented [SS17]: The legal basis is legitimate interest which does allow the datasubject to object. You may wish to look at the specific DPA schedule restricting the right to object in regards of research and how you could apply the conditions attacted to it

Commented [SS18]: It may be helpful to think about some limited consultation

Decision

Does legitimate interest apply?

Does legitimate interest apply to this processing? *Include additional detail to supplement the reasoning of the balancing test decision made in above question*	Due to the research focusing on the public interest, legitimate interest applies for this processing when weighing the benefits against the possible impacts on data subjects' rights and freedoms. However, to mitigate against the risks of the data subject not being informed of this processing or not expecting the processing, Smart Data Foundry will endeavour to: - Make public-facing announcements and discussions on research projects and outcomes, including naming data providers it works with - Ensure all research proposals are tested against the criteria of meeting 'research in the public interest' as well as Smart Data Foundry's missions - Utilising only data as it necessary, ensuring that by default that: o data utilised is effectively anonymised o Recording the necessity where effective anonymisation may not be possible o Ensuring separation of duties between data preparation and data users where utilising data more likely to be personal data (I.e. linked data or data containing free text or identifying characteristics) o Ensuring additional safeguards are in place when researching data which includes special category data
Completed by *(Name/Title)*:	Adarsh Peruvamba – Data Manager

Date:	10/03/2022

SMART DATA FOUNDRY

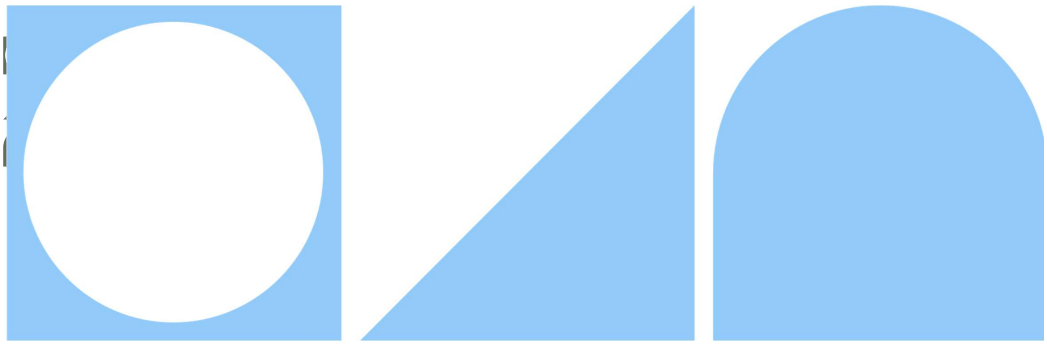
SMART DATA FOUNDRY

SMART
DATA
FOUNDRY

CAPTIONS

Sub heading
Body caption copy

Sub heading
Body caption copy



Effective Anonymisation Assessment – National Data Utility

Dataset in question: FOIA s.44 - Prohibition on disclosure

What elements of the data are identifiable considered against the below criteria?

Description of dataset involved *Include detail on fields, type of data, and other data (including reference data) that may be utilised in conjunction with this dataset*	FOIA s.44 - Prohibition on disclosure [redacted] with consistent pseudonym. Age band, sex, partial postcode, total credits and debits, breakdowns of credits and debits into FCA categories (currently 6 for credits, 7 for debits) and min, max, final balance. Pseudonym is generated and held by FOIA s.4 – Smart Data Foundry does not have access to the keys of index that would allow reidentification. For projects currently utilised, this dataset is not linked with any other dataset beyond reference data (post code mapping). While this data is stored in the same area as other datasets, the is no combining or linking of these datasets performed, and there are no shared pseudonyms.
Singling out *Are there data points that allow you to differentiate one **individual** from another, where you can isolate some or all records about an individual in the data you process?*	Yes – it is possible to identify an individual with a particular level of credits and debits, including detail within categorisation that provides a broad age band and general location. However, it is not possible to identify the individual as a 'named' person or retrace the data back to an individual.

Commented [SS1]: DO you (SDF) hold the key or index allowing for reidentification of the individual on the basis of the pseudoidenifier or it this held by the originating financial institution

Commented [PC2R1]: "pseudoidenifier" should be "pseudonym"

Unclear here if any other datasets are combined. If they aren't then that should be made clear here.

Commented [PA3R1]: Have added detail on this.

Commented [PC4]: Is there any risk of an internal adversary combining this data with other datasets?
If so, how is the risk mitigated?

Commented [SS5]: Would this allow you to reasonably identify the the individual as a named person or do you mean you can flag that the data sets will all l belong to person a who is still unknown to you?

Commented [PC6R5]: Would still be PD in both cases. The first case would be more identifiable.

Commented [PA7R5]: Cannot retrace back to an individual.

Commented [PC8R5]: I'm not clear on the response here. Your response is yes (singling out is possible), yet you have stated you cannot trace back to an individual.

Singling out does not have to identify by name. If you can individuate a person's record from others in a dataset, it is still singling out.

Chapter 2 of the draft guidance provides more detail on how this assessment should be approached.

"*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.*"

- consider the richness of the data and how potentially identifying different categories are.
- You also need to consider whether sufficient safeguards are in place to reduce this risk.

For example would the combination of attributes on credit/debit information and any other information in the database allow an individual record to be singled out? Would these combinations of attributes be unique, such that it would only apply to a single individual? If not, then how is the risk mitigated? E.g. small count suppression, aggregation of location/ financial data?
What tests are done to confirm no-one can be singled out in the data?

CAPTIONS

Sub heading
Body caption copy

Sub heading
Body caption copy

Linkability *Are there elements that could be present in multiple datasets and hence linked across other datasets to create a 'mosaic/jigsaw' effect, such as date of birth or IP address?*	No, we do not have identifiers that could be directly linked to another database; data of birth has been banded into an 'age band', and post code is partial. Data has been pseudonymised and provided with a reference field – however, we do not possess the key, this is with the original data provider. With the absence of date of birth or specific post code, as well as the lack of publicly available data on total debits and credits spent per person, there is no detail at a granular enough level that would be able to trace this data back to publicly available datasets. Smart Data Foundry does not have a legal gateway to retrieve the pseudonymisation key.
Inferences *Can details about individuals be inferred or predicted using information from various sources and making correlations between datasets, utilising this to categorise or profile individuals? This can include specific knowledge of others being utilised (such as family members or doctors who may have specific information available)*	Inferences from categorisation of income (such as benefits, salaried income and pensions) can be made perhaps on broad characteristics and profiling such as class. The sample of data also includes individuals with a high level of debits, which certain inferences on wealth can be made about. However, as this is a sample of data as well as the limited granularity of indicators such as age and location, an individual's identity cannot be inferred.

Commented [PC9]: This assessment should also consider publicly available information, not just data held within the company.

Commented [PA10R9]: added detail.

Commented [PC11R9]: This is better, what about other publicly available data, e.g. companies house?
It would be worth mentioning that the limited access controls would reduce linkability risk.

Again, it would be good to provide some information on any testing that was carried out on linkability of data sets (e.g. motivated intruder testing).

Commented [PC12]: Does SDF have any legal gateway to retrieve the key?

Commented [PA13R12]: added detail.

Commented [PC14R12]: All good

Commented [PC15]: Could an individual's identity be inferred?

Commented [PA16R15]: No.

Commented [PC17R15]: What about other sources that could be used for infer?
How do you deal with outliers, e.g. low/high values?
For example high earners or individuals with large debts?
Is there a risk of a high profile bankruptcy or wealth in the media being linked to someone in the data?
What is the size of the sample? (assume [redacted] from first para?).

Means reasonably likely to be used

Assess whether the processing is necessary for your purpose

Data and the environment in which processing will occur *What technical and organisational measures applied to control access to the data and reduce identifiability risk?*	For the full details on technical and organisational controls, see [Safe Haven/National Data Utility DPIA] Technical and organisational measures on controlling access specifically from an identifiability lens include: - The data will be hosted in a secure virtual machine environment at an ISO27001 facility, protected in transit by AES-256 level encryption. - All data entering the facility will transfer through an ingress quarantine area where it will be checked and further de-identified where necessary to mitigate confidentiality risks. Similarly, all outputs from the project will transfer through an egress quarantine area and be subject to disclosure control. - Only authorised project researchers will be allowed to access the virtual machine and the data, with role based access documented and logged by Information Governance and the service provider. All users require

Commented [PC18]: Do you use 5 safes principles?

	access to university VPN and then 2FA/multi-device authentication for the VM - this is enforced by default. - There are contractual controls in place for all researchers with significant penalties in the case of a user using the data to accidentally or deliberately re-identify individuals from the data if that is unreported to IG. Users are reminded of this when logging into the virtual machine, upon induction, and various risk management discussions. - No network access to the internet with data only entering via managed file transfer known as Serv U. This file transfer is managed by the information governance team, with data encrypted and scanned for malware/AV on entry. - Pseudonyms derived from utilising a hash function to map a real identifier to a hash value and then subsequently salted. The requirement of salting helps mitigate against the ability to back-compute a hashed ID and work out the hash in the remote chance the original ID is known. - Technical restrictions preventing any information or data being copied on or off systems. Copies of the data are also limited to project use case – all additional copies deleted once not used.
Context, scope and purposes of the processing *What is the sensitivity of the variables in the dataset?*	While none of the data is 'special category' data, the level of income and expenditure per person is data that is broadly classed and treated as 'private'. The context is to provide research output and aggregated dashboards on the changes in income and expenditure over a wide time period. Since the outputs are aggregated to a minimum checksum of 10 per sub-category, the sensitivity of the variables in the dataset are reduced in terms of disclosure.
Reasonable means available to a motivated intruder *Include thinking on motivation, competence needed, cost and time required, the available technologies, and legal gateways/likelihood of their use*	To consider the reasonable means available to a motivated intruder, we will regularly refer back to the technological environment described in the earlier section. **Motivation:** While the data can be considered valuable due to its exclusivity, the data itself does not contain sufficient personalisation to identify individuals on a large scale. However, as stated, the data is commercially valuable and made available for research in an exclusive basis due to this. If wishing to harm [redacted] as a data provider, this could also be part of motivation in exposing data of their customers. **Competence needed:** The competence required to penetrate the ISO27001 certified environment that the data is stored in is likely beyond the bar of 'reasonable competence and appropriate resources' due to the additional barriers of multifactor authentication, no connection to wider networks,

Commented [PC19]: Any contractual controls in place, e.g. penalties for re-id?

Commented [PA20R19]: added

Commented [PC21]: Sp.

Commented [PC22]: This all seems sensible. Would be good to steer SDF towards Ch4 of the anon guidance if they haven't already read it.

Commented [PA23R22]: Thanks - have based some detail off that wording but will generally link it as recommended controls to refer to.

Commented [PC24]: Aggregated?

Commented [PA25R24]: aggregated

Commented [PC26]: such that the risk of re-identification is remote?

Commented [PC27]: See point above regarding other datasets held by SDF

	and the data ingress/egress checks which ensure data cannot be exported from the environment. **Cost and time required with available technologies:** If access to the data is achieved, due to the pseudonyms being hashed and salted; even if an ID is prior known and recognised within the data (which is a remote possibility and not part of the motivated intruder test), the hash cannot be prior computed without knowing the salt. This is likely beyond the possible computational ability available to an intruder reasonably. With the above in mind, especially considering the relatively low level of sensitivity of the data and the lack of direct personal identifiers, it is a remote risk that a motivated intruder would succeed in gaining access to the data considering the likely available technology, competence and time required. **Motivated insider:** This discussion is focused however on an external user. However, once an individual is an approved researcher, there are means they can identify the data with if combined with external information or prior knowledge (for example, if they knew their own or another's exact amount of debits and credits and this was included in the FOIA s.44 - Prohibit [redacted] sample) – however the salting of the hash ensures that they would not be able to recompute the hash used for all records. There are legal agreements with data providers and contracts to help enforce against this. This is generally treated as a remote possibility. Researchers also undergo a disclosure check against prior criminal convictions before having access to the data environment.
Data disclosure and release *What data is being disclosed to whom, in what form, and in what stage?*	There are three stages to this: - Data is sent to the information governance team and so is available to view by the IG team at that stage - these then undergo data quality and identifiability checks. This data is already pseudonymised (without SDF having access to the keys) at this stage. The checks include removing the data of individuals who have opted out (prior informed by FOIA s.44 as well as removing fields/rows of data if accidentally including 'bad data', which could include accidental additional personal data theoretically. If the data requires further deidentification as a result of this assessment – for example the removal of a field increases risk without adding value to the project – this will be completed by the IG team at this stage. - The 'cleaned' or 'approved' data is then revealed to approved researchers (with access to data analysis tools such as PyPi and CRAN, among other standard data science analysis tools). Any additional software required by the data science team is pre-reviewed by the IG team prior to enabling access within the environment beyond the standard pre-approved toolset.

Commented [PC28]: Too vague. What techniques are used? How do they prevent re-id?
Consider the flow diagram on Ch2,

What about the cost, time taken to re-identify and available technology to re-id (computational power)

What level of competence would be required?

Commented [PA29R28]: have added some additional detail but open to suggestion re: what kind of specifics I can further add here with the context of the technical controls in place.

Commented [PC30R28]: What measures do you have in place to ensure the salt is appropriately secured and separated from the hash?

Are consistent hashes required for the processing?
If they are the salt would have to be shared as well as the hash value, therefore there is some risk of brute-force attacks. How do you mitigate against these?

Commented [SS31]: How much effort would this require?

Commented [PC32R31]: What information would they likely need to do this?

Commented [PA33R31]: Have added detail

Commented [PC34]: Remote risk?

Commented [PA35R34]: yes.

Commented [PC36R34]: If the risk is remote, then it could be considered effectively anon in the hands of the researcher.

Commented [PC37]: Is this always PD?

Commented [PA38R37]: In this specific example it is - will do the assessment against any new form of dataset we get, is the idea (or if a project involves multiple datasets combined/linked)

Commented [PC39R37]: The identifiability assessment should be repeated when any new dataset is introduced and over time due to technological progress. Ch2 of the draft anon guidance provides further detail.

	- Data is published to external stakeholders in a aggregated form which is checked by information governance. These checks include ensuring that all subdivisions of the output have a minimum checksum of 10, and that any fields with below that level are listed as NULL or removed. These checks also further ensure there is no accidental personal detail exported as a part of the data processing – such as a debit or credit level of an individual.

Risk Assessment and Mitigations

What techniques are agreed to be utilised to ensure re-identification is unlikely?

What are the identifiability risks listed from the above assessment?	Singling out – identification of debits and credit levels Inferences – class or employment inferences from data Motivated intruder – Motivation due to commercially valuable data. Motivated intruder – access to tools when approved for research
What anonymization techniques are to be utilised prior to making data available for further processing to mitigate the risks identified? *Utilising the Smart Data Foundry De-identification guidelines or otherwise, please list all mitigations in additional to organisational and technical measures listed above*	**Singling out – identification of debits/credits levels** Due to the lack of availability of the pseudonym key, the salted hash used as a pseudonym, and the lack of identifiable fields against other sources of data (I.e. age band instead of DoB, first half of postcode rather than full), it is unlikely that access to the data will give the ability to identify a named individual. It is acknowledged that even with this understood, the data is 'personal data' as it is possible to identify credits and debits of an unnamed individual. The controls ensuring that wider disclosure of this data is aggregated prior to view helps mitigate this particular aspect. The other control in this regard is the environment itself – unless specifically given access, the likelihood of circumventing controls on access to the technical environment are reasonably remote to mitigate the risk of the data. **Inferences – class or employment inferences from the data** The purpose of the database is to measure the change in debits and credits across a period of time. The research questions based on the dataset will focus on this detail, making inferences based on income and expense. Data minimisation – ensuring that it is a sample of FOIA s.44 - Prohibit [redacted], truncating location data and banding ages, and not including further categorisation that necessary for research –

Commented [PC40]: What techniques used for aggregation? What tests are done on the identifiability of the aggregated data?

Commented [PA41R40]: added some detail on this

Commented [PC42R40]: I think this answers some of my questions above.

Commented [SS43]: How and when is the psudoidentifer removed?

Commented [PC44R43]: What about the risk of singling out from a combination of indirect identifiers?
Is it still possible to individuate without the linking identifier?

Commented [PA45R43]: have added some detail here - it is remote due to the lack of granularity in location or age. But yes it is still possible to individuate to a degree - i.e. pseudonym x has abc data, pseudonym y has def data. Is this quite what singling out means...? I feel like this is the area that's hardest for me to understand from Ch2 - surely any data that lists row x = person x has the potential to single out, is that correct?

Commented [PC46R43]: As per my comments above. Could a researcher perform a query such that it would return a result that relates to a single person? Has any stress testing been performed on the data to confirm this is not possible?
It's not necessarily just a 1:1 relatiosnship between a row and person. Could a combination of attributes when queries, identify someone?

	helps mitigate the level of inferences possible from this level of data. **Motivated intruder – access to tools as a researcher** It is acknowledged that once a researcher is given access to the data within the specific environment, they have access to data in a form that may be able to identify individuals if they had prior knowledge of the individuals credits and debits on a monthly basis. This is recognised to be a remote possibility. The salted hashes also ensure that the hash utilised to compute the pseudo ID cannot be computed for the dataset as a whole retroactively. The controls on egress of data - I.e. the fact that IG have to check the detail prior to egressing data – also helps limit damage in case of misuse. There are also penalties in case of reidentification not being revealed. Contracts with [redacted] also ensure we will inform them in case of accidental or deliberate reidentification. **Motivated intruder – access to commercially valuable data** There is sufficient motivation for an intruder to attempt to access this data. The mitigations in this area include the broad information security controls listed above, which make the possibility of a successful intrusion remote.

Commented [PC47]: DO you have any mitigation measures in place for this risk?
E.g. noise addition on outputs?

Commented [SS48]: Im not sure that the detail of these mitigations in clear enough for me as an uninformed reader you may want to draw them out a bit more

Commented [PC49R48]: Agree, this is very little detail provided here, just vague statements.

This section needs to have more detail on the techniques used and rationale for using them.

Commented [PA50R48]: Have elaborated on the vague statements.

Commented [PC51R48]: See above point on rare cases.

Decision

From this assessment, is the information effectively anonymised?

Is the data effectively anonymized following the implementation of controls discussed? *Include additional detail to supplement the reasoning of this test*	From the section recognising the elements of the data, we have recognised that the dataset provides the ability to single people out (even if not named), as well as a general inference on wealth of the individual. From the section discussing the means likely to be used, It is recognised that the technical measures in place make the success of a motivated intruder in accessing the data to be remote. As for the access to the data for our researchers, the main control is contractual penalties as well as the salting of hashes used as pseudonyms to ensure wider re-identification is a remote possibility. From the risks and mitigations section, we can infer that while 'singling out' and 'inferences' both make this data personal data, it is recognised to be unlikely to be identifiable. Furthermore, when considering the data and its technical/organisational evnrionement, and the competence/technology required to obtain and re-identify the data, the identificability risk can be considered sufficiently remote. As a result, I would propose that while these controls are active, the data is **effectively anonymised.**

Commented [PC54]: Ok, this contradicts some of the comments made above regarding singling out.

Commented [SS52]: I would be interested to see how you put this together to clarify your decision

Commented [PC53R52]: Agreed, this section needs to provide an overview of the previous sections and provide any additional information not provided above.

Commented [PC55]: Can you provide more clarity on how the hashes and salts are handled?

Commented [PC56]: I think there needs to be a clear distinction between the identifiability of the data in the hands of SDF vs identifiability of the data in the hands of the researcher.

It is possible that the data is PD from the perspective of SDF but due to the controls, it may not be in the hands of the researcher. I think this needs to be made more clear in this section with stronger justification if you think the latter is true.

Completed by *(Name/Title)*:	Adarsh Peruvamba – Data Manager
Date:	08/04/2022

SMART DATA FOUNDRY

SMART DATA FOUNDRY

SMART
DATA
FOUNDRY

CAPTIONS

Sub heading
Body caption copy

Sub heading
Body caption copy