London Strategic Insights Tool for Rough Sleeping (SIT / SITRS)

Review of Information Governance in the Pilot (Phase 1) and Recommendations for Phase 2

October 2023

Table of Contents

Summary

Overall phase 1 of this pilot project was a real success. The project team completed the design, development and delivery of a Minimum Viable Product (MVP) solution joining up key data sources and providing valuable new insights into rough sleeping journeys in just 13 weeks. Rough sleeping services are extremely complex so being able to develop an operational system in such a short period of time is a huge cause for celebration. Particularly as this ambitious project on such an important issue for Londoners as homelessness, is now set to be rolled out more widely to become a pan-London initiative.

The following comments from participating organisations really illustrate the success of this phase:

"I love the connection between H-CLIC and rough sleeping data. This gives us insights we didn't previously have and it's so useful to have it all in one place. The Pan-London view is really helpful in enabling us to compare and benchmark our borough across the London average"

"Overall the Strategic Insights Tool will enable us to take a really strategic approach to accommodation commissioning"

"Previously we have only known what happens after our clients immediately leave our services, not what happens further down the line. The SIT enables us to put together a fuller picture of their journey."

Before the pilot stage was approved, a working group formed of Information Governance for London (IGfL) representatives across London boroughs, as well as key project stakeholders from participating service providers and the Greater London Authority (GLA), worked together to develop a Data Protection Impact Assessment (DPIA) on the project and to agree a Data Sharing Agreement (DSA) between participating organisations. In the course of this, they requested that a review of the DPIA be conducted between the pilot phase and proposed wider rollout.

That review was conducted during September-October 2023, and its findings are detailed throughout this report. A few of the key findings include:

- The pilot offers no obvious reason to reconsider the lawful basis conditions for processing. Through the delivery of the MVP solution during this phase, it has given participants increased confidence that the insights available to them through this tool can really support them in delivering effective and improved services to individuals experiencing rough sleeping. Feedback

from users and a wider group of interested organisations has been really positive.

- The pilot stage also offered no cause to reduce the scope of data being processed. The dataset was reviewed at the end of phase 1 to determine whether it was necessary and proportionate to continue to collect the data outlined in the original DPIA. A data checklist has been produced which includes reasons for why the data fields are required (either for the probabilistic matching, to produce the insights available within the tool in order to achieve the intended benefits, or for planned further development of the tool).
- Phase 1 resources encouraged service providers to review their privacy information and several phase 1 organisations made changes to ensure the clarity of privacy information provided to individuals, including the participation in research projects. We have seen a number of phase 2 organisations do the same in preparation for participating within the project. This has been a really positive impact of the project.
- The contractor (Faculty AI) has taken effective measures to ensure the security of the personal data being shared as part of this project. As a result, no security incidents have occurred. However we have had to take further measures to ensure organisations only submit data in line with the minimum data set requirements as at the beginning of the project some organisations shared data that was outside of this scope. Appropriate measures have been taken by the contractor and the project team to mitigate any risk and ensure this doesn't happen again. This is detailed further below.
- The data submitted was of sufficient quality for the purposes of the pilot and certain data cleaning / standardisation processes were implemented in order to further improve the quality of the data. Going forward it is important to support organisations to improve their data quality as part of the ongoing commitment to this project, to ensure the outputs of this tool are as accurate and valuable as possible.

Recommendations

<u>To improve data security:</u>

1. Provide organisations submitting data with a data checklist which clearly outlines the scope of the data request to reduce the likelihood of organisations submitting data that has not been asked for and is not used as part of the tool.
2. Automate the deletion of raw files where data falls outside the scope of the data request.

<u>To improve data quality:</u>

3. Provide participating organisations with resources and support to improve their data collection processes (detailed below).

To increase the likelihood of the programme delivering its intended benefits:

4. Deliver a workshop to phase 1 users to help them fully understand how to get real value out of the tool and work together to develop practical use cases.
5. Develop an ongoing forum for users to get together and share best practice, develop new use cases, and raise any ongoing issues.
6. Develop a 'trust centre' page in the tool to instil confidence in the data.

Other:

7. Issue a revised DPIA and DSA reflecting findings and amendments from phase 1.
8. Review the project again after the wider rollout is complete and the product enters a further development phase (considering additional data sources and functionalities).

Purpose

Has the purpose for processing remained the same?

The purpose for processing remains the same. Homelessness in London, including rough sleeping, continues to rise. Collectively, London seeks to implement a data-informed policy to address homelessness, resulting in improved collective action that cuts across institutions, services, and sectors. However, London has historically lacked a system that enables us to leverage the extensive existing data to achieve these aims. Rough sleeping data is held across multiple different systems which prevents users from understanding the full picture of the needs and journeys of the rough sleeping population. This makes it difficult to make strategic decisions about how best to support these vulnerable groups. The purpose of the project was to develop a system which allows join-up of data, thereby delivering the required intelligence and insights.

The purpose of the project remained the same throughout delivery of phase 1. It is a research project which aims not to change the delivery of support to a specific individual, but to inform and improve service delivery at a cohort, or population level.

Phase 1 focused on the development of an MVP solution of the Strategic Insights Tool for Rough Sleeping (SITRS). The SITRS is a new tool that will give decision-makers in GLA, London Councils, Local Authorities and homelessness

service providers, a clearer view of rough sleeping in their local area, through merging and integrating multiple key sources of data across the rough sleeping landscape:

- CHAIN (CHAIN records of London rough sleepers)
- In-Form (In-Form instances from service providers that work with rough sleepers across London)
- H-CLIC (Homelessness Case Level Collection submissions from London local authorities)

This means that for the first time, users of this tool are able to see the aggregated journeys of rough sleepers over time, as they show up through touch points in multiple systems, which include statutory homelessness applications; contacts with housing outreach officers, as they are seen bedding down; and interactions with service providers, who are commissioned to support them through various services.

Through the use of this tool, users are able to get actionable insights on how support can be improved, to make rough sleeping rare, brief and non-recurrent. By seeing the aggregated journeys of rough sleepers through various combined systems, users of this tool can better map the history and journeys of rough sleepers through their interaction with homelessness services. For example, by seeing the different inflows and outflows of rough sleeping by different boroughs, Local Authority or pan-London commissioners can make educated decisions about the effectiveness of different services and support, while forecasting and pre-empting rough sleeping trends over time.

Local Authorities can also see the inflows of rough sleeping into their borough, by seeing previous statutory housing applications that rough sleepers may have made, across London Local Authorities.

Service providers can see pan-London aggregated stats to understand the comparative benchmarking of certain rough sleeping services, as well as seeing how different boroughs compare in terms of aggregated rough sleeping journeys.

We have generated a list of example questions that users are able to answer through the SITRS. A second list of example questions specific to service providers has also been included:

Example SITRS user questions
1. Generally, is rough sleeping increasing or decreasing? 2. Which accommodation services do people go to after sleeping rough? 3. Which accommodation services support people into long term sustained and settled living? 4. Where do people go after interacting with particular accommodation services? 5. Before sleeping rough in my area, had people made a homelessness application, and where? 6. Where do people who sleep rough in my area go next? 7. What are the long term outcomes for people who have slept rough in my area? 8. What interventions lead to which outcomes for different groups? 9. What are the demographic characteristics of different rough sleeping cohorts? 10. What future demand can I expect?
Example service provider user questions
1. Is there more or less demand for our services in certain boroughs or across London? 2. Is there a business case for a local authority to commission additional services externally? 3. What services are particularly effective at targeting a certain cohort at a particular stage in their homelessness journey? 4. Are our services shaped and guided by the experiences of those rough sleeping? 5. Are our services effective at reducing rough sleeping?

Are the same organisations involved? Explain any change

The organisations involved in phase 1 of the project continue to be part of the next phase. The significant change in phase 2 is the onboarding of the remaining 29 London boroughs not involved in phase 1, and a further 7 service providers. This expansion will be the main focus for phase 2.

By making this a truly pan-London initiative, we will have greater join-up of data and richer insights into the journeys of rough sleepers through the rough sleeping ecosystem across London which will enable support to address homelessness to be further strengthened.

Has your funding source changed?

Funding was secured ahead of the pilot phase kick-off for development costs and the first year of running costs. It has been co-funded by London Housing Directors, the Greater London Authority (GLA) and the Department for Levelling Up, Housing and Communities (DLUHC). A commitment was also secured from London Housing Directors to continue to fund running costs of the tool while it is operational. This evidences the commitment to reducing homelessness and rough sleeping across London.

There have been no changes or additional funding sought at this stage. Additional funding will need to be sought in later phases for future developments / transformation work as this is scoped.

Have internal or external stakeholders asked for additional or different requirements?

No additional or different requirements have been requested by stakeholders for implementation at this stage. We have developed a live product roadmap for future feature enhancements and developments and to incorporate continuous feedback from users.

Phase 2 will largely focus on expanding the rollout to a larger number of organisations, alongside some small developments that were part of the phase 1 scope but had to be deprioritised due to the time constraints of the MVP build.

Any larger development or transformation pieces will be fully scoped out, with appropriate information governance (IG) input and review as part of future phases of work.

If you expect the project to move to a new phase or become business-as-usual, what needs to happen to make that possible?

The project is due to progress to a new phase which focuses on expansion of the tool to other organisations as listed above.

In preparation for this rollout, Phase 2 organisations were requested to fill out a technical survey to detail information on:

- their current data collection processes
- the system they use for rough sleeping data
- coverage of current privacy information supplied to clients
- and the current lawful basis conditions used for processing data on their clients.

Following submission of these survey responses, the project team are conducting follow-up conversations to brief the stakeholders on the project requirements. Information on the data that will be requested is being supplied prior to making a formal data request (this cannot take place until post DSA signature) so stakeholders understand the scope of the request and offer them the opportunity to raise any questions. Process for data upload is also being discussed and preferred means of future uploads (manual vs automated). These conversations are being frontloaded to allow any concerns to come to the fore and be dealt with at the earliest opportunity in order to de-risk the onboarding of such a large number of organisations.

No immediate concerns have been raised to date. A number of the service providers who have already been engaged commented on how thorough the phase 1 DPIA/DSA are and how much they appreciated being consulted at an

early stage. The IG guidance provided has also enabled them to update their privacy information to individuals to cover the scope of this project.

Necessity, Proportionality, and Benefits

Can you still justify that the processing is necessary for the stated purpose(s)?

As stated in the original DPIA, the reasons people enter rough sleeping are varied, so the treatments to address rough sleeping are equally as varied. This research project aims to better understand those reasons by mapping and connecting the continuum of rough sleeping. Previously there has been no one system that exists to achieve this rough sleeping data solution, so developing a new solution was necessary in order to leverage existing quality data sources in the short-term and identify data gaps to be addressed in the longer term.

Given the complexity of experiences with rough sleeping, it was necessary to build up a comprehensive dataset which could be explored to try and build up an accurate picture of these journeys which would be valuable for senior decision-makers, policy makers and commissioners within participating organisations to help them improve rough sleeping services. Throughout the course of the project, the delivery team worked closely with pilot partners to understand some of these complexities and develop the insights accordingly. This enabled the minimum dataset requirements to be considered throughout phase 1. An updated version of the minimum dataset is available in the Appendix.

Service providers are asked to provide data on **all** clients i.e. not just those with a CHAIN ID. That is because we are matching with both CHAIN and H-CLIC. Any unmatched data is not used within the visualisations as the purpose of this tool is not to duplicate existing processes and just show organisations their own data, but to show where their data is matched with other data sources. Some service provider organisations asked that their unmatched data is deleted from the database (back-end) and these requests were complied with appropriately. Subsequently it was decided that all unmatched data from all service provider organisations should be deleted and this will be the approach taken forward, both to ensure consistency of the insights, and to ensure data minimisation. This will be an automated process.

Following rollout of the MVP to pilot organisations, a 5-week usage testing period was conducted to determine whether the solution was achieving the expected benefits. **All** pilot organisations confirmed that their expectations are being met and that the tool does what they thought it would do. One stakeholder noted *"my main expectations were being able to say where our data crosses over with the*

other data sources and understand those interactions and this project has really successfully done that. It's particularly useful for understanding what is happening with our accommodation services and to know how long people are staying in accommodation."

Is the processing proportionate? Is there a way of achieving the same or similar benefits whilst processing data in a way that is less intrusive to an individual's privacy?

Data has historically been siloed across the rough sleeping ecosystem. Due to the complexity of journeys those individuals experiencing rough sleeping go through, the only way to truly understand those complex journeys and determine which interventions are working or not working, is through the creation of a system that joins up that data.

Processing of data has been limited to a minimum dataset to enable delivery of the desired outcomes. All data outputted in the tool is aggregated and anonymised and aimed at improving support at a cohort / population level which minimises the intrusion at an individual level.

During the course of the usage testing, it was discovered by a user that in some instances where multiple filters were applied to the data it was possible to get the number of individuals showing in a particular insight down to 1 or 2. Given that this tool is not supposed to present identifying information, the project team considered in depth how to mitigate possible re-identification risks. The project team sought professional expertise on the matter and considered the key question "**can a user work out something new about an individual?**". In summary, it was determined that if a user filtered to 2 individuals, and they already knew who one of them was, there would be a 100% chance that they would learn something new about that person e.g. their sexual orientation, where they moved to after rough sleeping in X location. This puts that person's private data at risk and therefore the recommendation is that we limit the view in line with ONS best practice. This means that any visualisation where there is an output of 5 or lower, we show "equal to or less than 5".

Are you achieving or on track to achieve the stated benefits? Do these still balance positively against the privacy intrusion?

A set of prioritised user requirements were defined as part of the discovery work led by Bloomberg Associates. These are set out below alongside how the developed solution is meeting those requirements.

Prioritised user requirements (Bloomberg Associates)	Component of the user interface which meets this need

1. Understand the common barriers preventing those sleeping rough to improve access to the support they need and move off the street	This need is met via delivery of insights across the SIT. There is an enhancement to be made to the MVP of the tool by way of its analytical capability and its ability to draw insight from the visualisations. This is an item within the workplan.
2. Understand what happens to rough sleepers after they move on to short-term accommodation, to understand any needs for improvement to services	Directly met through the "Accommodation services" page where there are a number of visualisations which show the destinations of rough sleepers after moving in and out of certain accommodation types to understand performance of different interventions.
3. Understand what happens to rough sleepers after they move on to long-term accommodation or other solutions (e.g. reconnection), to understand any needs for improvement to services	Directly met through the "Accommodation services" page where there are a number of visualisations which show the destinations of rough sleepers after moving in and out of certain accommodation types to understand performance of different interventions.
4. Track statutory offers and outcomes of statutory interventions to understand effectiveness of supports and providers	Directly met through the "Movement and Housing Options" page which captures eligibility for statutory duties, including individuals who went on to sleep rough.
5. Identify trends and emerging issues and promptly act, commissioning a solution; and segment the rough sleeping population into different cohorts according to their housing status/needs to better tailor services	Trends are surfaced throughout all pages in the SIT. Users are able to filter the data to understand how these trends change across different segments of the rough sleeping population. The dataset has been segmented according to their housing status.
6. Understand the pathways into homelessness and take action to reduce the factors that cause and contribute to rough sleeping through preventative or diversion services	Partly surfaced when we map the journey of those that have made a housing application and whether they have slept rough afterwards.

Whilst a reasonable amount of personal data is being processed, the outputs in the tool are anonymised and are aimed at a cohort / population-level. This reduces the impact of the privacy intrusion on individuals. The intrusion can be balanced against the positive benefits being delivered, through the provision of data to local authorities and service providers to minimise rough sleeping and its impacts on individuals.

Lawful Basis and Fairness

Has any of the applicable legislation or statutory guidance changed? How does that impact the system/process?

The applicable legislation and statutory guidance remain the same. This is detailed in the DPIA.

Do the original lawful basis conditions still apply? Has your justification for processing changed and how?

The parties use different lawful basis conditions to process the personal data but these have remained the same.

For the local authorities, their original processing of the data matches the lawful basis for all parties as joint controllers for this project, which is public task for personal data, and substantial public interest (Schedule 1, Part 1 DPA 2018 - safeguarding of children and of individuals at risk), and research (Schedule 1, Part 2 DPA 2018 - research) for special category data.

For the service providers they use:

- Article 6(1) (e) public task or Article 6(1) (f) legitimate interests
- Article 9(2) (g) substantial public interest or Article 9(2) (j) archiving and research

UK GDPR Article 5 states that, "...further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation');". Additionally, the ICO Guide to the General Data Protection Regulation (GDPR) states that, "If your new processing is for research purposes, you do not need to carry out a compatibility assessment, and in most circumstances you can be confident that your lawful basis is likely to be either public task or legitimate interests."

This project is a research project which is deemed 'not incompatible' with the original purposes for processing, for any of the bases used by all parties. Nothing in the pilot would reasonably change any of the above. The justification for processing therefore remains the same for phase 2 and the project team have been working closely with organisations coming on board in phase 2 to ensure the lawful basis conditions they use for processing data on their clients is in line with the above expectations.

Is the processing still considered fair to data subjects? Have you had any complaints?

As this is a research project, data subjects have not specifically been notified of this project, but all parties have privacy notices that cover this type of data use. Throughout Phase 1 of the project, and for the duration that the Strategic Insights Tool has been live and operational among pilot organisations, no complaints have been received from data subjects. The project team developed a user testing and feedback process to check in with users regularly about their use of the tool, focused around a set of 'system testing' and IG-related questions. All organisations have been asked if they have received any complaints and to date there have not been any. This will continue to form part of the continuous feedback approach.

Data

Did you use all the data you planned to use?

Not all of the data which was shared was used such as:

- Alcohol use
- Substance misuse
- Relationship status,
- Pregnancy status,
- Prison history,
- Current mental health concerns,
- Medical needs
- Domestic violence
- Care Leaver history
- Entitlement to welfare benefits

Phase 1 of the project was focused on developing an MVP solution so the project team had to be strategic about meeting the needs of the various user groups and ensuring the core requirements were met. Due to the tight timeframe of the 13-week build period, and delays experienced in receiving data from pilot users, we had to de-prioritise some of the original requirements.

Whilst the above data has not been used to support delivery of the MVP, it is required for planned feature developments which were not able to be incorporated into phase 1 due to the above mentioned constraints. Planned work included in the work plan involves the development of additional filters e.g. a "Pregnancy" or "Substance misuse" filter, meaning users are able to filter to these groups of people and understand how they interact with services. Secondly, the development of visualisations specific to the data e.g. "What is the movement of those with substance misuse?". These are important enhancements to build in as we know these categories have a significant impact on homelessness and the

types of interventions that would be appropriate and effective. It was determined that other features / filters were of higher priority in the MVP phase, however SIT users would benefit from understanding this information and this does form a part of the product work plan. Furthermore, if the data were to be deleted from the SIT environment, removed from future data requests, and deemed out of scope going forward, it would require significant resource to collect this information. The delivery team would have to work with participating organisations to amend the reports they provide (both now for Phase 1 users, and then for both Phase 1 and 2 in the future). Regardless of whether this is automated or manual, it would still utilise a significant amount of both the providers' time and delivery team resource. At the very least, it would require a 60 minute 1:1 conversation with c.47 organisations in the first instance, likely needing more time with the majority to iterate. As a lot of this data is currently captured by participating organisations in free text format, we also need the data to be able to explore its usability as we remain in development phase. The tool doesn't currently support free text but we could allocate technical resource to building this functionality if we determine the data to be usable, or work with organisations to convert it into a more structured format.

The unused data from phase 1 users is retained in the SIT environment but is not released or visible to users in any way. Retaining this data and continuing to collect it from phase 2 organisations will enable us to conduct the above planned development work and therefore we recommend it continues to form part of the data request.

In order to meet the needs as defined by both Bloomberg Associates (during the discovery phase) and the Phase 1 user community, it was necessary to expand the minimum dataset to build in more granularity of the data requested. The majority of expansions were regarding H-CLIC and CHAIN data.

- CHAIN: We built on the minimum dataset by requesting more granular data for each of the events. We requested time and location information for events data where it was asking if someone was seen bedding down (i.e. when and where, enabling us to build a much richer picture of the journey of that individual). Building these journeys are dependent on knowing the start and end dates of different event types - this information was not captured to the required level of granularity in the minimum dataset.
- HCLIC: There were a few columns which weren't well-defined. We therefore requested eligibility, whether someone had priority need, which type of duty the applicant received, and which type of service they got during the duty.

These amendments have been reflected in the updated minimum dataset attached in the Appendix. This will be used going forwards for the phase 1 organisations and for the new phase 2 organisations.

Can you reduce the amount or sensitivity of the data?

- Sensitivity: We can't reduce the sensitivity of the data we request and receive because we need that for the probabilistic matching process which is key to accurately merging the datasets. However PII is not included in the merged dataset - this is anonymised and the outputs that appear in the user interface are also anonymised.
- Amount: We cannot remove fields without reducing the usability of the tool, or hindering planned development, specifically to add in more filterable categories e.g. of the people that have abuse problems and how have they moved through the system.

Can you anonymise or pseudonymise the data?

We cannot anonymise or pseudonymise the data that is requested and received from participating organisations as it is required to match individual records across datasets. Users are only able to view anonymised aggregated outputs and steps have been taken to ensure that re-identification is not possible, as outlined above.

Do you need or want extra data? Why is this necessary and what would it allow you to do?

It is recommended that participating organisations improve their data quality by recording more granular information on the start and end dates of events e.g. accommodation stays as this data is key in mapping users journeys throughout the system. We have had to ask for more granular details, as described above.

We have not used all of the data within scope of our requests to produce the visualisations with the tool in phase 1. This data gets filtered out when we do the cleaning (between the raw -> transform stage i.e. the first part of the matching / merging process). The quality of this data is uncertain as we did not need to leverage it as part of MVP delivery. However it will be evaluated when the work to incorporate this data is commenced.

The project team are in the process of determining what phase 3 of the project will look like, including bringing in additional data sources to further build the view of rough sleeping journeys, as well as looking at specific data science opportunities and use cases. It is therefore likely that additional data will be sought in the future. This will allow us to produce full end-to-end journeys of those who are experiencing rough sleeping to better understand common barriers and routes that cause people to jump between the streets and different accommodations services and ultimately help identify where exactly the system is

failing people. Additional capabilities could also include resource management i.e. the ability to forecast demand for services to help understand and plan the best use of resources in line with expected demand, and providing intelligence for casework support through understanding and planning optimal intervention and placement strategies for certain groups of rough sleepers which ensures the prevention of rough sleeping.

These use cases require further scoping to determine the additional data that will be required to enable them. However, it's clear that there is a huge amount of potential for future developments of this tool which have the ability to transform how rough sleeping services are delivered across London.

Data Subjects

Are the data subjects (individuals) the same?

The data subjects are the same. For phase 2 it will also include clients associated with the additional 39 organisations that are being onboarded to the tool.

Have you adequately explained the processing to them?

As this is a research project, it is not a requirement to detail this specific project in privacy information, though some may wish to. However, it has been recommended to phase 1 and phase 2 organisations to update their privacy information as appropriate and ensure they are abiding by best practice.

In phase 1, an organisation updated their privacy notice and noted the following: *"For us, having access to the detailed resources, sound legal advice and support that was provided by the IG Lead for LOTI really gave us confidence and assurance to make the necessary changes to our policy and privacy notices, be part of the SITRS pilot and also be in a far more robust IG position for all future research partnerships and projects. Invaluable."*

As part of preparation for phase 2, to date three service providers have also updated their privacy notices to align with the project scope. One noted *"We have reviewed our Privacy Notice in relation to the guidance provided by LOTI and the recommended lawful bases. Our main lawful basis for general processing is already legitimate interest, and we have updated some existing references to research in the Notice in order to ensure clarity."*

The CHAIN team stated "*We are confident that the project is fully compliant with the requirements of data protection legislation, and that sharing of data for the*

purposes of the SIT is in line with CHAIN's existing data protection agreement and privacy notice." They have also informed all inputting services into CHAIN about participating in this project to ensure transparency about how the data is being used.

The pilot has demonstrated excellent information governance due diligence and best practice and has enabled a number of organisations, particularly some of the homelessness service providers, to improve their overall IG approach.

Did they understand the processing? How did they react? Did any individuals complain or ask for their data to be deleted or for the processing to stop?

During phase 1 of the project, and for the duration that the Strategic Insights Tool has been live and operational among pilot organisations (08/09/23) there have been no complaints from individuals and no requests for data to be deleted or for the processing to stop. This will continue to be monitored and complaints will be responded to appropriately.

One CHAIN inputting service did ask to see the DPIA in place for the project. This was made available to them along with supporting information on the project and the individual was satisfied with this.

Did you, or do you need to, change the way you tell individuals about the processing of their data?

As stated in the original DPIA, following UK GDPR Article 5, research is 'not incompatible' with purposes for processing so additional privacy information is not required for this project. However, all parties have been recommended to review their privacy information and adapt as necessary to reference research, and in the case of housing charities and similar, to describe sharing data with local authorities for the purposes of reducing homelessness.

Evidence of this has been detailed above.

Data Quality

Was the data of sufficient quality to allow you to meet your objectives?

From a project team perspective, it is hard to fully determine the quality of the data and its impact as we did not know how many matches to expect.

The quality of the data was different from different sources e.g. CHAIN compared to In-Form. CHAIN had low quality names at times; we had descriptions of names

of places where they were found instead of the name, descriptions of the tents they were sleeping in instead of the name. For birthdays within CHAIN, sometimes they would just record the year not the full DOB. Both of the above are factors that might inhibit our ability to match.

We believe that an additional benefit of this project will be that participating organisations take action to improve their data quality, in particular their data inputting / collection processes. There already seems to be an appetite for this as some organisations have asked the project team if we are able to support this. As a result, the technical team are aiming to develop some recommendations that can be shared with organisations. It is yet to be determined whether we would be able to support on a 1:1 basis at this stage due to the scope of the rollout in phase 2, but we want to ensure we provide resources and opportunities for data quality improvement within participating organisations. We plan to incorporate this in a planned workshop with phase 1 users. We are also seeing examples happening in practice, with one borough stating *"I can now see things in our own data that we have started to improve because of this project. The 'length of stay' information will be particularly useful to us so we are contacting our providers now to ensure recording is done properly to get the most out of the data"*.

Were you able to match data correctly to a high degree of accuracy?

Data was matched with a very high degree of accuracy. Client records from each database are matched and merged using probabilistic matching.

A probabilistic matching algorithm, implemented by the data science and engineering team at Faculty, uses unsupervised machine learning techniques to try to identify any matches between records across systems. For example, the algorithm may detect that a person who has previously completed a statutory homelessness application within a certain Local Authority, as seen in a local authority's H-CLIC uploads, has appeared in bedded down contacts by housing outreach officers, within the CHAIN system. In this instance, records relating to this person between H-CLIC, where their statutory homelessness application would be, and CHAIN, where their rough sleeping outreach contacts are recorded, would be associated with one another, and merged into a single rough sleeping "journey".

If conflicts arise between the original records, a prioritisation mechanism selects the information from the most reliable source. However, such conflicts have little impact on the final result, because they tend to affect fields like the name or phone number which are of no use after matching. This might not be as straightforward as just matching on the person's name, as typos, fake names, and other reasons, might mean that the name values differ between systems. As a result, the complex matching algorithm considers a number of different factors, including "fuzzy matching" between names, and other columns such as contact details, to score any potential matches.

The algorithm only associates records that meet an 85% probability of being a match. This gives us a high level of confidence and includes as few false positives (cases where we've accidentally matched two different people) in the final matched dataset as possible. The current level of recall is 91%, giving the model a high degree of accuracy. The 9% that is missing was due to quality issues e.g. we cannot confidently say that two records belong to the same person (they had different names or a lot of missing fields). Recall is a measure of how many relevant elements are detected i.e. true matches. 91% recall means that 91/100 true matches are correctly identified. Whilst this represents a high level of accuracy, it should be acknowledged that the risk here is that we miss 9/100 matches and numbers subsequently appear lower in places where they should be higher. To achieve higher recall, the data to be matched would need to be fuller and of better quality. Whilst this should be acknowledged, it's important to note that the SITRS exists to show trends on a system and population level. It is not a substitute for published data and reports. Data in the SITRS may be slightly different to that found in published, static, reports. This is due to differences in processing and timing of data preparation. The recall percentage will also vary as we ingest / match new data dependent on the quality of that data.

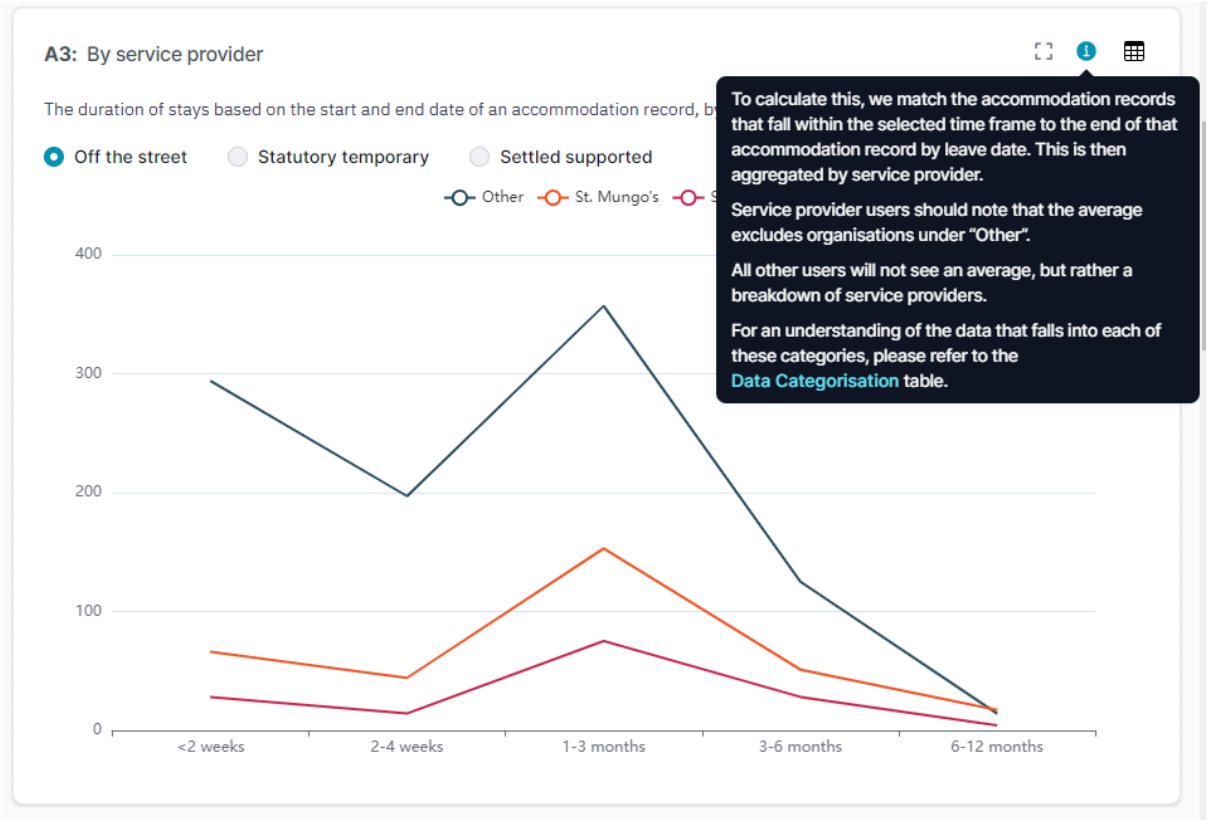
It's worth noting that the evaluation was only done for In-Form records from certain service providers (not all of them had CHAIN IDs for everything) and we didn't have any labels for H-CLIC records so we couldn't validate the matching.

Do you need to make changes to the data or the process to improve data quality?

Those that are collecting this data are doing so for different purposes i.e. not for the SITRS. The data is therefore expected to be of varying quality. To take an example, for CHAIN, they could improve the data collection process by leaving the name field blank and have a description of the person instead of polluting the name field with that description (this was done a lot of times). CHAIN users should also not fill in fields with data that is not relevant. Ensuring this rule is consistently applied, even only for one database with multiple users, would be a simple way of enhancing the data quality. An example of this is allowing a standard way of incorporating multiple phone numbers in a field; each user does it in a different way.

For the purposes of the MVP the data quality is deemed to be sufficient, however there have been comments from users expressing uncertainty about being able to trust the data. In part, this can be attributed to the fact that the tool is relatively

new and users need to get more familiar with it and the way that the visualisations have been developed (i.e. which data / combination of data they rely on). To mitigate this, we built in 'Tool Tips', illustrated in the below image when clicking on the 'i' button. This shows how each of the insights have been calculated and any associated limitations. Users have also been provided with a full user guide, which gives a detailed breakdown of each individual insight available within the tool.



For users to gain full confidence in the insights, we recognise that data quality is a fundamental part of this. A planned output of this project is some data quality improvement recommendations. This will increase accuracy and confidence in the outputs provided in the tool, particularly as we bring in more data and develop the tool further.

We also plan to run a workshop with phase 1 users to help them understand how to get the best value out of the tool and develop some practical use cases, with the hope this will also develop their confidence in the data. Another planned action for the coming weeks is to build a 'trust centre' page into the tool to help further build confidence in the data so organisations feel comfortable utilising the insights for decision-making.

<u>When data quality issues were discovered, were the parties able to make appropriate changes/updates to data within their systems?</u>

Due to the tight timescales and scope associated with the pilot phase of the project, we didn't have time to work with the organisations to make significant changes related to data quality issues. However, when data quality was low, we conducted various cleaning activities to standardise the data before conducting the matching, for example:

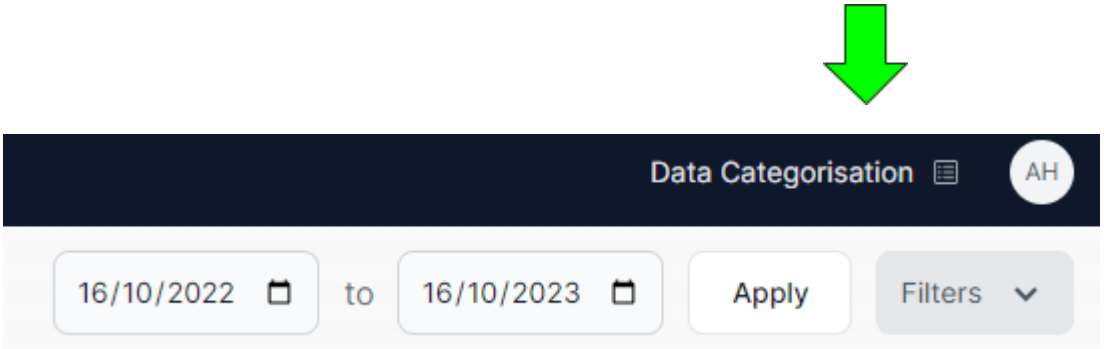
- We parse phone numbers (convert them to a different format to ensure consistency)
- We removed names that seemed suspicious e.g. they were too long or contained digits
- We parsed national insurance numbers (convert them to a different format to ensure consistency)
- We put all names in lowercase to ensure consistency across records
- We also remove duplicates

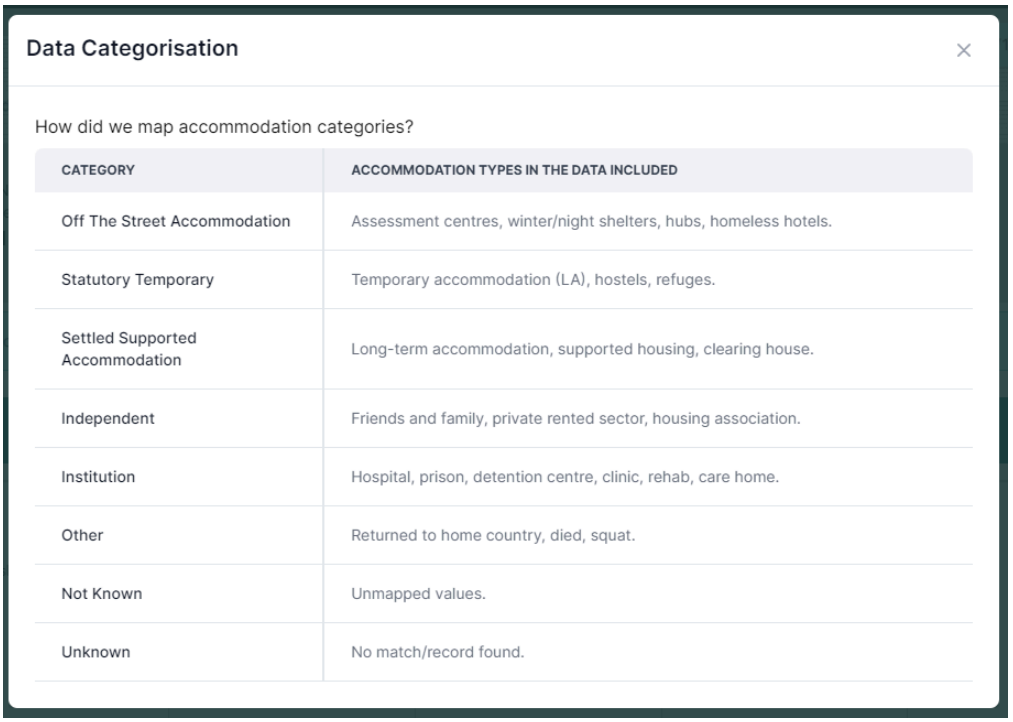
Furthermore, when organisations submit data to be ingested into the tool, these reports may look slightly different depending on the organisation and system they used. A custom mapping model is used in order to standardise all data submissions.

Were inappropriate assumptions made and what happened?

In general, we made appropriate assumptions in the absence of regular stakeholder input. When in doubt, we went back to the stakeholder to validate our assumptions. For example, there were 10s of different accommodation categories which needed to be mapped to higher level categories for the purposes of the SIT, or sometimes there was more than one start date associated with a rough sleeping event recorded in CHAIN. Assumptions on these matters were tested and validated with stakeholders across the pilot partner landscape. We leveraged the expertise of LOTI / GLA / London Councils / Beam to ensure that we were interpreting the data in the right way in the local government context.

For the mapping example given above, the data categories have been included in the tool itself so that users are clear on which specific accommodation types are included within each overarching category. It means users can quickly access and check this information while using the tool and interpreting the insights and mitigates the risk of incorrect assumptions being made about what the data is telling them. This is illustrated in the below image:





Data Categorisation

How did we map accommodation categories?

CATEGORY	ACCOMMODATION TYPES IN THE DATA INCLUDED
Off The Street Accommodation	Assessment centres, winter/night shelters, hubs, homeless hotels.
Statutory Temporary	Temporary accommodation (LA), hostels, refuges.
Settled Supported Accommodation	Long-term accommodation, supported housing, clearing house.
Independent	Friends and family, private rented sector, housing association.
Institution	Hospital, prison, detention centre, clinic, rehab, care home.
Other	Returned to home country, died, squat.
Not Known	Unmapped values.
Unknown	No match/record found.

Data Transfer and Security

Did data collection go according to plan?

Data collection was significantly delayed throughout Phase 1 owing to stakeholder availability. The delivery team did not receive all of the required data for the MVP until Week 13 (final week) of the project.

Because CHAIN and In-Form data schemas were different, and the delivery team only had the minimum dataset to work with, it also wasn't clear how columns in the minimum dataset would map to data collected by Phase 1 organisations. This meant that initial requests for data were vague. We had to request more data / iterate on what was provided multiple times based on significant data exploration in order to build the MVP.

This meant that the data collection process was more complex than expected during Phase 1. However, it has enabled us to put a much more robust process in place ready for Phase 2 rollout. We have developed a 'data checklist' for Phase 2 engagement. We are supplying this data checklist as part of initial engagement conversations prior to making a formal data request (after the DSA has been signed) so that organisations can prepare the necessary report(s) ready for transfer. This will enable us to be much more specific when requesting data from

Phase 2 organisations and make the data collection process more efficient, as well as going some way to mitigate risks around timelines.

Did the transfer mechanisms work and was data transferred securely?

Data was transferred to the delivery team in one of two ways:

- Phase 1 organisations uploaded a .csv file via a secure Faculty-hosted environment, Frontier. This is an internal system that is used for secure data transfer with clients. Members of the technical delivery team would then explore the data.
- Phase 1 organisations would share data via a secure messaging service (either Mimecast or Egress). This data would then be transferred to Frontier, enabling the technical delivery team to explore the data.

Both methods ensured the secure and efficient transfer of data from all pilot organisations.

Did data remain secure in data storage? Did storage locations or mechanisms change?

Data remained in secure data storage throughout Phase 1. Storage locations and mechanisms did not change once loaded into the SIT environment (where the tool is securely hosted). When we ingested the first cut of data from Phase 1 orgs, data was stored on an encrypted EFS (Elastic File System) drive. This raw data is then securely moved to the SIT environment and stored in an Amazon S3 bucket (a secure AWS cloud-based storage resource). It is then loaded into an AWS RDS Postgres instance.

Did the access controls work? Did you need to change who had access to the data or how?

Only members of the technical delivery team had access to the raw data and merged dataset. Credentials are securely managed by use of a password vault. Additional restrictions to only permit access via connection to a secure VPN is also enabled.

Users during the pilot phase can only be added to the tool by a Faculty admin.

We have opted for passwordless authentication. Passwordless logins are more secure than traditional passwords as they use a second factor of authentication that is more difficult for attackers to compromise. In the case of the SIT, users must authenticate via a ‘magic link’. In practice, this means that users follow the web link to access the site where the tool is hosted, they are prompted to enter their email address (associated with being an approved user) and they then

receive a ‘magic link’ to their email address. This link is specific to the associated account and expires after 5 minutes.

We developed a SIT user access model for the front-end / user interface which is as follows:

- Permissions model:
    - **Admin:** Access to all matched data and all views (LOTI, GLA, London Councils, Homeless Link)
        - NB: This role has also been assigned to named users from Bloomberg Associates for a limited time for testing purposes in accordance with their consultancy work on behalf of the GLA.
    - **Local Authority:** Access to Greater London, the subregional view to which the LA belongs, and only the borough-level data that the LA has provided. LAs also have access to all data that is matched to data that they have provided.
    - **Service provider:** Access to Greater London view, and only data within the subregional / LA views that has been matched to data provided by the SP. No access to other service providers’ data - this is all collated into an ‘other’ category where visualisations are broken down by service provider.

We worked with GLA and London Councils to determine the access model. During our testing phase all users confirmed that they have access to the data they should have access to and do not have access to anything they shouldn’t have.

During this early phase of the project we have chosen not to make any of the data publicly available. However a consideration for future phases is which parts of this data we may wish to publish. We also recognise that users may utilise the insights from the SIT in ways which could end up in the public domain (for example if an LA uses some of the information as part of a committee report). As the outputs are all anonymised there are no data protection concerns.

The anonymity of outputs also allows organisations to share outputs and their use of the outputs with other partners or contractors without data protection concerns.

Data Sharing

Did you have a suitable Data Sharing Agreement in place with all relevant parties?

A Data Sharing Agreement was developed and agreed using the IGfL working group approach. The DSA was signed off by all parties. A revised DSA is being put in place for phase 2 of the project and all parties must sign this in order to participate in the project and before any data is shared with the project team to be ingested into the tool.

Has the DSA work as expected?

The DSA successfully facilitated data sharing among all pilot organisations.

Do you need to make any changes and how will you do this?

The DSA is being revised in line with updates to the DPIA. New participating organisations for phase 2 will be added to the DSA.

Retention

Are the stated retention periods appropriate?

The Phase 1 DPIA stated that the period for personal data retention in the project will be 5 years, to allow data matching over time, with the aggregate, anonymised output data retained for a longer period. A key goal of the Strategic Insights Tool for Rough Sleeping is to better understand what happens to rough sleepers across their journeys into, during, and after they sleep rough, which will inform policy and service decisions. For most people who experience rough sleeping, the crisis is singular, short, and requires little intervention. For a few, time spent living on the streets is an intermittent occurrence amid other life challenges over a more extended period of time. These few long-term or episodic cases are the ones likely to need more services.

Being able to see the full user journey for those with higher service needs requires a longer timeline. The advent of rough sleeping is often a confluence of events over many years, from losing a job, developing a health problem, and then finally, perhaps, the loss of a supportive family member or friend. Periods during which individuals have not been engaged with the organisations do not necessarily indicate a positive move away from the streets, and may, for example, result from time spent in prison or rough sleeping in another geographical location. Time on the street can be short, but it can also last months or years, followed by time in hostels or other temporary arrangements. This cycle can stretch over five years, or more. Being able to see journeys over a longer period of time will be immensely useful for more complex, and more service needing individuals.

As part of the phase 1 review we have also considered what would be an appropriate 'trigger' for the 5 year retention period. It has been decided that this will be based on an individual rather than on events. So if an individual is still 'live' in the system, i.e. they have had any contact/interaction/event in the past 5 years, the data will be retained, but if there have been no contacts it will be deleted. There are some additional nuances with this in relation to accommodation types and whether there is an end date associated with them. For example if an individual is marked as in 'settled' accommodation, if there have been no further interactions for 5 years, even if there is no end date, we class this as a positive long-term outcome and consider the data to have reached the disposal point.

The 5-year retention period has been reviewed and is deemed to still be appropriate at this stage. It is also aligned with the way the DLUHC rough sleeping indicators are defined; if an individual hasn't been seen rough sleeping for 5 years, they stop treating the individual as an existing rough sleeper, and this means that if they return to the streets 5 years and 1 day after their last rough sleeping episode, they are treated as a new rough sleeper.

Did secure disposal/destruction of the data happen at the end of the retention period?
We have not reached the end of the retention period to comment on the disposal/destruction of the data.

As this project aims to build up journeys of people experiencing rough sleeping, the insights will become more valuable over time. It is therefore necessary to retain the data for the stated period.

Considerations have been made for the process of destroying the data when it comes to the end of the retention period.

Can you automate retention and destruction?
Yes, retention and destruction can be automated very easily. This action is included in the product workplan and is planned to be completed before Faculty handover the tool to Homeless Link.

Incidents

Did any data protection or data security incidents occur? Why did these happen? What have you done, or what can you do to reduce the likelihood or severity of future incidents?

Some phase 1 organisations provided unrequested data, this is described further down.

Otherwise, no data protection or data security incidents occurred during the pilot phase of the project. Various actions have already been taken to reduce the likelihood and severity of potential incidents. There has also been consideration about what more could be done.

- Likelihood - what have we done
    - Enforced at rest encryption of data stored in our database, and enforced encryption in transit for clients connecting to our database.
    - Implemented secure network firewalls to ensure connections to the database can only take place from trusted IP ranges
    - Database credentials are managed in a secure vault
    - Buckets are encrypted with a managed encryption key
- Likelihood - what can we do
    - Move the database to a private subnet and introduce a jump box or VPN such that no connections to the database go via the internet, only via encrypted connection to one of the two things mentioned here. TBC whether we have the time / resource to do this at this particular stage.
- Severity - what have we done
    - Removed un-required PII from the database where required
- Severity - what can we do
    - Introduce a more granular database roles-based access model

Risks

What were the main concerns?

Main risks identified were around:

- Data retention period
- Ensuring cybersecurity and due diligence processes
- Data quality and matching
- Access controls vs audit controls
- Suitable privacy information provided to clients

Did any of the anticipated risks occur?

None of the anticipated risks materialised into incidents. The risks identified were largely based on the fact that when the initial information governance processes were done, no contractor had been appointed and there was a significant degree of uncertainty as to what the finished product would look like. Having conducted this initial phase of the project, and with an operational MVP solution in place, we can be much more confident and assured in many of these aspects.

For example, we have a robust matching process in place which is producing a high degree of accuracy. We will of course have to continue to monitor this as we are set to onboard a large number of organisations and therefore run a significantly larger amount of data through the matching algorithm.

We have determined appropriate access controls to ensure the personal data is only accessible by a very limited number of users. This gives us confidence in the security of that data.

Our approach to privacy information has been hugely successful, with a number of organisations updating their privacy information to be more transparent to their clients and improve the overall robustness of their IG approach.

Data quality remains an ongoing consideration, particularly with the rollout of the SIT to a much larger number of organisations.

Did an unconsidered issue occur? Did you discover new risks and how did you/do you plan to reduce, tolerate or mitigate them?

One issue that was not considered but did occur, was a few organisations transferring data that was not in the scope of the original data request. However, this was identified and dealt with immediately.

This data was originally ingested into the database as per the pipeline, however all data that was shared that we did not ask for has been deleted from the database. This was done by deleting the fields in the code itself. We then used a tool to automatically reflect this in the database. The next step here is automating the deletion of the raw files when we receive new data from Phase 1 users (and Phase 2 when the time comes). This is important in the event that users once again share data that falls outside the scope of the data checklist. This is being implemented ahead of the rollout to Phase 2 users. We are also considering how we can work with data providers to ensure they undertake appropriate activity to restrict what they send us.

We stressed to Phase 1 users that for future uploads they must only share data that we have asked for. This is made easier now that our request is solidified in the form of a data checklist and this will hugely mitigate the risk going forwards.

No further incidents have occurred since the first data transfer and we have made a point of emphasising to new phase 2 organisations that they must only send data that falls within the scope of the request. They are also being provided with the data checklist ahead of the formal data request to ensure they have adequate notice to prepare a report as per the request.

Overall we are confident that appropriate mitigations have been put in place to ensure that this is avoided in the future.

Another issue that arose when users started testing the live tool was a possible re-identification risk. As we are dealing with a relatively small population, when numerous filters are applied it is possible in a small number of circumstances to filter the data down to an output of 1 or 2. As detailed above, the project team will mitigate this risk by limiting the visualisations to not display where there is an output of 5 or lower, in line with ONS best practice. Instead we will show “equal or less than 5”. We believe this significantly reduces the risk of potential re-identification.

Are you able to reduce the risks levels? For example, you can lower the risk for privacy information if during the processing you became more confident that individuals understood what was happening to their data.

As alluded to above, we are confident in the approach taken to privacy information. Several organisations have updated their privacy information to cover the scope of this project and we feel comfortable that there is an appropriate level of transparency from participating organisations.

DPO Advice

The project complied with the actions stated in the DPO advice for the phase 1 DPIA. This report to review the processing at the end of phase 1 was the main request.

Recommendations

The project recommends that a revised DPIA and DSA is produced with the intent to seek agreement for these from phase 2 parties and progress to phase 2.

Close Down Review

N/A because the project is moving to phase 2.

Appendix

Phase 2 minimum dataset requirements / data checklist

Required fields from all sources (H-CLIC coding in brackets)	Required fields from H-CLIC only	Fields that were part of the original minimum dataset and are to be used for further feature developments
Unique client ID (1.1, A1.2)	Eligibility (1.3, 1.6)	Staying safe plan
Client name (A1.4, A1.5)	Application date (1.22)	Disabilities
DOB (A1.3)	Priority Need (7.5)	Relationship status
Preferred language (not in H-CLIC)	**Duty information (sections 4, 5, 7, 8)**	Immigration status
Ethnicity (1.7)	Duty type	Pregnancy status
Gender (1.24)	Duty start date	Employment / economic status
Nationality (1.9)	Duty end date	Armed forces history
Sexual orientation (1.8)	Location of duty event	Prison history
Phone number (not in H-CLIC)	Duty activity	Current mental health concerns / substance / alcohol misuse
Email (not in H-CLIC)	Reason for duty ending	Current medical needs
NI Number (A1.6)	Duty outcome / destination	Domestic violence services needs
NHS Number (not in H-CLIC)		
CHAIN ID (not in H-CLIC)		
Accommodation start date (9.2, 9.3)		
Accommodation end date (9.4, 9.9)		
Accommodation type (9.5)		
Accommodation location (9.8)		
Reason for leaving / accommodation end reason (7.7, 7.8)		
Departure destination / outcome		