Data Sharing and Linking Service

Analysis of responses to the Technical Consultation on the design of Data Sharing and Linking Service

Data Sharing and Linking Service (DSLS)
November 2013

Contents

Acknowledgements

We would like to thank all those who participated in the consultation

Executive Summary

Background

In order to inform the design and development of a new Data Sharing and Linking Service (DSLS) a consultation on the 'Technical Design of a Data Sharing and Linking Service' was carried out during the spring of 2013.

Aimed at anyone with an interest in data linkage (ie. those currently doing or facilitating data linkage; those currently carrying out or facilitating research using data linkage; or those with a general interest in data linkage issues), the written consultation asked nine questions on four main topic areas on a defined set of proposals for a potential new data sharing and linking service. A total of 22 formal responses were received. Responses were received from individuals and organisations, both data users and data subjects with over half of respondents coming from the existing data research community.

Responses

The majority of respondents were supportive of the service as proposed with most comments referring to the detail of the operation of the service rather than having any in-principle objections. Almost three-quarters of respondents considered the service to adequately protect security, confidentiality and privacy.

Most respondents (around 55%) came from the research community and many were already involved in data linkage within Health. Around a third of respondents were data subjects with the rest coming from a data custodian background.

Principle concerns were raised over the enforced use of the entire package of end-to-end services, with a significant minority indicating that the service should support a more modular approach to allow the use of alternative third-party components such as linkers, data pre-preparation services, and (primarily) additional safe havens. Researchers were particularly concerned about being forced to use the DSLS safe haven environment, both geographically and in terms of not being able to employ the specific platforms and tools which might otherwise be available to them. These concerns were mainly expressed by the researchers already involved in linkage.

Research respondents also raised concerns about how the service would contribute to improving data quality and that the 'separation of functions' model would make it difficult giving feedback to users as well as being inefficient

The other main issue of concern for this group was around the 'create and destroy' approach, under which the created linked and any interim datasets are destroyed at the end of each project. Concern was expressed around the resource (time and effort of the DSLS team, the researchers and the data controllers) required to duplicate extraction, cleansing and creating of data products under this model. The majority of the data research community respondents favoured creating a data warehouse and retaining the linked data.

Those supporting this approach commented on the need for mechanisms to ensure the destruction of datasets, and the fact that data controllers will retain far greater control of their data with specific data sharing agreements for each project.

Most respondents were supportive of the use of the spine to perform record matching and the proposed 'Read Through' process. Any comments were generally relating to specific problems of execution rather than any fundamental issues relating to whether or not a spine should be created.

The majority of other comments were specific to particular issues, processes or concerns and did not represent significant numbers of respondents.

1. Introduction

Following consultation in 2012 on the aims, benefits and challenges to data linkage in the Scottish Government consultation document ‘A Scotland-wide Data Linkage Framework for Statistics and Research’, a number of proposals were developed. One of which was development of a Data Sharing and Linking Service (DSLS) to deliver the following functions:

1. Leading development of data linkage IT and expertise
2. Development and maintenance of methods for ‘Read Through’ between different individual referencing systems
3. A linkage service: conducting approved, within and cross-sector data linkages where necessary and efficient
4. A trusted data-exchange service
5. Development and maintenance of a 'population spine'
6. Co-ordination and support for any 'satellite' data linkage units/safe havens that continued to function in other bodies (for example ScotXed)

This new national service was proposed as a collaboration between Scottish Government, National Records of Scotland (NRS) and NHS National Services Scotland (NSS). The technical consultation on the design of such a service launched in March 2013.

The consultation sets out a technical design for the proposed new service based around expanding existing technical infrastructure. It is collaborative and is designed to deliver a proportionate approach to security of the data to be entrusted to the Service.

The Consultation paper was aimed at anyone with an interest in data linkage; those currently doing or facilitating data linkage; those currently carrying out or facilitating research using data linkage; or those with a general interest in data linkage issues. Anyone was able to comment.

The written Consultation paper was available and open for comment during the period March to June 2013. A number of organisations and individuals currently involved in data linkage or with a particular interest in the topic were contacted and invited to comment. Electronic distribution of the Consultation was assisted by the Administrative Data Liaison Service (ADLS). A number of hard copy print versions were posted out to contacts. In addition, four small-scale discussion events were organised and hosted by members of the DSLS team in Edinburgh, Glasgow, Dundee and Aberdeen to give an opportunity for respondents to discuss issues with the DSLS team to before they completed their individual consultation responses.

Objectives of consultation

Hearing the views and engaging in discussion with people and groups interested in data linkage is very useful in better understanding the issues and areas of concern around this topic. The proposals contained in the Consultation paper describe possible ways of designing the DSLS. The consultation exercise is an opportunity to inform interested groups and individuals of the proposals, to develop discussion around their potential value and appropriateness and to hear general views and opinions of groups and individuals interested in the topic.

The views and suggestions detailed in consultation responses are analysed and used as part of the decision making process, along with a range of other available information and evidence.

While particular responses and comments may usefully inform the design process, consultation exercises cannot address individual concerns and comments, although we hope to continue to engage with all respondents and are keen to continue a dialogue during the ongoing development of the DSLS.

Overview

The Technical Consultation on the Design of the Data Sharing and Linking Service set out a proposed design for how the previously agreed functions could be delivered by the DSLS and contained four main questions. Each question asked for views on different aspects of the proposed design with each main question then subdivided into a further two or three more specific questions around that aspect. The questions are given below;

Q1. The Data Sharing and Linkage Service is intended to add capacity and speed up the data linkage process.
1. To what extent do you think this proposal will achieve this?
2. How could the proposal be improved to deliver this better?

Q2. The proposal is intended to provide better technology and to improve methodology and processes to improve the linkage experience for users.
1. To what extent do you think this proposal will achieve this?
2. How could the proposal be improved to deliver this better?

Q3. The proposed approach to 'Read Through' using the linking population spine is intended to avoid a need to maintain actual linked datasets and to improve the efficiency of linking.
1. To what extent do you think this proposal will achieve this?
2. Are you content that the proposals for the linking population spine are appropriate and secure?
3. How could the proposal be improved to deliver this better?

Q4. The proposed approach intends to deliver both the benefits of efficient linking of data and the privacy of individuals by managing projects in a flexible manner that is proportionate to the risks associated with them.
1. To what extent do you think this proposal will achieve this?
2. How could the proposal be improved to deliver this better?

Respondents (included those who attended the events) were invited to answer all questions, noting responses on the consultation form, and then returning to the project team. Participants were asked to categorise themselves as belonging to one of three defined data 'role' types; Data Custodian, Data User or Data Subject.

A total of 22 formal responses were received; 19 completed consultation forms and three free-form communications.

Production of the Analysis Report

Analysis of the responses to the Consultation Paper was conducted internally during autumn 2013. A final version of the Consultation Analysis Report was published in November 2013.

Next steps

This report will now be considered alongside information from other consultations and research in this field as well as additional detail on related initiatives around the use of administrative data for research and statistics, in particular the Administrative Data Research Centres and Farr Institute. Taken together, this evidence will form the basis of a revision to the model for delivery of the DSLS functions in a collaborative and efficient manner

Responses to the Consultation

The report and all non-confidential responses to the consultation are available from the NHS NSS (ISD) website: http://www.isdscotland.org/Products-and-Services/eDRIS/DSLS-Consultation/

Print copies of the Consultation Analysis Report are available upon request by contacting:
Gerry Donnelly
Data Sharing and Linking Service (DSLS)
Area 2/1/1
Ladywell House
Ladywell Road
Edinburgh
EH12 7TF

T: 0131 314 4 312
E: gerald.donnelly@gro-scotland.gsi.gov.uk

Any questions on the consultation or on the DSLS more generally should also be directed to Gerry Donnelly using the contact details given above.

Structure of Consultation Analysis Report

The Analysis Report is structured as follows:

1. Introduction
Background and context to the project and the objectives of the consultation.

2. Responses to consultation questions
This section presents an analysis of responses to each of the four main questions asked in the consultation paper and their accompanying sub-questions. A summary of responses is included, as well as a listing of any additional comments or suggestions, by question.

3. Responses by respondent group
This section describes the profile of respondents and categorises them by interest group.

4. Free form responses
This section lists the comments made by respondents not using the requested Consultation Form.

Annex 1. Consultation Document
The Technical Consultation Paper on the Design of the Data Sharing and Linking Service.

Annex 2. List of Respondents
The list of respondents.

2. Responses to consultation questions

Question 1

The Data Sharing and Linkage Service is intended to add capacity and speed up the data linkage process.

Summary

10 respondents were broadly in agreement, with a further two indicating partial agreement on the question of whether the DSLS would add capacity and speed up the data linkage process. Two respondents broadly disagreed with the statement whilst four respondents were unsure, or felt there was insufficient information provided to answer. One respondent did not directly address the question in their response.

Of note, three respondents observed that proposal would ensure security, privacy and/or confidentiality. Concern however was expressed by four respondents over a unified/centralised service stifling competition and innovation, reducing capacity and being unable to adapt to increasing demand. The project-specific ("create and destroy") model however was additionally concerning to another respondent in the repetition of process/effort that this would entail.

Several other comments were made relating to how the service could operate to ensure that it did indeed add capacity and speed up linkage. There were also various comments relating to providing a portal of some kind to share information about all aspects of the linkage process.

Question 1.1 asked to what extent does the proposal achieve this?
19 respondents contributed to Question 1.1.

The following comments were made by more than one respondent:

- the proposal would ensure security, privacy and/or confidentiality
- the service will rely on commitment and resource from Data Controllers
- concerns over a unified/centralised service (which would variously stifle competition and innovation, reduce capacity and/or be unable to adapt to increasing demand)
- other concerns were that they could not judge whether the proposal would add capacity and/or speed, as insufficient information was provided regarding current and/or intended capacity.

In addition, the following more specific comments and suggestions were made :

- Improvements will be dependent on ongoing monitoring and sufficient resourcing
- The time spent cleaning and curating data shouldn't be underestimated
- A single organisation would be more efficient
- Strongly supportive of using a standard anonymisation process
- Strongly supportive of not creating a data warehouse
- Access to personal data should require explicit consent
- Concern that anonymity will not be able to be maintained
- The public should be better informed over public usage of personal data
- The service should function as a first contact point for new researchers
- The service will create additional jobs in the sector.

Question 1.2 asked how the proposal could be improved to deliver this better.
15 respondents contributed to Question 1.2.

The following comments were made by more than one respondent:

- that third-party services could be used to supplement the service, including a broader range of "safe havens" spread across the country for ease of access

- suggested implementing a thin information gateway or website to share experience and improve consistency
- suggested implementing standardised Data Sharing Agreements (DSAs) (rather than separate DSAs for each project)
- suggested standardising and simplifying the accreditation landscape
- suggested that some form of persistent data warehouse (either of payload data, or of standardised personal data and keys) would be a more rational model
- highlighted the benefits of improving the quality of source data
- emphasise that pro-active work with data controllers around in-principle sharing agreements and standardising data structures would help speed up linkage.

In addition, the following comments and suggestion were made :

- Standards need to be established to manage the approval process
- Indexing and Linking should be separate roles, rather than separate organisations
- Read through linkage keys should be held only by the data source/controller
- Researchers and data controllers should interface directly on data quality issues (without being reliant on DSLS as a trusted third party)
- Synthetic data could be created to allow less secure data access
- Transparent goals should be set, regarding capacity and speed
- A platform should exist to share strengths and weaknesses of different data
- Researchers should have a single point of contact to the service
- The service needs to be speedy, and accessible at a reasonable cost.

Clarification was requested regarding:

- Whether linkage is to be exact, probabilistic or clerical
- How sub-national reporting is to be achieved, preserving anonymity
- The existing infrastructure and performance metrics associated with it
- The capacity of the proposed process to support users
- Whether the proposed process will include prioritisation mechanisms.

In addition to comments made in direct reference to the questions, further more general comments and observations were made under this section:

- The debate should have a wider audience
- Data can never be truly anonymised
- Complex processes may slow the delivery of research
- Success will be dependent on good inter-organisational communication
- Apparent public concerns relating to persistent storage of data may relate only to identifiable data.

Question 2

The proposal is intended to provide better technology and improve methodology and processes to improve the linkage experience for users.

Summary

Eight respondents were broadly in agreement/supportive, and a further 2 indicated partial agreement of the proposal providing better technology and improving methodology and processes to improve the linkage experience for users. Five respondents broadly disagreed with the question. One respondent did not directly address the question in their response.

Of note, one respondent singled out the creation of a safe haven as a major positive for researchers without existing access to appropriate infrastructure or security. However, a lot of critical comment was made on the both the project-specific model of importing datasets (in comparison with models in which created and/or imported datasets were retained) and the service-centric model of using target data (in comparison with models in which researchers/institutes directly access the target data themselves).

A considerable number of comments and suggestions were made on how to improve methodology and processes in order to improve the linkage experience for others all of which pointed to the DSLS being required to continue to be develop and grow its expertise.

Question 2.1 asked to what extent this proposal will achieve this.
16 respondents contributed to Question 2.1.

The following comments were made by more than one respondent (generally by the research community):

- Various concerns over use of the proposed approach to the safe haven being restrictive to researchers and being difficult to access if centrally located, although one respondent singled out the creation of a safe haven as a major positive for researchers without existing access to appropriate infrastructure or security
- A number of unique comments were made in relation to the inefficiency of the create and destroy model for linkage with in general researchers against this while data subjects tended to favour this approach
- A number of unique comments were made in relation to the service-centric model, specifically from researchers who felt it was important that local institutions still had a role to play and a more flexible approach was required.

Some specific comment relating to use of a central safe haven :

- The safe haven is unlikely to deliver tools and support on a par with academic and research institutes
- Insufficient resourcing of the safe haven will impair the user experience
- A network of accredited safe havens, offering different facilities and methods, should be supported
- Long-term R&D activities on data should be supported within the safe haven
- The DSLS safe haven should not be required for projects in which security or reliability concerns are not high
- The service must be competitively priced for researchers.

Unique comments relating to the 'create and destroy' model (in comparison with models in which created and/or imported datasets are retained):

- Re-linkage of data used in earlier projects will waste significant time and resource both around the technical aspects of linkage and around negotiating access to data
- Many research projects may need data to be revisited for up to 20 years
- In order to be able to recreate historical linkages, frequently-used data controllers may have to store hundreds of project-specific data extracts
- In order to later revisit linked data, both the original and "cleaned" (and linked) data may need stored (problems may have arisen during cleaning/matching).

Unique comments relating to the service-centric model where the Service acts as the trusted third party processing data on behalf of the data controllers and researchers in comparison with models in which researchers/institutes directly access the target data themselves)

- Use of a core team will improve co-ordination, quality and provision
- Dedicated research co-ordinator roles should help users
- The service must be accessible and not too expensive (including requirements to become “approved”)
- The service will facilitate access and linkage to other data beyond ISD and NRS
- The proposal is cross-organisational, and may not deliver a coherent service
- Conflicts of interest may arise if the partners organisations offer a research service, as well as being host to a number of national datasets
- Data controllers/providers may come to exclusively favour this service
- Mechanisms should exist to share domain knowledge between this service and other service providers
- The “plugging in” of a range of third-party infrastructure/services should be accommodated within the model
- If DSLS carries out all curation, researchers will be restricted to DSLS approaches.

Clarification was requested regarding:

- How subjects could be contacted to recruit them for research
- How project-specific (e.g. survey, or sample) result data would be linked to the national dataset.

Question 2.2 asked how the proposal could be improved to deliver this better.
14 respondents contributed to Question 2.2.

The following comments were made by more than one respondent:

- That third-party safe havens (analytical environments) should be supported within the model (including research institutes).

In addition, the following comments were made by single respondents:

- That research support and analysis support should not be included in the model: a second respondent indicated that third-party data analysis functions should be supported within the model
- That the model should support data warehousing
- That the model should also deliver a training and advice service
- That the service would be more coherent as a single, accountable organisation
- That the service will not be scalable without greater automation or use of third-party components.

Clarification was requested on:

- How the spine will be created
- How non-matching records against the spine will be handled
- How the environment will be maintained and updated
- Available contingencies for when analysis requirements surpass the facilities available
- How large data objects (genome, video, etc.) will be handled
- How specific external (e.g. survey, or sample) data can be linked to the model.

In addition to comments made in direct reference to the questions, further more general comments and observations were made under this section:

- The linking of data may compromise the anonymisation of individual data sets
- Policies should be evidence-based, and based on more debate and dialogue
- Individuals may not have given consent to allow their data to be linked for further analysis, and therefore there is a risk that doing this through the DSLS could unfair and/or illegal.

- There is a reputational risk to claiming to anonymise medical data as record level data can never be truly anonymised
- As data controllers and sources are not standardised, unique linking methods will need determined for each project
- Components developed for the service might have a wider application (e.g. for secure intra-government file exchange)
- Greater engagement should take place with the international community
- Develop techniques to maximise linkage quality, including pre-assessment of source datasets
- Information on previous match rates and potential biases should be made available to researchers (and used to improve tools and methods)
- A framework should exist for adding technologies (e.g. third party GUIs, unique APIs) into the analysis environment
- Some mechanism should be developed to automate re-processing/re-transformation of data, in the event of repeated data requests or requests for additional fields post-analysis (to mitigate the repetition otherwise required in a project-specific model)
- The service should continually evaluate new software
- Methods need developed to protect the result of indexing/matching from being disclosive
- If DSLS offered a secure aggregation service this would enable data release with a lower level of required security.

Question 3

The proposed approach to 'Read Through' using the linking population spine is intended to avoid a need to maintain actual linked datasets and to improve the efficiency of linking.

Summary

Six respondents agreed that the proposed approach to Read Through and the population spine would improve the efficiency of linkage. A further four indicated partial agreement, with three respondents indicating they were unsure or felt there was insufficient information available to answer.

Of respondents agreeing that the proposals improved efficiency of linkage, four believed them to be appropriate and secure with the remaining two giving qualified support. Four respondents believed the proposal to be secure, but inappropriately so. One respondent didn't believe the proposal to be secure, with a further three saying they were unsure or had insufficient information to take a view.

Various additional suggestions were made on how the security of the spine and read through could be improved and several comments highlighted that there would be a maintenance overhead associated the spine and read through that shouldn't be underestimated.

Question 3.1 asked to what extent the proposed approach would achieve this.
18 respondents contributed to Question 3.1.

The following comments were made by more than one respondent:

- Highlighting that 'Read Through' could be problematic if the datasets concerned had changing data
- Highlighting that 'Read Through' would be dependent on the (ongoing) support and resource of data controllers
- Expressed doubts relating to maintenance mechanisms for the spine
- Expressed concern about whether post factum QA on matching operations would be possible against the spine.

In addition the following comments were made by single respondents:

- Use of Read Through keys against the spine will bypass anonymisation (ie data is pre-anonymised)
- Use of Read Through keys against the spine will allow inference as to whether an individual exists in a particular data source: the indexer therefore becomes a data holder
- Use of the spine leaves greater control with data controllers (than other methods).

Clarification was requested on:

- How data quality improvements would impact probabilistic matching.

Question 3.2 asked if respondents were content that proposals for the linking population spine are appropriate and secure.
17 respondents contributed to Question 3.2.

Comments were generally related to issues around implementation of security and the linking service itself with no specific themes raised by more than one respondent although several respondents commented on the operation of security and on the linking process,

Unique comments on implementation of security :

- Dummy linking IDs must be project specific
- Education of users on security is most important, with disciplinary procedures.
- Penetration testing should be carried out

- Given that complete anonymisation is not possible, security measures should be based on the perceived restorer
- No entities outwith the linking service should have access to the spine
- All participants should have to explicitly consent to the sharing of their data
- Proposals should be tested against current principles, to ensure that the process is not disproportionately difficult for approved research
- Care must be taken regarding the release of data which could be combined with other available data to become disclosive
- Release of data by data controllers would be dependent on a fuller understanding of both the spine and security.

Unique comments specific to the proposed linking process and the ability to recreate datasets using keys:
- Submitted data from data controllers must not be used to update spine details
- Large numbers of copies of the spine, or a complete history of changes to the spine, must be maintained to enable recreation of historical linkages, and to enable matching against out-of-date data sources
- Full project data sets including the spine must be retained, to order to address audit, verification, peer review
- Insufficient consideration has been given to dataset reproducibility
- The population spine should use source identifiers as part of the identifying data
- Matching rates should be monitored as an indication of the effectiveness of spine maintenance over time.

Clarification was requested on:
- Whether clerical matching will be used
- Whether details regarding the success or otherwise of matching will themselves be disclosive
- How the spine is created
- How the spine is maintained/updated
- What control mechanisms will be in place over access to linkage IDs
- What research might take place, using this proposed service.

Question 3.3 asked how the proposal could be improved to deliver this better.
16 respondents contributed to Question 3.3.

The following comments were made by more than one respondent:
- That some form of data warehousing should be considered
- That linking IDs should be project-specific, for security.

A number of additional comments were made relating to security and appropriateness and the 'Read Through' process and the population spine.

Additional comments relating to security and appropriateness:
- No data release should be made without explicit consent by the subject
- Care must be taken that backups/archives don't compromise the disposal of data
- Security measures should be re-assessed based on the profile of perceived attackers
- Security measures should be re-assessed against current principles.

Additional comments relating to the 'Read Through' process and the population spine:
- Source ID number should be used for matching, where available and appropriate
- DSLS will need to retain copies of the full datasets, for audit as well as reproducibility and verification
- A mechanism is required to reproduce the actual matching process, to resolve/investigate any issues
- The proposal will not work when underlying data source are changing
- The proposal is reliant on data controller input and resource
- Researchers should be allowed to retain copies of their project data

- Data controllers could hold the required linkage keys, rather than the service
- Create SLAs with data controllers to ensure they guarantee high levels of consistency if linkage is repeated
- Improvements will be possible once the process is live
- Repetition of linkage means data queries must be repeated
- Repetition of linkage means complex data operations must be repeated
- No meaningful comments can be made prior to testing.

Clarification was requested on:

- The proposed matching process to be used against the spine
- What matched/non-matched records would be delivered to researchers
- How unmatched records (not on the spine) would be handled
- How the spine would be archived to support repeat linkage
- How the spine is to be created and updated
- Who exactly has access to the spine.

In addition to comments made in direct reference to the questions, further more general comments and observations were made under this section:

- The proposal needs a much wider audience, to receive an open and thorough critique
- The project-specific model means the same data queries will have to be dealt with for each project
- The project-specific model means that derived fields and those with complex conditions will need recalculated/recreated for each project. Any reduction in preparation time will require clean and verified datasets.
- If "Read Through" works, are there material differences from maintaining a data warehouse?

Question 4

The proposed approach intends to deliver both the benefits of efficient linking of data and the privacy of individuals by managing projects in a flexible manner that is proportionate to the risks associated with them.

Summary

Seven respondents agreed that these proposals would deliver benefits of efficient linkage and privacy of individuals by managing projects in a flexible and proportionate manner. Partial or qualified agreement was given by a further four respondents. Six respondents disagreed , with one other saying they were unable to comment until the service was actually established.

Interestingly, when asked to suggested ways of improving these proposals (delivering benefits of efficient linkage and privacy of individuals by managing projects in a flexible and proportionate manner), this section elicited the fewest responses. Some respondents suggested that the substance of this question had been addressed in previous responses, with others giving a number of other more general comments or suggestions on the service.

Question 4.1 asked to what extent the proposed approach would deliver this.
18 respondents contributed to Question 4.1.

The following comments were made by more than one respondent:

- Noted that the proposal was too complex
- Noted that the consideration given to confidentiality was too high.

A number of additional comments were made:

- Very supportive that each project is treated individually, with a specific DSA
- The proposal potentially breaches privacy, data protection, ICO and DP legislation, as the data will reveal the subjects' identities
- Only the indexers should have access to the spine
- Operators and processes need independently monitored and made accountable to the proposed PAC.
- Researchers will be reluctant to invest in cleansing, analysis and transformation of data if they are unable to retain the results (i.e. through accrual of data)
- The "R&D" role implies that the process might substantially alter in future from the one being consulted on.
- Would delivery organisations within the service have privileged access to data?
- Insufficient attention has been made to the mutation of datasets, and consequent problems with reproducibility
- The project-specific approach will result in considerable, avoidable duplication of resource
- Long-lived research programmes would be facilitated by long-lived data assets
- Restricting the service to the DSLS safe haven is a risk to service users
- As the service sees all project datasets, it could create links between them.

Clarification was requested regarding:

- Whether any QA/QC would take place on linkage results (and by whom)
- How sub-national, geographic analysis will be performed while retaining privacy
- Better explanation (and better diagrams) around the anonymisation and linkage processes
- Controls restricting access to the spine (how are individuals chosen/vetted)
- Mechanisms to ensure that datasets are destroyed on time
- The physical realisation of the service
- Management of the relationships among the indexer, service provider and data controllers
- Data controllers should be able to restrict the scope of data released based on conditional matching against external data sources, to restrict the volume of data released
- The legal definition of "data ownership"

- Whether the same restrictions will be placed on the core team as are placed on researchers?

Question 4.2 asked how the proposal could be improved to deliver this better.
14 respondents contributed to Question 4.2.

This question elicited the lowest number of responses. The following comments were received:

- Multiple respondents indicated that the service should allow the use of third-party safe havens
- Several respondents indicated that some form of persistent data (warehousing) would benefit researchers.

Clarification was sought regarding:

- Security measures in place, including compliance with standards
- The potential for members of the public to sit on the steering group
- How secure geographic referencing will be delivered
- The consideration given to the control of inferences that may be made from whether a record was matched (or not)
- Generation and maintenance processes for the spine
- A greater explanation of the role of the gateway and co-ordinators in proactive interaction with data controllers.

In addition to comments made in direct reference to the questions, further more general comments and observations were made under this section:

- The project design work should be completed before data controllers will agree to release any data
- Systems (procedures) need to be in place to handle any future data leaks
- A mechanism should exist for individuals to opt out of data sharing/linking
- Researchers should have to post a summary of each project – this would inform the public, reduce duplication and encourage data controllers
- Assessment of data quality should sit primarily with the researchers
- Researchers and data controllers should work together to determine and enhance data quality
- Local Government pilot applications would be useful
- Standards should be set in advance relating to details and sensitivity required in the payload data
- Data source-specific IDs should not be required (if the transfer method is trusted)
- If agreed by data controllers, a single anonymised ID should be used across the project (this would enable data controllers and researchers to communicate directly about quality issues)
- Ensure that data security is as strong as possible
- The service should carry out data cleansing, and feed back results to data controllers
- Details of projects should be published, including aims, objectives, proposed linkages and the public benefit.
- If anonymisation has been effective, then lower levels of security can be applied to the linked data (than to the spine and index data)
- Greater separation of function is required, through increased independence of function and incorporation of additional architecture(s)
- More specific determination should be carried out of the required data to be delivered to the linking stage (rather than simply as a function of indexing)
- There is no reason why the linker and safe haven roles should not be joined
- Data quality processes should involve data controllers and researchers, operating securely within the safe haven.

3. Responses by respondent group

Overview

The consultation paper asked respondents to self-identify themselves in what capacity they considered themselves to be submitting comments as one of three defined data 'role' types (Data Custodian, Data User or Data Subject).

Of the 22 formal responses received, 19 consultation forms were returned with three free-form communications submitted. Of these, six respondents identified themselves as 'Data Subjects' and three as 'Data Custodians'. One respondent categorised themself as a 'Data Processor' and so when added to the 11 responses identified as 'Data User' this gave the majority of respondents (12/22) self-identifying as primarily part of the 'research community'.

The following abbreviations are used to represent the different communities of respondents, in the discussions below:

DC	Data Custodians
DR	The Data Research Community
DS	Data Subjects

Notes:

1. The three free-form respondents were assigned to one of these communities on the basis of the content of their responses
2. A number of points were raised in response to several consultation questions with some concerns not relate directly to the questions asked. Comments can therefore usefully be grouped together by subject topic, derived from the range and scope of the responses received
3. The analyses below are grouped by these topics and points are identified by the communities making them, in order to highlight not only the areas which attracted most feedback but also to identify whether concerns are particular to certain communities.

Analysis by Community

Service Overview

The majority of respondents were wholly supportive of the service as described in the proposal, or offered either neutral comments or qualified support . Only two respondents (both from the Data Subject community) were specifically opposed to the service:

Community	**Wholly Supportive**	**Neutral/Qualified Support**	**Against the Proposed Service**
Data Custodians	2/3 (67%)	1/3 (33%)	-
Researchers	7/12 (58%)	5/12 (42%)	-
Data Subjects	2/7 (29%)	3/7 (43%)	2/7 (29%)
Total	**50%**	**41%**	**9%**

Perceived cross-community strengths of the service were reported as:

	DC	**DR**	**DS**	**Total**
Adding capacity	2/3 (67%)	4/12 (33%)	2/7 (29%)	36%
Potentially speeding up the process	1/3 (33%)	2/12 (17%)	1/7 (14%)	18%
Improving methods and processes	1/3 (33%)	2/12 (17%)	-	14%

General cross-community concerns were reported as:

	DC	**DR**	**DS**	**Total**
Based upon existing technologies, so unclear what technical benefits would be delivered	1/3 (33%)	4/12 (33%)	2/7 (29%)	32%
Over-complexity (leading to lack of engagement, slow service delivery or inefficiencies)	2/3 (67%)	2/12 (17%)	-	18%
The service must be reasonably priced		1/12 (8%)	1/7 (14%)	9%

4. Free form responses

In addition to the 19 respondents whose input has been described above, a further three respondents submitted text or letter-based responses, contributing 15 unique comments:

- Absolute anonymisation is not possible
- Individuals must positively opt in to programmes using their data for research
- Individuals must give “fully informed consent”, including awareness of the increased possibility of inference through linkage
- Being supportive of the proposed service
- Data sources should not be exclusively Scottish
- The service should aim to attract as many external users as possible
- Clarification is required regarding how third parties may be convinced that results are not fabricated.
- Processes should be seamless with eDRIS
- The process should use the existing Safe Haven network (including a model whereby gateway functions sit locally)
- Project-specific linkage undermines adding new variables on an ongoing basis
- Research data should be made available in line with emerging “open access” policies
- Proposals for data storage and archiving need further developed
- Concern that governance will be disproportionate, unclear and inconsistent
- Comprehensive public engagement is required to demonstrate benefits
- Contributing datasets should be research-enabled (cleaned, harmonised and with standard metadata available).

The content and context of the three responses support identification of these three respondents as one Data Subject and two Data Researchers.

Technical Consultation Paper on the Design of the Data Sharing and Linking Service

This consultation paper is aimed at anyone who may wish to carry out or facilitate research using data linkage

In particular it is expected that the following groups are most likely to be interested in the technical nature of the paper and would wish to respond

- Any researchers or statisticians wishing to use data linkage
- Data controllers and staff involved in sharing or improving the quality of data used for statistics and research
- Anyone involved in or interested in data linkage, data security and privacy within an analytical environment

However, this is a publicly available document and responses are welcome from anyone.

1.

Contents

2.

Background

Data Linkage Framework

“Joined-up Data for Better Decisions: A strategy for improving data access and analysis” describes the framework and national strategy for improving data access and analysis to answer important questions for Scotland through legal, ethical and efficient data linkage. It has stated the Vision of what this will look like;

> “Our vision for the future is one where evidence of what works in delivering positive outcomes for all of Scotland is delivered quickly and efficiently with minimal burden on front-line services.
>
> By improving the ethical and legal governance arrangements, and the technical capacity to securely and efficiently link statistical data, we will enable the research needed to inform policy decisions.
>
> Scotland will be recognised the world over as a hub of innovative and powerful statistical research, attracting investment and job creation”

The Framework defines data linkage as the joining of two or more administrative or survey datasets to greatly increase their value for analysis. The Framework is exclusively interested in linkage for statistics and research purposes to understand groups or populations as opposed to identifying or directly impacting on any individual

The Data Sharing and Linking Service is one of five related elements in the Framework alongside the already established National Steering Group and Guiding Principles and a national Privacy Advisory and Ethics Committee and an Information Gateway to support users through the process. These are in development alongside the Service and there will be further consultation on the emerging details of these.

The Service will deliver additional capacity and technical expertise to enable linkage to take place more quickly, securely and efficiently and to enable new research to be undertaken.

Benefits of Data Linkage

The benefits of data linkage that will be delivered through the realisation of this vision are many and can broadly be summarised as.

- Speeding up cycles of improvement through the delivery of a higher quality cross-sectoral evidence base to inform public policy and strategic planning, spending and delivery decisions.
- Maximising the value of existing data to develop efficient and reliable methods of producing statistics, including better statistics at sub-national level.
- Allowing relatively low cost longitudinal research to be conducted both retrospectively and prospectively.
- □Increasing the capacity to robustly evaluate programmes, by providing the potential to answer far more sophisticated research questions than is currently possible.
- □Improving the quality and consistency of data, through general feed-back loops following linkage activities.

Barriers to Data Linkage
In order to deliver these benefits there are issues to be addressed. In particular the following barriers must be overcome

- In many areas it will be necessary to first improve the quality and consistency of existing administrative data systems to deliver data that is capable of being linked
- While some sectors benefit from existing high quality facilities for data sharing, linkage and analysis, access to these is not uniformly available and the existing capacity is limited
- There is considerable variation in the interpretation of the legal and regulatory environment and data controllers are often unsure whether they can legitimately make data available for linkage purposes.

Summary

This paper proposes the design of a new Data Sharing and Linkage Service for Scotland that will:-

- add capacity and speed up the process of data linkage to deliver more data linkage projects with a particular emphasis on linking data across sectors
- provide better technology and improve methodology and processes to deliver maximum efficiency and expertise in the linkage process
- work in collaboration with other existing data linkage centres to share the benefits of increasing efficiency and expertise
- deliver a highly secure environment that will allow linkages and analysis that have not previously been done for reasons of security to take place
- Contribute to improving the quality of data sources used in analysis

It will do this by working in collaboration with existing and emerging centres to provide new data linking capacity for areas with limited existing resource and will also focus on linking data across sectors. It will aim to harmonise its activity with other centres and seek to simplify the linking environment for users.

It will be delivered through collaboration between NHS National Services Scotland (NSS), and National Records of Scotland (NRS) and operated within the linkage framework under the direction of a Governing Board. It will be quick to evolve to take account of emerging developments in a rapidly expanding data innovation landscape.

It will deliver linkage by separating functions to ensure that at each point in the process only the minimum required data is used and share, to maximise security and greatly reduce risk to privacy. This means that there is organisational and physical separation between stages of the linking process. Personally identifying information (such as names) are used for the minimum amount of time and access to such information is strictly limited. All movement of data between stages is by secure file exchange.

The main components will be

- An **Indexing Service** provided by NRS at their offices in Edinburgh that will handle all personally identifying data. It will match data provided by data controllers to a linking population spine to allow names etc to be replaced using anonymous keys before passing to the Linking Service. The indexers will also maintain methods of efficiently re-creating prior linkages called ‘Read Through’. (See section 5)

- A **Quality Assurance and Linking Service** will be procured. This service will have no access to personally identifying information. It will join datasets using the keys created during indexing and make these available to researchers via a Safe Haven ensuring only the data agreed is shared
- An **Analysis Safe Haven** provided primarily by NHS NSS. Researchers will access the linked data by visiting a secure setting, by visiting other safe settings across the country that will have secure access to the Safe Haven or in certain circumstances via a secure remote connection. The Safe Haven is a secure, high specification, research environment that allows researchers to analyse the anonymised linked data, but allows only non-disclosive results, such as statistical tables, to be removed. The proposal is to build upon the infrastructure, already developed for the Scottish Informatics Programme (SHIP). This provides substantial new computing and analytical power, along with state of the art analytical tools.
- An **Information Gateway and Research Coordinator Service** will be developed to provide single point of entry to the Service, to share information about data linkage and to provide excellent customer service by providing one-to-one support from initiation to completion through the Research Coordinators
- A **Core Team** based in Edinburgh that will manage and coordinate all elements of the linkage service and aim to help align the work of the various other linkage centres and safe havens across the country, manage the secure file transfer service, coordinate and commission research and development into data linkage, lead on the assuring security across the Service and provide a data quality improvement support service

The operation of the Service will be enabled by policies and processes based on the Guiding Principles that will ensure that security and protection of privacy are at the core of all the work of the Service and recognising the high value of such research in the public interest of the population of Scotland. And while the descriptions given in the proposal relate to linking data on individuals, the same principles apply to data on other entities like businesses and addresses.

It is proposed that the Service will start operating by December 2013.

Based on this proposal, respondents are asked to provide answers to the consultation questions in order to inform the design and delivery of the Service

Consultation Questions

Q1 The Data Sharing and Linkage Service is intended to add capacity and speed up the data linkage process

To what extent do you think this proposal will achieve this?

How could the proposal be improved to deliver this better

Q2 The proposal is intended to provide better technology and to improve methodology and processes to improve the linkage experience for users

To what extent do you think this proposal will achieve this?

How could the proposal be improved to deliver this better

Q3 The proposed approach to ‘Read Through’ using the linking population spine is intended to avoid a need to maintain actual linked datasets and to improve the efficiency of linking

To what extent do you think this proposal will achieve this?

Are you content that the proposals for the linking population spine are appropriate and secure?

How could the proposal be improved to deliver this better

Q4 The proposed approach intends to deliver both the benefits of efficient linking of data and the privacy of individuals by managing projects in a flexible manner that is proportionate to the risks associated with them

To what extent do you think this proposal will achieve this?

How could the proposal be improved to deliver this better

Functions of The Data Sharing and Linking Service

The Data Sharing and Linking Service will provide the following functions

- Lead development of data linkage IT and expertise, generating capacity for more and better data linkage for research and statistics across Scotland
- Develop and maintain methods for read-through between different individual referencing systems, and support the development and maintenance of a 'population spine'
- Provide a secure and confidential linkage service: conducting approved within and cross-sector data linkages where necessary and efficient, delivering improvements on the existing range of services available to potential users of linked data.
- Provide a trusted and secure data-exchange service
- Provide support and encourage co-ordination across the network of data linkage facilities and safe havens that already exist, with a focus on collaboration in procurement and use of ICT and sharing of developments, good practice and methods for linkage
- Provide support and guidance on the development of linkable local and national sources in order to enhance the quality of strategically important data resources being shared and linked for statistical research purposes.

These were agreed following public consultation[1] along with the requirement that

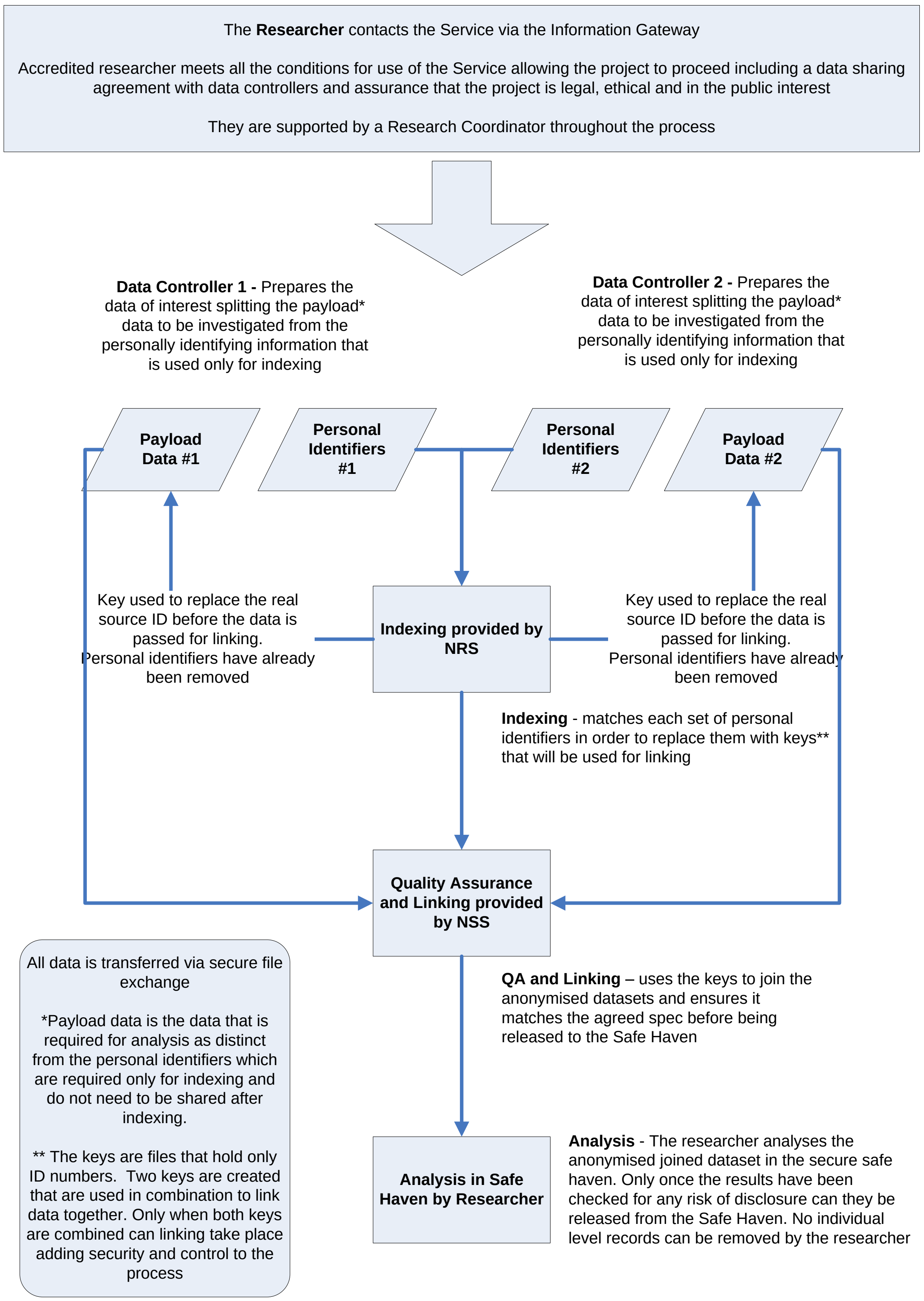
- the already established data linkage infrastructure delivered through the Scottish Informatics Programme (SHIP)[2]and the under development Beyond 2011[3] linkage infrastructure should form a crucial element of the technical capacity and
- the Service should be designed to meet the requirements for a Scottish Administrative Data Research Centre to enable linking and analysis of shared data from the devolved and UK Government departments as recommended by the Administrative Data Taskforce[4].

The proposal is therefore based on supplementing the existing linking centres in SHIP and Beyond 2011 with dedicated new resource and is being designed to be able to deliver flexibly to meet the requirements of the ADT as well as a wide range of other users. It will be developed in an iterative manner aiming to continually improve the experience it provides to its users. The advantages of the approach are

- Allowing the significant investment that has already been made in world class linkage and analysis infrastructure in Scotland to be scaled up to meet the requirements for additional capacity.
- Adding capacity and resource to the existing infrastructure as opposed to delivering completely new infrastructure avoids duplication and is quicker and more cost effective to deliver
- Providing opportunities for synergies and coordination between existing centres rather than adding additional complexity to the field
- Being agile to respond to new and changing requirements from its stakeholders

[1] More information on the Data Sharing Framework can be found at http://www.scotland.gov.uk/Topics/Statistics/datalinkageframework
[2] More info on SHIP can be found at http://www.scot-ship.ac.uk/
[3] More info on Beyond 2011 can be found athttp://www.gro-scotland.gov.uk/beyond-2011/index.html
[4] More info on ADT can be found at http://www.esrc.ac.uk/funding-and-guidance/collaboration/collaborative-initiatives/Administrative-Data-Taskforce.aspx

The Data Linkage Model



The **Researcher** contacts the Service via the Information Gateway

Accredited researcher meets all the conditions for use of the Service allowing the project to proceed including a data sharing agreement with data controllers and assurance that the project is legal, ethical and in the public interest

They are supported by a Research Coordinator throughout the process

Data Controller 1 - Prepares the data of interest splitting the payload* data to be investigated from the personally identifying information that is used only for indexing

Data Controller 2 - Prepares the data of interest splitting the payload* data to be investigated from the personally identifying information that is used only for indexing

Payload Data #1

Personal Identifiers #1

Personal Identifiers #2

Payload Data #2

Key used to replace the real source ID before the data is passed for linking. Personal identifiers have already been removed

Indexing provided by NRS

Key used to replace the real source ID before the data is passed for linking. Personal identifiers have already been removed

Indexing - matches each set of personal identifiers in order to replace them with keys** that will be used for linking

Quality Assurance and Linking provided by NSS

All data is transferred via secure file exchange

*Payload data is the data that is required for analysis as distinct from the personal identifiers which are required only for indexing and do not need to be shared after indexing.

** The keys are files that hold only ID numbers. Two keys are created that are used in combination to link data together. Only when both keys are combined can linking take place adding security and control to the process

QA and Linking – uses the keys to join the anonymised datasets and ensures it matches the agreed spec before being released to the Safe Haven

Analysis in Safe Haven by Researcher

Analysis - The researcher analyses the anonymised joined dataset in the secure safe haven. Only once the results have been checked for any risk of disclosure can they be released from the Safe Haven. No individual level records can be removed by the researcher

Initiation of the Research Project

Researchers and **Data Controllers** are provided with support and guidance to help them reach agreement to take forward projects that are legal, ethical and in the public interest

An Information Gateway will be developed to support data controllers and researchers to initiate and plan their projects and will provide them with dedicated one-to-one support throughout their research project. This will support them in moving efficiently through this phase and also in addressing issues of data sharing; enabling new data sources to be included in projects. The Information Gateway will be delivered through the UK wide arrangements recommended by the ADT with single Information Gateway providing services across all four nations. The details of the UK Gateway are under development, following which there will be further consultation on the arrangements for Scotland.

Prior to any linking activity taking place researchers must meet certain conditions to allow them to access the Service. This is in order to ensure that privacy of individual data subjects is protected throughout the project.

Key conditions for access to the service are:

- that researchers must secure a data sharing agreement with all data controllers involved, describing the project and including a privacy risk assessment and the rules that will be applied to the data being used, in particular describing the ownership of data throughout the project
- that researchers must have the status of approved researcher based on an external assessment by an appropriate accreditation body if they wish to carry out analysis themselves within the Safe Haven
- that researchers must provide evidence and external assurance that the proposed project is ethical, legal and in public interest.

Projects will be varied and the Service will be flexible in how it delivers the needs of data controllers and researchers. During the initiation stage of a project, all parties will agree the design, to ensure that security and efficiency underpin the project.

Stage 1 –Data Preparation by Data Controller

Data controllers receive help to improve the quality and value of their data and retain a key role in ensuring their data is shared correctly

Each data controller prepares the data as required in the data sharing agreement. Details of the ownership of the data throughout the process will be set out in this agreement. They supply a file containing only personal identifying information for indexing to NRS and receive back a 'source key[5]' from the Indexing Service. They use this to replace their own IDs with the new dummy source ID to be used for linking and pass this to the Service for quality assurance and linking along with the payload data[6]. The data owner releases only data that has been agreed.

[5] The Keys are files that hold only ID numbers and are used in combination to link datasets together in place of personally identifying information.

[6] Payload data is the data to be analysed which is distinct from personally identifying information which is used only for linking and is not needed for analysis. If, for example, a researcher was interested in school attainment the payload data would be the actual exam results and not the names or candidate numbers of the pupils involved which are only required during indexing

Throughout the process data controllers will be supported by the Core Team to improve the quality of their data, helping to speed up linkage projects and increasing the value of their data to research outcomes
At all times data is transferred between the bodies involved by Secure File Exchange.

Stage 2 – Indexing

Indexing ensures that personal information like names and addresses is kept separate from the rest of the process and is removed as soon as possible after it has been used for producing the keys which will be used to replace the personal information in the rest of the process

The Indexing Service will be based at NRS in Edinburgh and will be experts in data linkage, and will add capacity, speeding up the linkage process and contributing to the development of new technology and methodology.
.
The Indexing Service will receive a file from the data controller, containing only the personal identifying information such as name, address, date of birth and / or gender that is required for linking, as specified in the Data Sharing Agreement, specific to that project. The Indexing Service then creates a source key that allows the data controller to replace all the personal identifying information and source IDs with a new dummy source ID. There is no need for any personal identifying information or real source IDs to be used from this point forward.

The Indexing Service also produces a second, linking key that contains the dummy source ID and the linkage ID (described below) which they pass to the team in NSS to allow them to join all the data sets involved in the project.

To create the source and linking keys the Indexing Service will maintain a ‘linking population spine’ which holds the name, address, date of birth and gender of everyone in Scotland along with the a linkage ID unique to each individual. Access to this spine will be strictly controlled and limited to those individuals carrying out Indexing. It is not shared with the Data Controllers, or users of the data. The Indexing Service matches the data provided by the data controllers to the spine allowing them to allocate the linkage ID. More information on the linking population spine can be found in section 4

With the agreement of the data controllers, the Service and data controllers can jointly maintain these keys in order to re-use them to make the linkage process more efficient in future projects. This is called ‘Read Through’ (further details are given in Section 5).

The indexing function will be established based on sharing the secure data matching infrastructure already in place in NRS.

Stage 3 – Quality Assurance and Linking

The **Quality Assurance and Linking** stage adds to the quality of the process by providing a final quality check on the linked data before release to the researcher. It facilitates Read Through and adds additional security by ensuring only the agreed information is placed in the Safe Haven

The Quality Assurance and Linking stage will be provided by NSS in Edinburgh.

For each project staff at NSS will receive the anonymised 'payload' data from the data controllers via a secure file exchange. The payload dataset includes only the data of interest to the researcher. They will check the data matches the agreed specification and will then use the keys to join the datasets together.

At no point does the Linking Service have access to personally identifying information, which has already been removed at the Indexing stage of the process

This step allows any issues in the data supplied by the Data Controllers to be rectified, ensuring that researchers receive only the best quality data in the Safe Haven.

As a final security measure the linking ID is replaced with a new project ID that is unique to this project before the data is placed in the secure Safe Haven for the researcher to access.

Stage 4 – Analysis in the Safe Haven

Within the **Safe Haven, Researchers** are provided with a top class research environment that also strictly controls access to the anonymised data further ensuring the security of the data being shared by data controllers

The Safe Haven will provide researchers with a high security research facility. It will be a state of the art environment including analytical software, ample processing power and the anonymised linked data sets all housed in a high security central location provided by NSS. It will be accessed primarily from a secure location at the NSS offices in Edinburgh, where the routine analyses of NHS data already takes place. In addition the Service will develop additional access points sited within other suitable secure settings making it easier for researchers to access the environment in a highly secure manner[7].If Researchers meet additional security criteria they may be able access the Safe Haven remotely through a secure connection.

The Safe Haven which will be based on and aligned with the SHIP Safe Haven already in operation and providing Health researchers with substantial new computing and analytical power and state of the art analytical tools. The Service Safe Haven will deliver an equivalent environment to deliver these benefits to all researchers.

Only anonymised datasets containing the agreed data will be held and only for the duration of the project. No personally identifying information will be placed in the Safe Haven. Each researcher will have access only to the data deposited for them.

[7] Access points will be placed in other secure academic and government safe settings across the country. For example the Health Informatics Centre at Dundee University already has a secure safe setting where an additional terminal connected to the safe haven could be sited. Similar safe settings exist across the country where the Service will locate access points.

Once the anonymised data is available in the Safe Haven researcher will use the analysis tools provided to analyse the data and produce the outputs they need entirely within the secure and closed environment. The researcher has access only to the systems within the Safe Haven which is a stand alone secure facility. If remote access is used, it will deliver an equivalent level of security as physically visiting one of the Safe Haven access sites.

These outputs are then checked for any potential risk of disclosure before being removed from the Safe Haven. No data can be removed from the Safe Haven by the researcher, only the analytical results, which have undergone disclosure control by the staff of Service.

The Service will also offer an in-house analytical service that can be commissioned to carry out analysis on behalf of customers.

The Safe Haven is an efficient approach to providing a centralised high specification, high security environment. It opens up opportunities for analyses requiring cutting edge infrastructure and high levels of security that would be prohibitively expensive for researchers to deliver locally. It will be delivered by NSS who will use their expertise in delivering the SHIP infrastructure to commission a technical solution to meet the analytical and technical requirements of the Service. This may be an internal solution or may be through a technical partner working under contract to NSS as in SHIP. The decision on where the Safe Haven will be located will be based on ensuring the highest security standards are met. Ensuring the security of data within the Service is our paramount concern.

Why this model?

This model has several advantages compared to the alternatives of developing a completely new linkage centre specifically to deliver the Service or using an alternative configuration of the elements available amongst the delivery partners

- Delivering indexing and the safe haven in separate organisations as opposed to a single new centre delivers clear separation of functions which is a key security requirement for the Service
- Both NRS and NHS NSS are trusted public sector organisations with long experience of handling sensitive data and with excellent track records of doing this appropriately and securely and in the public interest
- Both organisations have a national scope and excellent existing resources and infrastructure that can be scaled up to meet the needs of the Service alongside development projects in place to further enhance this infrastructure
- Locating the Indexing service in NRS builds on the work on the Beyond 2011 project to explore alternative approaches to delivering Census information, in particular development of the population spine required for indexing can be delivered jointly with other work in NRS relating to population estimates
- Government departments are key data providers for the Service. It is therefore important that they understand that the security arrangements in place in the Service meet their needs for sharing identifiable information. NRS have existing infrastructure that meets standards used widely across government which makes it easier for other government departments to understand.
- The safe haven being delivered by NHS NSS will be a world class research environment based on and aligned with the SHIP linkage infrastructure. It will provide excellent security meeting the strict data security requirements of the NHS in Scotland

Experience of Users and Data Controllers

Data controllers understand and trust the linkage environment and controls provided by the Service, giving them assurance that their data will be handled securely and allowing them to support research using the information they hold

Researchers can access the secure world class research environment and receive expert support guidance throughout their projects, allowing them to complete new analysis to the benefit of all

The Service will provide data controllers and researchers with;

- a single point of contact to help them develop their idea into a viable project
- assurance, guidance and support to allow them to quickly agree that data can be shared
- a secure environment and processes that they trust to allow them to share data efficiently and safely
- rules and processes to ensure that data ownership is clear throughout the project and flexibility to meet the ownership requirements of each individual project
- assurance that everyone involved in the project is appropriately trained and accredited
- clear separation of functions to ensure that at each point in the process only the minimum required data is used and shared
- an indexing team who are experts at matching data and carry out this role on their behalf
- a strictly controlled research environment that minimises the risk of any breach of privacy during analysis allowing research that may have otherwise been impossible
- world class analytical tools and expert support to help deliver excellent outputs from research
- dedicated support by an individual research coordinator throughout their project helping them to agree a proportionate and efficient plan, negotiate the sharing and linking process, overcome barriers and access the technical expertise available through the Service
- a flexible service that can adapt to meet the unique requirements of each project

The Linking Population Spine

The Linking Population Spine provides a consistent ID number that can be used to replace personally identifying information during the linking process

The Linking Population Spine is a dataset that holds the name, date of birth, gender and postcode for everyone in Scotland at given points in time along with a linkage ID associated with each individual. It will be held and maintained by the Indexing Service in NRS.

It is needed to facilitate the indexing part of the data linking process by providing the consistent ID that is used to link between different sources.

It will be created and maintained by combining data from existing population registers and updated versions will be created on a regular basis and supplied to the Indexing Service, who will hold this in their secure systems. The precise details of its construction will be developed over the coming year.

The Linking Population Spine will be used for indexing for statistics or research by a small number of named and trusted individuals. It will not be accessible to the linking team or by any users or data controllers.

During the Indexing process each source data set is matched to the spine and two 'key' files are created. The first is used by the data owner to replace their real source ID with a dummy source ID before passing it to the Linking Service. The second key file is passed to the Linking Service and contains the linkage ID from the spine and the dummy source ID. The Linking Service then replace the dummy source ID with the linkage ID for each dataset involved and then joins them using the linkage ID.

The Linking Population Spine is therefore important as it allows the linkage ID to be consistently allocated across sources. It is also important that access to it is strictly controlled to mitigate privacy risk.

3. Efficiently Re-creating Linkages – 'Read Through'

Read Through can be used in certain agreed cases to efficiently re-create linked datasets, avoiding the need to permanently maintain these and to speed up the process of linking in general

The Service will not maintain a 'data warehouse'. It is being designed based on the principles that any new linked data sets will be held within the Service only for the lifetime of each specific project, as specified in the project's data sharing agreement.

However, there are benefits to being able to efficiently re-use the linking keys in order to recreate research datasets for further investigation or secondary analysis. This can avoid the need to maintain the actual datasets,

To achieve this, the Service proposes to develop 'Read Through' arrangements for maintaining linkage keys that could be re-used.

In doing this security and privacy are paramount.

Data controllers' permission must be in place and they would be required to permanently maintain one of the keys in their own system. This ensures that recreations of linkages cannot take place without their active agreement and participation.

It will be entirely up to the data controller to agree to take part. It is anticipated that where data is regularly involved in linkage, controllers will find this approach beneficial. For example to facilitate an annual National Statistics publication or where personally identifying data, such as names, is not readily available, making indexing more difficult.

It will also be beneficial where academic researchers are required to make research datasets available to peers on request after publication of results. To allow datasets to be recreated by this proposed method reduces the need for personal identifying information to be shared repeatedly for indexing, as the previously created anonymous keys can be used again.

The process is that the linking key is held by the Indexing Service for as long as the data controller wishes to participate in Read Through. The data controller also maintains their separate source key. Only when these two keys are combined can the data from the source system be linked to other data sets within the Service

This would only happen if the standard conditions for using the Service had been met i.e. data sharing agreements are in place, the project has been agreed to be legal, ethical and in the public interest etc.

Administrative datasets will change over time, therefore arrangements will be made with data controllers and the Indexing Service on how to maintain these keys in the most efficient and secure manner moving forward. Depending on the circumstances this may require regular maintenance of the keys, or indexing only of the changes to datasets when a new linkage is required. In some circumstances it may be that Read Through is not the most appropriate approach for a given dataset.

Core Team

The **Core Team** facilitate the work of the Service and leads collaboration with others involved in data linkage

In order to coordinate the work of the Service and seek out opportunities for collaboration, alignment and harmonisation with other linkage centres, a **Core Team** based with the partner organisations in Edinburgh will be established. This team will administer the secure file transfer service, coordinate and commission research and development into data linkage, lead on the assurance of security across the linkage service and coordinate a data quality improvement service in order designed to facilitate linkages.

Secure File Exchange Service

Secure File Exchange allows data to be moved securely and efficiently

Secure File Exchange arrangements will be put in place to allow data to move securely and quickly between the separate systems. The Secure File Exchange will provide security levels consistent with the data being transferred and will have the capacity to handle the large files it is anticipated will be used in the Service

All data will be moved only via the Secure File Exchange. No data will be transferred between stages of the process by email or by physically moving discs or USB drives. Secure File Exchange is both more efficient and more secure.

The Service will have access to existing file exchange tools in Scottish Government, NHS NSS and NRS to make up its Secure File Exchange Service. The range of tools gives the Service the flexibility to ensure specific requirements around data security or capacity associated with a particular project can be met. The service will be administered from within the Core Team.

Data Quality Improvement Service

The **Data Quality Improvement Service** will provide expertise and tools to users to help them improve the quality of the data they hold

A significant barrier to linking can be a lack of quality of some source datasets. A Data Quality Improvement Service will be established within the Core team to work with data controllers to both help improve data for linkage and provide improved data for their own business purposes

The Data Quality Improvement Service could become involved to help address quality issues that have been identified during linkage, or could be commissioned prior to any linkage by data controllers who have independently identified quality concerns they wish to address with a view to allowing linkage to take place in the future. In the case of strategically important gaps in the linkage landscape the Data Quality Improvement Service may directly approach data controllers with the offer of help in order to make datasets accessible for linkage.

The Data Quality Improvement Service will be established based on good practice in data quality improvement, data management, data validation and collection. The team will provide resource, tools and guidance to Data Controllers that will help them to automate data collection, allowing validation to be built in from the start and to collect data at lower levels to facilitate more and better analysis and linkage, and also that will help them to accurately document, validate and maintain their existing data

Research and Development

A dedicated **Research and Development** function will deliver strategic development across all aspects of data linkage, continually increasing the capacity to do more and better analyses

As well as immediately delivering live linkage projects, in order to maintain and advance our international position at the forefront of data linkage the Service will deliver discovery and creation of new knowledge about data linking technology, methodology, process and analyses of linked data. It will use this to deliver ongoing efficiency improvements and so deliver the maximum capacity for new linkages

In collaboration with stakeholders this team will deliver a research plan describing the programme of strategic research the Service will ensure is taken forward within this workstream. This will deliver both technical advances to improve our ability to carry out linkage and associated analysis as well as delivering strategic cross cutting linkage projects designed to deliver improvements to the evidence base and create follow on opportunities for additional new research.

The team will work in partnership with data controllers, academic colleagues and other linkage centres to identify, develop, deliver and share new technology, methodology and linkages. The research plan will be developed with these partners over the coming year.

Funding

The Service will seek to harmonise its funding and pricing with other centres across the UK and will seek out additional core funding to enable it to improve its offering to users

The service will work in an integrated way with the eData Research & innovation Service (eDRIS) established by NHS NSS as part of SHIP, and will share much of the infrastructure that has already been developed for that. eDRIS will continue to deliver linkage within health while the Service will deliver non-health linkages. It is also intended that the Service will be part of the UK data linkage service proposed by ADT. The Service will therefore develop a pricing model aligned with both of these.

It will seek out core funding from major users and providers of data including Government bodies and academic funding councils in order to minimise the direct cost to users.

Full details will be developed over the coming year taking account of the views of users of the Service with the core objective that the service will deliver projects based on the public benefit they will deliver rather than the ability to pay

4. Governance

The Service will be governed within the Linkage Framework and will develop corporate structures that will encourage cooperation while ensuring the core function of adding new linkage capacity is prioritised

The Service will not be a new public body. It is a partnership between NHS NSS, SG and NRS. The resources of the Service will be managed independently from these sponsoring bodies with formal agreements being put in place to govern these arrangements.

The work of the Service will be directed by a governing board and through them will report to Scottish Ministers.

The Service will also integrate with the UK data linking arrangements proposed by ADT and it is anticipated that members of the governing board will sit on the UK level governing board. Governance procedures will be put in place to ensure that there is equity of access to the Service for all stakeholders. From the researcher perspective these procedures will be proportionate and clear.

Not all data linkage activity will happen through the data sharing and linkage service, reflecting Scotland’s leading position on this. It is up to linkage organisations across Scotland to maintain high standards around security, ethics and privacy for the linkage work they do. However, the Service will promote and share good practice.

From the user perspective the Service, SHIP and B2011 will share the information gateway with projects being directed as appropriate. The Information Gateway will ensure that this is an easy process for users and will provide excellent customer service to move projects efficiently from initiation, through the actual process of data linkage to making data available in the Safe Haven for researchers to access. Details of the Information Gateway will be developed over the coming year in consultation with stakeholders.

Delivering Data Linkage Securely and Efficiently

With security at the centre of all of its activity the Service aims to create an environment where research that could not previously have taken place due to privacy concerns can be completed and deliver public benefit.

In practice it will require a proportionate approach to managing data to ensure that the restrictions on use and access are proportionate to the privacy risk associated with the data.

The service will

- Deliver policies and procedures to ensure that appropriate technical and organisational measures are taken against unauthorised or unlawful processing of personal data and against accidental loss or destruction of, or damage to, personal data. These will put the Data Linking Guiding Principles into practice.
- Deliver technical and physical infrastructure to minimise security being compromised and will meet all relevant standards based on a risk assessment of the data to be handled. An assessment of compliance with standards will be produced and regularly refreshed.
- Ensure roles and responsibilities of all parties are understood and agreed. The security incident policy will describe the responsibilities of users and staff in relation to a security incident along with the sanctions that may be imposed relating to a security incident
- Limit access to accredited researchers who will have completed specified data security and privacy training to ensure they are able to carry out their responsibility to deliver privacy
- Ensure compliance with the security and privacy rules through a data security role within the core team and external audit procedures
- Deliver a data sharing agreement describing how data will be managed throughout the linking project and in particular clearly specifying data ownership at all points of the projects.

5. Data Ownership

The Service will act on behalf of its users and meet their requirements around data ownership, but it will not be a ‘data warehouse’ that maintains linked datasets indefinitely

The control of data within the Service will be governed by project level data sharing agreements based around the principles that the Service will act on behalf of the data owners specified in the agreement and that data will only be held by the Service for as long as is required to complete the project. **The Service is not a data archive and will not hold research datasets beyond the lifetime of the project.** It will though deliver ‘Read Through’ arrangements to allow datasets that have been destroyed to be recreated quickly, if required and agreed by all parties.

It is intended that the Service (through NRS and NSS) will act as a processor on behalf of data controllers but each project is unique and will require a specific agreement setting out how data will be controlled. The key requirement though is that data ownership is clearly agreed and implemented throughout the project and the whole process is legal, ethical and in the public interest.

Each project data sharing agreement will require the signature of all parties in order to clearly show the ownership of the source and linked research datasets and to specify the end point of the project and hence the destruction arrangements for the data. Linked datasets may be jointly owned by different data controllers.

Annex 2. List of respondents

Alex Stobart
Carolyn Watson
Anonymous
David Bell
Anonymous
Scottish Council on Human Bioethics
Department for Work and Pensions
Emily Jefferson
Grampian Data Safe Haven
Anonymous
Falkirk Council
Jeremy Wickins
Anonymous
NO2ID
College of Life Sciences, University of Dundee
WithScotland
Anonymous
Anonymous
Centre for Data Linkage, Curtin University (Australia)
Glasgow City Council
MRC/CSO Social Public Health Sciences Unit
Generation Scotland