Research to evaluate and explore the implementation of an indexing first approach to data linkage

Sarah Cummins and Esther Lewis (Data Architecture, Location and Integration)

February 2026

Contents

Glossary

ADR-UK	Administrative Data Research UK
AI	Address Index
BI	Business Index
CDLS	Common Data Linkage Strategy
CI	Classifications Index
CROW	Clerical review Online Widget
DALI	Data Architecture, Location and Integration
DI	Demographic Index
DIMS	Demographic Index Matching Service
GI	Geography Index
GLADIS	Generalised Linkage for Administrative Demographic Index
HCLIC	Homelessness Case Level Information Collection
HMRC PAYE	His Majesty's Revenue and Customs Pay-as-you-earn
RTI	Real Time Indicators
IDAM	Integrated Data and Methods
IMS	Index Matching Service(s)
LFS	Labour Force Survey
MOJ	Ministry of Justic
MQD	Methodology and Quality Directorate
NMC	Nursing Midwifery Council
ONS	Office for National Statistics
PACAC	Public Administration and Constitutional Affairs Committee
PCSS	Population, Census and Social Statistics group
QUAIL	Quality Analyser for Interpreting Linkage toolkit
RDMF	Reference Data Management Framework
VOA	Valuation Office Agency
XIA	Cross Index Association

1.0 Key Messages

1.1 Purpose

This paper presents proposed research into an "indexing first" approach to data linkage, aligned with the Common Data Linkage Strategy (CDLS).[1] The research will help to ensure linked data products developed by ONS are fit for purpose by providing an evidence base to support data linkers to determine suitable data and projects for indexing. This paper sets out:

- the role of the Reference Data Management Framework (RDMF) in data linkage, and outlining the indexing first approach to data linkage in ONS;
- the purpose and aims of the indexing first research and the associated research questions;
- the research approach, including an overview of existing collaborative work;
- proposed research programme to explore research questions.

1.2 Recommendation

The paper recommends the following to progress the understanding of the implementation of an indexing first approach to data linkage:

- Work closely with colleagues in Methodology and Quality Directorate (MQD), RDMF engineers and subject matter experts to align and agree research plans.
- Continue to research the quality and efficiency differences between bespoke and indexing approaches to linkage, to inform the implementation of the indexing first principle.
- As research findings emerge, share with the linkage and research communities, as well as publicly, to promote knowledge exchange and support the continuous improvement of our approach.
- Discuss iterative recommendations from the research findings with relevant stakeholders, incorporate into the existing methods and procedures to improve the quality of the indexes and matching services, and use to inform triage processes.
- Develop a RAG status system for prospective data to indicate their suitability for indexing.

[1] The strategy is in draft and will be refined following the findings of this research. It provides a common mission for teams working on data linkage in the Data Architecture, Location and Integration division in ONS.

- Develop a Quality Assurance framework for indexed data based on the findings of this research programme.
- Further refine and develop our recommendations as the research progresses.

1.3 Key Asks of MARP

We are looking for input from the group in the following areas:

- Provide feedback on the indexing first research methodology to ensure we can be confident in the validity of our findings;
- The group's feedback on proposed research projects for providing an evidence base for the implementation of the indexing first approach;
- Suggest if there are any other options for exploring the implementation of an indexing first approach ; and
- The group's appetite for reviewing updates on this research.

2.0 Executive summary

The Office for National Statistics has a long history of utilising data linkage to facilitate research. Historically, linkage has been carried out using bespoke linkage methods which are designed based on data quality and customer requirement. However, an ever-increasing demand for the provision of linkable data has generated the need for a new, more efficient approach to data linkage services. A key enabler of these efficiency opportunities is the "indexing first" principle outlined in the CDLS, to deliver on the aims of the published ONS Data Strategy. Key to this is the development of the Reference Data Management Framework (RDMF) and its associated matching services. The RDMF comprises of 5 indexes representing people, businesses, addresses, geographies and classifications. A matching service is being developed for each index (excluding the geography index) using generalised matching algorithms, which enables records from a data source to be linked by automated and non-bespoke methods to the index and assigned an index ID. These matching services are in production and being used as proto-types. This process, described in this paper as 'indexing', enables indexed datasets to be joined on the index ID. The CDLS sets out that indexing and bespoke linkage are two available ways to link datasets together. It is up to us, as expert data linkers, to decide which methods best suit each linkage request. However, it takes an 'indexing first principle' whereby an indexing approach should be considered first prior to using a traditional bespoke approach as it can deliver linked data faster using less resource.

The indexing first principle aims to link data efficiently and more consistently and encourages the reuse of indexed datasets in multiple linkage projects (where appropriate).

This allows bespoke linkage services to focus resource on linking or indexing with poor quality data or on projects with specific quality requirements.

The indexing first principle is a key part of the vision for linkage services at ONS and we, as experienced linkage experts, aim to support and refine it by gathering evidence to ensure that products linked through RDMF IDs are fit for purpose. This includes gathering robust evidence to identify the most suitable data and populations for indexing and instances where bespoke methods should be used. The research aims to clarify when and how indexing can be applied most effectively, offering practical guidance to support the successful adoption of an indexing approach.

This research has become even more timely and relevant since the findings of the Deveraux review and the commentary from Public Administration and Constitutional Affairs Committee (PACAC) concerning the quality of outputs from ONS. Data Linkage supports the production of official statistics as well as research and analysis projects that all have different definitions of acceptable linkage quality. Analysts will sometimes accept lower quality data for scoping and exploratory research, within an acceptable range of quality, if they can access it more quickly In light of this review and the closure of the Integrated Data Service programme, the CDLS is going through a review and this research will inform indexing going forward. Producing high quality linked data continues to be at the core of the CDLS, but it will also increase its focus on communicating the quality of linked data to users so that they can make their own decisions about its usability for their specific project.

As data linkage analysts within the Data Architecture, Location and Integration (DALI) division, we have proposed a research programme to inform the adoption of the indexing first principle. The research aims to explore:

- Evaluate the differences in outputs between linkage methodologies employed in equivalent conditions, by comparing bespoke and indexing solutions to linkage regarding linkage accuracy, characteristic representation, efficiency of the process and usability of the linked dataset.
- Establish an evidence base and decision pathway for types of data, populations or research projects where indexing will not be an appropriate linkage method to meet the linkage or research requirements.
- Establish best practice guidance for the adoption of the indexing first principle, including quality assurance, and communication to users.

By leveraging existing team resources through allocated time for research and development, the research methodology has been strategically tailored to produce actionable and impactful results. We plan to use a range of data from existing bespoke linkage or indexing

projects in DALI to perform quality comparisons between methods. We will also be making use of relevant existing work across ONS, for example the Methodology and Quality Directorate (MQD) and Population Statistics.

Through this research we aim to develop evidenced-based recommendations for how and when indexing should be used to deliver suitable linked data products for research use. The breadth of the evidence and recommendations provided will increase and evolve over time as the range of data available to test and compare the different approaches increases.

Given the importance and prevalence of data linkage in the ONS Data Strategy, we are seeking input from MARP on the fitness for purpose of the research to inform the adoption of the indexing first principle. The results of this research programme and progress will be submitted to future sessions.

3.0 Introduction

3.1 Background

There is an ever-increasing demand for linked data to unlock new and valuable research insights and statistics. However, we recognise that linkage is one of the hardest problems in data – reliably joining datasets is difficult and time consuming. Data linkers spend large amounts of time cleaning and standardising the data and designing methods to maximise linkage quality. Every linkage should undergo quality assurance through clerical review to estimate precision and recall. However, it is not always possible to do for larger volumes of indexed data due to resource constraints.

At ONS, the value of linked data has been recognised for many years and the organisation has been a pioneer in the field – the Integrated Data and Methods (IDAM) hub in MQD lead the development of new linkage methodologies and provide methodological assurance, the Data Linkage and Integration Hub (DLIH) use assured methods to provide data linkage solutions to customers, and our clerical matching teams provide accurate manual linking and support quality assurance processes. Whilst we are constantly building on this track record of expertise, we also see that the demand for linked datasets is only increasing. For example, we have an ever-growing demand for linked data through partners like Administrative Data Research UK (ADR-UK) to provide linked data to researchers. However, our capacity to provide bespoke, point to point linkage between individual datasets cannot scale with it. For this reason, ONS has created a methodology to enable scalable linkage and has invested in building the Reference Data Management Framework (RDMF). This new approach makes use of ONS's unique access to reference data to build data spines of people, businesses, addresses, geographies and classifications.

Beyond being a solution for scaling our linkage capability, indexing has other benefits too: it can produce linked data that is fit for purpose in a timely way, eases pressure on

resource for certain policy making areas (particularly in urgent or crisis policy areas that cannot wait for outputs of a bespoke linkage process) and allows for greater access to administrative data for researchers and analysts across government and in academia. This approach to providing data for statistical processes is in line with other groups, such as Statistics Under Pressure, that aim to meet user needs in a timely way.

3.2 RDMF Indexes and Index Matching Services (IMS)

The RDMF consists of a series of indexes: the Demographic Index (DI), Business Index (BI), Location Index (which combines the Address Index (AI) and Geography Index (GI)) and Classifications Index (CI). Each index is uniquely constructed. The DI and BI are built through an initial process of data linkage for core admin data sources, with records clustered into individuals or businesses and assigned an index ID. The AI and CI are constructed using data products (e.g. AddressBase) and reference data, with an index ID representing a unique address or classification code. The GI is made up of geographical layers through a series of lookup tables. The indexes have varying levels of maturity in terms of their development and assurance.

The term 'indexing' in this paper refers to the process of linking other, non-core datasets to the index, through linkage of a source ID (such as NHS number for people, or PAYE for businesses) or a combination of fields that should uniquely identify the data subject (such as names, date of birth and addresses for people, or company names and addresses for businesses). Once data has been 'indexed', the index ID value can be returned for each matching record. This ID (referred to in the rest of this paper as an RDMF ID) can be used as a common identifier between datasets that have been linked to the same index, thereby being used as a mechanism to link datasets.

Datasets that have a common unique identifier (e.g. NHS number) are currently put through an indexing pipeline to attribute an RDMF ID. The data that comes out of this output is anonymised and ready to be joined to other indexed datasets.

The DI, BI, AI and CI also have an Index Matching Service (IMS), which is a generalised linkage algorithm, that links non-core data to the index where there is no common unique identifier. The IMS links on a combination of fields that should uniquely identify the data subject. For example, the Demographic Index Matching Service (DIMS) employs deterministic match keys followed by probabilistic linkage through Splink, an implementation of the Fellegi-Sunter method, on names, address, date of birth and sex, to return an index ID to each linked record in a dataset where possible.[2] These IMS are at

[2] https://moj-analytical-services.github.io/splink/
Accessed 17/12/2025

different levels of maturity with many still being at a proto-type stage. They all offer scalable linkage services that protect privacy by replacing identifiable information with RDMF IDs. However, they do not all yet provide robust measures of linkage quality.

Data can also be indirectly indexed. This happens when data is linked to another dataset that already has an RDMF ID. The ID can then be associated with both records in the link. This has happened in several cases in bespoke linkage projects and is way of making linked data reusable. It is a possible solution to indexing data that traditionally needs a bespoke solution and that might not perform well in the Index Matching Services.

Like all data, the indexes are not free of error. The quality of the indexes will affect the quality of indexing and data linked via RDMF IDs. Error can compound over these different processes. This research, and ongoing research elsewhere in ONS, aims to better understand the nature of this error so that we can develop standardised measure of quality and communicate them to users of linked data effectively.

3.3 Cross Index Association (XIA)

In addition to RDMF indexes and IMS, the development of Cross Index Association (XIA) products is in progress. XIA products consist of a reference between two indexes. For example, the XIA table between the DI and BI relates to people and the businesses they are associated with (e.g. own or work for) and has been built initially through the linkage of Pay As You Earn data, connecting employees and businesses/organisations. A XIA table has also been developed between BI and CI, to attach a SIC/SOC code to businesses and some work has been undertaken to create a DI/AI XIA table using valid from/to dates on DI addresses for core DI datasets. There is also an embedded XIA between the AI and GI as part of the Location Index product. It's worth noting that whilst the development of XIA 'products' are continuing, indirect XIA can occur when a dataset has been indexed to multiple RDMF indexes. Research may be required to understand the consistency between direct XIA products and indirect XIA.

3.4 Scalable Linkage

This indexing-based linkage method enables a dataset to be indexed once and then joined to multiple datasets using a shared, row-level identifier—transforming a traditionally manual and time-consuming process into something that can be done quickly with a simple join. Indexing is also privacy preserving by replacing sensitive information with RDMF IDs.

The techniques employed within the IMS are designed to accommodate a wide variety of data sources, offering a flexible and generalised approach rather than being tailored to specific data types. This differs to the traditional, bespoke approach where linkers would

review and discuss research aims to tailor the methods used to provide the best possible output for the researcher.

Indexing opens up new possibilities for the provision of linked data by providing a faster linkage solution to the customer. However, understanding how this generalised approach affects linkage accuracy and potential bias is essential to ensure the resulting linked data is fit for researchers and statisticians.

3.5 Indexing first principle

The "indexing first" principle is a central pillar of the CDLS, promoting the use of RDMF indexing solutions for data linkage where appropriate, while recognising the value of bespoke and clerical methods when RDMF indexes are not appropriate for the data or project.

Our research aims to strengthen understanding of when RDMF indexing is most effective, which would support more consistent and confident use of indexing. At present, decisions about a dataset's suitability for RDMF indexing often rely on expert judgment. Building a stronger evidence base will enhance these assessments and help optimise linkage outcomes by choosing an appropriate linkage method for the requirements.

To gain insight into the impacts of the adoption of the indexing first approach, a research programme has begun with the intention of exploring the efficacy of the indexing approach to linkage and generalised linkage algorithms through IMS. The research will focus on linkage using the Demographic, Business and Address Indexes taking particular interest in linkage quality (precision, recall and characteristic representation) when compared to bespoke methods. Furthermore, we will examine the trade-off between quality and efficiency. The programme is being progressed by using existing team resources, with ongoing discussions with MQD and RDMF colleagues to explore opportunities for additional support.

The research aims to:

- Evaluate the differences between linkage methodologies, by comparing bespoke and indexing solutions to linkage regarding linkage accuracy, characteristic representation, efficiency and usability of the linked dataset.
- Establish an evidence base and decision pathway for types of data, populations or research projects where indexing will not be an appropriate linkage method to meet the linkage or research requirements.
- Establish best practice guidance for the implementation of the indexing first principle, including quality assurance, and communication to users.

4.0 Methodology

4.1 Research Approach

To make the most of our limited resources and avoid duplication, we will draw on existing ONS research that can enhance our understanding. Where additional evidence is needed, we will prioritise current linkage and indexing projects. Our research will follow an iterative and continuous approach, enabling us to deliver timely insights, build a robust evidence base, and refine our understanding as new information becomes available. There may be variation in the exact methodology used to link and quality assure each project as evidence because of this approach, however we are still confident that the insights gathered are valid.

Data that has been linked by generalised methods will be put through robust quality assurance processes to build an accurate picture of the quality and coverage of indexed datasets. We will try to ensure that the quality assurance methods are consistent between indexed and bespoke linked products to make the results directly comparable, considering the accuracy, characteristic representation, efficiency and usability of the linked dataset. Slight differences in methods will be inevitable because of the nature of the data under study and any pre-processing that has been undertaken prior to receiving the data.

4.2 Research questions

We have identified several research questions that need answering to inform how an indexing first principle should be adopted.

We have identified three research questions which are not covered by existing research. We plan on answering them as part of the indexing first research.

1. **What are the differences in linkage quality and efficiency between bespoke linkage and linkage via indexing using generalised approaches?**

In answering this research question, we will explore the linkage quality of bespoke linkage approaches versus linkage via indexing using generalised linkage methods (IMS and indexing pipelines), focusing on linkage accuracy (precision and recall) and characteristic representation. We will also be exploring the trade-off between quality and efficiency of the opposing approaches.

The findings from this research question will:

- Provide evidence on the suitability of indexing approaches for different types of data and research projects.

- Inform the development of a triage to support decisions on the linkage method for new requests.

2. What population groups are underrepresented in indexed data?

Underrepresentation of certain population groups in indexed data may be caused by under coverage in the underlying index or bias within the methods used to index data. As well as exploring which groups are underrepresented in indexed data, we will explore the causes of underrepresentation.

The findings from this research question will:

- Improve understanding of the coverage of the RDMF indexes and any underrepresented groups, to help inform further development of the indexes and IMS.
- Provide evidence on the suitability of indexing for certain population groups.
- Inform the development of a triage to support decisions on the linkage method for new data linkage requests.

3. What are the differences in quality and efficiency between alternative approaches to indexing?

Whilst comparisons to bespoke linkage will focus on the indexing approach using generalised methods via IMS, there are several different ways to index a dataset. For example, using an indexing pipeline where a common unique identifier exists (such as NHS Number). For this research question we will explore comparisons between different indexing methods and the impact of these different approaches on quality and efficiency of the linked product.

The findings from this research question will:

- Provide evidence for the best practise implementation of indexing approaches for different types of data and research projects.
- Inform the development of a triage to support decisions on the linkage method for new requests.
- Give real examples of the differences in linkage quality between methods to illustrate the trade-off between quality and efficiency to customers.

We have also identified four further research questions that are currently out of scope of this research. For questions 4, 5 and 6, we have identified existing research work that will

provide some evidence (see section 4.3). For question 7, we are aware MQD are investigating the cumulative effect of error types in the DI on downstream outputs.

4. **What is the linkage quality of the RDMF indexes and Cross Index Association tables?**
5. **What are the coverage gaps in the RDMF indexes and Cross Index Association tables?**
6. **How do updates to indexes impact the quality of data linked via indexing?**
7. **What is the impact of differing levels of linkage error and bias on statistical analysis of linked data?**

Ask to MARP: Does the panel agree with this research approach

Ask to MARP: Does the panel feel that the indexing first research programme is focusing on the right questions (questions 1, 2 and 3)?

Ask to MARP: Does the panel think these research questions will help to understand the quality of linkage through an indexing approach to inform decisions about the usability of indexed data for linkage for prospective users?

4.3 Existing research

The **Integrated Data and Methods (IDAM) hub in MQD** are undertaking a programme of research to provide an initial methodological and quality assessment of RDMF including the development of methods to identify and measure error, which will provide evidence for research questions 4, 5, 6 and 7. An overview of this work was presented to MARP in May 2024 and summary of relevant projects is given below:

- Methodological evaluation and baseline quality assessment of the Demographic Index (DI), Business Index (BI), Address Index (AI) and Classifications Index (CI) and their associated Index Matching Services – currently work on the AI and CI are in their infancy, whilst evaluation of the BI and DI is more advanced.
- Method creation, development and research on error types within the Demographic Index (DI) – this currently consists of developing an approach to estimate the probability of False Positive and False Negative Clusters in the DI, with work planned on uncertain clusters and data measurement error. Understanding the quality of clusters in the DI is an important step towards understanding how error propagates through to the indexing and joining of datasets through to outputs. This work is also applicable to the BI but is not currently resourced.
- Generalised Linkage for Administrative Demographic Index (GLADIS), consisting of research on best practice generalised methods and development of an automated

algorithm for the Demographic index Matching Service (DIMS). Note that whilst GLADIS is in development, the indexing first research is taking place on an interim DIMS solution (also known as 'proto DIMS'). An evaluation of GLADIS is also planned through comparing the indexing of different types of datasets through GLADIS and proto-DIMS. This evaluation will be a valuable addition to our evidence base, providing more examples of the quality of indexing achieved through generalised methods. Possible outcomes of this evaluation include one method outperforming the other, no significant difference between the two, or methods from one being used to enrich the other,

- Development of the Quality Analyser for Interpreting Linkage (QUAIL) toolkit, which consists of research into and development of a code pipeline for the quality assurance of indexed data, to provide the stratification and sampling for clerical review, and subsequent calculations for estimating precision and recall. QUAIL is designed to work with outputs from GLADIS but could be adapted for use on other outputs. Once released, QUAIL and associated learnings will be integrated into indexing first research methodology.
- A Validation and Assurance Framework for RDMF has been developed and delivered. It provides minimum controls needed for assuring the quality of RDMF Indexes and IMS that DGO are currently working on.

Teams in the Population, Census and Social Statistics (PCSS) group are looking at administrative data-based alternatives to Census and survey methods for measuring population and migration statistics. The DI is used as a basis for producing the Statistical Population Dataset (SPD), where rules are applied to DI clusters to ascertain whether they are part of the usually resident population. A large volume of research in this space looks at population coverage of the SPD and may identify key gaps in DI coverage, contributing evidence to research question 5.

RDMF data engineers and architects in DALI have started work to identify the impacts of updates and changing versions of the core indexes on the linkage of index products, contributing to research question 6. This work is in its infancy but will be necessary for effectively adopting an indexing first principle. They are also working on how the indexes can contribute to ONS's Statistical Population Dataset (using the DI) and the Statistical Business Register (using the BI).

We will work closely with colleagues in MQD, PCSS and DALI to review and integrate findings from these research programmes into the outcomes and recommendations of the indexing-first research.

4.4 Research programme

To supplement existing research, data linkage analysts within the DALI division are proposing a research programme to compare the quality of bespoke linkage methods and indexing approaches across a wide range of data sources. The research plans to makes use of existing linkage projects that have used either bespoke linkage methods or indexing.

Existing bespoke linkage projects

There are many current and historical linkage projects which use bespoke methods to link data. We plan to re-link some of these datasets using indexing approaches. The data will be indexed to the relevant RDMF index using the most appropriate method(s) based on data linkage analysts' expertise and current knowledge about indexing methods, and subsequently joined on the RDMF ID. Once this additional linkage has been completed, we will compare the quality of the bespoke-linked and index-joined datasets, using clerical review to estimate and compare error levels. We will also explore representation of population subgroups between the two linked datasets.

We will aim to select research projects which include a range of data sources, with coverage of different minority populations and variation in the availability and known quality of linkage variables. Table 1 provides an example of the linkage projects that have the potential to be included in this research.

Table 1: Examples of possible research projects using existing bespoke linkage projects

Bespoke linkage project	Population	Indexing approach	Research aims
Homelessness Case Level Information Collection (HCLIC) linked to Census 2021	Individuals who are homeless or at risk of homelessness who have approached a local authority for help	Demographic Index Matching Service	• Understand linkage quality and efficiency of linkage via bespoke vs. generalised indexing methods • Explore representation of specific sub-populations who have frequent contact with public services but low quality variables within indexed data
Nursing and Midwifery Council (NMC)	Registered Nurses and Midwives	Demographic Index Matching Service	• Understand linkage quality and efficiency of linkage via bespoke

Register linked to Census 2021			vs. generalised indexing methods • Explore representation of sub-populations within indexed data
Longitudinal population study participants linked to HMRC PAYE RTI	Responding survey participants to longitudinal population studies	Demographic Index Matching Service	• Understand linkage quality and efficiency of linkage via bespoke vs. generalised indexing methods when creating longitudinal inked data.
Annual Birth Registrations linked to VOA property attribute data	Births registered in England and Wales each year	Address Index Matching Service	• Understand linkage quality and efficiency of linkage via bespoke vs. generalised indexing methods

We will include more examples from the Address Matching Service and Business Matching Service as these become available.

Existing indexing projects

There are also many existing linkage projects within DALI that are already using the indexes in the development of data linkage products. Data can be indexed using several approaches:

- Indexing using an IMS;
- Indexing using a source ID variable (such as NHS number) via indexing pipelines;
- Indexing using bespoke methods;
- Where an index ID has been inferred by linking to another indexed data source (indirect indexing);
- Where a Cross Index Association (XIA) table has been used.

We plan to carry out additional quality assessment (where required) of indexed data, using clerical review to estimate error levels and exploring the representation of population groups within indexed data compared to the source data. Where availability of linkage variables allows, we will also carry out additional indexing of data using alternative approaches to compare the quality of different indexing methods. For example, where data

has been indexed using a source ID variable such as NHS number, we will also index the data using an IMS.

Again, we will select a range of indexed datasets to include in this research, ensuring wide representation of different population sub-groups and variation in the availability and known quality of linkage variables. This will enable us to explore variation in the quality of data indexed using different methods and make recommendations about the suitability of indexing approaches for different data.

As DALI are commissioned to index new datasets our research plans may evolve to ensure we are using the datasets that are most relevant to the research. Table 2 provides an example of the indexed datasets that have the potential to be included in this research.

Table 2: Examples of potential research projects using existing indexed datasets

Indexed dataset	Population	Linkage variables	Indexing approaches	Research aims
MoJ Prison and Probation data	Individuals who have engaged with prison or probation services	Personal identifiers	• Demographic Index Matching Service	Explore representation of specific sub-populations within indexed data and data with distinct quality problems and a lack of distinguishing variables
Annual Birth Registrations (1993-2022)	All births registered in England and Wales each year	NHS Number, personal identifiers	• NHS number • Demographic Index Matching Service	Understand linkage quality of different indexing approaches
Diabetes prevention programme data	Patients enrolled on NHS England diabetes prevention programme	NHS Number, personal identifiers	• NHS number • Demographic Index Matching Service	Understand linkage quality of different indexing approaches
Labour Force Survey (LFS) /	Responding survey participants	Personal identifiers	• Demographic Index Matching Service	Assess linkage quality of indexing to inform suitability of indexing for survey data with

Transformed LFS				complex schema and low response rates.

Ask to MARP: Does the panel think that these proposed research projects have sufficient opportunity and coverage to answer the in-scope research questions?

Linkage quality assessment

We plan to assess linkage accuracy through clerical reviewing a sample of links and rejected candidate pairs to produce estimates of precision and recall. Precision is a measure of the accuracy of the matches that have been made and recall is a measure of the proportion of matches that have been made from all the possible matches. In linkage, there is a trade-off between these two types of error.

Stratification of the links and rejected candidate pairs may need to be tailored to the linkage methods or IMS used.

Where probabilistic scores are available, scores will be grouped into 5 to10 equal buckets to create false positive clerical samples (precision), and candidate pairs below the accepted threshold for links will be used to create false negative samples (recall). Where data has been linked deterministically, 5 to 10 match key groups with a similar likelihood of error will be used for false positive clerical samples, and we will run a probabilistic algorithm on residuals to sample and stratify by match weight for false negative clerical samples.

Sample sizes will be determined by Statulator using a confidence level of 95%. Sampled record pairs will be run through the ONS' Clerical Review Online Widget (CROW) tool for review using a team of experienced clerical matchers. Precision and recall for the entire population will be derived using total estimated errors. This is the sum of multiplying the error rate with the number of record pairs for each bucket and then aggregating up to the entire population. Confidence intervals are the calculated to reflect the variance and standard error in this estimation.

Aside from precision and recall, we will also be evaluating characteristic representation in linked data. This will give an indication of whether certain characteristics are less likely to link using the approach in question. We will be evaluating this using ONS' proportional discrepancy tool. Positive proportional discrepancies convey overrepresentation of the characteristic evaluated, while negative proportional discrepancy scores convey underrepresentation. For example, a proportional discrepancy score of negative 0.05 suggests the linked data have only matched 95% of the expected number of matches given the overall match rate. Whilst we can measure under and overrepresentation of

characteristics in the linked data, this does not directly evaluate linkage bias. Underrepresentation in the linked data may indicate RDMF coverage limitations (if evaluating an indexing approach), data source coverage differences or bias in linkage methods, which means that careful interpretation of proportional discrepancy is required. We have chosen not to measure linkage bias directly due to the additional clerical matching burden this would create.

The characteristics examined will depend on the data sources in question. Generally for linkage of people, we would evaluate age, sex, ethnicity or country of birth (if present) and geography as a minimum. For businesses, we would evaluate age, geography, business type/sector, status and size (if variables are available to do so). For addresses, we may consider variables such as geography and address type.

Ask to MARP: What are you views on this quality assurance approach?

Population sub-groups

As well as population-level data, we are interested in data for population sub-groups to help expand our understanding of how well sub-groups are covered by the relevant RDMF index and how well the population sub-group link using a generalised method. This will help to develop a knowledge bank of types of data with associated indexing performance. The types of data we are interested in include population sub-groups that are not well captured in admin data or may be more likely to contain data quality issues. Of particular importance are minority or vulnerable groups due to the sensitivity of the data to biases in a linked product. We are also interested in the effects of linking a comparably small data source to an RDMF index due to the increased likelihood of making false positive errors. Examples of population sub-groups of interest include data for prisoners, refugees, the homeless, small sample surveys, third-sector businesses and farming households/addresses.

Other considerations

There are other factors that may need to be taken into consideration when evaluating the differing linkage approaches. For example, we will be recording the length of time and volume of resource required for each approach. This will help gauge the quality versus efficiency trade-off. Another factor that will need to be considered is the requirements for running the IMS, e.g. in terms of data variables or quality or the ability to deal with many-to-many links. This will affect whether the IMS can be used at all.

Note where DIMS is referenced, we are referring to the existing DIMS solution (also known as ‘proto-DIMS’).

Ask to MARP: Can the panel provide feedback on whether the proposed research is appropriate to the research questions

5.0 Next steps

Indexing is being adopted as a linkage approach in the Office for National Statistics and we expect this research to continue indefinitely to develop and refine the indexing first principle. The research enables us to develop and test assumptions and processes to evidence the impact of the indexing first approach and provide assurance the linked products produced are suitable for research use. As such, we plan to develop a knowledge bank to store information on the indexing or linkage performance of different types of data and data subjects.

However, a limitation identified in the research plan is that the projects outlined are heavily person based, with limited projects on businesses or addresses. We acknowledge that this is a significant gap and aim to engage more with the business index and address index leads to identify further projects in these areas.

As research evolves, we plan to make recommendations on changes to linkage and indexing methods and procedures. This will help ensure that the development of the indexing first principle is evidence-based. We also expect to use research findings to develop a more detailed triage process for deciding the best approach for indexing or linking data.

We intend to share findings and recommendations with various stakeholders. The purpose of sharing the information varies from detailing specific project findings, relaying recommendations to providing information on how the recommendations should be implemented in the linkage process. It will also feed into work to be transparent about known quality of the RDMF by disseminating the findings of this work to potential users of RDMF.

6.0 Conclusion

The "indexing first" research programme is currently in development, building on existing work to explore the impact of this approach. Its continued evolution will be essential for a comprehensive evaluation of indexing approaches.

Advancing our understanding of this approach will play a key role in supporting the ONS to implement the data linkage pathway effectively. Applying research findings and recommendations will help ensure high-quality data outputs, reduce the risk of pipeline re-runs, and support more efficient processes. Crucially, it will empower decision-makers with the evidence needed to select the most appropriate linkage pathway for each data source. This will enable the ONS to consistently deliver linked data through the most suitable route,

supported by relevant quality assurance, while maintaining a balance between timeliness and accuracy to benefit analysts and outputs.