Methodological evaluation and quality assessment of the Reference Data Management Framework (RDMF) – an overview of the approach

Version number:	1.0
Date:	9 May 2024
Authors:	Michael Cole, Tom Hunter, Rosalind Archer, Leah Maizey, Josie Plachta

Key Messages of Paper

Purpose

The Reference Data Management Framework (RDMF) is a framework to manage the production of reference data indexes and services to match user-submitted data against those indexes, covering people, businesses, and location. It is an important piece of statistical infrastructure within the Office for National Statistics (ONS), with large programmes of work such as the IDS (Integrated Data Service) and Future of Population and Migration Statistics (FPMS) making use of it. As such, quality assurance of the RDMF is critical. In line with the Code of Practice for Statistics (the Code), presentation of our quality assurance approach to this panel will be part of the independent external quality assurance measures. This paper sets out:

- a brief description of the RDMF;
- the risk to the quality of downstream statistical production when the quality of the underlying indexes and matching services is not known or not communicated;
- a proposed framework of controls for assuring the quality of the RDMF to mitigate that risk; and
- a brief introduction to the methodological pieces of research that are currently in development as part of this programme of assurance work, which will be brought in detail to future Methodological Assurance Review Panel (MARP) sessions.

Key asks of MARP

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

- Whether the proposed model of quality assurance is fit for the purpose of providing users with the information required to make use of the RDMF to produce statistics of known quality;
- Whether there are any further aspects or measures of quality around the RDMF that the panel would like to see added to our quality approach;
- The panel's appetite for reviewing updates to this proposed model of quality assurance; and
- The panel's initial feedback on research projects in this space that are already underway, in anticipation of receiving full papers in future sessions.

These are summed up in the following questions:

Question 1: is this proposed model of quality assurance for the RDMF fit for the purpose of ensuring users have the information they need to be able to produce high quality statistics using the indexes and matching services within the RDMF?

Question 2: the validation and assurance framework will be kept up to date as the research landscape and understanding of user needs evolves - what scale of updates, single or cumulative, does the panel believe would warrant returning for refreshed external review?

Question 3: are there any additional aspects of the measurement or communication of quality regarding the indexes or matching services within RDMF that should be added to the proposed validation and assurance framework?

Executive Summary

The Reference Data Management Framework (RDMF) is a framework to manage the production of reference data indexes covering people, businesses, classifications, and location, and build and improve matching services for each index. The indexes integrate data from a variety of sources to construct their reference data. These indexes are intended to be used to facilitate better use of data for statistics by making reference data more easily accessible for statistical analysts. The RDMF is a general purpose product which will support analysts within the ONS and across government, as well as researchers outside of government.

As a piece of statistical infrastructure that will underpin a variety of statistical production activities at the ONS, as well as be made available across government and to researchers via the Integrated Data Service (IDS), assuring the quality of the RDMF is vital. If the quality of the RDMF is not fully known or not sufficiently well communicated to users, there is a risk that it may not be fit for purpose. Analytical outputs which use the RDMF in this scenario could lead to inaccurate conclusions being drawn, potentially leading to decisions being made based on poor quality evidence.

Methodologists in the Methodology and Quality Directorate (MQD) at the ONS have proposed a framework for understanding and assuring the quality of the RDMF. This framework is made up of a set of controls, which are policies, activities, processes, or outputs which together help to mitigate the above risk. These controls cover topics ranging from requiring the production of user guidance to support good practice in using the RDMF, to the production of linkage quality metrics to communicate the statistical quality of the indexes and performance of matching services.

Delivery of the controls will be completed at different levels of maturity. The depth and breadth of the assurance provided by the framework and its individual controls can increase over time as their application to the RDMF and its constituent indexes and matching services is better understood. This increase in maturity will be delivered incrementally. In line with the Code, the framework also includes multiple lines of assurance for each control, up to and including review by external experts to ensure a high degree of assurance, as is proportionate for a programme of this impact.

Supporting controls in the framework, we present overviews of current research projects. We present QUAIL, a project for sampling from highly integrated demographic datasets to support clerical evaluation and generation of linkage performance metrics. We also present GLADIS, a system for automating linkage to the Demographic Index. In addition, we present a research project for identifying False Positive and False Negative Clusters, which are different types of error peculiar to highly integrated data.

Given the importance of understanding and assuring the quality of the RDMF, we are seeking input from the panel at this early stage on the fitness for purpose of our framework for quality assurance of the RDMF, and the suitability of the research portfolio for providing quality information about the RDMF. The results of this research programme and progress against the controls we introduce in this paper will be submitted to future MARP sessions, which will constitute a solid foundation of statistical quality for the RDMF.

Assuring the quality of the RDMF

Introduction

Reference Data Management Framework

The Reference Data Management Framework (RDMF) is a framework to manage the production of reference data indexes covering people, businesses, and locations. This includes creating, and iteratively improving, matching services for each index which link user-submitted data to the indexes at scale. RDMF also maintains a history of the changes to reference data, supporting longitudinal analysis. As mentioned, the aim of RDMF is to make reference data and linked data easily accessible for statistical analysis by providing self-service reference data to customers, reducing the waiting time to access data.

The Indexes are the Demographic Index (DI), Business Index (BI), Classifications Index (CI), and the Address Index (AI) & Geography Index (GI) which are combined into a Location Index (LI). They are produced from administrative data sources to provide coverage across England, Wales, and Northern Ireland. These indexes are updated on a regular basis and historic records are retained to provide a longitudinal aspect. Further, the indexes contain a deidentified layer, which is the layer made available to analysts, where the Personally Identifiable Information (PII) is replaced with a Unique Record Identifier (URI).

The RDMF Index Matching Services (IMS) are made up of DIMS, BIMS, CIMS and AIMS, which correspond to the indexes listed above. These matching services allow the Unique Record Identifiers (URI) derived from the relevant index to be assigned to an incoming dataset by way of matching a variable, such as business name, address, classification code or personal attributes, that is present on both the index and the dataset. Such PII is then removed from the incoming dataset, leaving only the URI and the dataset's non-disclosive attributes. This process, known as 'Indexing', enables customers to:

- Link two or more datasets by using the Unique Record Identifier (URI) to join them;
- Joining datasets with RDMF data products on the URI for enriched analysis;
- Utilise Cross Index Association (XIA) - a method of joining deidentified records across indexes, such as joining an individual on the Demographic Index to their employer on the Business Index or a company on the Business Index to an Address (and related geography) on the Location Index; and
- Use linked datasets without seeing PII data, as they have been de-identified by the RDMF process.

RDMF Data Products are a collection of deidentified outputs, derived from Demographic, Business and Location Indexes, sometimes interjoined via XIA. These data products will become a core component in the production of statistics within the ONS and will be made available to analysts across government and researchers as part of the Integrated Data Service (IDS), aiding better analysis for the public good. As a general-purpose piece of statistical infrastructure, the RDMF will be required to support a wide variety of use cases such as providing the sampling frame for business surveys, providing the reference data that will support the more timely production of population and migration statistics, and facilitating the answering of complex research questions that require highly integrated data. To be able to make use of the RDMF most effectively, users must have access to the guidance on its best practices, and quality information to support their analysis.

The Code requires that the quality assurance of statistics is proportionate to the nature of quality issues. The RDMF has high-profile and high-impact use cases, such as being used to

directly produce official statistics, underpinning key outputs like the Admin Based Population Estimates (ABPE) derived from Dynamic Population Model (DPM) in the Future of Population and Migration Statistics (FPMS), and being made available to researchers across and beyond government via the Integrated Data Service (IDS). The scale and impact of potential quality issues within the RDMF makes a detailed level of quality assurance proportionate under the Code. The variety of use cases also means that the quality information about the RDMF must be sufficient in both content and availability to facilitate users in determining how best to use the RDMF and account for its quality.

Integrated quality measurements within existing RDMF processes.

The building of an index within the RDMF can range in complexity from the purchasing of a high quality reference dataset to a series of deterministic and probabilistic linkages that integrate a variety of admin data sources. A description of the build process for the Demographic Index, one of the more complex processes, has been taken to a previous MARP session (Methodological Assurance Review Panel, 2023). As part of creating the indexes and matching services, proxies for quality and quality processes are already being measured within the index and matching service development teams. Outlined below is a brief description of the quality measurements taking place as a part of routine operations – not all indexes and not all matching services in their current form log measurements of quality as they are run, so this list is not exhaustive across indexes and matching services.

Business Index

The Business Index is a statistical register of businesses operating within the UK. Datasets from HMRC, Companies House, and the Financial Conduct Authority are linked using either deterministic exact matching or Fellegi-Sunter probabilistic linkage depending on whether the guaranteed uniqueness of records appearing in Companies House can be leveraged. Additional logic such as which combinations of sources correctly indicate that a business exists and is active, and the order in which records about a potentially living business arrived, determine whether links or individual records are added to the Index to represent a business. Information about the data as it flows through the linkage process is captured throughout. For example, counts of records added to the index of currently living businesses, records removed as closed businesses, and counts of records that have characteristics that clerical review has previously identified as difficult to link are all tracked. Further, those difficult to link records are set aside and submitted for clerical resolution as part of the daily update process.

Demographic Index

During the process of building the Demographic Index (DI) from a variety of data sources, simple information such as row counts are logged during the process to ensure no serious technical issues have been encountered. When updating to new versions of existing data sources, QA checks are limited to checking for consistency with previous versions in typical data quality measures such as rates of missingness and counts of unique values. Once the DI has been built with this new data, counts of how many records have or have not been linked to existing persons in the DI. When a new data source is added to the DI, more detailed clerical review is performed to provide an assessment of how well it has integrated and to tune the linkage method.

Address Index Matching Service

The Address Index Matching Service (AIMS) is a tool used to index addresses generated by administrative processes or submitted by users against the AddressBase database of every postal address in the UK, Isle of Man and Channel Islands. The development team have

implemented a series of baseline performance tests that are run every time the reference data or AIMS code is updated. It runs seven datasets of varying sources and levels of known true data through AIMS to produce performance measurements that target different aspects and areas of AIMS performance. For example, one dataset has been created by taking valid addresses from the reference data and exchanging words to make another valid address, which surfaces how well AIMS considers word order when matching addresses, whereas another contains addresses generated by the general public and indexed by a predecessor address matching tool to benchmark “real-world” performance.

A score has been developed to codify different outcomes of the linkage of a single address: the true match was the top candidate addresses provided by AIMS, the true match was in the top five candidates, no candidate addresses were returned, and the true match was not present in the list of candidate addresses. The frequencies of these outcomes are calculated across each test dataset where the true outcome is known. In addition, they are cross-tabulated against their equivalent from previous versions of the Matching Service, indicating exactly how these different outcomes have changed as a result of updates.

Classifications Index Matching Service

The Classifications Index Matching Service (CIMS) generates quality information as part of its routine operations. It treats input data differently depending on whether or not it is supplied with truth data. Without truth data, as would be the case with customer data, CIMS returns confidence scores at the microdata level and overall match rate, the proportion of predicted classifications with match scores over the user’s selected threshold. With truth data, CIMS returns a range of quality metrics including precision for all records, precision for all records with predictions that pass the user-selected match score threshold, precision and recall broken down by specific classifications, and the confusion matrix at the match score threshold. In addition, the team working on CIMS has been developing an automated sampling methodology which uses the above precision by label results in combination with the distribution of residual confidence scores to target the weakest regions of data for clerical support.

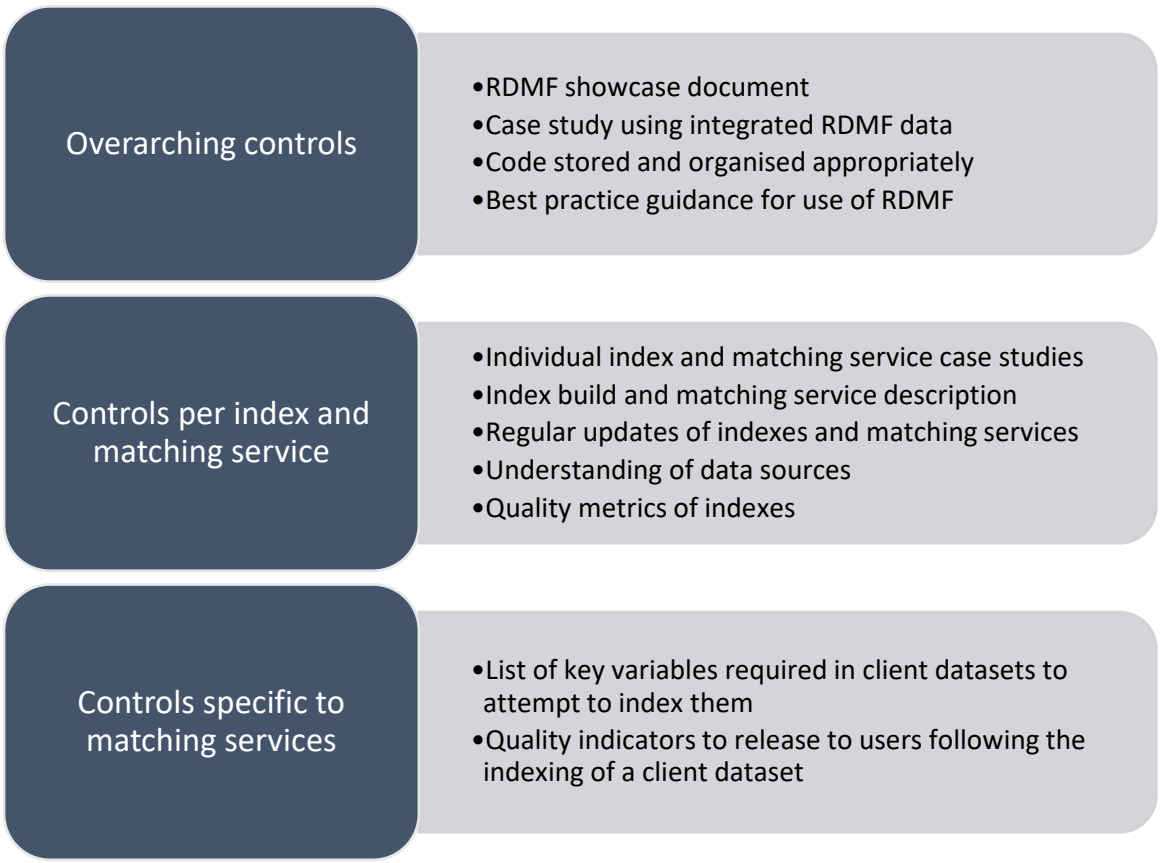
The proposed approach to quality assurance

Inherent to the use of RDMF is the risk that it may not be fit for purpose, that users will not have sufficient quality information to use the indexes and matching services effectively to make reliable statistics, and that the benefits and limitations of the RDMF will not be understood by users. Statistics based on the RDMF without appropriate knowledge of and accounting for statistical quality could lead to poor decisions based on them, posing a further risk to decision making across government as well as reputational damage to all users of the RDMF including the ONS and external researchers. To mitigate these risks, methodologists have proposed a framework of controls and assurance processes for the RDMF. Here, a control is an embedded activity, process, policy, or output which helps to mitigate the risk by generating information about the quality of the RDMF or ensuring that it is transparently communicated to users. These activities, products and policies include technical work such as quantifying measures of bias and uncertainty in the reference data indices and matching services, as well as provision of communication materials such as the creation of case studies and user guidance. Responsibility for the development and implementation of controls is distributed between methodologists, teams leading the overall RDMF programme, and teams developing individual indexes and matching services. Assurance of these controls will be provided internally by methodologists, externally by experts such as the Methodology Assurance Review Panel where appropriate, and finally by the RDMF governance board.

The ONS approach to the assurance of the RDMF is to treat the validation and assurance framework as a set of requirements which support user understanding and prevent poor use of the RDMF. These requirements can be met at different levels of maturity – both in terms of the depth of quality assurance increasing with the maturity of individual controls and the breadth of assurance increasing as more controls are completed. Incremental completion and improvement upon the controls will result in the maturing of the quality assurance of the RDMF. This includes where appropriate external assurance by external experts such as that of MARP. A first iteration of a control output may only contain surface level information and plans for further research, a mature iteration may include deeper investigations, additional information such as best practices, and will be informed by user feedback and priorities. Specific suggestions for higher levels of maturity will be suggested below with each control where appropriate.

The validation and assurance framework

The proposed validation and assurance framework currently consists of 12 controls which aim to mitigate the risk outlined above. These controls have an owner and different lines of assurance within the ONS. The first line of assurance sits within the teams that are developing the indexes and matching services. Each control is ultimately the responsibility of the RDMF governance board. Between those stages, assurance will be provided by methodologists in the Methodology and Quality directorate, outside of the teams developing the indexes and matching services, which will produce evidence for review by MARP in further papers.



Overarching controls

The first four controls are overarching across the RDMF as a project. The paragraphs following this table expand on the controls described in it.

Table 1: controls which apply across the entirety of the RDMF, not individual indexes or matching services.

	Control	Key outputs	Control owner
1.	Showcasing of the RDMF as a product	A RDMF description document	Overarching RDMF
2.	A RDMF case study on cross-index association (XIA)	MVP (Minimum Viable Product) documentation highlighting benefits of XIA	Overarching RDMF
3.	Access to RDMF algorithms in GitLab	Gitlab code artefacts	Index and matching service developers
4.	User guidance to demonstrate best practices in use of RDMF as a product.	Demonstrations with non-sensitive or synthetic data	Overarching RDMF

The first control is to clearly describe to users what is the overall statistical value of RDMF and convey any limitations on use due to quality, balancing selling points against usage and showcasing methods to use quality metrics effectively. It would be assured by the RDMF programme and by methodologists outside of the RDMF programme. It will also include a high-level summary of the overall quality level of the RDMF.

The second control is a communication piece to describe how RDMF is created, demonstrate benefits and statistical disclosure control risks to data suppliers, and create a common understanding of the RDMF among leadership within and beyond the ONS by demonstrating how cross-index association is done. The case study that has been chosen initially is an analysis of the ethnicity of company directors by region in the UK, demonstrating how the links between the Demographic, Business and Location Indexes can be leveraged to answer more complex research questions. More mature iterations on this control would include identification and communication of potential quality issues within the RDMF, which would inform future research directions to measure the impact and scale of these issues. Further, the cross-index-association could be extended to explore or simulate quality problems in the RDMF. As the understanding of the quality of individual indexes and matching services improves, quantitative research into the propagation of reference data error through XIA will also be undertaken. Assurance of this case study will be done by the RDMF Design Authority.

The third control is to ensure that the code responsible for the creation of the RDMF is stored in an appropriate version control system, Gitlab being the main system used by the ONS. This facilitates transparency and auditability, and generates a clear record of processes and changes made over time. Having code available in a git repository is required to support the provision of assurance outside of the development team. In line with best practice and guidance, code should be cleanly written and structured, follow existing Reproducible Analytical Pipelines (RAP) standards, and have appropriate governance such as change logs and periodic reviews. Review of these standards will be provided by methodologists.

The fourth control addresses the component of the overall risk that users may not understand the contents or benefits of the RDMF. It does so by providing guidance to users on how to use and leverage the benefits of the RDMF. It is also another opportunity to highlight the limitations on usage and interpretation of outputs derived from the RDMF. Assurance of this user guidance will be provided by methodologists.

Controls applying to each index and matching service

These five following controls apply to the individual reference data indexes and matching services produced by the RDMF.

The indexes form the spine against which the Matching Services link user-submitted data to facilitate large scale data linkage. The complexity and methods employed to build the indexes and perform matching within the matching services are heterogeneous and demand bespoke assessment of their quality, ranging from simply defining what items feature in the index to a complex series of deterministic and probabilistic linkage methods which integrate multiple admin data sources (Methodological Assurance Review Panel, 2023).

Table 2: controls which apply to each index and matching service individually.

	Control	Key metrics and artefacts	Control owner
5.	Case studies for each index and matching service	Description of measures used in case study to measure quality	Index and matching service developers
6.	Index build and matching service description	Index build process flow and design documentation	Index developers; methodologists for Demographic Index
7.	Regular updates of indexes and matching services	Schedule and roadmap for additions and changes to Indexes and matching services. Evidence-based update schedule and process for drift	Index and matching service developers
8.	Understanding of data sources	List of data sources used to create index or match against it	Index and matching service developers
9.	Quality metrics produced as part of index release note	Measures for bias (error) and uncertainty e.g. false negative, false positive clusters; Counts of records across source by year; cluster-level characteristics	Index developers

The fifth control, case studies for each index, would be used to measure and understand the statistical quality of outputs derived from an index or made via matching service. These case studies would be focused on the use of single indexes or matching services, complementary to the second control which highlights how the indexes and matching services can be leveraged in concert. This would give concrete examples to users of how index and matching service quality can affect an analysis and what actions they might take, or analysis methods they could use, to account for it.

The sixth control ensures that the statistical processes of building an index or indexing user supplied data are well understood and documented. This documentation will include

descriptions of input data, a change log, decisions made during the build or indexing process and justifications thereof alongside an assessment of, or recommendations for a methodology to assess, the impacts of those decisions.

The seventh control is to publish a schedule and roadmap for future development, ensuring transparency around the trajectory of the indexes and matching services. The frequency of updates will be determined both in response to user needs and by evidence from methodological research into data drift in these domains. This schedule would help inform users about how up-to-date the indexes and matching services are and whether specific required features will be added in the future, which will help users identify whether or not the relevant index or matching service is fit for their purposes.

The eight control is to investigate and document the quality of the data sources feeding into the RDMF. This will cover two types of quality: that which supports the operation of linkage such as the presence of key variables with correct levels and harmonised definitions; and the statistical quality such is its bias, coverage, completeness, uniqueness, and error rate. The relationship between data source quality and statistical quality of indexes and outputs of downstream statistical production is not yet known and will therefore be the subject of further research.

The ninth control in the indexes and matching services section is the production and dissemination of appropriate quality metrics released alongside indexes and matching services. As the construction of reference data indexes and the indexing of user data is fundamentally understandable as a data linkage problem, well understood linkage evaluation metrics such as global match rate, precision and recall can be used to deliver a baseline level of performance measurement. However, this domain is complex. Indexes themselves can involve integrating many admin datasets of varying quality, and linking to them using a matching service involves unseen user data of completely unknown quality. The responsibility for the research into these methods sits with methodologists, and creates an ongoing requirement for access to clerical assessment.

Measuring the statistical quality brings several critical benefits:

- Allowing users to make better decisions about how to use the indexes and avoid misuse, which will help minimise wasted efforts and avoid the publication of poor quality statistics and the formation of poor quality policies informed by them;
- Allowing users to interpret the statistical quality of their downstream analyses based on the strengths and limitations of the indexes;
- Connecting the statistical quality of the indexes more fully to statistical measures of quality for outputs, which will be developed alongside the case study work in control five; and
- Giving a baseline from which to make developments and improvements to an index by generating evidence for what changes will improve the statistical quality of an index or matching service the most.

Without these, it is not possible to say how useful downstream statistics are. The measures of quality that will be produced include measurements of uncertainty (such as precision and recall) adapted or analogised for highly integrated datasets formed from many linkages, analysis of false positive and false negative clusters, and match rates. These measures will also be checked for bias, investigating whether there are regions of the data that perform better or worse than others and ensuring discrepancies are presented to users. The availability of quality metrics will improve as the maturity of this control evolves. These measures are vital

for informing the communication of uncertainty in statistics based on the RDMF, such as Admin Data Based Population Estimates using the DPM as part of the FPMS. Methodological recommendations and research into suitability of quality metrics will be reviewed incrementally as they are developed.

Controls applying to each matching service

The following controls are specific to matching services, and do not apply to the reference data indexes.

Table 3: controls which apply only to the RDMF matching services.

	Control	Key metrics and artefacts	Control owner
10.	Key variable list (for client data coming in, to be indexed) to ensure high quality matching	Link to metadata model	Matching service developers
11.	Matching Service Quality indicators	Quality scores released to users following matching and indexing of a client dataset	Matching service developers

The tenth control is a technical requirement to be able to effectively use a matching service. If a matching service is expecting to be able to link people based on their name and postcode, users must be aware of the expectation that they will provide these key variables to use the matching service. This knowledge is also crucial for operational planning within the ONS as a measurement of the resources required to perform that indexing. These requirements will be supplied by data engineering teams within RDMF and assured by methodologists in MQD.

The eleventh control bridges the gap between understanding the quality of the reference data indexes and the output of the matching service as the counterpart to control number nine. Understanding the quality of the indexing process will support users in the effective analysis of indexed data, ensuring that indexing error is considered in their approach. It will also integrate with the understanding of the quality of the index. As a baseline, the quality indicators for indexing would be performance metrics for indexes such as match rate, precision, recall and F1 score and breaking it down by characteristics of the data. Further, methodologists would also investigate the ability of the matching service to cope with imperfect or outright false data, such as presenting the change in quality indicators as data errors are deliberately introduced to test sets. As this understanding matures, quality measures will be produced to help users understand the quality achieved in the indexing of their specific dataset, not just how the matching service performs in omnibus or representative testing.

> Question 1: is this proposed model of quality assurance for the RDMF fit for the purpose of ensuring users have the information they need to be able to produce high quality statistics using the indexes and matching services within the RDMF?

Updating the framework

The twelfth control in the validation and assurance framework is to ensure it is kept up to date and fit for purpose. Regular review of the controls to reflect maturing understanding of user needs and behaviours, technical understanding, and the changing operational landscape is essential to ensure efficient use of time and sufficient provision of assurance. This control is to be owned by methodologists. Changes to the framework will be agreed with the RDMF programme. Substantial updates such as adding new quality dimensions to the framework will be reviewed where appropriate.

> Question 2: the validation and assurance framework will be kept up to date as the research landscape and understanding of user needs evolves - what scale of updates, single or cumulative, does the panel believe would warrant returning for refreshed external review?

Current methodological research – developing the maturity of quality metrics

As presented, the proposed framework requires supporting research to be completed for all controls to reach higher levels of maturity. We present an overview of three research projects currently under active methodological development which will advance the maturity of controls 9 and 11 which refer to providing quantitative measures of quality.

1) False Positive and False Negative Clustering

The Demographic Index (DI) clusters records across several admin sources and years into clusters, which are assigned an "ONS id", this id is who the DI believes is a person. In November 2022 MARP reviewed (Methodological Assurance Review Panel, 2023) a paper that described the design of the DI and recommended a list of research to measure its quality. Since then, we have completed some of this research and have developed a better conception of DI error, which informs our current work and future plans.

We are focussing on pursuing quantitative measures of error in DI in accordance with control 9, and we have chosen to start with the simplest ones, rather than tackling trickier research questions such as longitudinal error, or error across geography or characteristics. The reason for this is that the simpler errors are more tractable, and we believe that they are a first necessary step towards solving the more complicated questions.

Overall, we conceive of DI error as having three components: clustering error, coverage error, and data measurement error. Currently, MQD is focussing on clustering error, and have broken this error into three sub-types:

- False Positive Clusters (FPC), where records for more than one person are mistakenly clustered into one ONS id;
- False Negative Clusters (FNC), where records for a single person are mistakenly spread across more than one ONS id; and
- Uncertain Clusters (UC), where the quality of the data does not allow either an algorithm or a clerical reviewer to resolve records into a cluster without error.

At present, we are developing a simple estimation method for FPC, where we are stratifying ONS ids according to how likely they are to contain this error. Through clerical investigation and input from domain experts, specific variables have been identified as having an

association with clustering errors. We are stratifying based on these variables, and variables not used in this stratification will be inspected in the results for any further associations. Clerical review will be used to label the false positive clusters, and the proportion of this error per stratum. Then we will use bootstrapping to produce estimates of uncertainty per proportion.

This work will initially be evaluated as a proof of concept using feedback from the DI quality research community and the results of clerical review. The resulting measure will be a stratum flag for every ONS id, which will allow users to allow for the impact of FPC error on their outputs. This ONS-id-level approach is important, because users are likely to choose a portion of DI for their analysis, such as only ONS ids for a specific year.

This work is exploratory and experimental, heavily reliant on expert opinion in lieu of existing data and results. The path to maturity in this research will cover improvement of stratification, and optimising our use of clerical review data.

Later this year we hope to begin work into a similar estimation methodology for FNC, and one for UC in 2025.

We expect that the evidence base for this work will be obtained through clerical review. Clerical review allows identification of error types in DI and is instrumental for developing our definitions and methods. Furthermore, reviewers identify the correct way to group records; this creates labelled data, which could be used to test the DI build and demonstrate the impact of design changes on clustering, or even as a basis for training a machine learning model to automate error detection and estimation. A key part of our stakeholder work is making the case that clerical review and work to test the DI should be prioritised and resourced. This also ties in with the seventh control in the assurance framework, as an understanding of the rate of data drift will inform the expected lifetime of any algorithm deployed to automate this work before review.

Besides clustering error, we are scoping a methodology for data measurement error, which we aim to also bring to MARP in the future for review. We have not been asked to consider coverage error, as this is the purview of the population statistics transformation team in ONS; however, we would like to see this error also handled as part of DI quality, rather than considered as part of constructing secondary datasets such as the Statistical Population Dataset (SPD), which will inform the DPM in the FPMS programme. To support a methodology for coverage error, we expect that ongoing linkage of high-quality data to DI will be required, such as a coverage survey or an independent admin data source.

2) Quality Analyser for Interpreting Linkage (QUAIL)

Understanding the quality of linkages within, and to, the RDMF is a crucial requirement of assuring the RDMF. Some of the indexes are constructed by linking more than two data sources, making measuring quality using typical metrics such as precision and recall difficult. Challenges increase as we link data to these indexes. QUAIL is a project that aims to provide a recommended package of researched methodologies and tools to facilitate the quality assurance process of data linked to the RDMF in demographic scenarios. The objective is to provide automated and generalisable methods to minimise user interaction and ensure wide-ranging application.

The research MQD is conducting through the QUAIL project will contribute to the eleventh control in the validation and assurance framework – Matching Service Quality Indicators. Understanding the quality of the output of data linked to the DI is vital to ensure user's apply

appropriate analyses and make sensible interpretations. Our initial research focuses on two key areas:

1. Stratified sampling

Clerical review is the manual process of evaluating the match status of links and potential links to estimate the incidence of false positives (links that have been made incorrectly) and false negatives (missed matches), respectively. The incidence of these different types of error are used to compute two key quality metrics: precision and recall (link accuracy and proportion of true matches identified, respectively). The anticipated size of datasets linked to the DI renders it unrealistic to review all links and possible links for error estimation; sub-sampling is thus required. There is no standardised sampling approach for this purpose. Traditional methods require an understanding of the error expected within the data, which is difficult without intricate knowledge of the datasets and methods involved in the linkage. This is impractical for generalised use and can lead to the potential for under- or over-sampling and uncertainty and inaccuracies in error estimates. The QUAIL project aims to develop methods for stratifying links to create representative samples while ensuring an appropriate number of links are reviewed, optimising resource efficiency.

Due to links and potential links containing different types of error we are undertaking an evaluation of stratification methodologies. We aim to recommend an appropriate generalisable approach that groups data in a way that minimises error variance between links within a stratum but maximises the variance between strata. Sampling from a range of strata will ensure reviewed links are representative of all types of error with more accurate evaluation than is possible with simple random sampling. MQD are currently comparing the utility of static, percentile, and adaptive granularity thresholding methods. For this purpose, we are initially exploring the use of match score, a likelihood measure of match status based on probabilistic linkage, before the potential for a multivariate approach.

MQD are currently researching methods to ensure that we sample sufficiently from strata, ensuring adequate sample sizes for accurate estimation while minimising the risk of under- and over-sampling. To ensure generalisability, MQD are researching Bayesian methods with pre-set precision and recall priors to provide samples of adequate size for effective clerical assessment. Different priors could be selected based on the needs of the analysis of the linked data (e.g., to assess if precision is at least 95%) or available clerical resource. We are also investigating alternative methods whereby the data itself can be used to inform the prior (expected error) before sample size is computed. Methods currently under review by MQD include the use of Beta distributions and Markov Chain Monte-Carlo in combination with Proportional or Neyman Allocation methods.

Methods for testing our stratified sampling options are under development, with plans to leverage data which has been clerically matched to a high standard, as well as the creation of synthetic datasets of varying quality and variable distributions under consideration.

2. The creation of a Precision and Recall tool

To increase efficiency and accuracy of quality metric computation, MQD have created a proof-of-concept Precision and Recall tool. Currently based on Frequentist stratified sampling methods, the tool automates data entry post-clerical review and computes precision, recall and confidence intervals. The tool will be adapted if Bayesian methods are recommended.

Collectively, the QUAIL package aims to streamline the clerical process. Reducing manual intervention will reduce the burden on clerical reviewers while simultaneously enhancing the accuracy, reliability, efficiency and reproducibility of quality reporting.

Future QUAIL research is to be determined through stakeholder discussion, but we anticipate expansion of the quality insights the package could provide. This includes, but is not limited to, the provision of additional summary metrics and bias analysis, the investigation of conflicts in data linkage, cluster-to-cluster and longitudinal linkage quality assurance, as well as guidance on how this information can be used. While QUAIL is targeted at person data linkage, there are potentially lessons and methods learned here that can be used in the other indexes, or for use in cross-index association linkages.

We will be submitting a subsequent paper covering the methods themselves and our evaluation methods in detail for scrutiny at a future MARP session.

3) Linkage Methods in the Demographic Index Matching Service

The Methodology and Quality Directorate are currently developing a prototype data linkage methodology for linking datasets to the Demographic Index. This prototype is referred to as GLADIS (Generalisable Linkage of Administrative Demographic Index Service).

While the prototype is being created using research-based methods where parameters are tested on different datasets to maximise and balance precision and recall, the underlying methodology relies on standard linkage techniques including the Splink implementation (Linacre, 2022) of the Fellegi-Sunter model to ensure a timely delivery is possible. Following on from this basic method, additional research will be commenced to ensure GLADIS contains the best suitable methodologies. This will include comparing the deterministic and probabilistic approach currently used with alternative methods. It is also of vital importance that we understand more about how a single method operates when applied to a variety of datasets, as is the service aim of DIMS. MARP assurance will be sought on this future research.

This work is attempting to deliver a single linkage pipeline resilient enough to perform fit for purpose linkage on a broad range of datasets with minimal manual intervention. Other areas that will be explored include:

- Requirements for appropriate test data to investigate generality of linkage method;
- The diagnostics which can be used to review dataset quality and whether the data is fit-for-linkage, threshold recommendation for what is fit-for-linkage and what is not;
- Assurance on our research into non-typical research methods and how their performance is compared, such as Goldstein's Scalelink, Maximum Entropy, Neural Networks, and Decision Trees;
- Propagation of errors from the Demographic Index itself when reporting linkage quality of GLADIS, and how to communicate this to users;
- Use of alternative variables such as previous surname or postcode in GLADIS; and
- Handling non-standard data structures such as longitudinal datasets or episode-based data in GLADIS.

Current research in RDMF development team – “Indexing First”

The Data Linkage Hub in the Data, Growth and Operations (DGO) directorate are undertaking a programme of research to explore the impact of using generalised linkage methods or ‘indexing’ via the Reference Data Management Framework (RDMF). Whilst this idea of linkage via indexing to the RDMF is conceptually well understood, further research is needed to understand the quality implications and practical implementation of use of this as a data linkage mechanism and the situations for which indexing would not be suitable to link data. An understanding of the coverage, representation, and quality of data linkage products will help researchers understand the populations they are able to explore and how to interpret findings from linked data. This will contribute to the maturity of control 4 in the validation and assurance framework, guidance and best practice, as well as controls 9 and 11.

Through the research we intend to explore how different subsections of the entities within the indexes are represented and how data linkage influences their inclusion or not. Therefore, the purpose of the work is not to focus specifically on one subsection of the population but rather explore how those different groups are represented and how best to ensure they are included and represented within data linkage processes. For example, within the demographic index we are unsure how well individuals with refugee status may appear in the data and over time. From the specific data sources, we are aware of issues with re-location and duplicate ID assignments as well as name spelling and consistency errors. As such, we will work with the team in ONS to understand the type of methods they have applied to ensure they capture people through their bespoke linkage and seek to learn more about the quality of the links being made. We will support a linkage to the demographic index generally and in bespoke methods to see how many of the same links can be identified. Therefore, building an understanding of how the population are represented on the demographic index but also the type of linkage needed to ensure a good research sample can be collected and used by analysts. The results will feed into general linkage methods as there is likely lots which can be learned about linking this population but also indicate if generalised methods are appropriate for linking this type of data.

The proposal of this research is to use existing projects to compare bespoke and general methods. The bespoke methods used are exactly those which are being used or would be used to complete a bespoke linkage service. Depending on the research question, data provided and desired quality the methods used will be a combination of deterministic and probabilistic methods. Both methods use personal identifiers to find links within the data.

Indexing, in theory, will be completed using a generalised linkage method. This is a combination of the methods outlined above, however there is minimal modification to the application of the method. The method is currently used within ONS and is being reviewed by the data linkage methodology team.

Included as part of the linkage process is the use of clerical review. This enables us to determine the linkages success by reporting against quality measures which will be an important part of the comparison of methods. It also enables researchers to explore where and how methods may not be performing as expected i.e. why are a sub-set of entities being missed by the method. Depending on who and what tools are available, records will be sent for clerical review where links found in the automated methods are presented to a human to establish agreement or not. This is done using the personal information within the records. An individual will review either using the Clerical Resolution Online Widget (CROW) or the Clerical

Matching System (CMS). Both tools sit within ONS' existing cloud platform so data is not transferred out of the secure space.

Alongside this, information will be gathered from other areas of ONS that have done relevant research. This will involve collating research papers and gathering evidence from other teams to understand their methodology and results. It may also lead to additional work to accompany the teams' findings and fill in gaps. This could be in the form of additional quality checks or supplementing the work by running a comparison with a different method.

Conclusion

We have proposed a framework of controls and assurances which together act to mitigate the risk that the RDMF will not be fit for purpose or not understood by users. To mitigate the risk, the framework must have sufficient coverage over all dimensions of quality and communication. If the framework has sufficient coverage and the controls are themselves of high enough standard, then the framework will be fit for purpose. Applying the framework given those conditions should mitigate the risk that users will not have sufficient information to use the indexes and matching services effectively to make reliable statistics and that the benefits of the RDMF will not be understood by users.

A pragmatic approach to assuring and communicating the quality of the RDMF is to take an iterative approach, with controls increasing in maturity with further research over time.

> Question 3: are there any additional aspects of the measurement or communication of quality regarding the indexes or matching services within RDMF that should be added to the proposed validation and assurance framework?

Future engagement with MARP

It is our intention to bring papers to future MARP sessions which will present the methods developed to support the controls outline in this paper. This programme of research, with external assurance provided by the expert review of the MARP panel, will fit together within the framework to form a strong foundation of quality for the RDMF. These future papers will include but are not limited to:

- Verification that the validation and assurance framework has been applied for each index or matching service;
- Future quantitative measurements of quality such as estimates of False Positive and False Negative Clusters, with potential extension to the development of automation with machine learning;
- The research methodology used to design the QUAIL and GLADIS projects; and
- Significant changes made to the validation and assurance framework.

Bibliography

Linacre, R. a. (2022, August). Splink: Free software for probabilistic record linkage at scale. *International Journal of Population Data Science, 7*(3). doi:10.23889/ijpds.v7i3.1794

Methodological Assurance Review Panel. (2023, February 9). *Methodological Assurance Review Panel – Agenda & Minutes 8 November 2022 – UK Statistics Authority*. Retrieved from United Kingdom Statistics Authority: https://uksa.statisticsauthority.gov.uk/publication/methodological-assurance-review-panel-agenda-minutes-8-november-2022/