Framework for Assuring the Quality of Administrative Data for Use in the 2021 Census of England and Wales: A Working Paper.

Steven Dunstan and Sara Correia

Contents

1. Introduction

1.1 Quality Assurance Approach

Administrative data present great opportunities for Census statistical production, but their potential can only be realised if their strengths and weaknesses are fully understood in relation to the ways in which they will be used in Census production. Hand (2018) identified several statistical challenges posed by administrative data including, among others, the need to "develop detectors for particular quality issues", the need to "construct quality metrics and quality scorecards for data sets", the need to "audit data sources for quality" and the need to "be aware of time series discontinuities arising from changing definitions" (2018, p. 10).

As a first step to meet the above challenges, this paper proposes a Quality Assurance (QA) Framework for assessing the quality of administrative data to be used in the 2021 Census of England and Wales, based on a qualitative scoring system. By applying this framework, quality issues will be detected and specific metrics applied systematically, so that acceptable thresholds can begin to be developed. The framework aims to provide confidence that the administrative data are of sufficient quality to be integrated into the Census design as intended. In addition, this framework will be useful for the QA of administrative sources for use in statistical production beyond the Census.

The current approach for the QA of the 2021 Census of England and Wales (ONS 2020) identifies two broad stages: Assurance of Processes (process quality) and Validation of Estimates (output quality). In a traditional Census, the quality of the Census data is assured through the robustness of the processes. However, given the different nature of administrative data collection and the varied ways in which administrative data will be integrated into the 2021 Census design, this framework covers the assessment of what is sometimes referred to in the literature as "input quality" (e.g. Eurostat ESSnet KOMUSO 2016), i.e. the quality of administrative sources and data. The framework therefore complements the existing Census QA strategy.

The framework covers the assessment of single sources and is not intended to cover the changes in quality due to the transformation of administrative data for use in the Census (e.g. through data cleaning, integration and linkage, mapping, editing, imputation, modelling, etc.). However, the information gained through the application of the framework will inform the necessary transformation stages to incorporate administrative data into the Census design.

Drawing on existing literature (Groves et al. 2004, Zhang 2012), the framework distinguishes broadly between measurement errors (in the measurement of variables or characteristics such as age, gender, etc) and representation errors (in the identification of population units or objects) within the sources used. Zhang (2012) also distinguished between the quality of single sources as provided by administrative authorities and the quality of transformed and/or integrated sources, after processing by a statistics organisation. The framework focuses on guiding the statistics producer in developing an in-depth understanding of the quality of single sources.

The framework includes two stages of assessment: Source Quality and Data Quality. Within each stage, key data quality dimensions are outlined, along with criteria / indicators to enable the assessment of each dimension. The framework aims to capture the key elements of the Office for Statistics Regulation (OSR) Quality Assurance of Administrative Data (QAAD) standard (UKSA 2015) and draws on international best practice (UNECE 2018, Daas et al. 2009, and the work of the UNECE Taskforce on *the Development of Guideline for the Assessment of Quality of Administrative Sources for use in Census*) (UNECE 2020).

1.2 The Structure of the Paper

Section two provides a summary of the purposes for which administrative data will be used or considered for use in the Census, along with the principles that will guide decisions on whether an administrative source is fit for purpose. The application of this framework will help determine whether these principles can be met.

An overview of the framework is given in section three. This includes an outline of the key stages of assessment within the framework; details of the quality dimensions that an administrative source is assessed against; along with an explanation of how the associated quality metrics and indicators should be used for the assessment.

Section four outlines the Input quality assessment in more detail, including the Source Quality and Data Quality stages. Source quality relates to the information that can be obtained about the source without access to the data itself, through conversations with the data supplier and via a review and assessment of metadata. The Data Quality stage relates to the information that can be obtained following the receipt of the administrative data supply. This stage includes analysis of the data, along with comparisons with other sources.

Section five draws some conclusions and recommendations.

1.3 Feedback

The framework is a working document and by no means finalised, we therefore welcome feedback, suggestions and ideas from the Expert Panel to aid its further development.

2. Use of Administrative Data in the 2021 Census

The 2021 Census design and its use of administrative data is still evolving in light of the Census rehearsal results and the current COVID pandemic. Nonetheless, the aim is to use administrative sources in the following key areas:

- The Census Address Frame: to provide confidence that the Frame will be fit for purpose, with adequate coverage;
- Field Operations: to improve the collection process and validate field outcomes;
- Processing and analysis: to support editing and imputation, coverage adjustments and general quality assurance; and
- Census Outputs: to produce new and enhanced outputs including on education achievement, number of rooms and income statistics.

The following principles will guide decisions on whether an administrative source will be used in the Census. The application of the QA Framework will help producers of statistics to determine whether these principles are met:

i. The administrative data must add value to the design set against associated costs and any risks to quality across the dimensions (relevance, accuracy and reliability, timeliness, accessibility and clarity, comparability and coherence). This could be with respect to:
    a. Improvements to the efficiency of the operation (reduced costs / time), supporting reductions in respondent burden and improvements in response.
    b. Improvements to the quality of Census outputs, with respect to the quality dimensions.
    c. Enhanced or new Census outputs

ii. The use of an administrative source can be clearly explained to users, including the necessary assurances on fitness for purpose.
iii. ONS is confident that the administrative data will arrive on time (punctuality) and to the required specification (as agreed with the data supplier).
iv. ONS is confident with respect to stakeholder acceptability of the use – e.g. the necessary privacy impact assessments, ethical assessments have been completed and the use has been communicated to the suppliers and public.

It is proposed that the intended use of each administrative source, along with the anticipated / expected outcomes is outlined in some detail (where possible) prior to the application of the framework. This will help inform how the framework is used and the level of assessment needed. Once the framework has been applied, the information gathered can then be used to determine whether an administrative source meets the guiding principles above and is therefore suitable for the proposed use. Annex I provides a basic template for recording this information.

3. Quality Assurance Framework Overview

3.1 Quality Stages

This framework adds Input Quality to the existing Census 2021 QA strategy, in order to assure the quality of administrative data sources. This comprises two stages: the Source Quality stage and the Data Quality stage. These stages can easily be mapped against the Census production, as they precede the current Assurance of Processes and Validation of Estimates stages:

- Source Quality: The assessment of the quality of the source through communication with the supplier and the available metadata. This includes, but is not limited to, determining that the source meets the requirements of the Census design, it's accessible and that adequate institutional requirements are met.
- Data Quality: The quality assessment of raw administrative data as they are received by ONS for use in the Census. It includes the cleaning and harmonisation of source data; checks against metadata/requirements; validity and completeness checks.

Within each stage, indicators and metrics are proposed to assess the source against key data quality dimensions. The following section provides an overview of these data quality dimensions.

3.2 Quality and Dimensions for Assessment

The quality of statistical output is broadly defined as 'fitness for purpose' (e.g. ONS 2013; Eurostat 2013), i.e. the statistics produced are adequate to meet the user's needs. Fitness for purpose requires statistics to be "based on appropriate data and methods" and "not materially misleading" (UKSA 2018). Furthermore, there is broad agreement that there are multiple dimensions when deciding whether source data and resulting statistics are fit for purpose and that sufficient information should be provided to users so they can come to their own conclusions on whether a statistic is fit for their purposes.

ONS (2013) has adopted the quality dimensions set out by the European Statistical System (ESS) for the assessment of the 2021 Census outputs: Relevance; Accuracy and Reliability; Timeliness and Punctuality; Accessibility and Clarity; Coherence and Comparability (Eurostat 2013, 2016). Several of these can be directly translated into data quality dimensions, as outlined in Table 2 below. In addition, there are a number of data-specific quality dimensions which will ultimately impact the quality of the outputs (e.g. the Institutional Environment).

Table 2 – Data quality dimensions defined

	Data Quality Dimension	Definition
SOURCE QUALITY	Metadata Availability & Quality	The extent to which easily comprehensible metadata are available to develop a full understanding of the administrative data source and how the data are collected and processed by the data supplier.
	Relevance	The degree to which the administrative data source meets the needs of the Census. It includes the extent to which the source covers the target population and required characteristics, as well as the timeliness of the data and the ability to perform linkage as required by the Census design.
	Data Accessibility and Stakeholder Acceptability	The ease in which the ONS can obtain the administrative data source, set against any restrictions (legal and those imposed by the supplier, including costs); the ease of data transfer and receipt; and stakeholder acceptability.
	The Institutional Environment	The organisational factors affecting the data supplier's capacity to supply data to the quality expected and to the agreed timetable (punctuality). Including the strength of the relationship, previous experience, existence of formal agreements, risks associated with the status of the supplier and the supplier's quality standards.
DATA QUALITY	Validation & Harmonisation	Adequate data validation/harmonisation arrangements are in place upon data-transfer to ONS, ensuring consistent processing across Census use cases. The data meets the expectations set out in the above Source Quality assessment.
	Dataset 'Linkability'	Adequate linkage variables are available, including either common unique identifiers, or a combination of variables which will enable linkage.
	Data Accuracy & Coherence	The accuracy and coherence of the data supplied matches the requirements of the specific Census use-case for which it will be used.
	Population Coverage	Comparisons with alternative sources reveal acceptable levels of coverage/under-coverage. This may be explored through prior feasibility research but needs to be confirmed on (re-)supply.
	Supplier Communication	Where queries regarding the data arise post-supply, there are adequate mechanisms in place to ensure their resolution.

3.3 Quality metrics and indicators

As well as identifying which dimensions should be assessed within each of the quality stages, this framework suggests indicators/metrics which will enable their assessment for the purposes of the 2021 Census.

A qualitative scale is suggested to signal when the acceptance criteria for each metric/indicator is considered to have been met. However, overall acceptance that the source is good enough will

depend on each Census use-case and thus case-by-case thresholds must be established by the relevant production teams.

Design-thinking and professional judgement will be required to establish what thresholds to accept, especially as for most criteria there is no scientific consensus. This is an area for further development, where methodologies to arrive at consensus may be considered[1]. Nonetheless, for the purposes of this framework, the emphasis is on adequately justifying and documenting the assessment of quality throughout to guide the decision-making process.

The annexes provide tables with the set of indicators in full, to be used as a checklist / template to be completed when undertaking the assessment at each of the quality stages.

1 Methods which might be used to develop expert consensus include the Delphi technique (e.g. Dalkey and Helmer 1963) and Interpretative Structural Modelling / Interpretative Ranking Process (e.g. Janes 1988).

4. Input Quality

4.1 Source Quality

The Source Quality stage refers to the assessment of the quality of the administrative source, based on the information obtained from the data supplier and the available metadata. The stage precedes the Data Stage, which covers the assessment once actual data are received. The Source stage is key for determining whether ONS should proceed with an acquisition (or continue with a re-supply of the data). The key data quality dimensions (as defined above) for assessment at this stage are:

- **Metadata Availability and Quality**
- **Relevance**
- **Data Accessibility and Stakeholder Acceptability**
- **The Institutional Environment**

Each dimension is covered in more detail below, with descriptions about the areas for assessment. Annex II provides an associated template and checklists for making the assessment, which draws on the checklists developed as part of the Eurostat Methodologies for an Integrated Use of Administrative Data (MAID) in the Statistical Process Project (ESSNet MIAD 2014, Deliverables B1 and B2) and the frameworks developed by Iwig et. al. (2013) and Daas et al. (2009).

4.1.1 Metadata Availability and Quality

Without complete, accurate and comprehensive metadata, it may not be possible to understand and assess the administrative source against the intended use for the Census. Besides the importance for internal process and quality management, metadata are vital for providing access to the data (where applicable) and providing the users with the knowledge necessary to interpret and use the data and outputs. The framework, therefore, includes metadata availability and quality as a key data quality dimension. The following key aspects are considered and assessed under the framework:

i. Clear and meaningful descriptions covering the units or objects within the administrative source and any relationships between the objects. The objects could, for example, be people, transactions or events. This information is required to assess how comparable the objects are against the Census target population set;
ii. The attributes / measurements that are recorded against the objects, including any definitions, concepts, classifications used. The time reference periods, frequency of updates and whether timestamps are included. This information is important to assessing conceptual alignment with the Census;
iii. The process and rules which determine which units or objects are included in the administrative source. Including any legal requirements placed on the subject and the administrative authority (e.g. tax payments and collection). This information is key to assessing over and under coverage;
iv. How the administrative data are collected, validated, processed and quality assured. Including any changes that are made to the data via edits, imputation or other adjustments. Also, any changes that will be made before data are made available (e.g. encryption, perturbation, aggregation, top-coding). This information is key for assessing the accuracy of the data;
v. Any difference affecting *i to iv* over time or space – this could be as a result of changes to, or differences in procedures or legislation. This is important in assessing the impact on quality against the Census usage;

vi. Details on the data structures, formats, size and the complexity and relationships within the data. Also, information on the lag between the reference period of the data and when they can be provided to ONS. This information is important to ensuring the datasets can be accessed and used within ONS's infrastructure in a timely way (also see Section 4.1.3 on Data Accessibility).

Annex II, Table 1 provides a checklist / template for assessing the completeness and quality of the metadata against key quality indicators (covering the key aspects outlined above). The user should provide evidence against each of the quality indicators, so that an assessment can be made using the basic scale provided in the final column of the table.

In order to obtain the relevant metadata about data sources the following processes are in place at the ONS:

- Secondments (e.g. ONS has carried out secondments at Home Office, NHS Digital and HMRC).
- Face-to-face and telephone meetings with suppliers.
- Statistical Quality Working Groups (e.g. there is an SQWG with NHS Digital, whereby researcher and data experts meet to further develop their understanding of NHS data sources).
- Reviews of existing metadata.

ONS is also developing qualitative methods for working with suppliers to develop an in depth understanding of collection procedures and the impact on quality.

4.1.2 Relevance

The availability of comprehensive and accurate metadata covering the key aspects outlined in section 4.1.1, supports the assessment of whether a data source is relevant against the identified use for the Census. This includes whether the administrative source covers the population units of interest for the Census and whether the measurements/characteristics of these units align with the concepts Census requires. The assessment should cover:

i. Alignment of the administrative data objects/units with the Census target units;
ii. The coverage of the set of objects in the administrative data against the population of interest for the Census;
iii. Conceptual and definitional alignment against the required census variables / attributes;
iv. Timeliness – i.e. the time lag between the reference period in the administrative data and when the data can be made available to ONS (considering any lags between an event or phenomenon occurring and being captured in the administrative source);
v. The impact associated with how the data supplier collects, processes and adjusts/imputes and quality assures the data on the quality of the data available to ONS;
vi. An assessment of the impact of changes over time and space to the administrative source that could impact on the quality of the data.

Annex II, Table 2 provides a checklist / template for assessing these key aspects, via a series of quality indicators against the census needs. The user should provide evidence against each of the quality indicators, so that an assessment can be made using the basic scale provided in the final column of the table.

4.1.3 Data Accessibility and stakeholder acceptability

It is important to assess the ease in which ONS can access and use an administrative source for use in the Census, accounting for any restrictions and challenges with the receipt of the data. Including public and supplier acceptability of the use of their data. The assessment should cover:

Restrictions on data access and use

These may include legal and privacy restrictions, affecting what data ONS can receive and how it can be used. Common restrictions affecting the accuracy and usefulness of the data include, disclosure treatment applied by the supplier (e.g. suppression of records or variables, encryption of identifiers, perturbation, banding and top-coding of variables), as well as rules on disclosure methods that ONS must apply (according to supplier agreements) affecting the Census outputs. An assessment should be made against the impact of such restrictions on the use of the administrative source for the Census.

Ease of data transfer and receipt

It is necessary to assess whether it is feasible to receive and ingest the administrative dataset(s) into ONS's system, set against any cost and timeliness constraints. This may include costs imposed by the supplier, or costs incurred by ONS in developing capability to receive and process the administrative data. An understanding of the data models, formats, schemas, file sizes, software and hardware used by the supplier for transmitting the data is necessary to assess ease of transfer and receipt (also see Section 4.1.1 on metadata).

Stakeholder Acceptability

It is necessary to assess the level of acceptability of ONS's use of an administrative source for the Census. If either the data supplier, or the data subjects (often the public) are against the use of their data, there is a risk. For example, it could change the way the public interact with the Census and / or the administrative source – directly affecting quality (and more generally may impact ONS's reputation). The Framework suggests the use of Data Privacy Impact Assessments (DPIAs) and Ethics assessments, as tools and indicators of public acceptability, where applicable. For the assessment of data ethics, the UK Statistics Authority's Ethics Self-Assessment Tool[2] and the advice of the National Statistician's Data Ethics Advisory Committee (NSDEC)[3] are key resources. The use of an administrative data source (set against the public benefit) should also be recorded (and accessible) on the ONS website, along with the necessary assurances on data security and privacy.

ONS already has well established processes in these areas, including corresponding DPIAs against all relevant data acquisitions (as required under GDPR).

Annex II, Table 3 provides a basic checklist / template covering the quality indicators for assessment. The user should provide a detailed description against each indicator (where this is relevant against the planned use), on which a basic assessment can be made using the scale provided.

4.1.4 The Institutional Environment

ONS is completely reliant on the administrative authority to collect, process and deliver the administrative data to the quality expected and to the agreed timetable (punctuality). ONS is also reliant on the information the data supplier provides about the data (see Section 4.1.1 on metadata)

---

[2] https://uksa.statisticsauthority.gov.uk/about-the-authority/committees/national-statisticians-data-ethics-advisory-committee/ethics-self-assessment-tool/

[3] https://www.statisticsauthority.gov.uk/about-the-authority/committees/nsdec/

and any foreseen changes to the data. It is, therefore, important to assess confidence in the data supplier's ability to meet these needs.

In practice, it is key to develop processes for managing the relationship with data supplier and ensuring a continuous dialogue. These should include mechanisms for the communication of:

1) changes in data collection including changes to the legal basis for the data, the concepts and classification, but also the processes and procedures for data collection and management and
2) mutually beneficial improvements to data quality.

This will ensure that any data quality concerns are brought to ONS's attention quickly, understood and any mitigating actions applied immediately to avoid any negative effect on the published statistics.

The following indicators are proposed for the assessment of the Institutional Environment:

i. The strength of the relationship with the supplier, including how responsive the supplier has been to ONS's queries and requests; the existence of regular and effective communication mechanisms;
ii. Previous experiences with the supplier, for example have previous supplies of data been subject to any unexpected errors or issues (e.g. delays in sending the data);
iii. The existence of agreements, such as Data Sharing Agreements, Memorandums of Understanding (MOU), including any legal contracts, which cover the key requirements, including timetables for delivery;
iv. The risk posed by the status of the supplier, including whether they are an established, stable, reputable organisation. Also, whether there is any legal or regulatory basis to the administrative function the supplier carries out;
v. The risk posed by the complexity of the supplier organisation and whether there are multiple organisations responsible for the administrative source (which could thus increase risk);
vi. Evidence of the supplier's quality standards and their application, including the principles, standards and guidelines adopted by the supplier for assuring quality.

Annex II, Table 4 provides a basic checklist / template covering the quality indicators for assessment. The user should provide a detailed description against each indicator (where this is relevant against the planned use), on which a basic assessment can be made using the scale provided.

4.2 Data Quality

The Data Quality stage refers to the quality assessment of the administrative raw data as supplied to ONS for use in the 2021 Census. This stage of QA involves developing tools to apply metrics, thus validating the raw data against the expectations based on the Source Quality assessment and the requirements of the Census design. Five data dimensions have been identified for this stage of assessment:

- **Validation and Harmonisation**
- **Dataset 'linkability'**
- **Data Accuracy & Coherence**
- **Population Coverage**
- **Ability to clarify data queries with supplier**

4.2.1 Validation & Harmonisation

Data structure and validation checks are needed to establish confidence in the data transfer and ingestion process. In addition, the application of office-wide harmonisation rules will ensure consistency of use. As such, data validation and harmonisation processes should cover:

i. Verification of data readability, i.e. that the files have not been damaged in any way or data lost in the transfer and ingestion process;
ii. Ensuring that the correct reference periods have been supplied along with all required variables/classifications for the purposes of the Census. This includes running basic checks on the variables and variable types received, assessing the completeness of dataset and variables, verifying that values are within expected ranges and ensuring duplicates are resolved;
iii. Data harmonisation rules are agreed and applied upstream from production teams to ensure consistency of processing of the data source across different census use cases. Such rules need to be developed based on production needs and as such this framework is not prescriptive. Nonetheless, examples of harmonisation processes include:
    a. All missing values are coded in the same way;
    b. Duplicates can be identified and are removed from dataset or otherwise resolved;
    c. Rules developed to check inter-variable coherence e.g. rule developed and applied where individuals' date of birth ≤ date of registration on data source.
iv. The results of data validation and harmonisation procedures are documented (e.g. through data profiles[4]);
v. There are no unexpected differences between current and previous supplies of the data source (e.g. based on compared data profiles);
vi. The data was supplied in a timely and punctual manner.

Where the expected variables are not supplied, the relevance of the data may be affected. In addition, if a re-supply or unexpected processing is required to make the administrative data useable, this may negatively impact on the reliability and timeliness. As such, it is important that basic validation is carried out before the data are made available to the Census production teams.

[4] Data profiles are semi-automated reports generated based on the running of pre-defined validation and harmonisation rules.

Drawing on previous literature (Daas et al. 2009; ESS Net MIAD 2014), Table 1 in Annex III provides key quality indicators/metrics for this data validation and harmonisation dimension. As noted through this guide, the annexed checklists should be used by project leads and the teams carrying out the quality assessments.

4.2.2 Dataset 'Linkability'

On receipt of an administrative dataset, it is crucial to test expectations of the ability to link it to other datasets, as required by the Census design. As such, specific *linkage validation checks* should be developed in light of Census linkage requirements and applied on receipt of the data. Broadly, these checks should cover:

i. That the expected linkage variables have been supplied;
ii. That the linkage variables supplied are of sufficient quality to be linked on. Three factors are key to the quality of the linkage variables including completeness, uniqueness and unbiased distributions.

The quality of linkage variables is particularly important where there are no common keys between the data source and the Census data. In such cases linkage may be performed based on several common variables (e.g. a combination of date of birth, name, sex, address etc.). In addition, different checks will be required based on the unit of linkage (i.e. personal level versus address level).

Furthermore, where linkage variables have been provided to ONS in a 'hashed' form (i.e. masked via a one-way algorithm to protect the privacy of the data subjects), it must be verified that the hashing performed by the supplier matches the hashing algorithm used at ONS. If not, it will not be possible to link the data supplied to other data sources, undermining the relevance of the data.

In addition, it will be necessary to agree with the Data Supplier which derived variables[5] should be calculated and how these should be standardised before hashing. This is necessary to overcome the loss of edit distance measurements that are commonly used in matching procedures, but which are of no use with hashed data.

When linkage variables are provided in a hashed form, quality assessment of the linkage via clerical review of matches is very challenging. It is therefore key that the data supplier agrees to also provide a sample of the records 'in-the-clear' (i.e. not hashed). Linkage quality measures including precision and recall can then be estimated by calculating the number of false positives (incorrect matches) and false negatives (missed matches) from this sub-sample.

Annex III, Table 2 provides key quality indicators/metrics for the linkability of the data, based on previous literature and best practice at ONS.

4.2.3 Data Accuracy & Coherence

The accuracy, coherence and coverage of the data supplied should be verified in detail, beyond automated checks, in relation to the specific Census use-case for which it will be used. As such, in addition to the automated checks described in section 4.2.1 above, detailed *(manual) validation and distribution checks* should be carried out. Some checks including completeness, longitudinal

[5] A derived variable is one which is not directly collected but can be calculated from the information in the data source e.g. age from date of birth or income from a range of income-related data

consistency and consistency across the geographies supplied relate to the internal validity of the data. Others ensure the data are consistent against other sources (external validity).

Completeness (or missingness) checks are straightforward, but should be carried out across demographic characteristics to ensure that the distribution of missing values is not biased. A percentage of over and under-coverage can only be computed through comparison with an alternative source considered a 'gold standard' for a specific (sub-)population. However, feasibility research and previous supplies of a specific data-source will provide an indication of the level of coverage expected. Finally, the presence of outliers in the data may be indicative of measurement or response errors in the data.

4.2.4 Population Coverage

Related to the above, coverage checks require the data to be linked to or compared with alternative sources (where possible). When linking to other sources, it is also possible to explore the levels of selectivity or bias in the data.

Such detailed checks are especially important in the post-COVID context as patterns of administrative data recording are very sensitive to operational changes by, or service demand on, the data supplier. In the future, indicators may be developed based on relative differences between the distributions of current and previous supplies.

Annex III, Table 3 provides key quality indicators/metrics for assessing data accuracy, coherence and population coverage.

4.2.5 Supplier Communication

Finally, even where an in-depth source quality assessment has been carried out, it is inevitable that Census production teams will run into issues when using admin data which will require clarification from the supplier. As such, it is important to assess in advance whether the necessary communication channels exist to enable this dialogue.

To enable this dialogue, the *Statistical Quality Working Groups (SQWGs)* are a key resource. The NHS Digital SQWG for example, which brings together production, data architecture and data acquisition teams at ONS, as well as representatives from NHS direct.

Annex III, Table 4 provides a quality indicator for assessing whether such communication mechanisms exist with the supplier.

5. Conclusion & Recommendations

The framework and associated checklists should provide a useful guide for the assessment of the quality of administrative sources for use in the Census. The framework aims to identify the key dimensions of data source quality and how to assess them, recognising that many factors affecting quality are difficult if not impossible to measure objectively. To this end, the framework provides a basic scale against various indicators to help assess whether an administrative source can meet the needs of the Census. Inevitably, each indicator will be developed over time, as knowledge of the data source accumulates and, ideally, the quality of the data supplied improves based on ongoing dialogue with suppliers.

It is recommended that further work be carried out to develop approaches for setting thresholds for each of the indicators proposed in this paper. Of course such approaches will depend on whether it is possible to transform administrative data (including through integration with other sources) to overcome limitations and make the data ready for statistical analysis. This may include the integration of multiple data sources to provide the necessary coverage for Census and to reduce measurement error (by selecting the highest quality variables where they are available from multiple sources). It also includes data mapping, editing, imputation and modelling to address measurement and representation errors.

As such, it is also recommended that further work be carried out to develop an error framework for multisource statistics (where multiple sources are used in the production of Census outputs), building on the work of Zhang (2012) and the Eurostat Essnet on the Quality of Multisource Statistics (2019). By focusing on single sources, the working framework proposed in this paper provides a useful guide for identifying and evaluating error associated with the input administrative sources. This is a key step before seeking to transform the potential of single sources through integration.

References

Daas P., et al. (2009). Checklist for the Quality evaluation of Administrative Data Sources. Available from: https://ec.europa.eu/eurostat/documents/64157/4374310/45-Checklist-quality-evaluation-administrative-data-sources-2009.pdf/24ffb3dd-5509-4f7e-9683-4477be82ee60

Dalkey, N. and Helmer, O. (1963). An Experimental Application of the DELPHI Method to the Use of Experts. *Management Science*. 9(3), pp. 458-467.

Eurostat. (2013). Eurostat Quality Assurance Framework.

Eurostat ESSnet MIAD. (2014). MIAD deliverable B2, B3 - Quality check list for the Source phase [Data]. Available from: https://ec.europa.eu/eurostat/cros/content/miad-deliverable-b2_en

Eurostat ESSnet KOMUSO. (2016). Checklist for Evaluating the Quality of Input Data.

Eurostat ESSnet KOMUSO. (2019). Quality Guidelines for Multisource Statistics, Available from: https://ec.europa.eu/eurostat/cros/system/files/qgmss-v1.1_1.pdf

Hand, D. J. (2018). Statistical challenges of administrative and transaction data (with discussion). *Journal of the Royal Statistical Society*, Series A (181), Part 3, pp. 555-605.

Iwig, B., Berning, M., Marck, P. and Prell, M. (2013). Data Quality Assessment Tool for Administrative Data.

Janes, F. R. (1988). Interpretive structural modelling: a methodology for structuring complex issues. Transactions of the Institute of Measurement and Control, 10(3), 145–154.

Groves, R. M., F. J.Fowler Jr., M. Couper, J. M.Lepkowski, E.Singer and R. Tourrangeau. (2004). Survey methodology, Wiley, New York.

Office for National Statistics (ONS). (2013). Guidelines for Measuring Statistical Output Quality.

Office for National Statistics (ONS). (2020). Approach and processes for assuring the quality of the 2021 Census data. Available from: https://www.ons.gov.uk/census/censustransformationprogramme/approachandprocessesforassuringthequalityofthe2021censusdata

UK Statistics Authority (UKSA) (2015). Administrative Data Quality Assurance Toolkit. Available from: http://www.statisticsauthority.gov.uk/assessment/monitoring/administrative-data-and-official-statistics/quality-assurance-toolkit.pdf

UK Statistics Authority (UKSA) (2018). Code of Practice for Official Statistics (Edition 2.0).

UK Statistics Authority (UKSA). (2019). Quality Assurance of Administrative Data (QAAD) toolkit. https://uksa.statisticsauthority.gov.uk/publication/administrative-data-quality-assurance-toolkit/

UNECE 2020. Taskforce on Measuring the Quality of Administrative Sources for Use in Censuses. https://statswiki.unece.org/pages/viewpage.action?pageId=263229070

Zhang, L-C (2012). Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica 66(1): 41–63.

Annex I - Template for the assessment of overall suitability of the administrative data source

Relevance for census design	Description	Assessment Scale
1.1	**Description of intended use.**	1 - Doesn't meet need 2 - Partly meets need 3 - Meets need
1.2	**Expected outcomes – efficiency.** Could include reductions in time, cost and respondent burden, or improvements to census response rates.	1 - Doesn't meet need 2 - Partly meets need 3 - Meets need
1.3	**Expected outcomes – quality.** May include improvements to the relevance, accuracy and reliability, timeliness and punctuality, coherence and comparability of Census outputs	1 - Doesn't meet need 2 - Partly meets need 3 - Meets need
1.4	**Outline any costs** associated with the acquisition or processing of the admin data that would need to be considered against the benefits realised	1 - Costs outweigh benefits 2 - Acceptable cost implications 3 - No cost implications
1.5	**Outline any risks** identified with the accuracy (e.g. could the admin data add more uncertainty, or could it require additional disclosure control affecting accuracy?); timeliness (e.g. could the use add to processing time?); accessibility (e.g. does the use of the admin data risk what detail can be published due to confidentiality concerns?); coherence (e.g. do the definitions, concepts align with census?), clarity (e.g. can we adequately explain the use?) and circularity (e.g. are we using the source across multiple stages of the design, which could cause issues of circularity?)	1 - High risk 2 - Medium risk 3 - Low risk

Annex II - Checklists / templates for assessing the quality of administrative sources at the source quality stage.

Table 1 - Checklist for the assessment of Metadata Availability and Completeness.

1. Metadata Availability and Completeness	Quality Indicator	Assessment Scale
1.1	**Administrative objects / units** Is the metadata sufficient to describe the objects / units? *Key information:* What definitions, methods, processes are used to identify and include an object in the source? Are there any laws or regulations that define the objects?	1 - No 2 - Partially 3 - Yes
1.2	**The administrative object set / population** Is the metadata sufficient to describe the coverage of the objects / units? *Key information:* Are there any rules, legislative or regulatory requirements, including penalties for non-compliance that may impact on the inclusion and exclusion of objects on the source? What methods and processes are adopted by the data holder to include new objects that meet the required inclusion criteria / definitions (e.g. registration procedures) and to remove objects that no longer align with the target population for the administrative source (e.g. deregistration procedures)?	1 - No 2 - Partially 3 - Yes
1.3	**Relationships between objects / units** Is the metadata sufficient to describe the relationships across objects / units?	1 - No 2 - Partially 3 - Yes N/A
1.4	**Attributes / characteristics of objects** Is the metadata sufficient to describe the attributes of the objects (i.e. the variables)? *Key information:* What concepts, definitions and classifications are used? Is there a difference between the data holder's ideal target concepts and what is actually achieved through their operational target measure used in the collection?	1 - No 2 - Partially 3 - Yes
1.5	**The collection process** Is there sufficient information about how the administrative data are collected? *Key information:* What is the purpose for collecting the data? Are there any legal obligations, targets or incentives (or lack of incentives) that could influence the quality of the data collection? Are any data items recorded by proxy and therefore not reported directly by the data subject (increasing the potential for misreporting)? Are there any incentives or disincentives on the data subjects to provide complete and accurate information to the data holder?	1 - No 2 - Partially 3 - Yes

1.6	**The reference period, frequency of updates** Is there sufficient information about the reference periods to which the administrative data relate, including details on the frequency of updates / interactions and whether the time an update is made is recorded? *Key information:* Are there any known delays between an event or phenomenon occurring and being captured in the administrative source (e.g. parents do not have to register a birth for several weeks on the birth register)? Are there any incentives or disincentives for a data subject to update their information as and when their circumstances / information changes on the administrative source (e.g. benefits or penalties for not doing so / or doing so)?	1 - No 2 - Partially 3 - Yes
1.7	**Processing carried out by the Data Holder** Is the metadata sufficient on the processes carried out by the Data Holder that could change the underlying data? *Key information:* Are data edited or imputed? If so, when and how, and is there an indicator on the data source to identify when an edit and imputation has taken place? Are there any rules, regulations or incentives on the data holder that may impact the way data are processed?	1 - No 2 - Partially 3 - Yes
1.8	**Validation and quality assurance processes** Is the metadata sufficient on the processes carried out by the administrative authority to validate and quality assure the administrative data? *Key information*: What procedures are in place to validate and check data on entry by the data holder? What procedures are in place to minimise or identify errors due to data processing? Do quality indicators exist on the datasets?	1 - No 2 - Partially 3 - Yes
1.9	**Changes over time affecting 1.1 to 1.8** Is the metadata sufficient on any changes that have occurred over time that could affect any of the earlier components (e.g. the objects, their attributes, coverage, etc)	1 - No 2 - Partially 3 - Yes NA
1.10	**Changes over space affecting 1.1 to 1.8** Is the metadata sufficient on any differences over space or domain (e.g. regions, collection and processing organisations within the administrative authority) that could affect any of the earlier components (e.g. the objects, their attributes, coverage, etc)?	1 - No 2 - Partially 3 - Yes NA
1.11	**Uniqueness of units and availability of linkage variables** Is the metadata sufficient on the availability and treatment of unique identifiers on the administrative source to be used for linkage?	1 - No 2 - Partially 3 - Yes N/A
1.12	**Data structures, formats, size and complexity of datasets and relationships.** Is the metadata sufficient on the data structures, formats, size and complexity of datasets and relationships?	1 - No 2 - Partially 3 - Yes

1.13	**Timeliness of supply** Is there adequate information on the lag between the reference period of the data and when the data can be provided to ONS?	1 - No 2 - Partially 3 - Yes

Table 2 - Checklist for the assessment of Relevance

2. Relevance	Quality Indicator	Assessment Scale
2.1	**Objects align with the Census target units** Do the administrative source objects or units align to the Census need? *(Note: in the case of misalignment with the Census units, is a transformation possible that could meet the Census needs?)*	1 - No 2 - Partially 3 - Yes D/K
2.2	**Required relationships between objects** Do the relational structures between objects in the administrative source meet the Census needs?	1 - No 2 - Partially 3 - Yes D/K N/A
2.3	**Coverage alignment with the Census** Does the coverage of the population units/object set meet the needs of the Census? For example, is there any evidence of under-coverage (objects that are missing from the source, but are part of the Census target population) and over-coverage (objects that are in the source, but are not part of the Census target population) that would impact the usefulness of the source? *(Note: In the case of coverage error, are there other data sources that could be used in combination with the source to overcome, for example, under or over-coverage in the source?)*	1 - No 2 - Partially 3 - Yes D/K
2.4	**Attributes of the units align with Census concepts / definitions** Are the administrative source concepts, definitions and classifications comparable with the Census needs? *(Note: In the case of misalignment with the census concepts, definitions and classifications, is a transformation possible to meet the census needs?)*	1 - No 2 - Partially 3 - Yes D/K
2.5	**Quality impacts of the collection process** (including any biases in the way the data are collected) Does the collection process suggest the quality of the resulting data will meet the needs of the Census?	1 - No 2 - Partially 3 - Yes D/K

2.6	**Data reference periods and frequency of update requirements** Are the reference periods, frequencies of updates adequate for Census needs?	1 - No 2 - Partially 3 - Yes D/K
2.7	**Impact on quality of data processing** *(including, editing, imputation and any treatment by the Data Supplier that would impact quality (e.g. disclosure treatments)* Does the processing carried out by the administrative authority suggest the quality of the resulting data will meet the needs of the Census?	1 - No 2 - Partially 3 - Yes D/K
2.8	**Validation and quality assurance processes** Are the validation and quality assurance processes carried out by the administrative authority adequate for the needs of the Census?	1 - No 2 - Partially 3 - Yes D/K
2.9	**Changes over time impacting quality** Are the administrative sources coherent and consistent enough over time to meet the needs of the Census?	1 - No 2 - Partially 3 - Yes D/K N/A
2.10	**Differences over space / domains that affect quality** Are the administrative sources coherent and consistent enough over space / domain to meet the needs of the Census?	1 - No 2 - Partially 3 - Yes D/K N/A
2.11	**Linkability of the data source** Are unique identifiers / linkage keys available at a high enough quality to meet the linkage requirements of the Census?	1 - No 2 - Partially 3 - Yes D/K N/A
2.12	**Timelines of supply** Can the ONS acquire, structure and process the data in time for the needs of the Census, based on the lag between the reference period and the date the Data Supplier can deliver the data?	1 - No 2 - Partially 3 - Yes D/K

Table 3 - Checklist for the assessment of Data Accessibility and Stakeholder Acceptability

3. Data Accessibility and Stakeholder Acceptability	**Quality Indicator**	**Assessment Scale**
3.1	**Restrictions on data access and use** Is there a risk associated with any legal or privacy restrictions that could prevent ONS acquiring and using the data for the intended purpose? *Describe the risk(s) and the impact*	1 - High risk 2 - Medium risk 3 - Low risk D/K

3.2	**Ease of data transfer and receipt** Are the data file structures, formats, schemas and the methods (and software / hardware needed) for the transfer and ingest of data acceptable to ONS?	1 - No 2 - Partially 3 - Yes D/K
3.3	**Ease of data transfer and receipt – Cost** Are there any potential costs (or significant resource implications) associated with the acquisition and processing of the data that could impact the feasibility and value of the supply? *Describe the costs and the impact*	1 - High cost 2 - Low cost 3 - No cost D/K
3.4	**Stakeholder Acceptability** Based on (where applicable) the completion of a Data Privacy Impact Assessment and Ethics Assessment (and any public / supplier feedback or engagement exercises) what is the perceived level of acceptability of the use?	1 - Low 2 - Medium 3 - High N/A D/K

Table 4 - Checklist for the assessment of the Institutional Environment

4. Institutional Environment	Quality Indicator	Assessment Scale
4.1	**Strength of the relationship with the supplier** What is the strength of the relationship with the Data Supplier? *Key evidence might include the responsiveness of the supplier to ONS requests. Whether they have been proactive in their communication. The existence of effective communication mechanisms.*	1 - Poor 2 - Average 3 - Good
4.3	**The existence of robust agreements covering ONS's requirements** Are Data Sharing Agreements in place that adequately cover ONS's requirements, whereby the supplier has committed to delivering the services needed?	1 - Not in place 2 - Partly in place 3 - In place
4.4	**Status of the Administrative Organisation** Is there a risk associated with the status of the Administrative Organisation to meeting the needs of the Census? *Key evidence could include whether there is a legal / regulatory basis to the supplier's existence / function. Whether their organisation is well established with a strong reputation.*	1 – High risk 2 – Medium risk 3 – Low risk D/K

4.5	**Complexity of the Administrative Organisation(s)** What is the risk posed by the complexity of the administrative organisation(s) in meeting the Census needs? *Consideration should be given to whether there are multiple organisations involved in the collection, processing and supply; the strength of the links between such organisations and the levels of coordination.*	1 - High 2 - Medium 3 - Low D/K
4.6	**Quality standards adopted by the Administrative Organisation** Do the quality principles, standards and procedures adopted by Administrative Organisation (or the lack of these) pose a risk to the supply of data to meet ONS's needs? *Consideration should be given to previous experiences (where applicable)– e.g. if the supplier has met ONS's quality requirements.*	1 - High risk 2 - Medium risk 3 - Low risk

Annex III – Checklists / templates for assessing the quality of administrative sources at the Data Quality stage.

Table 1 - Checklist for assessment of data validation arrangements and results

1. Data validation and harmonisation	Quality Indicator/Metric	Assessment Scale
1.1	Diagnostic tools or validation checks have been developed based on data requirements and the source quality assessment.	1 - No checks developed 2 - Checks under development 3 - Checks agreed and developed
1.2	Readability of datafiles – all the data provided can be accessed with no format or structural errors / all data items successfully loaded.	1 - Considerable errors/omissions in data transfer 2 - Moderate errors/omissions in data transfer 3 - No errors/omissions in data transfer
1.3	Data source has been standardised/harmonised by the Data Engineering team.	1 - Incomplete 2 - Ongoing 3 - Complete
1.4	Data validation and harmonisation process has been documented.	1 - Incomplete 2 - Ongoing 3 - Complete
1.5	Coherence between metadata descriptions and data supplied: *Does the analysis of the data versus the metadata reveal anomalies that put the data into question and could limit the use of the data or require seeking clarifications from the data provider (e.g. reference periods, variables supplied etc.)?*	1 - Supplied data does not match metadata 2 - Moderate differences between supplied data does not match metadata* 3 - No significant differences between supplied data and metadata*
1.6	Differences between new and previous supplies (based on data profiles).	1 - Significant differences between new and previous supplies 2 - Moderate differences between new and previous supplies* 3 - No significant differences between

		new and previous supplies*

Table 2 - Checklist for assessment of data 'linkability'.

2. Dataset Linkability	Quality Indicator/Metric	Assessment Scale
2.1	The correct linkage variables were supplied as required by Census needs.	1 - Linkage data requirements not met 2 - Linkage data requirements partly met 3 - Linkage data requirements met
2.2	The quality of linkage variables (e.g. unbiased distributions, completeness and uniqueness) is adequate to Census needs.	1 - Low quality 2 - Medium quality 3 - High quality
2.3	The hashing of linkage by supplier matches the ONS hashing algorithm. (where applicable).	1 – No/Not known 2 – Under review 3 – Yes
2.4	A sub-sample of data 'in the clear' has been provided to test linkage quality. (where applicable)	1 – Not provided 2 – Under request 3 – Provided

Table 3 - Checklist for the assessment of detailed data accuracy, coherence and coverage

3. Data accuracy, coherence and coverage	Quality Indicator/Metric	Assessment Scale
3.1	**Variable distributions** are Plausible with respect to known or expected relationships / ranges / distributions. *Given the results obtained, how do you rate the limitations of the data on the potential uses of interest?*	1 - High Impact 2 - Medium Impact 3 - Low Impact DK: don't know

3.2	**Levels of missingness** for key variables calculated as percentage of total records. *Given the results obtained, how do you rate the limitations of the data on the potential uses of interest?*	1 - High Impact 2 - Medium Impact 3 - Low Impact DK: don't know
3.3	**Variable missingness selectivity.** Assess if the objects with variable non-response have similar characteristics than objects with non-missing data. These should be calculated by object type (i.e., within each object set). Compare summary statistics for the objects attributes available within the data space of interest for objects with variable non-response and objects without variable non-response. Ideally, the variables used are in relation with the outcome or study variables of interest. Histograms or bar plots can be used. *Given the results obtained, how do you rate the limitations of the data on the potential uses?*	1 - High Impact 2 - Medium Impact 3 - Low Impact DK: don't know (e.g., unable to assess given the variables available)
3.4	**Element non-response** occurs when all the data are missing for the variables of interest for a given element. Valid zeroes do not count as missing. Calculate the percentage of elements (or units) with all data missing for all variables of interest. This indicator can be calculated separately for the main variables of interest. *Given the results obtained, how do you rate the impact of element non-response on the potential uses of interest?*	1 - High Impact 2 - Medium Impact 3 - Low Impact DK: don't know (e.g., unable to assess given the variables available)
3.5	**Level of unit non-response selectivity.** Compare summary statistics for the object type attributes available within the data space of interest for responding and non-responding elements (histograms or bar plots can be used). *Given the results obtained, how do you rate the limitations of the data on the potential uses?*	1 - High Impact 2 - Medium Impact 3 - Low Impact DK: don't know (e.g., unable to assess given the variables available)

3.6	**Calculate the percentage of elements that violated edit rules** for the range of acceptable values for a given variable of interest. These can also be calculated by object type (i.e., within each object set). *Given the results obtained, how do you rate the limitations of the data on the potential uses of interest?*	1 - High Impact 2 - Medium Impact 3 - Low Impact DK: don't know (e.g., unable to assess given the variables available)
3.7	**Percentage of outliers is calculated and assessed**. For variables of interest that do not have known values ranges, examine the distribution of the values to identify outliers. Consider using different outlier detection techniques such as the quartile method. Outliers should be flagged. These can also be calculated by object type (i.e., within each object set). *Given the results obtained, how do you rate the limitations of the data on the potential uses of interest?*	1 - High Impact 2 - Medium Impact 3 - Low Impact DK: don't know (e.g., unable to assess given the variables available)
3.8	**Variables - imputation by data source provider** Calculate the percentage of elements with values imputed by the data provider for the main variables of interest. These can also be calculated by object type (i.e., within each object set). *Given the results obtained, how do you rate the limitations of the data on the potential uses of interest?*	1 - High Impact 2 - Medium Impact 3 - Low Impact DK: don't know (e.g., cannot be told by the information available)
3.9	**Variables – rounding** Rounding can affect the value distribution. Is there any evidence that rounding occurred for the main variables of interest? This can be detected by producing summary statistics and plots or histograms. *Given the results obtained, how do you rate the limitations of the data on the potential uses of interest?*	1 - High Impact 2 - Medium Impact 3 - Low Impact DK: don't know (e.g., unable to assess given the variables available)
3.10	**Timeliness** (where time stamps are recorded can be assessed)	1 - High Impact 2 - Medium Impact 3 - Low Impact

	Given the results obtained, how do you rate the limitations of the data on the potential uses of interest?	DK: don't know (e.g., unable to assess given the variables available)
3.11	**Consistency of time and domain** (e.g. geography) *Given the results obtained, how do you rate the limitations of the data on the potential uses of interest?*	1 - High Impact 2- Medium Impact 3 - Low Impact DK: don't know (e.g., unable to assess given the variables available)
3.12	**Methods for determining under and over coverage** have been designed.	1 - No 2 - In development 3 - Yes
3.13	**Under-coverage** levels, established by comparison with alternative sources, are adequate to meet the needs of the Census use-case.	1 - Unacceptable 2 - Moderate 3 - Good
3.14	**Over-coverage** levels, established by comparison with alternative sources, are adequate to meet the needs of the Census use-case.	1 - Unacceptable 2 - Moderate 3 - Good
3.15	**Selectivity/bias of variables** based on comparisons with alternative sources. *Given the results obtained on the selectivity, how do you rate the limitations of the data given Census use-case?*	1 - Low 2 - Medium 3 - High

Table 4 - Checklist for the assessment of ability to clarify data queries with supplier

4. Ability to clarify data queries with supplier	**Quality Indicator/Metric**	**Assessment Scale**
4.1	Mechanisms are in place for production teams to seek clarification from supplier post data transfer.	1 - No 2 - In development 3 - Yes