Department of Energy & Climate Change

National Energy Efficiency Data-Framework: Making data available

Consultation document

Publication date: 21 November 2013

Closing date for comments: 21 January 2014

Any enquiries regarding this publication should be sent to us at EnergyEfficiency.Stats@decc.gsi.gov.uk.

This publication is available from our website at www.gov.uk/decc.

Contents

1. Introduction

The Government's commitment to "Open Data" has been set out in the Open Data White Paper (June 2012)[1] and reinforced in its commitment to the G8 Open Data Charter (June 2013)[2].

The benefits of making more data available are extensive. Data can be used in innovative ways to undertake research or produce tools and products which would not be possible without wider access to these data. This has benefits for Government, business and individuals.

DECC (and its predecessor departments) has a long history of making data available. Use has been made of administrative data (both private sector and from other Government departments). Matching different data sources together has further increased utility of data with benefits for a range of users and the wider public. DECC is now proposing to make further progress with Open Data through the publication of an anonymised dataset[3] of data from the National Energy Efficiency Data-Framework (NEED).

NEED was set up by DECC to provide a better understanding of energy use and energy efficiency in domestic and non-domestic buildings in Great Britain. The data framework matches – at individual property level – gas and electricity consumption data with information on energy efficiency measures installed in homes. It also includes data about property attributes and household characteristics.

In summary, it is proposed that two anonymised NEED datasets will be released:

1) **Public use (or training) dataset**: Approximately 20,000 records including information on energy consumption, energy efficiency measures installed in properties and property attributes. This dataset would be made available to all.
2) **End user licence dataset**: Approximately four million records including more variables than the public use dataset. It would be published in a slightly more restricted format; all individuals would be required to agree to an end user licence before having access to the data.

These two datasets would be samples of domestic properties in England and Wales. Data would be anonymised to prevent any individual household or business being identified. The data would be published in a format that could not be used for targeting specific households. It is envisaged the data would primarily be used by researchers looking at how energy is used in households, including the impact of installing energy efficiency measures.

This consultation invites views on DECC's proposals to publish these data. Views are sought on the proposed content and suggested approach to implementation. Consultation questions are set out in Section 5 but more general comments may also be provided.

---

[1] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/78946/CM8353_acc.pdf.

[2] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/207772/Open_Data_Charter.pdf.

[3] Data relating to a specific individual or property where the identifiers have been removed to prevent identification of that individual or property (directly or indirectly).

2. General information

Purpose of consultation

This consultation invites views on DECC's proposals to publish anonymised datasets containing record level data from the National Energy Efficiency Data-Framework (NEED). The proposed datasets contain records for individual households. Data would be anonymised to prevent any individual household or business being identified.

Views are sought on the proposed content and the suggested approached to implementation. Consultation questions are set out in Section 5 but more general comments may also be provided.

Issued: 21 November 2013

Respond by: 21 January 2014

Coverage

NEED includes data for domestic and non-domestic buildings in Great Britain. However, the coverage of different data sources varies. It is proposed that the anonymised datasets cover domestic properties in England and Wales only. We may seek to expand coverage of the anonymised dataset to include domestic properties in Scotland in future. Work on non-domestic NEED remains at development stage and as such no plans to release anonymised non-domestic data are being proposed at present.

How to respond

Please email responses and enquiries to: energyefficiency.stats@decc.gsi.gov.uk

Or by post to:

NEED Consultation
Statistics Team
Department of Energy and Climate Change
Area 6B, 3 Whitehall Place
London, SW1A 2AW

Consultation Reference: 13D/298

When responding please say if you are an academic institution, Government organisation, business, individual or representative body.

Additional copies

This document can be accessed from: https://www.gov.uk/government/publications?departments%5B%5D=department-of-energy-climate-change&publication_filter_option=consultations.

You may make copies of this document without seeking permission.

Other versions of the document in Braille, large print or audio-cassette are available on request. This includes a Welsh version. Please contact us under the above details to request alternative versions.

Confidentiality and data protection

Information provided in response to this consultation, including personal information, may be subject to publication or disclosure in accordance with the access to information legislation

(primarily the Freedom of Information Act 2000, the Data Protection Act 1998 and the Environmental Information Regulations 2004).

If you want information that you provide to be treated as confidential please say so clearly in writing when you send your response to the consultation. It would be helpful if you could explain to us why you regard the information you have provided as confidential. If we receive a request for disclosure of the information we will take full account of your explanation, but we cannot give an assurance that confidentiality can be maintained in all circumstances. An automatic confidentiality disclaimer generated by your IT system will not, of itself, be regarded by us as a confidentiality request.

We will summarise all responses and place this summary on our website at https://www.gov.uk/government/collections/national-energy-efficiency-data-need-framework. This summary will include a list of names or organisations that responded but not people's personal names, addresses or other contact details.

Quality assurance

This consultation has been carried out in accordance with the Government's consultation principles which can be found here: https://www.gov.uk/government/publications/consultation-principles-guidance.

If you have any complaints about the consultation process (as opposed to comments about the issues which are the subject of the consultation) please address them to:

DECC Consultation Co-ordinator
3 Whitehall Place London
SW1A 2AW
Email: consultation.coordinator@decc.gsi.gov.uk

3. Background

The Government is committed to making more data available more widely, the Open Data White Paper (June 2012) states:

> *In the realm of anonymised data, the Government Open Data and transparency agenda is also keen to encourage the better use of Government data for analysis.*

This is reinforced through the Government's commitment to the G8 Open Data Charter (June 2013) which highlights the benefits of open data:

> *We have arrived at a tipping point, heralding a new era in which people can use open data to generate insights, ideas, and services to create a better world for all.*

The benefits of making more data available are extensive. Data can be used in innovative ways to undertake research or produce tools and products which would not be possible without wider access to these data. This has benefits for Government, business and individuals. As summarised in the UK 2013 Open Data Draft National Action Plan[4]: "Access to information allows people to work together more effectively, collaborating with each other, with policy-makers and with service providers to improve governance, public life and public services to make more informed decisions."

To date, DECC has made good progress in these areas. Use has been made of existing administrative data; both private sector and from other Government departments. Matching different data sources together has further increased utility of data with benefits for a range of users and the wider public. DECC is now planning to make further advances with open data through the publication of anonymised datasets of data from NEED.

Development of The Framework

The UK has collected and published energy consumption data within the Digest of UK Energy Statistics since 1948. This has been produced and published at a national level and is based on aggregate information from energy suppliers. These headline national data are still important, but UK energy policy has required more detailed data at local level to help deliver and monitor reductions in energy use and emissions. DECC was looking for a low cost solution to meeting its requirements and found that the consumption data obtained from the existing administrative systems of the energy companies would meet the requirements. Data at individual meter point – for all properties in Great Britain, around 30 million electricity meters and 25 million gas meters - was first obtained in 2004. Data are now collected on an annual basis under the Statistics of Trade Act. This data collection provides the data for gas and electricity consumption statistics down to lower level super output area. The value and innovation of the approach to producing local area data was recognised by the Royal Statistical Society (RSS) in 2010 when this work received the RSS award for excellence in official statistics.

Future utility of the meter point data was considered at an early stage. Working closely with the energy industry and other key energy efficiency stakeholders, plans were established for a future data architecture suitable for matching the consumption data with other data sources. The National Energy Efficiency Data-Framework (NEED) is the means by which this has been achieved. Gas and electricity consumption data are matched, at an individual property level, with information about energy efficiency measures installed at the property, property attributes and household characteristics.

---

[4] https://www.gov.uk/government/consultations/open-government-partnership-uk-draft-national-action-plan-2013/ogp-uk-2013-draft-national-action-plan-from-open-data-to-open-government#introduction.

The Framework was first announced in the Heat and Energy Saving Strategy in 2009 and was developed by DECC, with support from the Energy Saving Trust (EST) and gas and electricity suppliers, in order to assist DECC in its business plan priority to "save energy with the Green Deal and support vulnerable consumers". It forms a key element of DECC's evidence base supporting DECC to:

- develop, monitor and evaluate key policies;
- identify energy efficiency potential which sits outside the current policy framework;
- develop a greater understanding of the drivers of energy consumption; and
- gain a deeper understanding of the impacts of energy efficiency measures for households and businesses.

The Framework

At the core of NEED is AddressBase the national standard for all buildings and addresses in Great Britain[5]. Each record within AddressBase has a Unique Property Reference Number (UPRN). Through address matching, this UPRN is assigned to each record in each of the dataset which forms NEED. This provides a reference key to join related records across different datasets.

Data in NEED cover the domestic (or residential) and non-domestic (commercial/industrial) sectors across the whole of Great Britain. However, different sources of data have different coverage and therefore it is proposed that at this time the anonymised dataset covers domestic properties in England and Wales only.

Table 3.1 summarises the main sources of data proposed to be used in the anonymised dataset. Section 4 provides more detail on the variables from each of these sources considered for inclusion in the final dataset.

Table 3.1: Proposed data sources to be used in the anonymised NEED dataset

Data	Source	Basis on which data is obtained for NEED
Energy Consumption	Energy suppliers and Gemserv/Xoserve/ Independent Gas Transporters	Obtained by DECC under the Statistics of Trade Act
Information on energy efficiency measures installed through government schemes (including EEC, CERT and CESP[6])	Gas Safe and EST (EST obtained data for the Homes Energy Efficiency Database from energy suppliers, CORGI and Fensa).	Obtained by DECC through data sharing agreements.
Information on Property attributes from Energy Performance Certificates	Landmark/Department for Communities and Local Government (DCLG)	Obtained by DECC through agreement with DCLG and in accordance with the Energy Performance of Buildings (England & Wales) Regulations 2012

[5] http://www.ordnancesurvey.co.uk/business-and-government/products/addressbase-premium.html.

[6] https://www.gov.uk/browse/benefits/heating.

Data	Source	Basis on which data is obtained for NEED
Area characteristics e.g. Output Area Classification, Index of Multiple Deprivation	Office for National Statistics DCLG Welsh Assembly Government	Open Government licence

For more information on all the data which is used within NEED see Annex A of the June 2013 NEED publication[7]. This annex also includes more detail on how the framework was developed and how the analysis sample used for DECC's publications was created.

More information on how DECC is addressing privacy issues relating to NEED can be found in the Privacy Impact Assessment (July 2013)[8].

Uses of NEED

NEED has already supported a number of DECC policies, with important consequences. For example, The Green Deal. NEED has been used to understand the reduction in consumption (and resulting reduction in energy bills) for households installing energy efficiency measures. To date NEED has looked at savings from a number of measures, including cavity wall insulation, loft insulation, installation of condensing boilers and solid wall insulation. The estimates from NEED were used to inform "in use factors" for the Green Deal.

NEED has also had a smaller, but still significant, part to play in a range of other DECC policies, for example, the Renewable Heat Incentive and Fuel Poverty. Data on consumption by property attributes, including the distribution of households' consumption, has been used to help DECC understand the likely under or over payment if payments for the renewable heat incentive were to be based solely on property attributes available in NEED. It has also informed fuel poverty analysis so there is a better understanding of actual consumption for different types of properties and households and therefore a better understanding on how policy options will impact on different households. Having this information enables DECC to provide better value for money and understand better the impacts of policy options, for both DECC and consumers.

Externally NEED outputs have been used by a wide range of interested parties, including local authorities, academics and energy suppliers. Examples of uses include:

- Used by Energy UK - with support from DECC statisticians - to create a comparison tool (www.comparemyenergy.org.uk) allowing households to enter information about their property and compare their gas and electricity consumption to typical consumption for a house with the same attributes. This tool, and the underlying data available on the DECC website, has led to use by individuals to understand potential for energy efficiency within their homes and as evidence when querying energy bills with suppliers.
- Used by energy companies and academics to validate and inform their own research and estimates.

[7] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/209090/Annex_A_-_What_is_NEED.pdf.

[8] https://www.gov.uk/government/publications/national-energy-efficiency-data-framework-privacy-impact-assessment. .

- Used by the Committee on Climate Change to inform reports, including recommendations to Government.
- Used by energy suppliers to act as an independent trusted source demonstrating the benefit of installing energy efficiency measures.
- Used by local authorities to help with modelling housing stock and understanding impact of installing energy efficiency measures.

DECC will continue to publish the aggregate outputs which have already proved valuable, but believe publication of an anonymised sample of data from NEED would further increase utility of the data by facilitating analysis of record level data by external researchers.

Stakeholders, including representatives of Local Authorities and academics institutions, have made us aware of their desire to access record level NEED data. Some examples of potential uses which we have been made aware of include:

- Creating and validating models
- Providing a tool for students to better understand energy use and efficiency
- Imputation of missing values in household surveys
- Modelling housing stock
- Understanding more about fuel poverty

We have engaged with a range of organisations to gain an understanding of their priorities for the dataset[9]. This discussion has fed into the proposals set out in Section 4.

[9] See Annex C for a summary of discussion at the NEED stakeholder event in May 2013 and priorities provided by DECC's NEED Project Board in October 2013.

4. Proposals

Summary

DECC is proposing to publish two record level datasets based on data from the National Energy Efficiency Data-Framework (NEED). The proposed datasets would contain records for individual households and would be anonymised to prevent any individual household or business being identified.

In summary, the two proposed datasets are:

1) **Public use (or training) dataset**: Approximately 20,000 records including information on energy consumption, energy efficiency measures installed in properties and property attributes. This dataset would be made available to all.
2) **End user licence dataset**: Approximately four million records including more variables than the public use dataset. It would be published in a slightly more restricted format; all individuals would be required to agree to an end user licence before having access to the data.

It is clear from feedback that there are many different purposes for which potential users would like access to record level data from NEED and the range of these requirements could only be met by the publication of two different datasets as proposed.

Some users considered that a more basic dataset would meet their requirements and they would prefer to forgo the additional utility of a more detailed dataset in order to have access to the dataset with no restrictions. While other potential users were happy to comply with any level of access requirements, and the additional time this may take, if it enabled them to gain access to more detailed data.

There was also significant difference in which variables were considered higher priorities by different users. A number of users had no requirement for geographic information, while others felt the needs of their organisation would only be met with information at lower levels of geography (e.g. Local Authority). The additional detail on geographic location would increase the risk of disclosure of an individual property and therefore must be balanced with the other information available on the dataset and the way the dataset is released.

The publication of two datasets allows DECC to meet a wider range of users' needs. The smaller public use, or training, dataset made available to all can be used for learning and more basic analysis. It makes a valuable training dataset, as its smaller size makes it more useable by students. It will also follow a similar structure to the end user licence dataset, allowing work to be planned prior to full analyse on the end user licence dataset. Alongside this, the larger end user licence dataset will provide greater utility for those who require more detailed data and are willing to take the time to meet associated access requirements.

For both of the proposed datasets, the data are a sample drawn from a larger dataset containing data for individual properties. The ICO anonymisation code[10] states that "there is a clear legal authority that where an organisation converts personal data into an anonymised form and discloses it, this does not amount to a disclosure of personal data". Information on how the records will be anonymised, to avoid disclosure of any individual property or business, is set out later in this section and will be done with reference to relevant guidance from the Information Commissioner's Office (ICO)[11].

---

[10] http://www.ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Practical_application/anonymisation_code.ashx

[11] See footnote 10.

Proposals for the dataset cover data sources relating to consumption, energy efficiency and property attributes in England and Wales. Proposals also cover some local area data sources already publically available under the Open Government Licence. Due to differences with data sources and data agreements for Scotland, DECC has taken the decision to exclude data for Scotland from the scope of the work to avoid delaying publication of a dataset. It is hoped that data for Scotland can be included in a future version of the dataset.

The rest of this section provides more detail on the proposed datasets.

Variables

While the proposed anonymised dataset is primarily based on data currently used in NEED, it has not been possible to use all the data sources which form NEED. Property attribute data collected by the Valuation Office Agency (VOA) and data from Experian (covering household characteristics) cannot be used due to legal and contractual restrictions.

DECC has instead proposed using information from Energy Performance Certificates (EPCs) for information on property attributes. DCLG has made a specific regarding these data in its Open Data Strategy (to make necessary changes to the legislative framework to facilitate more suitable access to the data held on its registers from April 2014)[12]. Inclusion of the EPC data in the NEED anonymised dataset would build on this commitment.

Use of EPC data has a number of advantages compared to VOA data, including the potential to use relevant information which is not available from VOA, such as Energy Efficiency Band. However, EPC data coverage is not as comprehensive as VOA data and has not yet been used in NEED. There is therefore more uncertainty about the quality of the data, and outputs from the anonymised dataset may not match DECC's published totals produced using VOA data.

DECC has not yet identified an alternative for household characteristics data which is not subject to restrictions preventing its use in an anonymised dataset. DECC will continue to look at the potential to include these and additional variables in future, including the possibility of modelling some key variables (e.g. tenure) itself rather than relying on Experian data in future.

Table 4.1 sets out all the variables for consideration in the proposed datasets. As part of the consultation, users are asked to give views on which variables they would most like to see included in the final dataset and why. Final decisions on which variables are included will be based on these responses balanced against the potential risk of disclosure of an individual property or business as a result of a particular variable being included.

To help focus responses to the consultation, the final two columns of Table 4.1 show the current level of priority for each variable in the public use and end user licence datasets respectively, assigned by DECC and based on feedback from users to date:

1) Priority – Variables which DECC intends to include in the dataset.
2) Important - Variables which will be included in the dataset if not prevented due to potential disclosure of individual households.
3) Under consideration – Variables which could be included in the dataset if inclusion does not lead to additional risk of disclosure or if a strong case is made for its inclusion at the expense of a variable in one of the other categories.

All proposals are subject to disclosure checking. Testing of the final dataset will take place once it has been produced. If there is deemed to be an unacceptable risk of an individual property or business being identified as a result of publication of the dataset then changes may have to

[12] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/254495/131031_2013_DCLG_Open_Data_Strategy.pdf.

made. For example, the published dataset may include additional banding of variables, or one or more variables may have to be dropped.

Table 4.1: Proposed data in NEED anonymised dataset

Variable	Source	Description	Public Use dataset	End use licence dataset
Gas consumption 2005 - 2012	Energy suppliers and Xoserve/Independent Gas Transporters	Gas consumption data for each year from 2005 to 2012. Data are for gas year e.g. 2012 data cover year from 1 October 2011 to 30 September 2012. Data are weather corrected. Includes "valid" records, those with a consumption value between 100 kWh and 50,000 kWh (approximately 75 per cent of records).	Priority	Priority
Electricity consumption 2005 – 2012	Energy suppliers and Gemserv	Electricity consumption data for each year from 2005 to 2012. 2012 data cover year from 27 January 2012 to 26 January 2013. Data includes valid records i.e. data between 100 and 25,000 kWh.	Priority	Priority
Economy 7 flag 2012	Gemserv	A flag indicating whether a property has a profile 2 electricity meter (economy 7). This does not necessarily mean the household are on an economy 7 tariff. Flag would only be included for most recent year.	Under consideration	Important
Energy Efficiency Band	Energy Performance Certificate (Landmark/DCLG)	A measure of the overall energy efficiency of a home. The higher the rating the more energy efficient the home is. The variable shows the bands from A - G: A is best, G is worst.	Priority	Priority

Variable	Source	Description	Public Use dataset	End use licence dataset
Environmental Impact Band	Energy Performance Certificate (Landmark/DCLG)	A measure of a home's impact on the environment in terms of carbon dioxide (CO2) emissions. The higher the rating the less impact it has on the environment. This variable shows the band from A - G: A is best, G is worst.	Under consideration	Under consideration
Property Age	Energy Performance Certificate (Landmark/DCLG)	Banded year of construction.	Priority	Priority
Property Type	Energy Performance Certificate (Landmark/DCLG)	Describes whether property is a house (detached, semi etc), bungalow or flat. Combination of Built Form and Property Type variables from EPC.	Priority	Priority
Floor area band	Energy Performance Certificate (Landmark/DCLG)	Floor area in metres squared.	Priority	Priority
Main heating fuel	Energy Performance Certificate (Landmark/DCLG)	Main fuel used to heat the property, e.g. Gas/Electricity.	Important	Important
Mains Gas	Energy Performance Certificate (Landmark/DCLG)	Whether mains gas is available in the property.	Under consideration	Under consideration
Loft insulated thickness	Energy Performance Certificate (Landmark/DCLG)	Banded loft insulation depth in mm.	No	Under consideration
Wall construction	Energy Performance Certificate (Landmark/DCLG)	Predominant wall construction e.g. sandstone, cavity, system built.	Under consideration	Under consideration

Variable	Source	Description	Public Use dataset	End use licence dataset
Cavity wall insulation installed	EEC/CERT/CESP (via HEED) & Warm Front	Yes/No stating whether cavity wall has been installed through a Government scheme.	Important	Important
Cavity wall insulation year	EEC/CERT/CESP (via HEED) & Warm Front	Year cavity wall insulation installed.	Important	Important
Loft insulation installed	EEC/CERT/CESP (via HEED) & Warm Front	Yes/No stating whether loft insulation has been installed through a Government scheme.	Important	Important
Loft insulation install year	EEC/CERT/CESP (via HEED) & Warm Front	Year loft insulation installed.	Important	Important
Solid wall insulation installed	EEC/CERT/CESP (via HEED) & Warm Front	Yes/No stating whether solid wall has been installed through a Government scheme.	Important – if sufficient records to avoid disclosure	Important – if sufficient records to avoid disclosure
Solid wall install year	EEC/CERT/CESP (via HEED) & Warm Front	Year solid wall insulation installed.	Important – if sufficient records to avoid disclosure	Important – if sufficient records to avoid disclosure
New boiler	Gas Safe/CORGI	Whether a new boiler has been installed – data available for boilers installed since April 2009 (data prior to April 2009 would be treated in the same way if they can be included).	Important	Important
New boiler install year	Gas Safe/CORGI	Year of installation of boiler– data available for boilers installed since April 2009 (data prior to April 2009 would be treated in the same way if they can be included).	Important	Important

Variable	Source	Description	Public Use dataset	End use licence dataset
Region	National Statistics Postcode Look-up	Nine regions in England, and Wales.	Under consideration	Under consideration
Local Authority	National Statistics Postcode Look-up	District, borough and city councils, Unitary Authorities and London and metropolitan boroughs.	No	Under consideration
Index of multiple deprivation	ONS/DCLG/WAG	The indices of deprivation identify the most deprived areas across the country. They combine a number of indicators, chosen to cover a range of economic, social and housing issues, into a single deprivation score for each small area. Dataset would include, English index of multiple deprivation 2010 and Welsh index of multiple deprivation 2011.	Under consideration	Important
Output Area Classification (OAC)	ONS	This classification groups output areas into clusters based on similar characteristics e.g. blue collar communities, city living etc.	Under consideration	Under consideration
Fuel poverty indicator	DECC local area fuel poverty estimates	England Only. Percentage of households in a lower level super output area which are considered to be in fuel poverty. Based on the low income high cost definition of fuel poverty.	Under consideration	Under consideration
Weighting variable	DECC	A variable providing values to weight the dataset to national total. Weighting based on information about property attributes (not occupants).	Under consideration	Under consideration

Public use/training dataset

Creation

The public use dataset would be made up of approximately 20,000 records. It will be selected from the population of households with an EPC which can be matched to consumption data ("matched EPC dataset"). This number of records is proposed because it allows a good distribution of records with respect to geography along with other variables relating to the property.

The structure of the dataset will be similar to the structure of the end user licence dataset to allow researchers to develop analytical strategies and understand data prior to accessing the larger end user licence dataset. It is also anticipated this dataset will allow students to interrogate and understand data relating to energy efficiency and energy consumption without needing to go through a registration process.

The dataset will be a random stratified sample of the matched EPC data, selected to be representative of the England and Wales housing stock. The stratification may be based on variables which are not published in the final dataset (e.g. if region is not included).

Anonymisation

The level of statistical disclosure will be assessed in a similar way for the public use dataset and the end user licence dataset. However it is anticipated it will be more straightforward for the public use dataset as the small number of records and high level of geography will offer protection against disclosure.

In summary, the following approach will be taken to prevent disclosure of an individual household or business.

A number of "visible" variables will be selected; these are variables which an intruder[13] would be most likely to use to identify a property in the data, such as property age and property type. If a property can be identified through these variables then non visible more sensitive information could be discovered.

Where the combination of these visible variables selected appears at least three times in the population (not necessarily in the sample itself), it is impossible for the specific property to be identified and therefore there is no potential for disclosure. In any instances where this is not the case other actions to prevent disclosure will be taken as required, including; record swapping; suppressing records or variables; and recoding variables.

This approach will ensure the risk of disclosure is minimised while maintaining utility of the data.

Release

This dataset would be available to all. It would be published as a CSV file available to download from the NEED webpages of the Government website. An associated meta-data file will also be made available. DECC will request users provide information on how the data are being used, but there will be no requirement to do so.

---

[13] Refers to a group or individual who wishes to identify people in the data or attributes relating to these people. Also known as an attacker they may or may not have malicious intent.

End user licence dataset

Creation

DECC's intention is to produce a dataset of approximately four million records for use by external researchers. As with the public use dataset, it will be selected from the population of households with an EPC which can be matched to consumption data ("matched EPC dataset"). A dataset of four million records is proposed as it will be sufficient to allow results to be produced for detailed breakdowns. This is the same size as the dataset used by DECC for the majority of analysis relating to domestic properties in NEED. A file this size will also help with data protection and processing speed.

The dataset will be a random stratified sample of the full dataset of consumption data matched to EPC data, selected to be representative of the England and Wales housing stock as far as possible. The stratification will be based on region, consumption band in 2012 and physical property attributes. Testing will be done to confirm that the sample selected provides representative figures for other variables, such as total and average electricity and gas consumption and number of properties with insulation measures installed.

Anonymisation

The same broad approach set out in the public use file will also be applied for the end user licence dataset; a number of key variables will be selected and disclosure control will be applied to these visible variables to ensure there is no potential disclosure of non-visible variables. However, there will be a slightly higher risk of disclosure with this dataset in order to increase utility. It is proposed that this additional risk will be tolerated because of the requirement for individuals to agree to the licence before gaining access to the dataset. All users wanting access to this dataset will have to agree to conditions of use, including:

- The requirement to notify the data service provider of any breach of the End User Licence;
- To use data only for purposes of not for profit research, teaching or personal educational development;
- To preserve the confidentiality of information pertaining to individuals and/or households where the information is not in the public domain;
- Not to use the data to attempt to obtain or derive information relating specifically to an identifiable individual or household, and not to claim to have obtained or derived such information;

In testing for disclosure, a number of things will be considered:

- Look for possible disclosure problems in tabulations of key variables – variables which an intruder would be most likely to use to try and identify individuals – such as low counts or rows/columns with many zeros.
- An assessment of variables which are already in the public domain – these may not require the same level of disclosure control as those which are not in the public domain, but will need to be carefully considered to ensure they don't lead to identification of properties and therefore lead to disclosure of other variables;
- The potential for identifying a geographic area by combining LSOA or geography variables (e.g. Local Authority, Index of Multiple Deprivation, Output Area Classification and Fuel Poverty Per Cent) and implications on identification of a household.

- Intruder testing – can properties be correctly identified by 'expert' intruders;
- How sensitive any potential disclosure is; DECC considers electricity and gas consumption to be the most sensitive of the variables proposed for inclusion and accordingly will take particular care with this variable.

As with the public use file, where there is deemed to be an unacceptable risk of disclosure the following will be considered:

- Record swapping – performed on regions;
- Record suppression;
- Changes to banding/recoding;

Efforts will be made to damage or change the data as little as possible with applying the above techniques.

Release

It is proposed that any individual wanting access to this dataset will be required to sign up to an End User Licence. The UK Data Archive[14] is being considered as the most likely option and is consistent with the way record level data from the English Housing Survey are released. This would mean all users are required to agree to the terms and conditions of use of data as described in the End User Licence (EUL): http://data-archive.ac.uk/media/381244/ukda137-enduserlicence.pdf.

The conditions of the End User Licence do not allow organisations to use the data for commercial purposes without obtaining permission. DECC would allow commercial organisations which assist in the delivery of Government policies to use the data under strict conditions. Any commercial organisations requesting the data on this basis would have to set up a specific agreement with DECC which sets out the requirements on it, including:

- To preserve the confidentiality of information pertaining to individuals and/or households where the information is not in the public domain;
- Not to use the data to attempt to obtain or derive information relating specifically to an identifiable individual or household, and not to claim to have obtained or derived such information;
- Not to use the data to attempt to obtain or derive information on a household's energy supplier
- Not to use the data to attempt to target individual households with measures.

Alternative vehicles for release of data under consideration include controlled access to data through the ONS virtual microdata laboratory[15].

[14] http://www.data-archive.ac.uk/.

[15] http://www.ons.gov.uk/ons/about-ons/business-transparency/freedom-of-information/what-can-i-request/approved-researcher-accreditation.html.

5. Consultation questions

DECC welcomes views on the following consultation questions. Please provide reasons for your responses wherever possible.

Consultation Question	
1.	**Do you agree DECC should release anonymised NEED data?**
Consultation Question	
2.	**Do you agree with the proposed approach to publishing two separate dataset for different purposes?**
Consultation Question	
3.	**In relation to i) the public use dataset and ii) the end user license dataset, what are your priorities for variables in the dataset?** **a) Do you agree with the priority variables set out in Table 4.1? If not, which of the variables listed do you consider to be priorities?** **b) Do you agree with the variables assigned as important in Table 4.1? If not, which of the other variables listed do you consider to be important?** **c) Do you agree that those variables listed as "under consideration" are less important than the variables listed as priority or important?** **d) Are there any variables included in the proposals which you think should not be included?** **e) Do you agree that inclusion of a lower level geography identifier is less important than a wider range of variables?** **f) Which lower level super output area data is most useful? Index of multiple deprivation, output area classification or percentage of households in fuel poverty?** **g) Would a weighting variable be useful?**
Consultation Question	
4.	**Proposed bandings for variables in the dataset are set out in Annex B. Do you agree with these proposals in relation to i) the public use dataset and ii) the end user license dataset? Please bear in mind that greater granularity of data will reduce the number of variables that can be included in the final dataset.** **a) Annex B sets out options for banding variables please let us know which you would prefer for each variable of interest to you.** **b) Are there any variables that can be banded further than proposed without significant loss of utility?** **c) Are there any variables which would no long be useable for analysis if the proposed banding – or one of the proposed options - is applied?** **d) For variables such as consumption and floor area, is it preferable to have bands of the same size (which may have to be larger) or more**

	detail in the centre of the distribution with larger bands at the extremes?
Consultation Question	
5.	**Do you agree with the proposed approach to anonymisation for** **i. The public use dataset; and** **ii. The end user licence dataset?**
Consultation Question	
6.	**Do you agree with the proposed approach to publication and access?** **i. Do you agree with the proposal for a smaller publically available dataset?** **ii. Do you agree with the proposed restrictions on access to a more extensive dataset?**
Consultation Question	
7.	**If you are a potential user, please tell us how you think you would use these data.**
Consultation Question	
8.	**Do you have any other comments on the proposals?**

If any individual would like to be added to the NEED stakeholder distribution list to receive updates about future publications and progress with the anonymised dataset then please let us know at energyefficiency.stats@decc.gsi.gov.uk or when you submit your response.

6. Next steps

We will issue a response to this consultation. Subject to the outcome of the consultation, we plan to publish the public use and end user licence datasets in spring 2014, following the implementation of changes by DCLG to the legislative framework to facilitate greater access to EPC data.

This proposed dataset will be the first publication of data from NEED at record level. If this publication proves useful, in future we will look to expand on the data available, increasing coverage to Scotland and including variables relating to household characteristics; two areas which are not covered in the proposals set out in this consultation.

Annex A: Glossary

Open data is information that is available for anyone to use, for any purpose, at no cost.

An **Anonymised dataset** is a dataset in which direct identifiers have been removed. Further protection may be required if indirect identifiers are present in the data.

A **Public use dataset** is typically record level data which can be accessed by any individual, with no restrictions on use. It will not contain personal data. It may be of more use as a training tool than for researchers.

An **End user licence dataset** will have more detail than one for Public Use. Users will have to sign an agreement with one of the requests being that no attempt will be made to use the data to identify any individual, household or organisations.

Anonymisation involves removing the direct identifiers from a microdata record. This term is used frequently when microdata are being protected. Direct Identifiers are variables which will enable a property to be identified with a high degree of confidence such as address.

Microdata are individual level data, for example data about individual people or households.

Indirect Identifiers are variables in a dataset that assist with identification of a household or property without directly referring to them. For example combinations relating to a property in a table could allow an individual to be identified with a great degree of confidence.

Direct Identifers are variables in a dataset will help an intruder easily identify an individual. These include Property Reference Number.

Disclosure Control refers to a number of techniques which can be applied to the data to limit disclosure risk. The most common techniques include recoding, suppression and rounding.

Disclosure risk occurs if an individual's confidential information can be ascertained either exactly or to within a defined narrow bound by an intruder with a high level of confidence. This risk can be mitigated by applying disclosure control.

Intruder refers to a group or individual who wishes to identify people in the table or attributes relating to these people. Also known as an attacker they may or may not have malicious intent.

Granularity is the level of detail provided in the data. High granularity refers to record level data or similar. Low granularity would be aggregated or summarised data.

Key variable is a variable which is commonly used in tabulations. If a large number of tables are produced they are likely to be linked via one or more key variable. By combining these tables an intruder may be able to identify an individual or associated attributes.

Visible variable is a variable which enables identification of an individual or other statistical unit by placing them in a certain category for particular key variables.

Lower level super output area is a geographic area made up of a number of output areas. Super output areas were designed to improve the reporting of small area statistics. Each LSOA contains between 400 and 1,200 households. There are 32,844 lower level super output areas in England and 1,909 in Wales.

The Energy Efficiency Commitment (EEC) set targets on energy suppliers to achieve improvements in energy efficiency by providing energy efficiency measures to households across Great Britain. The first scheme (EEC1) ran from 2002 to 2005 and the second (EEC2) ran from 2005 to 2008. EEC2 had a

requirement for at least 50 per cent of the target to be met in relation to Priority Group consumers, defined as those in receipt of certain income-related benefits and tax credits.

The Carbon Emissions Reduction Target (CERT) ran between 1 April 2008 and 31 December 2012 and followed EEC. It required all domestic energy suppliers with a customer base in excess of 250,000 customers (increased from 50,000 at the end of December 2011) to make savings in the amount of carbon dioxide emitted by households in England, Scotland and Wales.

The Community Energy Saving Programme (CESP) targeted households across Great Britain, in areas of low income, to improve energy efficiency standards, and reduce fuel bills. There were 4,500 areas eligible for CESP. Like CERT, CESP was funded by an obligation on energy suppliers and electricity generators.

Annex B: Detailed variable proposals

See Excel Spreadsheet published alongside this consultation: NEED anonymised dataset variable proposals including banding.xlsx.

Annex C: Summary of user input

C.1 Stakeholder Event Summary May 2013

Presentations from the NEED Stakeholder event in May 2013, along with the note reproduced below, are available on the DECC pages of the Government website at: https://www.gov.uk/government/statistical-data-sets/national-energy-efficiency-data-need-update.

Overall Comments

Feedback on NEED was positive with a particular interest in the progress being made with non-domestic NEED.

Users felt an anonymised dataset of any description would be beneficial, but there are also a lot of cases where the aggregate outputs already published - or planned - can be used without the need for the anonymised dataset, such as validating models, or understanding distributions. Not all users were aware of some of the tools and data tables that could help with these endeavours.

Consumption by multiple attributes (unformatted): https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/65986/6950-gas-and-electricity-consumption-multiple-attribute.xls.

Table Creator: https://www.gov.uk/government/statistical-data-sets/need-table-creator

Other data tables including headline consumption and distributions: https://www.gov.uk/government/publications/national-energy-efficiency-data-framework-need-report-summary-of-analysis-2013-part-1

Main uses

Attendees came from a range of spheres and had varied plans for how an anonymised dataset would be helpful. This included work with an interest in behaviours, physical properties of the housing stock and the ability to target energy efficiency measures. Specifics included:

- Creating and validating models
- Providing a tool for students to better understand energy use and efficiency
- Imputation of missing values in household surveys
- Modelling housing stock
- Fuel poverty impacts

There were also academic users who had resource for analysis and where open to being guided by Government priorities for policy development.

Variables

In general users wanted all possible variables. Specifically:

- Consumption
- Local authority
- Property type
- Property age
- Number of bedrooms
- Floor Area
- Tenure
- Number of occupants

- Measures installed

In addition to those currently held within NEED, users were interested in:

- Wall Type
- EPC band
- Heating fuel/Heating system
- Store height
- Fuel poverty flag
- Stay at home flag (e.g. indicator of those most likely to be at home during the day, pensioners, people who work from home, mothers with young children)
- ESRI mapping information on building direction
- Additional EHS variables
- a weighting variable - to understand the proportion of the population with particular characteristics

It was also felt that it would be helpful to provide totals so users can develop scaling factors. Information on uncertainty would also be valuable.

Banding/level of geography

Attendees where sympathetic to the need to have some banding in order to reduce the likelihood of users being able to identify a specific household within the data. However, there were differing views on the best approach.

Geography - Some users were happy with region as the most detailed geographic identifier, while other felt the dataset would not be able to meet their requirements unless data were available at lower level super output area (LSOA).

Number of occupants – banded by 1,2,3,4, 5 or more.

Consumption data – some users felt distribution (e.g. percentiles, perhaps deciles) was more useful than consumption bands. Another suggested approach was to have 100kWh bands for consumption in the centre of the distribution with larger bands at the extremes.

As a result of the divergent views of attendees each table reached the same suggestion that it would be best to produce two different datasets. One at a detailed geography with fewer variables and one with region as the most detailed geography and more detail in other variables. For some users it was actually post code level consumption data that would be most useful rather than the range of other data which is available through NEED.

Sample size

Views varied from a sample the same size as the English Housing Survey (about 16,000 for England) being sufficient to a request for all households in Great Britain. This depended on intended use.

Access

It was suggested that two different datasets could be made available with different levels of access; a fully open dataset which could be made available to all users and a more detailed dataset which is only available to registered users. It was felt that some users would be happy to comply with any access requirements in order to get hold of detailed data, while others would rather have less detailed data if it meant there were fewer requirements.

Users felt that there is still value in a dataset which cannot be linked to other sources, but that there would be even more advantages if data could be linked.

C.2 NEED Project Board October 2013

Extract from the minutes of the NEED Project Board on 14 October 2013

The meeting included a discussion on the proposed publication of an anonymised dataset. It was set out that the priority variables would be electricity and gas consumption for the latest year, energy efficiency measure installed (e.g. loft insulation) and year of installation and property attributes (property type, property age and floor area band). Attendees were then asked to mark their priorities for other variables with the following results:

Variable	**Number of votes**
Energy Efficiency Band	XXXXX
Gas consumption 05-10	XXX
Wall insulation	XXX
Index of multiple deprivation	XXX
Electricity consumption 05-10	XX
Output area classification	XX
Census outputs e.g. employment/income	XX
Main heating fuel	X
Economy 7 Flag	X
Urban/Rural flag	X
Loft insulated	
Double glazed	

Decisions on which of these variables are included in the final dataset will be based on these priorities in combination with disclosure considerations and feedback from the stakeholder event in May 2013.

It was also stated that banding of consumption may make analysis difficult so it would be better to have rounded figures if possible.