Opening up government data while maintaining data privacy

Caroline Tudor
Office for National Statistics
Segensworth Road
Titchfield, PO15 5RR, UK
+441329444730
caroline.tudor@ons.gov.uk

Philip Lowthian
Office for National Statistics
Segensworth Road
Titchfield, PO15 5RR, UK
+442075928640
philip.lowthian@ons.gov.uk

Keith Spicer
Office for National Statistics
Segensworth Road
Titchfield, PO15 5RR, UK
+441329444983
keith.spicer@ons.gov.uk

ABSTRACT
In this paper, we describe a UK approach to opening up microdata collected by government with examples of actual use-cases of anonymising datasets. We describe briefly the reasoning behind the Open Data movement and the challenges faced in trying to release data openly in practice. Several case studies are provided including that of the Department of Energy and Climate Change (DECC) public use file, and the microdata teaching file from the 2011 UK Census. The anonymisation approach mainly involves detecting quasi-identifier attributes in the data and then modifying the dataset to ensure relative anonymity based on those attributes. This approach is aligned with the principles of k-anonymity. It also involves intruder testing to simulate linking attacks, whereby friendly intruders attempt to attack the dataset and find vulnerabilities to further inform disclosure risk assessment.

Categories and Subject Descriptors
J.1 [**Computer Applications**]: Administrative Data Processing - Government , K.4 [**Computers and Society**]: Public Policy Issues – Privacy.

General Terms
Security

Keywords
Disclosure risk, open data, government data, intruder testing, k-anonymity, linking attack.

1. INTRODUCTION
This paper examines one approach the UK Office for National Statistics (ONS) recommends for opening up government record level data while maintaining data anonymity. Section 2 provides some background to the *Open Data* movement and how this has largely been enabled by technological advances allowing data to be processed and shared far more easily. We also describe what is meant by open data. Section 3 examines the impact of privacy attacks within an open data framework with reference to the *jigsaw (mosaic) effect* perhaps known more widely in the privacy literature as a *linking attack*. In section 4 we put this into the context of government data and set out the value that open datasets have as well as their limitations.

We usefully summarize a general approach to opening up micro-datasets. The approach broadly operates on the principle of *k-anonymity* such that information for any person is hidden amongst k-1 other individuals. Sampling as well as suppression and recoding of variable categories are used to achieve this. We describe *intruder testing* which is used to help inform risk assessment and appropriate selection of *quasi-identifiers*. Our overall goal is to try to achieve k-anonymity with a small k value (e.g. k = 2) to at least remove all uniques, and a weak k-anonymity with a larger k value (e.g. k = 3 or more) for subsets of records which are more vulnerable to attack. This procedure works well to satisfy real-world requirements for a balance of risk-utility. Section 5 describes application of our anonymisation approach to examples of open government datasets and the steps taken to minimize disclosure risk. Section 6 concludes with a short discussion including a summary of the challenges ONS have faced in creating useful open data.

2. OPEN DATA CONTEXT
Web-based technology has allowed increasing numbers of people to share and link data. Information disclosure can now be in digital form: downloadable from the internet and easily processible by computer. In 2008, the Open Data movement undertook to make more data public and accessible, particularly data collected by government, with the argument that this information collected on our behalf should be made freely available to hold government to account. Open data as a concept simply encompasses data that are made available by organizations, businesses and individuals for anyone to access, use and share[1] no matter where they are and what they want to do with the data. Advocates of the open data movement espouse innovative combining of datasets leading to improved citizen engagement and empowerment, being a driver of economic growth and leading to better delivery and efficiency of services. However the UK government, along with other participating countries, faces a number of challenges in order to transition towards open data, one of which is to reconcile the right to information with the right to privacy. While open data must be data that do not relate to an identified or identifiable data subject, achieving this in practice is difficult due to the linking attacks, also known as the jigsaw effect of comparing multiple datasets to eventually reveal disclosive information about one or more identifiable individuals. Two or more datasets each posing negligible disclosure risk in isolation, present an increased risk when the information from these datasets is pieced together in some way.

[1] http://theodi.org/guides/what-open-data

Traditionally UK government release many sensitive datasets at record level under licence; either End User Licence (EUL), Special Licence (SL) or within a safe setting. These have conditions attached which include signing up to a set of conditions on use of the data (EUL) to registration of the user and detailing the purpose of use (SL). There has more recently been a greater emphasis on releasing data with minimum restrictions for the user. As part of the government's commitment to the open data agenda, the UK National Archives developed the Open Government Licence (OGL) which enables and encourages free use of government information. The user is allowed to publish, adapt and combine with other data as long as the information is not *personal data*. Personal data means data relating to a living individual who is or can be identified either from the data or from the data in conjunction with other information that is in, or is likely to come into, the possession of the data controller (UK Data Protection Act, 1988). It is therefore a difficult balancing act between producing open data which are of some use and protected to a reasonable level even when combined with other data sources.

In this paper we discuss some real-world examples of how some government datasets have been made open datasets and thus made available publicly, taking account of privacy considerations, and the resulting limitations on such datasets. We discuss publicly released datasets from the department of Business, Innovation and Skills (BIS), Department for Energy and Climate Change (DECC), the 2011 Census microdata, and also take a look at the licensed survey datasets within the Office for National Statistics and how they were assessed for potential open release.

3. ASSESSING RISK OF LINKING ATTACKS ON OPEN DATA

In order for data to be released under OGL, it is first necessary to reduce the risk of identification. All explicit identifiers such as name and date of birth should be removed from the dataset. Once data have been de-identified so that it is no longer possible to establish links to particular individuals, data may be considered for release openly and used for a wide range of purposes. However there may be a small residual risk that identifiable data could be revealed. Sets of attributes may still be linked with external data to uniquely identify individuals in the population and are called *quasi-identifiers* as defined in [1].

The risk of a linking attack becomes more likely when many similar data are available, to a large number of people. As set out in [2] the risk of jigsaw identification/linking attack with the inclusion of anonymised databases in a transparency programme increases due to three reasons (adapted here in summary):

1. The very concept of open data precludes the possibility of withdrawing access to data if need be.
2. The amount of data on the web grows annually thanks to information on social networking sites and local press coverage.
3. Jigsaw identification is computationally complex. However dramatic increases in computer power have made this easier and complete future-proofing against such disclosure is almost impossible.

Crucially, the responsibility for preventing linking attacks lies with the releasing agency. According to [2], this depends on the nature of the information, the availability of other information, and the technology in place that could facilitate the process of identification. Determining the level of acceptable risk in open data according to these factors is therefore complex.

Privacy risk is regulated at the European level by an EU Directive[2] which states that to determine whether a person is identifiable, account should be taken of all the means likely to be reasonably used either by the controller or by any person to identify the said person. In the UK, the Information Commissioner's Office (ICO) - which is responsible for enforcement of the Data Protection Act (DPA) - released in 2012 its "Anonymisation: managing data practice protection risk code of practice"[3] online. This details how to release anonymised data with the caution that publication under an open government licence is a release to the wider world and carries more risk. The stance from the ICO Anonymisation Code of Practice[4] is for data providers to assess whether it is reasonably likely that an individual can be identified from the data and to consider what other data are available and how and why the data could be linked. It suggests that data providers should establish an auditable process for ensuring an adequate level of anonymisation. One particular assessment that the Code of Practice advocates is a test of whether an intruder might be able to achieve re-identification. This would be done by way of a 'motivated intruder test' as part of a risk assessment. In section 5 we describe how such a practical test provides useful additional information to support risk assessments of datasets, particularly in reference to linking attacks.

4. AN ANONYMISATION APPROACH TO OPENING UP GOVERNMENT DATA

In a government context, a primary purpose of open datasets is for teaching or as training datasets. These allow code to be tested and checked before using it on a more complete dataset released under end-user or special licence. These datasets permit researchers to get a feel for what the data may be like and to allow preliminary hypotheses to be formed. At present, most open government datasets are generally not suitable for research projects other than making initial speculations and for some simple tabulations.

Our anonymisation approach essentially implements variations of the k-anonymity principle as a way of cutting down the detail into a much reduced open dataset. The concept of k-anonymity was first introduced by [3] as a way of preserving privacy. In essence a release of data is said to have the k-anonymity property if the information for each person contained in the release is hidden amongst k-1 other individuals. In practice this means that any quasi-identifier in the released table must appear in at least k records. So if the quasi-identifiers are age and sex, then it will ensure that there are at least k records with 30-year old females for example. This property can be achieved by generalization and suppression. Generalization refers to publishing more general values which can be done by recategorizing age into bands for

[2] http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf

[3] https://ico.org.uk/for_organisations/guidance_index/~/media/documents/library/Data_Protection/Practical_application/anonymisation-codev2.pdf

[4] https://ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation

example. Suppression can be done by removing outliers or by providing only a sample of records. Sampling has the benefit that an intruder can never be sure whether a person is contained within the dataset.

In the context of real-world data, using algorithms to find optimal anonymous tables can be unpractical and even for optimal solutions, the distortion of the data can be too high leading to un-useful tables ([4]). The balance of risk-utility is absolutely crucial in practice for government data releases. [4] discusses an alternative known as *weak k-anonymity* which requires it to be enforced in just a subset of the records. This of course means that there is a possibility that those variables which are not controlled via k-anonymity can be used to identify someone. However this risk is generally small when interest in these variables is low and so is typically a more practical option. Our approach aims for k-anonymity enforced on the entire dataset with a small k value e.g. k=2 to remove all uniques as a minimum; and a *weak* k-anonymity (on a subset of records) with a larger k e.g. k = 3 or more, depending on the dataset and its particular vulnerabilities.

K-anonymity involves consideration of sets of attributes that can be linked with external information to re-identify the respondents to whom the information refers. As in [5] a data release is said to satisfy k-anonymity if every combination of values of quasi-identifiers can be indistinctly matched to at least k individuals. This information can be known only by linking the released data with externally available data. We make use of intruder testing to help refine the appropriate set of quasi-identifiers.

The general procedure that ONS advise here for creating open datasets is outlined as follows:

i. Assess dataset background
ii. Choose the key variables that might be used for identification (key variables described in [6])
iii. Consider how a dataset might be disclosive under intruder scenarios (see [7])
iv. Analyse variable combinations / quasi-identifiers ['uniques' analysis or similar]
v. Carry out a formal intruder test and refine steps above
vi. Generalize (recode variables) and/or suppress (reduce dataset/sample) to make suitable for public release to achieve properties of k-anonymity

(i) Dataset Background

Disclosure risk assessment should commence by talking to potential users of the data to understand the types of research that the data will be used for. This step is sometimes ignored, to the detriment of the final dataset. Main considerations might be the variables in which they are most interested and the level of detail that is needed, and particularly the required level of geography. It is also important to understand whether the original dataset is a survey sample, an administrative dataset or a census. General consideration must be afforded to any existing protection in the dataset due to an intruder's uncertainty as to whether an individual is actually present in the data.

(ii) Assess disclosure under intruder scenarios

Having done this preparatory knowledge gathering, the next stage is to consider how a dataset could be shown to be disclosive by looking at a number of intruder scenarios. An intruder (or attacker) is somebody who attempts to discover personal information about an individual, household or business in the dataset. This is most likely to occur if the intruder has some initial knowledge about a particular member of the dataset with respect to a number of variables known as key. For example an individual in the data could be a relative, neighbour or work colleague These intruder scenarios include combining the dataset with other data sources.

(iii) Analyse variable combinations

Based on the intruder scenarios that fit with a particular dataset, the procedure is then to select a set of key variables (five or six for each dataset) to form an *identification key* and tabulate combinations (also referred to as quasi-identifiers in the privacy literature) to create a series of two, three and four dimension tables. These combinations should be plausible i.e. likely to be similar to tables required by researchers.

In general, knowledge of the data should lead to a suitable range of combinations being selected. It should be noted that creating tables with a large number of variables will be counterproductive as patterns may emerge that would not be noticed by a researcher. Most records are unique if a large number of variables are combined. Instead, we should consider just a limited set of variables, the values of which are what an intruder is likely to know. These are the aforementioned identification key, the variables likely to be used by an intruder to identify the individual and then discover the rest of the information in the remaining variables.

Variable combinations which are rare or unique will indicate potential disclosure issues and variables will need to be recoded or excluded if required. This approach of looking for rare combinations of variables is similar to the more formal k-anonymity method described in [3] and [8].

(iv) Carry out a formal intruder test

This involves using "friendly intruders" to try and see if they can re-identify anyone in the dataset. These friendly intruders should have some background knowledge of the data (as a data user might) but should not be specialist hackers using advanced techniques. This is in consideration of the phrase "means likely and reasonably" as referenced in the EU directive and UK Data Protection Act. The intruder motives would not be malicious. They would not release their findings into the public domain but would feed back their finding to aid in the publication of a secure dataset. One of the main purposes of such a test is to try and capture what other information may be linked to the dataset by the intruder to attempt disclosure. Thus appropriate selection of intruders in terms of awareness of similar data sources and good penetration skills (able to search and analyze the data) are important to get accurate results. The information resulting from the intruder test may be used to refine the previous steps, particularly with regards to which variable combinations are considered to be quasi-identifiers and therefore utilized in the identification key. For a more detailed discussion of intruder testing, please see [9].

(v) Generalize and Suppress

The last stage of the process involves taking steps to minimize overall risk in the dataset. Our approach is to ensure that k-anonymity is achieved to a low k value, i.e. at least k =2 to ensure no uniques (or k = 3 to eliminate pairs) in the data. This is usually achieved by sampling. The next stage is to recode variable categories to reduce detail (generalization) so that weak k-

anonymity is achieved to a higher k value for a subset of the data where there are particular vulnerabilities.

5. EXAMPLES OF MAKING DATA OPEN

In this section we consider how the general procedure outlined in section 4 is applied in practice to three examples of open datasets produced by the UK government. One of these is a sample of microdata from the UK 2011 Census collected by the Office for National Statistics (ONS) while the second is a sample from an administrative dataset produced by the Department for Energy and Climate Change (DECC) on domestic gas and electricity consumption. The steps followed in order to produce these datasets are shown below. There are many similarities in producing these datasets but some important differences. There were fewer variables that could easily be recoded in the DECC data leading to less flexibility in reshaping the data. The DECC data is typical of a lot of datasets which contain a lot of information but not a lot of variables which can be recoded in a straightforward way. Most variables are dataset specific and any recategorization would reduce the utility significantly. An ongoing project to publish education data held by the Department for Business, Innovation and Skills (BIS) is also discussed briefly as our third example.

In all three cases, statistical disclosure control has to be applied to ensure that sufficient protection is given to avoid an individual, household, business or other statistical unit being re-identified. As detailed in section 4, data might be recoded and/or only a limited number of variables released. For a more general discussion of statistical disclosure control techniques that might be applied during this process, the reader is referred to [10].

5.1 2011 UK Census microdata

The 2011 UK Census is a rich data source with many published tabular outputs available from the ONS website. In addition to these tables, record level data are being made available to researchers; a teaching dataset at individual level was published in 2014. The data can be accessed through the link below. Note that more detailed datasets will shortly be available under more prohibitive licensing and access conditions. This dataset is a random 1% sample of records for England and Wales published to encourage a wider use of census data and as an introduction to these more detailed datasets.

http://www.ons.gov.uk/ons/rel/census/2011-census/2011-census-teaching-file/index.html

The broad approach was to use both sampling and suppression of variables to remove uniques/pairs and achieve k-anonymity to a k of at least 2 in the published database, and then recoding to further achieve weak k-anonymity to a larger k based on the most identifying variables which were age, ethnic group, industry, economic activity and religion.

Producing the Census Microdata teaching file

A small sample size is used in order to aid protection (among other factors such as imputation for non-response), since a potential identification might be uncertain because the intruder will have doubt as to who is in the sample (and who is not). An intruder may find a record which corresponds to an individual for whom they are searching, possibly somebody unique in the sample with respect to specific visible characteristics. However they cannot be certain as to whether this sample unique is the person they are attempting to find because of the small sample size. The individual who is unique in the sample will not necessarily be the person they are looking for and there is no certainty that they would be unique in the population.

Starting from a large dataset – containing most variables and some further derived variables, with all the standard categories – it would clearly be possible to identify an individual, either directly or indirectly from these data. Hence some work was necessary to create a dataset suitable for 'open' data and public release:

- Remove all direct personal identifiers such as Name, Address and Date of Birth. The released file will have to contain no information allowing identification of an individual or household so this is the initial step in producing the data
- Decide on the variables to include in the data. Only a subset of census variables should be present in this teaching dataset, including basic demographic information and those variables used in the most popular tables. The level of Geography is to be Region (9 Regions for England plus Wales). Other variables include Sex, Ethnic group, Country of Birth, Industry, Marital Status and Household composition.
- Identify the key variables. These are the variables (usually in combination) which are most likely to assist an intruder to identify an individual in the data. These variables are usually those that are in the public domain such as Sex, Age, Ethnic group or those which a friend, relative or work colleague might know such as Occupation, Marital Status, Hours worked/week along with more sensitive variables such as Health.
- Create tables from the 1% sample using combinations of the key variables. Any low counts could lead to an individual in the data being identified. Note that this is a sample so there will be considerable doubt if a unique combination of variables in the sample is equivalent to a unique combination in the population.
- Create the same tables from the population data. Look for unique or rare combinations
- The most identifying variables were found to be
  - Geography
  - Sex
  - Age
  - Ethnic group
  - Industry
  - Economic Activity
  - Religion
  - Country of Birth
- Recode some of these variables to protect the data.
  - Recode Age into 8 Categories
  - Recode Ethnic group from 16 to 5 categories
  - Recode Industry from 17 to 12 categories
  - Recode Economic Activity from 13 to 9 categories
  - Recode Religion from 10 to 8 categories
- Recreate the tables from earlier using the recoded variables. The results show many combinations with sample uniques but very few with population uniques.
- Swap a small number of records (include these population uniques along with other records) between Region.

- Intruder testing was used as confirmation that risk was reduced to an acceptable level based on the number of correct identifications (if any).
- Publish the Data as an open data microdata file

5.2 DECC – ENERGY DATA

DECC has published two datasets from the National Energy Efficiency Data Framework (NEED) One of these is a Public Use File (open data) to be discussed here (49,815 records). The other is a file released under End User Licence (4,086,448 records). Both datasets are based on samples of properties which have been assessed for an energy performance certificate (EPC). Variables relating to the property are included along with gas and electricity consumption values.

The same methodology was applied in producing these datasets. A link to the Public Use File is shown here.

https://www.gov.uk/government/publications/national-energy-efficiency-data-framework-need-anonymised-data-2014

In this example the broad approach was again to use sampling and suppression of variables to achieve k-anonymity across the entire published dataset with a low k value. Intruder testing is used to confirm which variables might be used for identification and the "Year of EPC Assessment" variable subsequently removed. Recoding of variables was also applied to achieve weak k-anonymity with a higher k, for a subset of the most identifying variables.

The process was as follows for the production of the open data microdata file.

- A consultation period with potential users of the data was set up. This gave an indication of the level of detail users expected in the output data.
- Direct identifiers and detailed geographical indicators were removed from the data.
- The most visible variables were selected as key variables. These are the variables most likely to be used by intruders in attempts to identify a property. These variables are shown below along with plausible intruder scenarios.
    - Property Type (for example detached or end terrace). This would be obvious to anybody walking past the property in many cases, although there could be some doubt. For example is a house a single property or has it been divided into flats?
    - Property Age. An estimate of this can be made, although it may not be correct. Specialist property knowledge could be required for an accurate estimate. If the exact date of construction was known and the variable published at this level of detail it would provide an ideal starting point for an intruder.
    - Floor area. The floor area band would not be easy to estimate from outside. A visitor to the property would have a much better idea of this value, although even then a correct estimate may not be easy.
    - Geography. At a lower level of geography there will be fewer properties thus making a correct identification more likely. This is to be taken into consideration when deciding whether to release the data at National, Region or Local Authority levels.
- Look at distributions of the visible variables both individually and in combination. Are there low counts at National, Region and LA levels? If a property can be identified as belonging to a particular combination, much additional detail including the approximate gas and/or electricity consumption could be determined. If combinations of these variables produce low counts then certain variables may require recoding. The response variable of major interest is gas / electricity consumption. Low counts in the bands would give some information about the property but possibly not too much. Look out for values at the top and bottom of the range which are highlighted in the consumption data. In combination with the visible variables they could require protection.
- As a result of this analysis both Property Age and Floor area size are recoded into a smaller number of categories to reduce the number of low cell counts. It was also decided that the data would be released at Region level and not Local Authority level.
- The actual gas and electricity consumption values are given additional protection by being rounded to the nearest multiple of 5. This ensures that the actual value is not released in the dataset.
- A small number of records were swapped between Regions.
- Intruder testing was carried out by post graduate students at Southampton University. A cash prize was offered for a correct identification. There were no correct identifications but as the 'year of the energy performance certificate' was considered to be of particular use by the 'intruders' this variable was removed from the published open dataset, although it remains in the End User Licence data.
- Publish the Data as an open data microdata file.

5.3 BIS – FURTHER EDUCATION DATA

The Department of Business, Innovation and Skills (BIS) is planning to publish an open dataset of Further Education Learning aims, Providers and outcomes. This is a large dataset with many millions of records. It was hoped that a number of 'essential' variables would be included in the published data. These include variables relating to the type of course and an outcome grade variable.

In this example we achieve k-anonymity for a low k by suppressing variables non-essential to the user, as well as weak k-anonymity for a higher k by removing entire records within certain regions.

The process was as follows:

- From the list of essential variables decide on which are most visible and therefore key variables.
  Age group (3 categories)
  Sex

Learning aim (equivalent to a detailed course description)
Delivery Provider (a college or a company)

- The Region in which the learning took place was used as a geography variable.
- Tables of combinations of the key variables resulted in many unique combinations. The data could not be published in the current form. There was a requirement that the learning aim variable was retained and the following approach was followed.
- Age group to be recoded into 2 groups. Sex was dropped from the dataset.
- Records with a Learning Aim with fewer than a pre-defined number of enrolments within a Region were excluded from the data.
- Records which were unique with respect to Age group, Provider and Learning Aim within a Region were removed from the data.
- Data are currently in the process of being distributed for intruder testing before final release as an open dataset.

5.4 Should all licensed data (EUL) instead be released under OGL?

Currently many outputs from the UK Data Service[5] are released under a more restrictive End User Licence (EUL). The EUL is a 'light touch' licence with users promising not to attempt disclosure and to ensure that any outputs passed on do not compromise the confidentiality of individuals. Users of the EUL should keep the data confidential and not attempt to identify organizations, individuals or households in the data. In practice these datasets are designed so that the possibility of disclosure is remote. On this basis, the ONS recently conducted an intruder testing exercise to see whether the EUL was too conservative and whether these data could potentially be released under OGL. The Labour Force Survey and Living Costs and Food Survey microdata were used as two example datasets for assessment (see [11]). These were interesting cases as the intruder testing assumed the intruder had *response knowledge* of who was in the sample.

The disclosure scenario of response knowledge was considered a reasonable possibility under an OGL since the data would then be available to a much wider audience who would not be signing to a set of agreed conditions, unlike with the EUL. It was subsequently found from the intruder testing that re-identification was possible for certain individuals. Response knowledge meant that intruders would have much wider knowledge of individual attributes beyond the limited set of quasi-identifiers used to achieve k-anonymity under traditional EUL intruder scenarios. The conclusion from this was therefore that the conditions of an OGL mean extra precaution should be taken with releasing government data and to make careful assessments on a case-by-case basis. Significantly reducing detail by limiting the number of variables and their categorical breakdown is paramount to reducing the additional risk that comes with releasing open data.

[5] http://ukdataservice.ac.uk/

6. SUMMARY AND DISCUSSION

This paper has discussed the approach the ONS has taken towards opening up government data. Broadly speaking, suppression of variables and sampling (in two of the examples) are used to guarantee k-anonymity for a low k value (generally to remove uniques as a minimum). As a second stage, recoding/recategorisation of variables is used to generalize the dataset so that weak k-anonymity is achieved for a higher k value for a subset of records that are more likely to be attacked. Intruder testing is used to help inform the process in consideration of external information that might be linked to the dataset. The three examples discussed demonstrate the limited amount of information that can be made available openly. The purpose of these open datasets is usually only for use as teaching or training datasets. We briefly discussed how more detailed datasets available under End User Licence were not suitable for open release in their current form.

We have shown with our examples the difficult balancing act between producing Open Data which are of some use and protected to a reasonable level so that they remain non-personal data. Increases in technology in the past ten to fifteen years have changed the data environment beyond recognition. The consideration of other publicly available data sources is virtually impossible with the continual addition of data on the web. Intruder testing goes some way towards testing this in a practical way but is dependent on using knowledgeable and skilled intruders. As mentioned in the ICO anonymisation code of practice, this should be carried out periodically as the risk of re-identification may change with time bearing in mind likely increases in computing power and as the public availability of data increases. Feedback on intruder testing has been mostly positive and some of the benefits of this approach are outlined in [9]. Benefits include learning which variables and which types of individuals might be vulnerable to attack, and perceptions of disclosure. These provide a practical feel for data controllers of the level of risk. Further work would be helpful in developing expertise further in undertaking intruder testing, including working towards establishing reasonable standards and guidance. These would include the methodology employed, the length of time reasonable for an intruder to attempt disclosures, the level of 'uncertainty' that is reasonable, and better advice on the use of external information. However the importance of theoretical and sound practical risk measures should not be forgotten since the use of intruder testing is very much a snapshot of risk specific to each intruder and the parts of the data they are given (e.g. intruders might only assess particular geographies local to them and which they are knowledgeable about).

It also follows from this that there is a need to establish how much effort is "reasonable", (as mentioned in the ICO code) and where to set the line of acceptable risk. What values of 'k' are acceptable in open data? Can we measure units of anonymity to help data controllers make a decision? There is a clear link here to the concept of differential privacy – how much extra we can learn from an individual being included in a database as opposed to not including them. The purpose of a statistical office is both to collect and disseminate statistical information that will aid policy and research and, generally, be for the 'public good'. Hence it is an unreasonable, and usually unattainable goal to aim for only releasing datasets that are zero risk or in these examples to obligate strong k-anonymity. Ultimately, there is a legal interpretation – what risks would it be reasonable to protect against, so that the data publisher has a defence.

The future of anonymisation is unclear given the ever increasing amount of information being made publicly available. Open datasets only add to the disclosure risk. Currently these data generally have poor utility for answering complex research questions. An alternative we mention very briefly here is the potential use of synthetic or modelled data in an attempt to move towards a much richer set of data retaining at least all primary properties of interest to the researcher. One may argue that synthetic data have little or no risk as they do not represent the original data. However the creation of synthetic datasets that are truly representative of the population is very much an art. There is a related cost-benefit argument to whether the idea of open data is sustainable given the amount of effort government agencies need to produce such datasets. It is also important to remember that due to lack of research value, many open government data are still released alongside other licensed datasets (as is the case with the DECC data for example) Work for the future is not only about how to make open data as useful as possible but must also address all associated wider issues: data privacy but also technical, legal, economic and policy issues.

7. ACKNOWLEDGMENTS

The authors would like to thank Mary Gregory from DECC and Johanna Hutchinson from BIS for allowing us to reference their datasets. The authors are also grateful for the reviewers' comments which helped to improve this paper.

8. REFERENCES

[1] Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M., *L*-diversity: Privacy beyond *k*-anonymity, ACM Transactions on Knowledge Discovery from Data (TKDD), v.1 n.1, p.3-es, March 2007

[2] O'Hara, K, Whitley, E and Whittall, P (2011) Avoiding the Jigsaw Effect: Experiences With Ministry of Justice Reoffending Data.(unpublished briefing paper) http://eprints.lse.ac.uk/45214/

[3] Sweeney, A.L.. K-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570

[4] Atzori, D.M.. Weak k-Anonymity: A Low-Distortion Model for Protecting Privacy. Information Security Lecture Notes in Computer Science Volume 4176, 2006, pp 60-71

[5] Samarati P., (2001) Protecting Respodents' Identities in Microdata Release, "IEEE Trans. Knowl. Data Eng., no 6, 1010-1027.

[6] Elliot, M.J., and Dale, A. Disclosure risk for microdata: Workpackage DM1.1 What is a key variable? *Report to the European Union ESP/204 62/DG III , 1998*

[7] Elliot, M. J., and Dale, A. Scenarios of attack: The data intruder's perspective on statistical disclosure risk. *Netherlands Official Statistics. Vol 14, Spring 1999, 6-10.*

[8] Sweeney, B. L.. Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 571-588

[9] Tudor, C, Cornish, G, and Spicer, K. Intruder Testing on the 2011 UK Census: Providing Practical Evidence for Disclosure Protection. *Journal of Privacy and Confidentiality*: Vol. 5: Iss. 2, Article 3, 2014
Available at: http://repository.cmu.edu/jpc/vol5/iss2/3

[10] Hundepool A., Domingo-Ferrer J., Franconi L., Giessing S., Schulte Nordholt E., Spicer K., de Wolf P., Statistical Disclosure Control; Wiley (2012)

[11] Elliot,M. Mackey, E., O'Shea, S., Tudor, C., Spicer K., EUL to OGD: A Simulated Attack on Two Social Survey Datasets. *CD – only proceedings of Privacy in Statistical Databases, International Conference*, Ibiza, Spain, September 17-19, 2014.