Report of the
Director of Policy & Resources

Decision to be taken after
5 February 2014

NORTH LINCOLNSHIRE COUNCIL

CABINET MEMBER

POLICY AND RESOURCES

DATA DE-IDENTIFICATION POLICY

1. OBJECT AND KEY POINTS IN THIS REPORT

1.1 To approve a corporate data de-identification policy

1.2 The key points in this report are as follows:

- The recent transfer of public health into the council has seen the need to formalise arrangements to protect the confidentiality of service user information and ensure it is protected through the use of appropriate anonymisation, pseudonymisation statistical or aggregation techniques
- The policy set out in appendix 1 has been developed to set out our arrangements for data de-identification.

2. BACKGROUND INFORMATION

2.1 It is a legal requirement that when service user data is used for purposes other than direct care i.e. secondary uses the service user data should not be identifiable unless otherwise legally required such as having obtained the service users consent or Section 251 approval. This is set out clearly in the Department of Health's document 'Confidentiality: the NHS Code of Practice', which states the need to 'effectively anonymise' data prior to the non-direct care usage being made of the data

2.2 Information governance legislation namely the Data Protection Act 1998 and the Human Rights Act 1998 require that minimum personal data is used to facilitate any particular purpose and that information obtained in confidence should not normally be used in an identifiable format without the permission (consent) of the service user concerned unless there is a lawful exemption to do so.

2.3 The council currently has an over-arching information management framework that sets out the principles for managing all information assets and the application of regulatory frameworks and standards for managing data and information across the council.

2.4 The data de-identification policy forms part of the information governance

policy framework and should be read in conjuction with both the internal information sharing protocol and the information security policy.

2.5 A number of techniques are used for data de-identification and these are set out below:

- Anonymisation or data masking involves stripping out obvious personal identifiers likes names, addresses, DOB and postcodes to create a new data set where no personal identifiers are present.

- Pseudonymisation involves the de-identification of data so that a coded reference of pseudonym can be attached to a record enabling the data to be associated with a particular individual without the individual being able to be identified.

- Statistical or aggregation is where data is displayed as totals. No data relating to or identifying any individual is shown, where there are small numbers in total these are often suppressed, grouped or omitted altogether

2.6 The data de-identification policy includes:

- Scope
- Definitions
- Roles and responsibilities
- Guidance on the techniques for data de-identification
- Transferring information
- Monitoring and review

2.8 Careful consideration must be given before sharing personal data and alternative datasets should be investigated before applying any data de-dentification techniques. In all cases approval from a senior manager should be sought before undertaking any de-identification action.

2.9 There are currently no corporate tools for anonymising, pseudonymising or re-identification of data, however service areas do currently redact data manually. Consideration therefore needs to be given to developing corporate tools to anonymise and pseudonymise data. We have recently been offered a unique opportunity of a 6 month free trial of the software used to anonymise the 2011 census data.

2.10 Following a trial this policy will be be subject to further review to incorporate the results of the trial and any shift in national thinking.

2.11 An elearning package on data de-identification is currently being developed and a series of workshops will also be available for staff to attend. In addition directorate information asset owners will be the champions for data de-identification and will be able to provide additional support where necessary.

3. OPTIONS FOR CONSIDERATION

3.1 Approve the data de-identification policy.

3.2 Approve the trial of the software used to anonymise the 2011 census data.

3.3 Do not approve the data de-identification policy and request changes be made.

3.4 Do not approve the trial of the software used to anonymise the 2011 census data.

4. ANALYSIS OF OPTIONS

4.1 Approving the data de-identification policy would provide a clear framework for protecting service user identifiable information. It would also ensure the council is complying with NHS requirements, the Human Rights Act 1998 and the Data Protection Act 1998. Additional benefits include data sharing, monitoring and surveillance of population, service users' needs and demands across the council.

4.2 Approving the software trial will enable us to evaluate appropriate use of automated anonymisation, pseudonymisation and re-identification processes. Following a trial this policy will be be subject to further review to incorporate the results of the trial and any shift in national thinking.

4.3 Not approving the data de-identification policy could result in data being unlawfully identifiable and expose the council to legal action for breach of legislation/policy.

4.4 Not approving the participation in the software trial would result in a missed opportunity to test some market leading software for free.

5. RESOURCE AND OTHER IMPLICATIONS (FINANCIAL, STAFFING, PROPERTY, IT)

5.1 A training programme is currently being designed to raise awareness of data de-identification techniques, and this will take varying forms to ensure it reaches all appropriate council employees. Where possible staff will be trained through an e-learning package, whilst those without IT access will have alternative training made available. No extra resources will be needed to cover this as the Policy and Resources directorate will lead and work with other directorates on this.

5.2 Failure to comply with information governance legislation can result in the Information Commissioner imposing fines of up to £500,000. In addition the reputation of the council would be affected as a result of any negative publicity.

6. OUTCOMES OF INTERGRATED IMPACT ASSESSMENT (IF APPLICABLE)

6.1 An integrated impact assessment has been undertaken and impacts identified have helped shaped the procedure. Both the policy and the integrated impact assessment will be kept under constant review.

6.2 The development of this policy goes some way to ensure compliance with NHS requirements, the Data Protection Act 1998 and the Human Rights Act 1998.

7. OUTCOMES OF CONSULTATION

7.1 Consultation has taken place with the Information, Improvement and Value for Money group, Caldicott Guardians, the Senior Information Risk Owner, information asset owners, internal audit and others as appropriate to develop this policy.

8. RECOMMENDATIONS

8.1 It is recommended the Cabinet member for Policy and Resources approve the data de-identification policy and authorise the six month trial of the 2011 census anonymisation & pseudonymisation software.

DIRECTOR OF POLICY AND RESOURCES

Civic Centre
Ashby Road
SCUNTHORPE
North Lincolnshire
DN16 1AB
Author: Rachel Johnson
Date: 15 January 2013

Background papers used in the preparation of this report:

North Lincolnshire Council Information Sharing Policy
North Lincolnshire Council Information Security Policy
Draft Anonymisation Code of Practice – ICO

DATA DE-IDENTIFICATION POLICY

Introduction

This policy outlines how North Lincolnshire Council will meet its legal obligations and sets out the approach taken within the council to provide a robust data de-identification framework for the current and future protection of service user identifiable information.

It is a legal requirement that when service user data is used for purposes other than direct care i.e. secondary uses, the service user should not be identifiable unless otherwise legally required such as having obtained the service users consent or Section 251 approval.

Section 251 of the NHS Act 2006 was established to enable a common law duty of confidentiality to be overridden to enable disclosure of confidential patient information for medical purposes, where it was not possible to have anonymised information and where seeking consent was not practicable, having regard to the cost and technology available.

The Data Protection Act 1998, the Human Rights Act 1998 requires that minimum personal data is used to facilitate any particular purpose and that information obtained in confidence should not nomally be used unless there is a lawful exemption under the Data Protection Act or Human Rights Act to do so.

Definitions

Personal Identifable Data – is information about a person that would enable the person's identity to be established. This might be fairly explicit such as intitials and surname or isolated postcode or items of different information which if taken together could allow the person to be identified. All information that relates to an attribute of an individual should be considered as potentially capable of identifying them to a greater or lesser extent.

Any of these items can be considered collectively or in isolation as identifiable information: Surname, Forename, Initials, Address, Date of Birth, Other dates (e.g. death, diagnosis) Postcode, Gender, Occupation, Ethnic group, NHS or Hospital Number, National Insurance Number, Telephone number or other local identifier.

Primary use – is the use of data that directly contributes to the safe care and treatment of a service user and includes diagnosis, referral and treatment processes together with relevant supporting administrative processes, such as letters, service user administration, managing appointments for care as well as the audit/assurance of the quality of the healthcare provided is considered primary use.

Secondary use – includes other uses of the data, that is the non-direct care usage referred to above, and is usually known as secondary uses. Examples of secondary uses are for preventative treatment, trend analysis, research, financial audit and the management of health care services.

Information processing – means obtaining, recording or holding the information or data or carrying out any operation or set of operations on the information or data, including:

- Organisation, adaptation or alteration of the information or data,
- Retrieval, consultation or use of the information or data,
- Disclosure of the information or data by transmission, dissemination or otherwise making available, or
- Alignment, combination, blocking, erasure or destruction of the information or data.

Anonymisation –involves stripping out obvious personal identifiers like names and addresses from data to create a new data set where no personal identifiers are present.

Pseudonymisation - is de-identifying data so that a coded reference can be attached to a record enabling the data to be associated with a particular individual without the individual being able to be identified.

Caldicott principles – All health and social care public organisations must appoint a Caldicott Guardian, of which the council complies. The guardian must ensure the six key principles are applied when using service user-identifable information. Compliance with these principles reduces the risk of breaches of confidentiality and breaches of legal requirements.

Scope

This policy applies to all North Lincolnshire Council staff who use service user data for secondary use purposes. The key principle is to ensure, as far as is practicable, that individual service users cannot be identified from data that are used to support purposes other than their direct care or to quality assure the care provided.

Data de-identification is about minimising the risk of re-identification because there is no guarantee of stopping re-identifcation occuring. The supporting guidance sets out the different techniques available for de-identifying data.

Roles and Responsiblities

The Director of Policy and Resources has overall responsibility for Information Governance within the authority and has corporate responsibility for ensuring that it meets its legal responsibilities. The Director of Policy and Resources also has a responsibility for the adoption of internal and external governance requirements which oversees the implementation of the data de-identification policy.

The Information Governance team has delegated responsibility for implementation and monitoring of the information governance framework and also for providing training and support.

The Information, Improvement and Value for Money Group are responsible for ensuring that this policy is implemented and that processes are developed, co-ordinated and monitored for data de-identification.

Information Asset Owners are responsible for championing this policy within their service area and providing support to ensure successful implementation.

Service managers are responsible for:

- Identifying all areas that need to be classified under this policy
- Nominating a member of staff to manage that classified area
- Identify staff that have a justified purpose to access service user identifiable information, and obtain authorisation to access information from the Caldicott Guardan
- Ensuring all staff have completed relevant training
- Organising the removal of staff rights to identifable information where access is no longer justified
- Maintaining a comprehensive register of those staff authorised to access identifiable information, including details of where access has been reviewed and removed e.g. on change of duties or cessation of employment
- Ensuring all systems holding service user identifiable information are accessible via robust log on and passwork mechanisms only and that these are allocated on a need to know basis.
- Ensuring this policy and any supporting standards and guidelines are successfully implemented into local processes and that there is ongoing compliance on a day to day basis. Any breaches or suspected breaches of confidentiality or information security must be reported for immediate investigation.

All staff who create, receive and use service user records have data de-identification responsiblities under the Data Protection Act and Information Governance requirements are responsible for:

- Ensuring they understand the requirements of this policy and its supporting standards and guidelines
- Ensuring that this policy and its supporting standards and guidelines are built into local processes and that there is ongoing compliance on a day to day basis
- Ensure all breaches or suspected breaches of confidentiality or information security are reported for immediate investigation
- Highlight areas of potential weakness to their managers immediately for appropriate corrective action.

Contractors and partner organisations must accept and comply with the councils responsibilities and guidelines for data de-identification under information

governance and data protection requirements.

Transferring information

Where is has been identified that information sharing is to take place with other organisations, information sharing agreements should be documented, agreed and signed up to by the Caldicott Guardians/SIRO's of the partner organisations to that agreement. Template agreements can be found in the Humber Information Sharing Charter.
Should sharing of de-identified data take place, the council has a legal obligation to ensure there is no risk of potential re-identification, should there be a potential risk this needs to be highlighted and an appropriate decision on whether to share or not needs to be taken and documented.

Monitoring and Review

This policy will be reviewed annually. Earlier review may be required in response to exceptional circumstances, organisational change or relevant changes in legislation or guidance.

A routine audit programme to monitor the adequacy of systems and policies will be developed.

Guidance on aggregation, anonymisation and pseudonymisation to support the authorities policy for de-identifying data

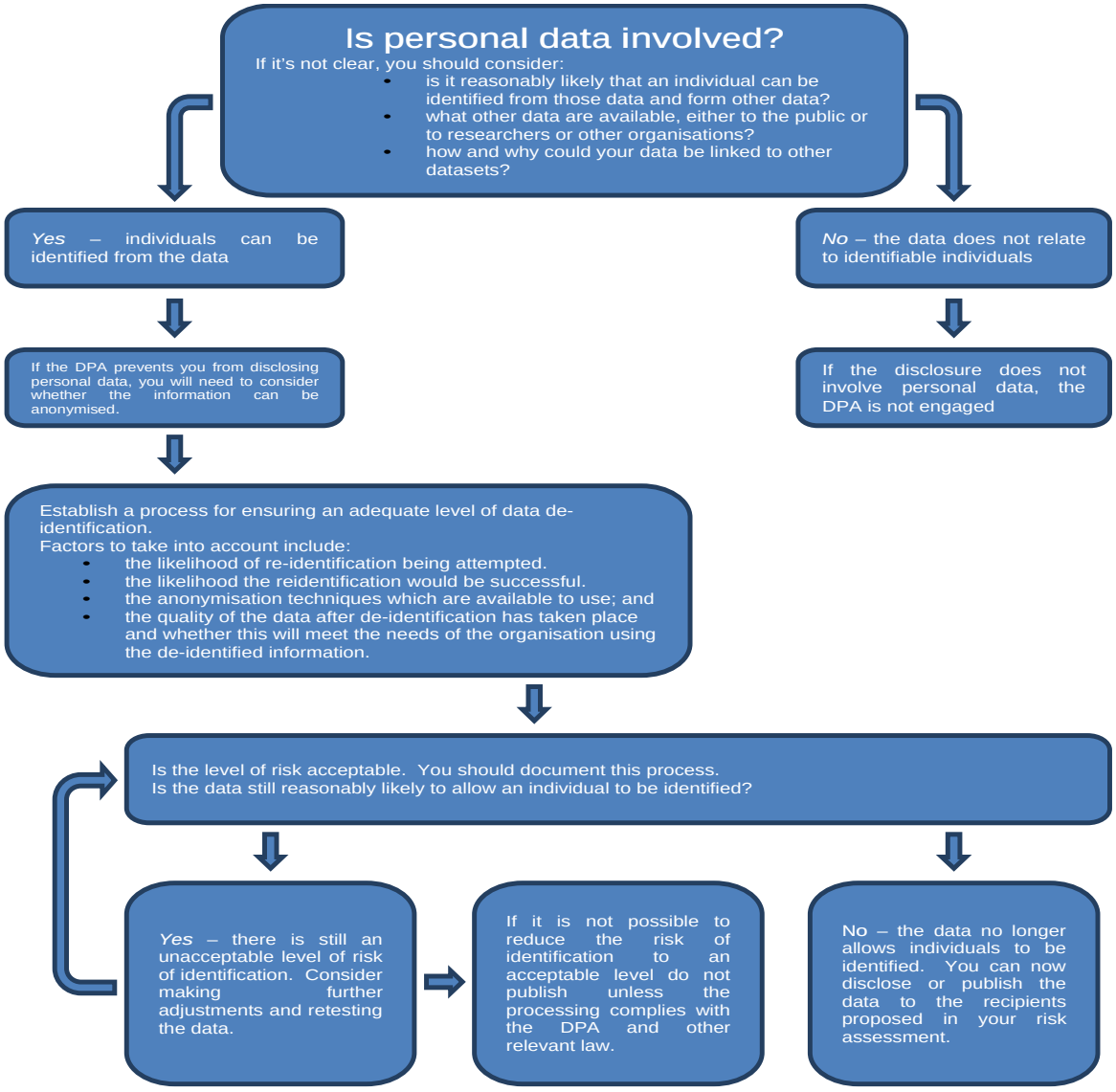
Introduction

How do you edit exempt information from paper and or electronic documents prior to them been released?

This document has been produced to provide guidance on editing exempt material from information held by the authority. Its purpose is to promote good practice across the authority when sharing information or releasing information.

This guidance is aimed at all officers who have responsibility for sharing and or releasing of information, in particular officers who deal with Freedom of Information Act (FOIA), Data Protection Act (DPA) and Environmental Information Regulations (EIRs) but not excluding sharing of information to third parties. Whenever we share, anonymise or pseudonymise data or records we should record the decisions on this, in particular when we redact data. Appendix C provides a simple form for recording redaction decisions.

Procedure for data de-identification



Data De-identification Methods

Aggregation

Data is displayed as totals. No data relating to or identifying any individual is shown, where there are small numbers in total these are often suppressed, grouped or omitted altogether

There are 7 variants to aggregation (examples are shown in appendix A):

I. ***Cell suppression*** – where data is from a sample then it may be inappropriate to release information that contains small numbers of individuals. Suppression of cells with small number for quality purposes act in tandem with suppression for disclosure purposes. Refer to appendix A for an example of cell suppression.

II. ***Inference Control*** – some cell values in statistical data can present a greater risk of re-identification. Depending on the circumstances, small numbers can either be suppressed (discussed above), or the values manipulated (as used in the method of perturbation – discussed below). Should a large number of cells be affected, the level of aggregation could be changed. For example, the data could be linked to wider geographical areas, or age bands could be widened.

III. ***Perturbation*** - This method is often employed by public agencies to enable them to provide information for statistical purposes without infringing the information privacy rights of individuals to whom the information relates. This technique of disclosure control is for tables or counts and involves randomly adding or subtracting 1 from some cells in the table or averaging bands of data.

IV. ***Rounding*** – rounding a figure up or down to disguise the precise statistic or value.

V. ***Sampling*** – when there are very large numbers of records available it can be acceptable to release a sample of records, selected through some stated randomised procedure. By not releasing specific details of the sample, data holders can minimise the risk of re-identification.

VI. ***Synthetic data*** – mixing up the elements of a particular dataset so that all of the overall totals and values of the set are preserved but do not relate to any particular individual

VII. ***Tabular reporting*** – a means of producing tabular (aggregated) data that protects against re-identification.

VIII. ***Derived data items and banding*** – a set of values that reflect the character of the source data but hide the exact original values. This is achieved usually by using banding techniques to produce coarser-grained descriptions of values i.e replacing dates of birth with ages or years, addresses by areas of

residence or wards, using partial postcodes or rounding exact figures so they appear in a normalised form.

Aggregation usually is a relatively low risk technique because the techniques used make data matching more difficult or impossible. The resulting data can be relatively rich but can lack granular detail but presents a relatively low re-identification risk.

Anonymisation

Anonymisation or data masking involves stripping out obvious personal identifiers like names and addresses from data to create a new data set where no personal identifiers are present.

There are three variants to anonymisation:

I. ***Partial data removal*** – this results in the data having some of the personal identifiers, e.g. name and or address being removed, however others like dates of birth etc, still remain
II. ***Data quarantining*** – this technique only supplies data to a recipient who is unlikely or unable to have access to the other data required to facilitate re-identification. This technique allows the disclosure of unique personal identifiers only as long as the 'key' required to links these to particular individuals is not disclosed.
III. ***Data masking*** – this results in the data primarily in hard copy haing a copy make but all personal data is ***redacted*** i.e. personal information, location information is crossed out using a marker that prevents the information being seen. Appendix C provides further guidance on redaction of information.

Anonymisation of data can be a relatively high risk technique to use, because the anonymised data still exists at an individual level. With the introduction of other data sets then there can be a high risk of re-identification, however this type of data is also relatively rich in data enabling better analysis.

The golden rule when anonymising data before you share or release it find out how this data is going to be used, and what other data will be used with it. If you have any doubt that re-identification of individuals would be apparent then seek management permission.

Pseudonymisation

The de-identifying of your data so that a coded reference or pseudonym can be attached to a record enabling the data to be associated with a particular individual without the individual being able to be identified.

Deterministic modification is a similar technique, meaning that the same original values are always replaced by the same modified value. This means if multiple data records are linked, in the sense that the same name occurs in all those records, the corresponding records in the modified data set will also be linked in the same way.

This enables certain types of data analysis to still be conducted without having the identifying data been shown.

When pseudonymisation techniques are consistently applied, the same pseudonym is provided for individuals across different data sets and over time. This allows the linking of data sets and other information which is not available if the Personal Identifying Data (PID) has been removed completely.

To effectively pseudonymise data it is advisable the following actions are taken:

1) Each field of PID must have a unique pseudonym
2) Where pseudonyms are to be used it is advisable they are of the same length and the original data and formatted on output to ensure readability.
3) Consideration needs to be given to the impact on existing systems both in terms of the maintenance of internal values and the formatting of reports
4) Where used pseudonyms for external use must be generated to give different pseudonym values in order that internal pseudonyms are not compromised
5) When we are sharing or releasing pseudonymised data carefully consider that you only share or release data items that are required. This is in line with DPA and Caldicott Guidelines
6) Pseudonymised data should have the same security levels as the original data

Pseudonymisation is also a relatively high risk technique and also shares the same strengths and weaknesses to that of anonymisation.

Manual redaction of information

A common error when redacting is using the wrong method especially when redacting an electronic file. This guidance provides a partial list of methods NOT to use:

1. Never redact the original or master version of an electronic or paper record. Redaction must always be carried out on a new copy.
2. Changing the font colour ie to white does make the words disappear, but they are still there.
3. All word processing programs retain a lot of hidden code. This hidden code could reveal anything that was contained in the file at any time, even text that has previously been deleted or changed, even if the file was re-saved.
4. Adobe Acrobat (PDF files generally use this software) does have tools that can black out areas including pictures and text. The problem is that people can easily remove them as well.
5. Ink-marking, using tape or paper to cover areas of a document to be scanned can still sometimes show enough information for someone to see what was assumed hidden.

Redaction must irreversibly remove the required information from the redacted copy of the document. This guidance provides some examples of ways to ensure that your documents may be redacted as you intended:

1. Redacting a Word-Processing File: As previously mentioned all word-processing programmes retain a lot of hidden code. The simplest way is to use a simple text editor (i.e. Notepad) to create the final redacted version of the document as a text editor (i.e. Notepad) does not save any hidden code.

   To do this, in your original document replace all the text that you wish to redact with the word **[REDACTED]**. Once all text has been replaced save the file as a new version. Copy all the text from the file and paste it into a text editor (i.e. Notepad) and save the text file. This text file is the document that can then be shared. Should you require reformatting the text file you can open in your Word Processing software (i.e. Word), this would be fine as a brand new blank file, DO NOT place the text into the original file.

2. Redacting a Scanned File (i.e. tiff, jpeg, gif, etc): This can be slightly trickier since you are modifying an image. When you scan a document more than likely the document will be saved as a PDF, this will be covered using method 3. The issue around redacting a scanned file is that the data which contains that image may not be fully removed or destroyed when using common software tools. A simple but not always ideal solution is once you have redacted on the image, print this image out and share the hardcopy.

3. Redacting a PDF file (scanned or converted): This is the most delicate and difficult as Adobe Acrobat by itself cannot redact a document using any of the built in tools. There are plug-ins that can do this but you would need to discuss with the authorities IT. You would be better considering printing out the document and using method 4.

4. Redacting a Paper Document: Before you scan the document ensure you complete the following.

   a) Is the Paper Document printed on both sides? In order to redact a Paper Document correctly you need it only printed on one side.
   b) Cut out (literally) all the text to be redacted and properly dispose of the clippings (confidential waste). This method is always 100% effective.

   OR

   Use opaque correction fluid (usually opaque correction fluid is 100% impenetrable by light but you will need to check your brand to be sure), tape or paper to cover over the sections to be redacted. A warning is do not use plain paper as the scanner may pick up images through the paper. Even black paper may allow light reflection.

OR

Use a photocopier that has redaction facilities. There are photocopiers that have the facilities to automatically remove marked out areas on a document. This provides a secure method of redaction.

Re-identification testing

It is good practice to use re-identification testing. This testing is a type of penetration test to detect and deal with re-identification vulnerabilities and involves attempting to re-identify individuals from an anonymised data set(s).

Firstly you need to take stock of the anonymised data that your department has published or intends to publish, share or release. Secondly you need to determine what other data – personal or not is available that could be linked to the anonymised data to result in re-identification (this can be difficult to do in practice as it may be impossible to determine what other information is available or easily accessible through the internet).

A penetration test should address the following:

- Should attempt to identify particular individuals and one or more private attributes relating to those individuals
- The test may employ any method that is reasonably likely to be used by an intruder
- The test may use any lawfully obtainable data source that is reasonably likely to be used to identify particular individuals in the datasets.

Assessing the re-identification of statistical data becomes more complex because the volume of publicly available data, if matched in particular way could result in re-identification.

Appendix A – Examples of de-identification methods

Example 1: Aggregation – cell suppression

Data suppression involves not releasing information for unsafe cells, or, if nothing else works, deleting individual records or data items from the file. If a table contains totals, it may be possible to calculate the value of a suppressed cell by subtracting the value of other cells from the total. At least one additional cell may also need to be suppressed to prevent identification.

A cell that is suppressed because it fails one of the confidentially rules is called a primary suppression cell. The suppression of other cells is called secondary suppress or consequential suppression.

In the example table below the cell containing the number of low income earners aged 50-59 was identified as unsafe cell using a frequency rule of 5. This cell could be protected by suppressing it, as shown by the 'X'.

It would however still be possible to work out the value of the cell by using the remaining values. For example, low income earners aged 50-59 could be calculated by subtracting the medium and high income earners from the total. This would require subsequent suppression to protect the unsafe cell, to accomplish this secondary suppression would be required. This secondary suppression would be referenced with "Y".

	Income			
Age group (years)	Low	Medium	High	Total
15-19	20	0	0	20
20-29	14	11	8	33
30-39	8	12	7	27
40-49	6	18	24	48
50-59	**X**	~~6~~ **Y**	14	23
60+	12	9	7	28
Total	64	55	60	179

Example 2: Perturbation – micro aggregation

The idea of micro-aggregation is to replace an observed value with the average computed on a small group of units. The groups will contain a minimum predefined number of ***k*** units. Here ***k*** is a threshold value and the partition is called a ***k-partition***. In order to obtain the micro-aggregates from a dataset you usually require that data in a meaningful order i.e. size.
For example if we had 16 individuals in a dataset and wished to form a 4-oartition, then the dataset would be grouped into segments of 4. For each group the average replaces the actual value so if we have 16 individuals and we wish to form 4

partitions. Firstly we need to sort the data in a meaningful way (i.e. Age) we then want to average salary for each segment

Original Data

Age	Gender	Post Code	Income	Expenses
22	F	SO17	£20,000	£1,000
25	M	SO17	£22,000	£1,000
25	F	SO18	£24,000	£1,500
30	F	S0018	£32,000	£1,850
30	M	SO19	£32,500	£3,500
35	F	SO19	£42,000	£1,500
35	M	SO20	£45,000	£1,500
40	F	SO17	£65,000	£2,000
40	M	SO18	£69,000	£3,500
45	F	SO19	£59,000	£3,000
45	M	SO20	£59,000	£1,500
50	F	SO17	£34,000	£1,000
55	M	SO17	£25,000	£500
60	M	SO18	£28,000	£500
65	F	SO20	£15,000	£500
65	M	SO20	£12,000	£1,500

Original Data sorted by Income

Age	Gender	Post Code	Income	Expenses
65	M	SO20	£12,000	£1,500
65	F	SO20	£15,000	£500
22	F	SO17	£20,000	£1,000
25	M	SO17	£22,000	£1,000
25	F	SO18	£24,000	£1,500
55	M	SO17	£25,000	£500
60	M	SO18	£28,000	£500
30	F	S0018	£32,000	£1,850
30	M	SO19	£32,500	£3,500
50	F	SO17	£34,000	£1,000
35	F	SO19	£42,000	£1,500
35	M	SO20	£45,000	£1,500
45	F	SO19	£59,000	£3,000
45	M	SO20	£59,000	£1,500
40	F	SO17	£65,000	£2,000
40	M	SO18	£69,000	£3,500

Segmented partitions require average income

Age	Gender	Post Code	Income	Expenses
65	M	SO20	£12,000	£1,500
65	F	SO20	£15,000	£500
22	F	SO17	£20,000	£1,000
25	M	SO17	£22,000	£1,000
25	F	SO18	£24,000	£1,500
55	M	SO17	£25,000	£500
60	M	SO18	£28,000	£500
30	F	S0018	£32,000	£1,850
30	M	SO19	£32,500	£3,500
50	F	SO17	£34,000	£1,000
35	F	SO19	£42,000	£1,500
35	M	SO20	£45,000	£1,500
45	F	SO19	£59,000	£3,000
45	M	SO20	£59,000	£1,500
40	F	SO17	£65,000	£2,000
40	M	SO18	£69,000	£3,500

Age	Gender	Post Code	Income	Expenses
65	M	SO20	£17,250	£1,500
65	F	SO20	£17,250	£500
22	F	SO17	£17,250	£1,000
25	M	SO17	£17,250	£1,000
25	F	SO18	£27,250	£1,500
55	M	SO17	£27,250	£500
60	M	SO18	£27,250	£500
30	F	S0018	£27,250	£1,850
30	M	SO19	£38,375	£3,500
50	F	SO17	£38,375	£1,000
35	F	SO19	£38,375	£1,500
35	M	SO20	£38,375	£1,500
45	F	SO19	£63,000	£3,000
45	M	SO20	£63,000	£1,500
40	F	SO17	£63,000	£2,000
40	M	SO18	£63,000	£3,500

Data set re-sorted by age (*k partition = 4*)

Age	Gender	Post Code	Income	Expenses
22	F	SO17	£17,250	£1,000
25	M	SO17	£17,250	£1,000
25	F	SO18	£27,250	£1,500
30	F	S0018	£27,250	£1,850
30	M	SO19	£38,375	£3,500
35	F	SO19	£38,375	£1,500
35	M	SO20	£38,375	£1,500
40	F	SO17	£63,000	£2,000
40	M	SO18	£63,000	£3,500
45	F	SO19	£63,000	£3,000
45	M	SO20	£63,000	£1,500
50	F	SO17	£38,375	£1,000
55	M	SO17	£27,250	£500
60	M	SO18	£27,250	£500
65	M	SO20	£17,250	£1,500
65	F	SO20	£17,250	£500

This method guarantees that at least units in the data are identical; however the loss about specific individuals is high.

Example 3: Anonymisation of a dataset

A company has collected the following data:

Name, address and date of birth	Period on Special Assistance	Body mass index (BMI)	Research cohort reference number
Mr Bates 69 Long Road Stevenacres 20-04-1974	1 yr 1 mth	12	1a5
Mr C Averil 65 Long Road Stevenacres 18-03-1968	5 yr 4 mth	13	2b4
Mrs C All 38 Long Road Stevenacres 16-02-1969	1 yr 9mth	17	3c3
Mrs A K Ward 75 Long Road Stevenacres 14-01-1962	5yr 1mth	20	4d2
Mr M Hunt 54 Long Road Stevenacres 12-11-1966	1yr 6 mth	15	5e1

In this example all of the collected data would constitute as personal data because all ther data items realte to identified individuals. If this information was to be disclosed to a third party in its form above then it would be subject to the data protection policies and principles. However redacting the above data set then the information could be released as below:

Name, address and date of birth	Period on Special Assistance	Body mass index (BMI)	Age Range	Research cohort reference number
	Less than 2 years	12	35-40	1a5
	Greater than 5 years	13	40-45	2b4
	Less than 2 years	17	40-45	3c3
	Greater than 5 years	20	50-55	4d2
	Less than 2 years	15	45-50	5e1

In creating this extract of the data it would provide individual results on a granular level and would not breach the data protection policies or principles as the purpose of the redaction process is to protect the individual. The redacted data set still holds

personal data in the hands of the original company because it still holds the full version of the original data, and they hold the key that is required to link back to the personal identifiers. Third party companies cannot do this because they don't hold any information in the extract that could allow the linkage to be made.

The company can provide information like:

> *40% of the individuals that have been on special assistance for less than two years had a Body Mass Index of over 12.*
>
> or
>
> *One individual out of five had a BMI of 20 having claimed special assistance for over five years.*

Even though the information above may relate to only one individual, it is still not personal data once disclosed, provided that no other organisation know the identity of the individuals they cannot be linked together. The statement above could have blurring applied and the information may look like this:

> *0-3 individuals out of between 3 and 5 had a BMI of between 13 and 20 having received special assistance for 2 – 6 years.*

Blurring the information in this way means that no-one could identify individuals but there is a debate about the usefulness and clarity of using this technique.

Example 4: Anonymising qualitative data

An interview was conducted with a child and carer. The original minutes taken can be converted into an anonymised form that still contain valuable information but does not identify the child.

Original text:

> Interview recorded: 15:30pm on the 10th January 2012
> Interviewee: Joe Done
> DoB: 9th December 2006
> Location: Department of Social Services, Long Road
> Child's report – I live on Westfield Lane so I walk to school each day. I live in a small flat with my mum, when I get home from school I like watching telly. I don't read as I don't like it but I like watching Harry Potter. At school my best subject is art and my teacher is Mr White. Mr White is nice. Neil and Chris would bully me each day but I told Mr White and they stopped it.

Anonymised text:

> Interview recorded: January 2012
> Interviewee ref: J2011D/44

School Year: Key Stage 2
Authority Ward: 17
Child's report – I live on {DN15 postcode] so I walk to school each day. I live in a small flat with my [parent], when i get home from school I like watching telly. I don't read as I don't like it but I like watching Harry Potter. At school my best subject is art and my teacher is Mr [Name]. Mr [Name] is nice. [Pupil 1] and [Pupil 2] would bully me each day but I told [the teacher] and they stopped it.

Sample of simple form for recording redaction decisions

Date of redaction	Document reference	Details of exempt material	Exemption applied	Justification for exemption	Reviewer's comments	Any other comment