Statistical Disclosure Control Protocol for Revenue

Updated January 2019

Statistics & Economic Research Branch

Table of Contents

1   Introduction to Statistical Disclosure Control

Confidentiality is central to Revenue's relationship with customers. Revenue's obligations to safeguard data are established by a range of legislative and administrative provisions designed to protect the rights and interests of citizens and businesses.

General provisions include:
- The Official Secrets Act
- The Data Protection Acts 1988 and 2003
- Revenue Code of Ethics
- Civil Service Code of Standards and Behaviour
- General Data Protection Regulation 2018

Specific to Revenue are the Revenue Code of Ethics and Section 851A, Taxes Consolidation Act 1997.[1] Section 851A formalises taxpayer confidentiality and provides assurance taxpayers that their personal and commercial information disclosed to Revenue is protected against unauthorised disclosure to other persons. Revenue is unable to share taxpayer data unless specific legislation exists to permit the exchange.

Revenue is publishing increasing statistics and research on tax receipts and costings of tax policy options.[2] In doing this, there is a need to balance the expectations of the public, researchers, Government and the Oireachtas for information with the requirements on Revenue from data protection rules and the need to protect the confidentiality of taxpayer information. Protecting taxpayers' data is central to all of our processes including publication of statistics.

This paper outlines the issues concerned with protecting the confidentiality of taxpayers and Revenue's Statistical Disclosure Control (SDC) methods for ensuring that the public interest in taxpayer information is balanced with managing data disclosure risks. It also outlines the key steps that a data provider must consider before disclosing statistical data. It is assembled from Revenue experience in the provision of statistics and from best practices in other organisations, in particular the Central Statistics Office and the Office for National Statistics.

SDC techniques can be defined as a set of methods to reduce the risk of disclosing information on individuals, businesses or other entities. Such methods are only related to dissemination and are usually based on restricting or modifying the data released.

---

[1] S851A was introduced by Section 77 of the Finance Act 2011.
[2] Revenue Statistics: https://www.revenue.ie/en/corporate/information-about-revenue/statistics/index.aspx.

2 Guidance and Key Steps

2.1   General Principles

Revenue publishes and disseminates a broad range of statistical data, on both a regular and ad-hoc basis. Revenue tends to publish tables that are seen to be **'low risk'**, aggregated statistics rather than individual or unit level data. However, with increasing requests for more detailed **'high risk'** statistical data, as well as the Government's Open Data initiative, it has become necessary to introduce some guidelines and SDC measures to protect against disclosing confidential data.

> **Low Risk Tables**
>
> When tables are presented at a high level of aggregation, e.g., National or Regional level, they are classified as low risk tables. In this scenario disclosure issues are less likely to arise.

> **High Risk Tables**
>
> When tables become more detailed and the counts of individual taxpayers in individual cells are small, the risk of identification increases and protection may be needed.
> There is also a risk of identification where cell totals are dominated by one single statistical unit, i.e., individual, household or business.

A first Revenue SDC protocol was published in 2016, which this current document updates. Although prior to 2016 no formal SDC measures were in place, generally a 'Common Sense' approach was used.
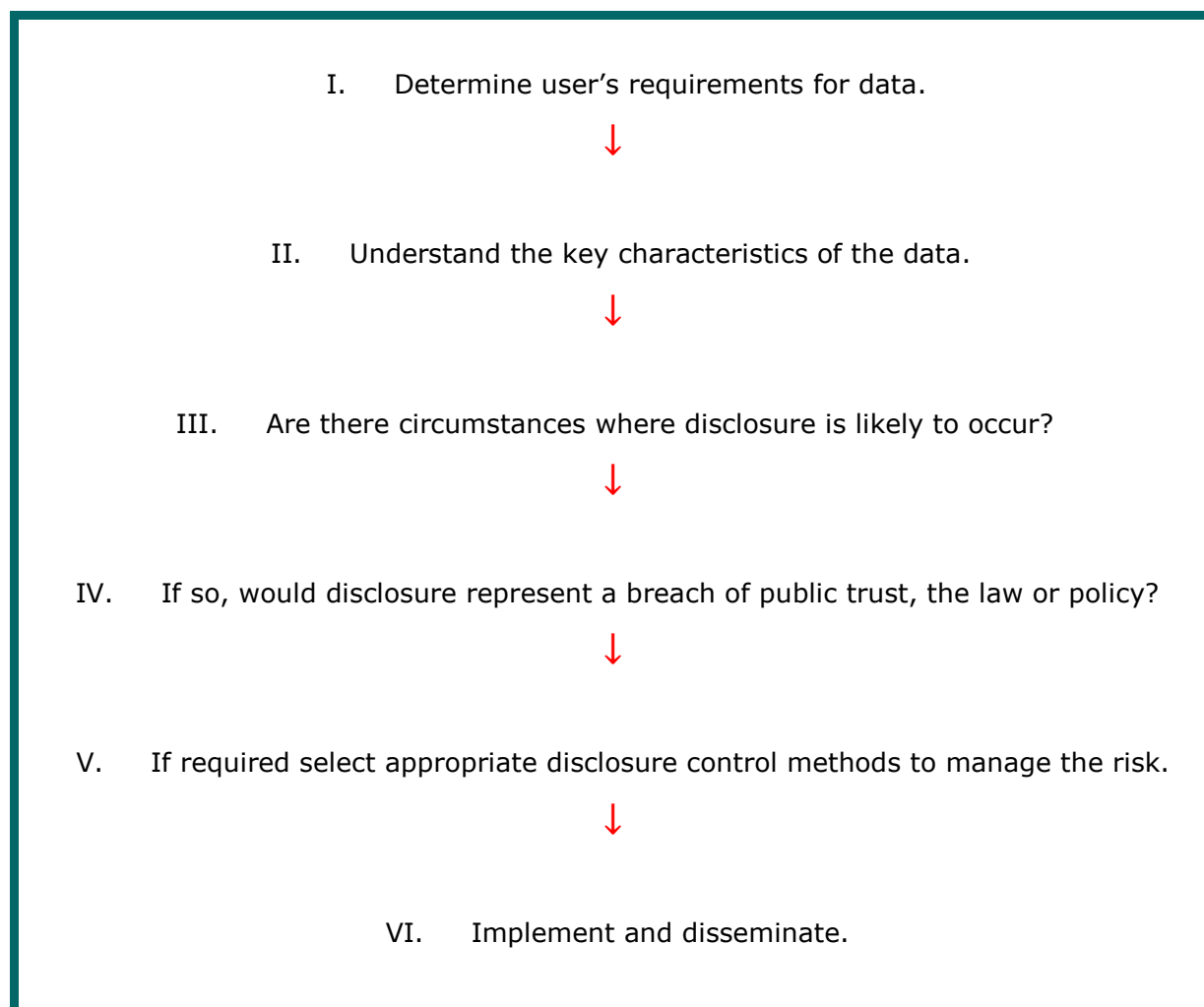
> The overall principle of Revenue's approach to SDC is that data or tables are condensed or redesigned when numbers (of taxpayers) are shown to be below 10 units.[3]

For much of the data collated by Revenue for publication or in other response to external queries, the risk of identifying individuals or companies will be minimal and no SDC methods necessary. Where there is disclosure risk, the principle of not disclosing information for less than 10 cases is applied. Some situations may be more complex. There is no one solution available for these instances but this protocol offers support.

---

[3] The use of a minimum number is a widely accepted approach in statistics offices and elsewhere, although the specific minimum varies, often between 3 and 30. Revenue's use of 10 is supported by, for example, http://neon.vb.cbs.nl/casc/ESSnet/GuidelinesForOutputChecking_Dec2009.pdf.

Every new request for statistics, whether for aggregated or case level data, should be assessed for disclosure risks. Guidance and key steps are provided in the table below. This guidance will allow data providers to develop their own confidentiality methods for different statistics.

2.2 Key Steps – Summary

I. Determine user's requirements for data.

↓

II. Understand the key characteristics of the data.

↓

III. Are there circumstances where disclosure is likely to occur?

↓

IV. If so, would disclosure represent a breach of public trust, the law or policy?

↓

V. If required select appropriate disclosure control methods to manage the risk.

↓

VI. Implement and disseminate.

2.3 Key Steps – Explained

I. Determine user's requirements for data

Statistical data should be designed according to the needs of the users: identify the main users of the statistics required, understand why they need the data and how they will use it. This is necessary to ensure that the design of the output is relevant and that the measures of disclosure protection applied has the least possible adverse effect on the usefulness of the statistics.

It is important to note in the Revenue context that data are collected for tax administration purposes (and to a minimum required level to reduce burden on the taxpayer). Transforming tax data into statistical data is often challenging and the limitations of Revenue data need to be understood by users. In assessing user requirements, the question of whether the proposed use of the data could or ought to be performed within Revenue should be considered. Whether alternative sources might be more appropriate (e.g. Central Statistics Office data) should also be considered.

II. Understand the key characteristics of the data

Typically statistical units refer to individuals, taxpayer units or businesses. It is important to assess which units are represented and identify any sensitive variables. Sensitive variables including names (companies or individuals), addresses, PPSN or other identifying numbers should not be disclosed. Beyond this there is a wide range of variables or statistics that may present risks depending on how these are presented. The source of the data may affect the need to protect confidentiality.

It is important to consider the characteristics of the tables and determine whether the table is deemed to be **Low Risk** or **High Risk** as noted earlier.

III. Are there circumstances where disclosure is likely to occur?

A risk assessment should be undertaken to develop suitable confidentiality protection. The risk assessment should include the nature of the variables and the structure of the table (the number of units in the table and their distribution). Disclosure risk is high when a table is designed so that there are cells in the table with **low frequencies** or when there are rows or columns where all the counts are in a small number of cells.

Below is an extract from a hypothetical statistical report. It shows only 2 cases that claimed a certain relief in a year. It is possible for one of the claimants to calculate exactly the amount claimed by the second case by subtracting their own amount from the total. They may add this to their own knowledge or to other sources of information to determine who the other taxpayer is, in addition to the amount of relief used. In the example shown below, SDC measures should be considered

Example where SDC measures are required

Relief Number	Number of Cases	Amount of Relief used
0001	2	323,586

In broad terms the following criteria[4] should be considered in assessing disclosure risk:

- Threshold Rule: Have all outputs at least 10 units underlying any cell or data point presented?

- Group Disclosure: In all tabular and similar output does any cell contain more than 90% of the total number of units in a row or column?

- Dominance: In any tabular or similar data, does the largest contributor to a cell exceed 50% of the cell total?

- Is the data within the dataset disclosive (e.g., contain PPSN, Name, Address or any other identifier)?

- Can the units making up the data or subsets be identified?

- Is the rank ordering of contributors known or inferable (is the largest/smallest/latest etc. contributor known or guessable)?

- Does the context of the data present a disclosure risk (e.g., level of geographic breakdown, detail of sectoral detail etc)?

- If maxima, minima or percentile values are sought, do they refer to single data points? Particularly relevant to maxima and minima values, which may be combined with other information and may lead to the indirect identification of taxpayers.

- If percentile breakdown sought (e.g., decile, ventile etc.) does each unit pass the Threshold, Group Disclosure or Dominance tests?

- Where it is determined that Graphs are permitted as an output can data points be identified with units? Are there significant outliers?

- Formatting of outputs should ensure that disclosive data are not embedded in charts in documents or in "hidden" columns/rows in spreadsheets.

IV. If so, would disclosure represent a breach of public trust, the law or policy?

Revenue depends on the cooperation and trust of citizens and businesses in Ireland. In maintaining this trust, sensitive personal records need to remain confidential. On the other hand, there is a legitimate public interest need to have access to statistical information.

As previously outlined, there are significant legislative and regulatory safeguards in place to protect Taxpayer confidentiality. In the Revenue case, the legislative requirements underpinning taxpayer confidentiality override any public interest access to Revenue data at a granular level.

---

[4] See Statistical Disclosure Control (2012): Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, Peter-Paul de Wolf : Wiley 211-223.

V. If required select appropriate disclosure control methods to manage the risk.

Each request should be evaluated and risk assessed in the context of the aforesaid obligations. This evaluation should inform any decision on whether or to what extent the requested data should be disclosed, and if so whether SDC methods should be applied before disclosure occurs.

Where required, SDC methods can be used to reduce the risk of disclosure by modifying risky data structure in a table. SDC methods include:

a. Table Redesign

Table redesign is recommended as the initial method of disclosure control but should be balanced against user needs and publication plans.

Method	Description	Advantages	Disadvantages
Table Redesign	Protect unsafe cells by grouping categories or ranges within a table.	Original counts in the data are not affected.	Detail in the table reduced.

Example of table redesign method used on a hypothetical statistical table

Range of Total Income From €	To €	Number of Cases	% of Total	Reduction in Tax €m	% of Total
-	10,000	28	2.30	0.01	2.62
10,000	12,000	19	1.56	0.01	1.85
12,000	15,000	48	3.95	0.02	3.54
15,000	17,000	34	2.80	0.01	1.81
17,000	20,000	59	4.85	0.02	4.81
20,000	25,000	131	10.77	0.05	9.70
25,000	27,000	41	3.37	0.01	2.26
27,000	30,000	85	6.99	0.03	6.01
30,000	35,000	160	13.16	0.05	10.54
35,000	40,000	108	8.88	0.04	8.01
40,000	50,000	205	16.86	0.09	17.47
50,000	60,000	119	9.79	0.06	11.85
60,000	75,000	101	8.31	0.05	11.02
75,000	100,000	55	4.52	0.03	5.98
100,000	150,000	17	1.40	0.01	1.70
150,000	200,000	4	0.33	0.00	0.53
200,000	275,000	1	0.08	0.00	0.13
>275,000		1	0.08	0.00	0.18
Totals		1,216	100	0.49	100

Numbers shown in red are below 10 units, so are deemed to be unsafe to publish.

Redesigned table

Range of Total Income		Number of Cases	% of Total	Reduction in Tax	% of Total
From €	To €			€m	
-	10,000	28	2.30	0.01	2.62
10,000	12,000	19	1.56	0.01	1.85
12,000	15,000	48	3.95	0.02	3.54
15,000	17,000	34	2.80	0.01	1.81
17,000	20,000	59	4.85	0.02	4.81
20,000	25,000	131	10.77	0.05	9.70
25,000	27,000	41	3.37	0.01	2.26
27,000	30,000	85	6.99	0.03	6.01
30,000	35,000	160	13.16	0.05	10.54
35,000	40,000	108	8.88	0.04	8.01
40,000	50,000	205	16.86	0.09	17.47
50,000	60,000	119	9.79	0.06	11.85
60,000	75,000	101	8.31	0.05	11.02
75,000	100,000	55	4.52	0.03	5.98
>100,000	150,000	23	1.89	0.01	2.54
Totals		1,216	100	0.49	100

Ranges from 100,000 and over have been grouped into one range for confidentiality reasons. Totals remain the same but detail of breakdown above 100,000 is reduced.

b. Cell Modification – Suppression (Primary and Secondary)

Cell Suppression is not generally used in Revenue currently, but examples of Primary Suppression and Secondary Suppression are shown below.

Method	Description	Advantages	Disadvantages
Primary Cell Suppression	Unsafe cells are not published. They are suppressed or replaced by a special character, such as 'x'.	An original count in the data that is not suppressed is not affected.	Most of the information about suppressed cells will be lost
Secondary Cell Suppression	Cells that are deemed to be safe are suppressed to avoid recalculation of un-safe cells.		Secondary suppressions will hide information in safe cells.

Example of Primary Suppression

Applying a frequency rule of 10, any cell with less than 10 contributors poses a disclosure risk. In this example, the age group 50-59 years has **4** contributors in the low income cell and needs protection. This cell could be protected by suppressing it, as shown by the **'X'**.

Age group (years)	Income			
	Low	Medium	High	Total
15–19	20	10	15	45
20–29	14	11	11	36
30–39	10	12	12	34
40–49	11	18	24	53
50–59	4 X	10	14	28
60+	12	11	12	35
Total	71	72	88	231

However, it would still be possible to work out the value of the cell by using the remaining figures. For example, low income earners aged 50-59 could be re-calculated by subtracting the medium and high income earners from the total **(28 minus 10 minus 14)** or by subtracting the number in each age group in the low income column from the total in the low income column **(71 minus (20+14+10 +11+12))**. Subsequent suppression is needed to protect the unsafe cell. In order to make the table safe, secondary suppressions are necessary.

Example of Secondary Suppression

In this table the cell with an **X** indicates the unsafe cell (the primary suppression). The cells marked **Y** would also be suppressed to prevent the unsafe cell being recalculated from the totals (secondary suppression).

Age group (years)	Income			
	Low	Medium	High	Total
15–19	20	10	15	45
20–29	14	11	11	36
30–39	10	12	12	34
40–49	~~11~~ Y	~~18~~ Y	24	53
50–59	~~4~~ X	~~10~~ Y	14	28
60+	12	11	12	35
Total	71	72	88	231

VI. Implement and disseminate

The final stage in the process prior to publication is implementation of chosen SDC methods and dissemination of the data to users.

The most important consideration when disclosing data is maintaining confidentiality. When applying disclosure control methods to statistical outputs, consideration must also be given to the relationship between risk and utility. Ideally the application of disclosure controls will result in an output dataset from which there is minimal risk to taxpayer confidentiality while maximising value to the user.

When data are published on the Revenue website, the conditions under which it may be re-used should be stated. It is clear what if any licence agreement (e.g., CC-BY licence) is applicable. The status of the data should also be stated, so where figures are provisional or preliminary an appropriate disclaimer clarifying this status should be appended.

In delivering/publishing data, especially where it has been determined that it is necessary to remove specific fields for disclosure purposes, it important that possible hidden fields and meta-data do not constitute a disclosure risk. For example, where it has been

determined that it is necessary to remove specific fields for disclosure purposes, the data in a cell/row/column should be deleted from the spreadsheet/table/document/etc rather than hidden.

A further consideration relates to Revenue data exchanged with the CSO.[5] Once received by the CSO, Revenue data may be accessible to relevant researchers under that Office's Research Microdata File (RMF) process.[6] RMF access to Revenue data follows CSO SDC protocols following consultation Revenue as laid out the Memorandum of Understanding between the two Offices. This may include a combination of CSO SDC protocols (e.g., removal of outliers such as the 50 largest groups for Corporation Tax datasets) and Revenue's SDC (e.g., no disclosure for less than 10 cases).

---

[5] A Memorandum of Understanding outlines the data cooperation between the two offices: https://www.cso.ie/en/aboutus/lgdp/legislation/memorandumsofunderstanding/mourev/.
[6] See https://www.cso.ie/en/aboutus/lgdp/csodatapolicies/dataforresearchers/researchmicrodatafilesrmfs/.

3 Conclusion

This guide provides a protocol to formalise certain procedures already implemented in the Revenue publications. It also documents why SDC are required in Revenue, should publication policies be questioned.

In general, the primary disclosure control approach adopted in Revenue is to not provide figures where numbers are shown to be below 10 units. This threshold applies to individuals, companies or groups depending on the statistics being produced. Where requests for information on individual units are received, these are not provided. This includes where information is sought, for example on the minimum or maximum in a dataset or a list of the largest amounts/claimants for an indicator.

Even where there are larger numbers of cases, if there is a degree of dominance (e.g., where the largest contributor to a cell or dataset accounts for 50% of the total), this also represents a disclosure risk and may not be suitable for release.

Any changes in disclosure control rules for a published statistic raises the issue of revisions to previous releases. In general new disclosure control rules will be implemented for future releases with the rules not being applied to past releases. An exception can be made in cases where the disclosure controls are altered to allow more data to be released. Here it may be feasible to re-release older data with greater detail.

Another factor to be considering when publishing updates is whether changes in the update (compared to the previous period) may itself lead to disclosure risks. This will be reviewed on an ongoing basis as new tables are published.

It should be noted that statistics published on the Revenue website are subject to Revenue's Re-use of Public Sector Information policy[7] which conforms to the Creative Commons Attribution Licence (CC-BY 4.0 International) standard. Under the PSI licence, the users of Revenue information must:

- Acknowledge the source of the information by including the following attribution statement:
  - o 'Information provided courtesy of the Revenue Commissioners under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence'.

---

[7] Available at: https://www.revenue.ie/en/corporate/using-revenue/reuse-of-public-sector-information/index.aspx.

and, where possible, provide a link to the licence.

- If using information from several providers and multiple attributions are not practical for the product or application, users must use the following attribution statement:
    - 'Contains Irish Public Sector Information licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence'.

Aside from published statistical tables, Revenue receive a considerable volume of requests for data or statistics – in the form of PQs, media queries, queries from other Government departments and members of the public. The same disclosure risks need to be considered for such requests and the approach to controls put in place as required.

The following disclaimer is added by Statistics & Economic Research Branch whenever releasing statistics (except with Revenue colleagues, the Department of Finance (including PQ responses) or the CSO):

"The information provided herein has been collated on the basis of the latest data available to Revenue's Statistics & Economic Research Branch at the time of response. These data may be subject to future update and adjustment. The information is furnished on the same basis as material published on our website, the terms of which are detailed on Revenue's Re-use of Public Sector Information webpage."

In general, Revenue do not provide data in spreadsheet format when responding to requests, instead the data is given in plain text in the body of an email. Where a spreadsheet is required, for example when dealing with large amounts of data, Revenue has designated "gatekeepers" who are responsible for assessing each spreadsheet against a set-list of pre-checks to ensure, for example, that there are no filters which have been applied to the data which can be rolled back, that there are no hidden rows/columns which contain data in breach of the SDC rules, etc.

When looking at published statistics, users should be informed that the dataset has been assessed for disclosure risk, and any methods of protection required have been applied.

Finally, it is hoped that publication of this protocol will provide a useful context and understanding for consumers of Revenue statistics.

Any queries on statistical disclosure can be addressed to statistics@revenue.ie.