EUROPEAN COMMISSION
EUROSTAT

Directorate A – Resources
A.5 – Methodology; Innovation in official statistics

Understanding output checking

Deliverable 1 of the Eurostat specific contract: Access to European microdata in Eurostat safe centre: automatic checking of the output

Eurostat Framework Contract N° 2018.0086

Ref. No.: ESTATMET2-000058-6000054239-REQ-01

Contractor:

GOPA Worldwide Consultants in joint venture with GOPA Luxembourg

Authors:

Elizabeth Green, Felix Ritchie, and Jim Smith

University of the West of England

May 2020

Contents

Abbreviations

AI	Artificial Intelligence
CIMES	Centralising and Integrating Metadata from European Statistics
DWAR	Discretionary within-activity review
EoAR	End-of-activity review
ESS	European Statistical System
MISSY	Microdata Information System
ONS	Office for National Statistics
OSDC	output SDC
PBOSDC	Principles-based output SDC
RBOSDC	Rules-based output SDC
RDC	research data centre
RDF	Resource Description Framework
RRSA model	Runners-repeaters-strangers-aliens model
SDC	Statistical Disclosure Control
SO	statistical organisation
SWAR	Seamless within-activity review

Acknowledgements

We are grateful to the Methodology Team at Eurostat and members of the ESS Expert Group on SDC for their extensive and detailed comments on versions of this document. All remaining errors and omissions are the responsibility of the authors. The views expressed in this document are those of the authors and do not necessarily represent the views of Eurostat or any other agency.

1 The purpose and action of output checking

1.1 The output checking process

When analysts produce statistics from confidential data, a residual risk exists that the outputs might breach confidentiality. For example, a table may disclose that an individual with unusual characteristics has a rare illness, or that a group of survey respondents all claim to have taken illegal drugs. Avoiding such a 'disclosure' is a concern of the statistical organisations (SOs) which provide the data.

The risk of such a disclosure increases with the sensitivity and identifiability of the data. The highest-risk data is typically made accessible to analysts through 'research data centres' (RDCs). These allow the analysts full access to view, manipulate and model the data in an environment controlled by the SOs. The secure RDC environment typically blocks analysts from accessing the internet, uploading or downloading data, sharing with others, or otherwise removing or re-identifying the data. In addition, analysts using RDCs undergo some form of data governance training (ADSS, 2016, appendix) to ensure that they follow procedures.

Nevertheless, despite the user training and the secure environment, there remains the possibility that an analyst may accidentally produce a disclosive output from this very sensitive data. Therefore all RDCs operate some process to check outputs. Historically, statistical disclosure control (SDC) focused on two topics: anonymisation of data, and protection of tabular statistics, and tools were developed to deal with these in SOs (muArgus/sdcMicro and tauArgus/sdcTable, respectively). However, in the last decade or so there has been a recognition (Ritchie, 2007) that SDC to cover the full range of analytical outputs requires a much wider range of techniques; these are often referred to as 'output SDC' (OSDC, to distinguish it from 'input SDC', the removal of identifying information from microdata before giving the microdata to analysts); tabular data protection is a subset of OSDC.

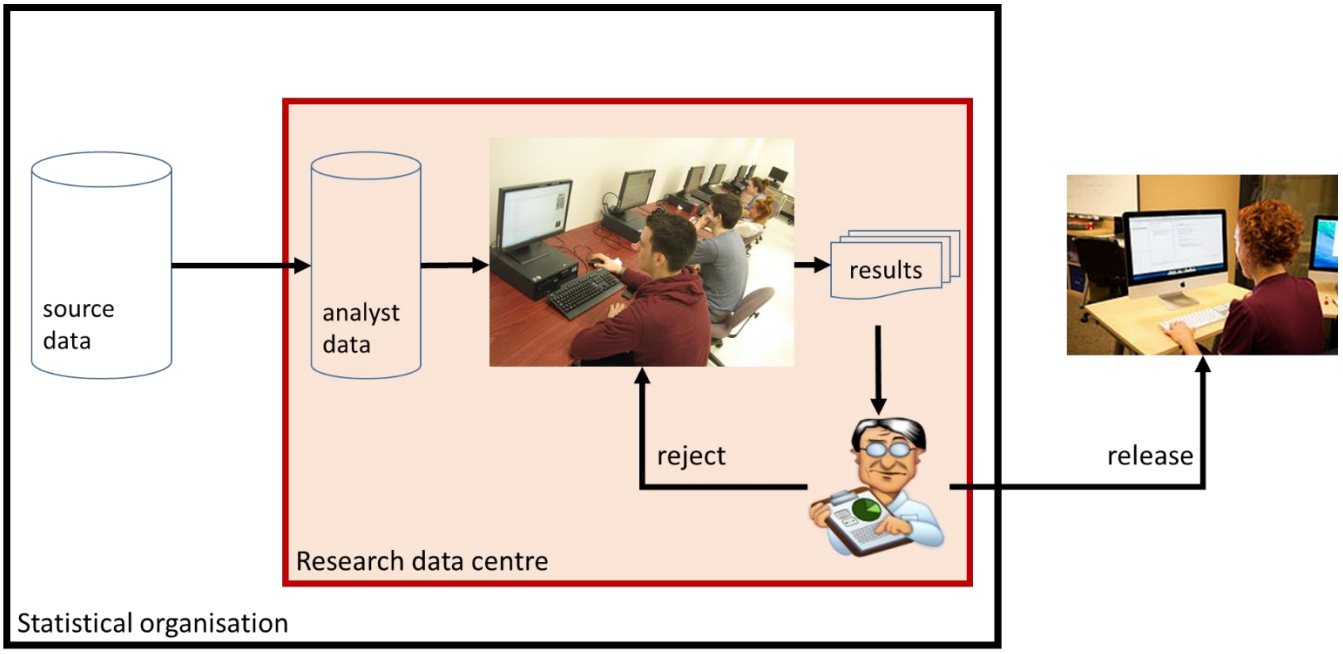
This terminology is not universal, for example, Statistics Netherlands uses 'output checking' to refer to both the statistical analysis and the operational procedure. For clarity, in this document we use 'OSDC' to refer to the collection of statistical techniques, and 'output checking' to refer to the workflow/operational procedures followed by the SO.

The RDC that Eurostat operates is a typical example:

- The researcher physically visits the RDC in Luxembourg
- The researcher has full access to data which has had some basic confidentiality protection applied
- Statistical results are produced within the RDC environment
- The RDC manager is notified of the results that the researcher wants released from the secure environment
- The RDC manager applies OSDC guidelines and then either releases the result to the researcher outside the RDC, or rejects the output as too risky for publication

See Figure 1 below. Most RDCs operate in this broad way, albeit with differences in whether the RDC is visited physically or remotely, whether all output is checked, and whether researcher can apply the necessary checks themselves and make release decisions.

Figure 1: Schematic representation of a research data centre



1.2 Statistical components of OSDC

Research outputs present a very low base level of risk: data are typically sampled, manipulated and transformed in ways that make the disclosure risk negligible in realistic environments. Researchers also tend to be interested in low risk outputs such as estimation; high-risk outputs such as tables are used to describe the analytical dataset and so are likely to have large numbers of observations.

Nevertheless, OSDC is used to reduce that risk still further by imposing restrictions which guard against errors and provide an extra margin of safety. This is acceptable to analysts because statistical validity (such as many observations, or lack of extreme values or outliers) is generally consistent with good confidentiality practice.

In this section we consider the underlying statistical issues.

1.2.1 Identification and association

Disclosure may take the form of (a) identity disclosure – that is, identification of an individual or group of individuals, and/or (b) attribution disclosure – the association of some protected information with that individual. As the latter is generally not feasible without the former, attention tends to focus on the identification of individuals by, for example, discovering that a combination of categories can be used to identify one individual.

Identification is normally required to be exact for a disclosure to take place; that is, a person or group is unambiguously identified. In contrast, the association does not need to be exact to be disclosive. For example, an exact disclosure would demonstrate that the richest person in a named village earned €765,432 in a year. An approximate disclosure would show that the richest person in the village earned between €740,000 and €780,000 in the year; whether this counts as a disclosure in practice would depend on the tolerance of uncertainty at the SO.

1.2.2 Primary disclosure

Primary disclosure is where a statistic directly allows inferences to be made about individuals from the information in that statistic.

Frequencies allow individuals or groups to be identified from the categories used for the statistic (e.g., a single company making electric wheelchairs in Belgium; all individuals in a survey identified as diabetic also have the fragile-X gene).

Magnitudes allow values to be ascribed to individuals or groups (maximum earnings in a village; minimum research and development spend in an industry). Magnitudes are also susceptible to dominance. For example, turnover and employment in a small town may be dominated by a large international company headquartered there, so that statistics are largely approximations of the values for that one big organisation. Dominance is almost impossible to assess simply by looking at an output, but requires detailed knowledge of specific values used to construct the dataset.

Primary disclosure is managed by simple rules. Minimum thresholds ensure that there are sufficient observations to be uncertain about any individual respondent. Even with minimum thresholds, if the contribution to a statistics is dominated by a few individuals, this may allow large respondent values to be approximated. Dominance is managed by the 'p%-rule': ordering observations 1..N with the largest first, there is no dominance if the sum of observations 3..N is at least p% of largest observation. Dominance can also be managed by an 'N, K' rule, where the largest N observations must contribute less than K% of the total. These are easy to calculate, but is often ignored by researchers.

One implicit problem of primary disclosure is the case of non-complete categories: for example, only display the percentage of male respondents and omitting the proportion of females. The same simple rules clearly apply, but the problem lies in noting that the set of categories is not complete. As Hundepool et al (2012) notes, the solution is to require analysts to produce a complete set of categories, but this is not necessarily enforced.

In the case of primary disclosure, automated output checking may have an advantage over human analysts. Checking for dominance is a deterministic procedure, as is ensuring that the full set of categories is accounted for. An automated solution may be able to overcome the lack of interest in these checks shown by human analysts.

1.2.3 Class disclosure

Class disclosure, also known as group disclosure, occurs when all or none of the observations fit into a particular category (some SOs define class disclosure as 'almost every/almost no observation' but this is a risk management definition rather than a statistical one). Rather than a specific individual being identified, it is enough to know that an individual fits into a group about which something is known. For example, if a study reports that all students surveyed have tried cannabis, then this is a disclosure for any student who can be identified as a survey respondent, without any other information.

Class disclosure breaches the general rule that good statistics and disclosure protection go hand-in-hand. The most valuable class disclosures are also likely to be the most sensitive; for example, a disclosure that no-one living in a poor ex-mining community earns over €500m per year is technically but not practically disclosive, and of little value. In contrast, a disclosure that no-one in the area earns over €20,000 per year is both an important statistical finding, and very informative about the people in that area.

Class disclosure can be dealt with by disallowing the publication of figures where 0% or 100% of the population can be put into one cell. However, this is complicated by the existence of 'structural zeroes'. These are cases where the expectation is that the cell is zero. For example, when tabulating qualifications of different age groups, it would not be disclosive to show that none of the 16-17 year olds have acquired a university degree. A blanket rule of "no 0% or 100%" could also create perverse incentives. For example, a statistics "0% of this group had a college education" could be blocked by "37% left school with no qualifications, and 63% achieved school qualifications but nothing higher" is much harder to pick up. For these reasons, class disclosure is the hardest primary disclosure risk to assess automatically.

1.2.4 Secondary disclosure

Secondary disclosure is where the statistic can be combined with other information to reveal identities or values. Examples of secondary disclosure include suppressed cells being reconstructed using totals or other statistics.

The major problem for secondary disclosure is differencing; that is comparing two similar statistics that only differ in a known way (one more observation, say, or the data being broken down into different overlapping categories). The difference between the two statistics can be used to compute both frequencies (differences in the number of observations in categories) and exact values (for example, by reverse engineering two means which differ only by a single observation.

Secondary disclosure is an unsolvable problem from a theoretical perspective (Ritchie, 2019b) as all possible past and future outputs are in scope for differencing. Even if secondary disclosure checks are restricted to an archive of N previous outputs, there remains a major computation problem. Taking a pragmatic approach to developing automated testing, it would be possible to check a new output against the archive items individually in linear time – that is to say that N pairwise checks would be needed. If it was felt that the new output would be compared against all possible differenced pairs of existing outputs, there would be $O(N^2)$ of these; comparing against triples of outputs would scale as N^3, etc. By contrast, Moore's Law suggests that computing power doubles every 18 months.

Hence, all SOs develop rules as to what constitutes a 'reasonable' set of outputs to check (for example, recent outputs by the same researcher, or outputs produced by the SO from the same dataset). This is a context- and technology-specific issue which is likely to change over time (Ritchie and Smith, 2018): a decision needs to be made about the likely growth rate of outputs to be checked (and in the archive), the risk of disclosure by differencing from 2 or more existing outputs, and the cost/availability of compute resource. Any automated output checking system therefore needs to allow for the subjective preferences of the SO.

A second differencing issue concerns the semantic (ie 'meaningful') characteristics of the data. Consider the properties of a variable called 'age': which represents

1. a transitive relationship (x>y and y>z means x>z).
2. a temporal relation (in a fixed population the people age over time)

Awareness of the transitive property of the age variable allows for automated reasoning about set inclusion. For example, if two outputs taken from the same population survey showed that (i) the maximum salary for people between 18 and 23 was €25,000, and (ii) the maximum salary for people between 18 and 24 was €35,000; then it can be inferred that someone in the group of 24 year-olds earned €35,000.

However, the property of being temporal means that if the two outputs above came from surveys X months apart, you would only be able to make *probabilistic* inferences. These would have to take into account the rates at which people aged, and the distributions of age and birth-months within a population.

Not all properties that might be coded as continuous variables describe transitive or temporal relationships. Road distances are not transitive, due to the existence of one-way streets; height has no unambiguous temporal component.

Frameworks for capturing this 'meta-data' about variables, and so making datasets 'self-describing', do exist. In the European official statistics context, the DataWithoutBoundaries[1] project developed two metadata systems (CIMES, MISSY) with detailed information on a large number of datasets. 'Meaning' was introduced by identifying synonyms within and across languages for descriptors, but full conceptual meaning was beyond the scope of the project. The internet standards consortium responsible for creating the 'Semantic web' designed methods such as the Resource Description Framework (RDF) for unambiguously storing some properties to facilitate automated reasoning. A fully described variable would allow checking by an automatic process which 'understands' the data; this could also include derived variables.

While noting that adding full resource descriptors to Eurostat's datasets is not feasible within this project timeline or budget, identifying how such information could be recorded will be useful for *future* funded work. Thinking along these lines may also be useful to inform *current* practice: for example, prompting the SO to make and record policy judgements about when the time-gap between surveys meant that 'temporal' variables can be excluded from differencing.

[1] https://cordis.europa.eu/project/id/262608/reporting

1.2.5 'Safe' and 'unsafe' statistics

Ritchie (2008) introduced the concept of 'safe' and 'unsafe' statistics (originally referred to as 'safe and unsafe' outputs, but the term was changed to avoid confusion). These are sometimes now referred to as 'low review' and 'high review' outputs in user training such as ONS (2019).

A 'safe statistic' is one where there is no meaningful disclosure risk in the statistic because of its mathematical form, an example is a linear regression coefficient (Ritchie, 2019a). Safe statistics may have some associated rules (for example, in the case of linear regression coefficients, that the explanatory variables do not consist solely of categorical variables with all interactions included as regressors). These rules need to be capable of being easily, quickly and reliably checked.

An 'unsafe statistic' is one where the mathematical form allows for primary or secondary disclosure without exceptional difficulty. For example, the mean is an 'unsafe statistic' as it is plausibly susceptible to primary disclosure (singe observation, dominance) and secondary disclosure (by differencing). For an 'unsafe statistic' to be released, it needs to be demonstrated that the output, in the specific instance being considered, has no significant disclosure risk.

Questions of primary and secondary disclosure only apply to 'unsafe statistics':

Type		Action		Decision
Safe statistic	→	Check rules	→	Release if rules are met
Unsafe statistic	→	Check primary and secondary disclosure	→	Release if non-disclosive

Safe statistics can therefore be checked automatically.

Bond et al (2015) provides a current list of safe and unsafe statistics.

1.2.6 Actual versus potential disclosure

The above discussion identifies potential disclosure based on implicit worst-case assumption that the identification of a sample unique is a breach of confidentiality. As noted in Hafner et al (2015), this is a sensible criterion in the context of methodological advance. However, in genuine analytical environments the worst-case assumption is difficult to justify; there is little support for the claim that a single observation or an empty cell definitely exposes confidential information about a known individual or groups.

Unfortunately, it is hard to develop rules as to what turns a potentially disclosive output into a non-disclosive one. Converting values into deviations from means, for example, might be an appropriate transformation in some cases, but not others.

SOs persist with the worst-case assumption for two reasons. First, as noted above, OSDC rules can generally be accommodated in genuine research; experience of research centres over the last fifteen years or so shows that a cautious approach does not normally have a significant impact on usefulness of outputs. Second, SOs are concerned about perceptions of confidentiality breaches: a single observation may be non-disclosive, but it could be mis-interpreted as disclosive and used to attack the reputation of the SO. For both these reasons, SOs are expected to be more cautious than the strictest statistical case justifies.

It has been argued (eg Ritchie and Smith, 2018) that growth in computing power, accessible data sources such as social media, and machine learning will increase re-identification possibilities in the future, and therefore SOs should take a position which is stricter than currently justified to prevent future breaches of confidentiality. It remains to be seen whether this will affect research outputs; given the unknowability of the future this is a pure judgment call for the SO.

1.3 Implications for output checking

The above discussion highlights areas of possibility and limitations for automatic output checking.

- Primary disclosure and 'safe statistics' are both areas where simple rules can give unambiguous and consistent clearance decisions
- Automated checking may be better placed to solve some of the process mistakes made by individuals
- Secondary disclosure is likely to involve, at minimum, a number of arbitrary rules about what to check
- Understanding the 'meaning' of variables may affect their perceived disclosiveness
- Assessing the *actual* disclosure risk of the outputs is probably beyond the scope of an automated procedure at present, except in very simple cases
- In summary, from a statistical perspective, automatic output checking will require subjective decisions reflecting the SO's perception of risk to be incorporated into the model.

2 Operational aspects

Alves and Ritchie (2019) note that output checking is not a statistical problem, but on operational problem. That is to say: the statistical issues discussed above come into play when faced with an output to evaluate, but the operational decisions that lead up to whether an output is reviewed by a human, a machine, or no-one, are much more important for both resource use and confidentiality protection.

Accordingly, this section considers operational aspects and how those may be reconsidered in the light of automatic output checking.

2.1 Theory

2.1.1 The RRSA model

Alves and Ritchie (2019) use a customer-segmentation model from management literature to separate outputs into four types:

- Runners: the bulk of outputs which can be dealt with simply automatic rules (for example, a simple table with a threshold rule, or regression output)
- Repeaters: a small but significant part of outputs which require some human intervention to interpret rules and guidelines (for example a scatter plot of residuals)
- Strangers: rare events which require statistical knowledge and should lead to the development of new rules and new types of runners/repeaters
- Aliens: output requests which are outside the scope of the service, such as asking for microdata to be released

Automated output checking should be expected to deal with runners, and may be able to deal with repeaters under certain conditions (perhaps via adaptive limits for tolerance, or, in the longer term, some form of machine learning). Automated checking should not be expected to deal with strangers or aliens, other than to highlight that something unexpected has been presented.

This establishes that automated checking cannot be a universal solution. Rather, its goal should be to reduce human intervention to the minimum necessary. This is not necessarily a bad thing. Knowing that human judgment can be exercised in exceptional cases is important to researchers, and helps to build the bridge of trust between researchers and the support team. The key to automated output checking is to ensure that the time of the output checkers is used most effectively, doing things that only a human can do, or that a human does best.

2.1.2 Training of analysts

The analyst translates confidential data into non-disclosive outputs. This is not a neutral process; the actions of the analyst affect the resource cost to the SO and the well-being of the analyst.

While the focus of analysts is to produce analytical results, they are also, generally, incentivised to produce non-disclosive outputs. Incentives are negative (producing disclosive outputs can put you in jail) and positive (outputs with no disclosure risks will be released more quickly than outputs which cause concerns).

All RDCs provide information to analysts on the importance of producing non-disclosive outputs, through documents, online courses, or face-to-face training (Green and Ritchie, 2015, appendix). The difficulty is that analysts, being human, are unwilling to acquire information unless the information is useful, interesting, or necessary (Green et al, 2017). Analyst training can easily focus on the ‘necessary’, which means that analysts view output checking as an obstacle to be overcome. Even training which focuses on the ‘useful’ (Eurostat,

2015) or on positive messages (ONS 2019) may be quickly forgotten if producing good output does not show any obvious advantages for the analyst.

2.1.3 PBOSDC vs RBOSDC

When the SO directly checks output (as opposed to letting researchers self-check), Ritchie and Elliott (2015) and Alves and Ritchie (2019) distinguish between two approaches to the output checking process. These have important implications for automated output checking.

Rules-based OSDC (RBOSDC) takes a strict position on the clearance – either something is cleared in accordance with a pre-defined rule, or it is not. If there is no rule, a rule needs to be defined. There are no exceptions. In the RRSA model this means that there are only runners (yes/no outputs) and strangers (reasons to develop new rules).

This system works well for official statistics; it clearly well designed for automatic output checking, and official statistics systems can make use of formal output-checking tools such as tau-Argus. The problem is that this is poorly designed for research outputs: (1) it is hard to define sufficient and sufficiently accurate rules; (2) each rule has to do two jobs, confidentiality protection and releasing useful results; and (3) it takes no account of variable transformations, subsampling, or subjective categorisation.

Principles-based OSDC (PBOSDC) uses rules-of-thumb rather than hard rules. All outputs are potentially allowed if the analysts can demonstrate that[2]:

I. The output is non-disclosive
II. The output is important to the analysts work
III. This is a rare request for an exception

Rule (i) is obvious. Rule (ii) ensures that the output checkers only spend time checking outputs that are worthwhile. Rule (iii) ensures that analysts do not abuse the system. If rules (ii) and (iii) do not apply, the rules-of-thumb are treated as hard rules. The advantages of this process are that: (1) rules-of-thumb can be simple as they do not need to cover all cases; (2) rules-of-thumb can focus on confidentiality, as usefulness is addressed through the exception process; and (3) output checkers can reject any analysis on the grounds that the analyst is abusing the system, irrespective of whether the output is non-disclosive or not. This is an extremely efficient process, as it embeds the analysts as part of the safe-output production process (Alves and Ritchie, 2019). The disadvantage is that this system requires active buy-in from the analyst, which in turn requires specialist training.

In the context of the RRSA model, runners, repeaters and strangers are allowed. Repeaters can include runners where an exception is being sought (e.g., tables with cell counts below the threshold). Strangers are acceptable, in their own right and as a reason to develop new runner/repeater guidelines.

Earlier it was noted that the runners fit directly into automated output-checking, the repeaters and strangers less so. The output-checking system therefore has implications for automation:

	Principles -based	Rules -based	Automatic?
Runners	Allowed	Allowed	Yes
Repeaters	Allowed	Not allowed	No
Strangers	Allowed	Not allowed but a new rule can be defined	No

[2] The ESSnet *Guidelines for Output Checking (Bond et al, 2015)* assume PBOSDC but do not explicitly describe how to operationalise it or train researchers. ONS (2019) provides a modern formulation and examples.

The organisations approach to flexibility therefore has an impact on the design of the tool. An organisation using a rules-based system should be able to adapt more easily to automatic checking. A principles-based organisation needs to allow a mechanism for human engagement, including requests for exceptions.

2.2 Organisation of output checking in ESS and other RDCs

This project is intended to provide a solution for the particular circumstances of Eurostat. However, in this section we consider practices in other organisations running similar restricted-access RDCs. The experience of other organisations in respect of procedures, user training, clearance regimes, and user expectations can provide insights into the feasibility of the non-statistical elements of the solution.

This is not an exhaustive review, but as part of this report we reviewed presentations supplied to Eurostat, plus web searches and our knowledge of other facilities. We have also used results from a small survey of RDCs carried out in 2016 (and which may be out of date). Unless otherwise described, the examples are taken from the National Statistical Institute of that country mentioned. Information is only taken from sites or presentations in English, and so is likely to have missed much of the local guidance.

All the RDCs follow the model depicted in the diagram in Section 1.1. Most offer social data; a smaller proportion offer business data. However, they differ in many practical respects. We considered

- Charges
- SDC regime
- Researcher training, accreditation and guidance
- Expectations of research users and RDC managers

2.2.1 Charges

There is little consistency over countries in the approach on charges. Portugal, the UK (all RDCs, since 2012), Slovenia, Eurostat, the European Central Bank and Australia have no charges for access or clearance. In other countries, organisations charge based on

- Dataset preparation
- The number observations and/or variables
- The number of times the facility is visited
- Provision of physical facility by the user community

It is not clear whether the aim of charging is to cover the marginal costs of provision of a service, or to cover the total cost of the service across all users. For example, in Germany, the fee is intended explicitly to cover the full costs of user support; the Slovakian fee covers the set up cost, not outputs; and the UK fee (charged 2004- 2010) was an arbitrary day rate unrelated to actual costs.

Where output is charged for (not all organisations do this), this appears to be targeted at the marginal cost of the service. Some organisations charge for the volume of output. The Netherlands does this but gives each user an initial ‘free’ allocation, and also offers a zero-cost ‘light’ output option if the researcher can produce something that can be checked in less than half an hour.

France and NZ both charge based on the amount of time spent checking outputs; in other words, directly costing the checking process. However, both give a generous allocation of ‘free’ outputs before charging kicks in. In France, 20 outputs with an average time of 30 minutes-worth of checking are allowed before charging is brought in. In New Zealand, researchers can also buy extra clearance time, but the basic allowance is very large so that most users are not expected to hit the limit. In Germany, the fee for the data access is €250 per year per ‘statistic’.

Although researchers presumably would prefer to have no cost for access or clearance, no countries indicated that user attitudes to costs significantly affected their views of the service.

2.2.2 Training, accreditation, and guidance in procedures

2.2.2.1 Prerequisites

Many organisations require a basic statistical knowledge but not all – the UK for example does not explicitly require it. Often, it is implicit in the application form. For Eurostat, the research competency of the organisation is assessed; this is then applied to all staff from that organisation who apply. In Belgium and Bulgaria, researchers explicitly need to demonstrate a basic prior understanding of SDC.

2.2.2.2 Training

ADSS (2016) highlighted a wide variety in training practices amongst 12 RDCs from around the world:

Is training provided on security awareness?	9 Yes 1 Optional
How is the security training delivered?	5 Face-to-face 3 Online course 1 Online guide
Is training provided on using the system?	7 Yes 4 No
Is training provided on statistics eg aspects of data linkage?	1 Yes 4 Optional 4 No
Is refresher training required?	1 Yes 3 Not recent 7 No
Are researchers trained in checking output for disclosure risk?	8 Yes 3 No

Data taken from survey of 12 RDCs in Europe, N. America, Oceania.
Source: ADSS (2016, Appendix). Not all questions answered by all.

Looking at Member States (MSs), this variety persists. While most organisations have guidelines for researchers, face-to-face is rare. In Europe, only the UK and Dutch appear to run regular mandatory face-to-face training in using the facilities and in SDC, although (as at February 2020) Bulgaria is investigating this. Eurostat and Canada both have one-to-one training sessions with researchers on their first arrival. Finland runs training and continuing development sessions, but these are not mandatory.

For some organisations, formal 'training' is interpreted as formal education, such as a University degree in statistics.

The Dutch RDC asks users a question on security and disclosure control at log-in (a wrong answer means login is delayed) to ensure that users get some regular reminders of the operations of the facility, SDC and good data management.

2.2.2.3 Accreditation

Accreditation of researchers as 'safe' (in whatever way this is interpreted) varies widely. In Hungary, the completion of the application process is also the accreditation. In Slovakia, the completion of the application process plus evidence of statistical qualification is the basis for accreditation. In the UK and Australia, persons are formally accredited as the result of simple checks on qualifications and mandatory training. In Finland, Canada and the European Central Bank, the signing of a 'pledge of secrecy' completes the accreditation; this is backed up by training in Finland, stringent personal background checks in Canada, but neither in the ECB.

Eurostat is unusual in that the research organisation (such as the university) is the accredited party. The individual researchers need to apply for access for a specific project, but the organisational accreditation means that the project is being assessed, rather than the individuals (as long as they belong to an accredited organisation).

2.2.2.4 Timing

Where a person is accredited as a result of training, the accreditation generally has a fixed life – for example, five years in the UK. Some countries (e.g. Hungary) have a 'once accredited, always accredited unless blocked' model. In Canada the pledge of secrecy is for life, but so it appears is the accreditation.

2.2.2.5 Guidance

Every organisation provides some written guidance in procedures and in SDC. There is much variation. For example, Hungary produces extremely detailed SDC guidance tailored to its researcher. In contrast, for SDC guidance UK RDCs generally refer users to training material and online resources, rather than bespoke publication.

2.2.3 SDC regime

2.2.3.1 Responsibility

There are four types of output checking

- Check everything (for example, Australia, Bulgaria, Eurostat, France, Germany, Hungary, Netherlands, Portugal, UK)
- Check random outputs (not currently practised in SOs checked for this report)
- Check everything for inexperienced researchers otherwise random (for example Denmark)
- Check nothing and trust researchers (ECB, Belgium, Hungary, Sweden)

In the last case, the assumption is that researchers have read, absorbed and applied the guidance provided.

Generally, the researcher is supposed to provide non-disclosive output; the purpose of the RDC team is to check for safety, not create safe output from researcher outputs.

In Denmark, researchers are required to run tau-Argus on produced tables to check for SDC; it is not clear whether they are required to provide proof that this has been done.

2.2.3.2 Rules vs principles

As noted above the SDC regime has an impact on tool design and functionality, as well as training expected of users and output checkers. In the survey in ADSS (2016), SOs were equally split between whether principles-based (PBOSDC) or rules-based (RBOSDC) was the dominant model. However, PBOSDC has increased in popularity generally since 2016, and in Europe in particular where information about international practices tends to disseminate quickly.

Most European RDCs now claim to be principles-based, with France being an exception. However, this does hide some differences between countries (and within countries such as the UK which operate multiple RDC models). For example, the 'rule-of-thumb' in the Hungarian research guidance appears to relate to population versus sample differences rather than the crude 'good enough' measure used in the UK.

2.2.3.3 Authority for decisions

Most countries have their own written guidance for researchers visiting their RDCs. Three main authorities are cited for the assertions and recommendations made:

- European Statistics Code of Practice
- Brand et al (2010, sometimes referred to as the ESS Expert Group report; incorporated into Hundepool et al, 2012), and its updated version Bond et al (2015, sometimes referred to as the DwB report); either may be referred to as *Guidelines for Output Checking*
- National legislation

Some (eg Netherlands) also direct users to the original academic papers, both those underlying *Guidelines* and more recent work.

The popularity of the European Statistics Code of Practice is odd, as it only contains very general statements and does not give detailed guidance. This may be why *Guidelines* is far and away the most common source cited for practical problems.

The Eurostat (2015) guide for safe microdata use[3], including SDC, is generally not referenced although anecdotal evidence suggests that this is a popular resource for researchers. ONS in the UK is unusual in not having its own guidance; it simply points researchers to Eurostat (2015), *Guidelines*, and academic papers. However, non-ONS RDCs in the UK have written guidelines (Griffiths et al, 2019), again drawing heavily on *Guidelines*.

2.2.4 User expectations and the purpose of SDC

Some MSs (such as Hungary and the UK) explicitly point out to researcher that SDC is about confidentiality and not research quality. For most countries this is not discussed in the materials available.

Most countries also assume responsibility for the confidentiality of output if it has been checked. Where output is not checked, researchers take responsibility for any errors. In some cases (for example, Sweden) the researchers are responsible for secondary disclosure, while ECB researchers are asked to assume responsibility for any possible negative outcome, including accidents, in perpetuity.

2.2.5 Summary and relevance

It is clear that Eurostat is not a major outlier in any of its practices, except in relying on organisational accreditation. This is helpful as it suggest the Eurostat solution might have wider validity, and may also be something that researchers from non-Eurostat RDCs will be able to relate to.

Researchers come from other organisations with other perceptions. Often training is not seen as priority and there is relatively limited knowledge or information about SDC. The tool is therefore going to be most effective if it can be run with little or no training.

Some organisations check little or no output. For these, incentivising researchers to work with the tool (for example, by producing nicely formatted Excel output) will be key.

Finally, we note that resource requirements differ enormously between SOs – from a handful of requests a year in some facilities, to thousands of requests in others. For some organisations, it is a vicious circle: a small number of requests for use means that there is limited return from investment in training in scalable solutions; concerns over the cost of checking therefore encourage limited use.

This is a specific concern for Eurostat, which aims to improve access through remote RDCs but fears substantial demands on resources. It is also aware that remote access from users' own machines both reduces the demand for clearance, but allows the user to avoid some SDC checks.

An automatic tool may allow RDCs to break this cycle. However, take-up of a tool is likely to be affected by whether the tool requires substantial investment, or whether it can easily be integrated into existing practices. Eurostat's desire to move to remote access also requires that the tool is viewed positively by researchers.

[3] https://ec.europa.eu/eurostat/web/microdata/overview/self-study-material-for-microdata-users

3 Feasibility of automated output checking

Given the above discussion, this section considers the feasibility automated checking.

3.1 Desirable characteristics

For an automated output checking process to be both effective and successful, it ideally meets the following criteria:

Characteristic	Rationale	Importance
Perceptions and usability		
User acceptance	Users not accepting procedures and finding ways to avoid them is a major risk factor for data holders	Essential
Data owner (MSs) acceptance	The tools should be seen to provide an effective checking mechanism	Essential
Data centre manager acceptance	The tool should have the active approval of the data centre manager	Essential
User training	Necessary user training should be minimised	Important
User preference	If users choose to use an automated tool over alternatives because it is convenient for them, then this promotes user acceptance and facilitates user training	Desirable
Language independence	Researchers use a range of coding languages (R, Stata, Python, SPSS etc). An ideal solution would be able to deal all the likely languages being used without having language-specific versions	Desirable
Statistical characteristics		
Flexibility	If the organisation follows a principles-based system, then it is necessary that users can ask for exceptions to be considered In an ad hoc system, it is not clear if this is needed as, in theory, rules are followed rigidly.	Essential
Primary disclosure	Must be addressed	Essential
Secondary disclosure	The RDC's business rules for what to check must be followed	Essential

Characteristic	Rationale	Importance
Effectiveness	An automated tool should be able to replicate business rules/statistical limits reliably	Essential
Comprehensiveness	An automated tool should be able to deal with all types of output, even if the assessment is "I am not able to judge this"	Essential
Consistency	Identical outputs should produce identical clearance decisions; similar outputs should produce similar decisions	Essential
Implementation		
Scalable over users and outputs	The tool continues to operate effectively as the number of users and past outputs grows- i.e. the problem of 'combinatorial explosion' is managed	Essential
Maintainability	The tool should not require regular operational adjustments	Essential
Ability to handle auxiliary information	The tool can adjust to rules set up to target specific datasets or users	Essential
Final responsibility	The data centre manager needs the ability to overrule automatic decisions	Important
Need for auxiliary information	The tool does not require a complex, perhaps dataset-specific, set-up process for additional datasets to be added	Important
Avoiding technological lock-in	The tool can be produced using different technologies, ideally open-source	Important
New or changing rules	Changes to the rules can be implemented without major disruption to the tool	Important
Skill level in data centre	The tool can be managed and maintained without extensive specialist knowledge	Desirable
Ease of deployment	The tool can be added to data centre systems relatively easily	Desirable

3.2 When should automated checking take place?

3.2.1 End-of-activity review (EoAR)

EoAR occurs when the analyst, having completed the statistical processes, presents the results to the output checker for review (in some RDCs, the analyst and output checker are the same person). Output checkers use the rules and their own knowledge of the data and/or statistics to make decisions about release. Output can be checked for differencing by comparing with information stored about previous releases.

This practice is almost universally adopted by RDCs. The advantages are

- There is no setup required other than training of the checkers and putting in place systems to allow users to submit outputs for review
- Humans can make subjective assessments of risk when considering differencing for example
- Humans familiar with the data can make quick judgments about disclosure risk
- Humans are good at understanding what is being presented to them

The last two advantages are the reason there is currently no feasible automatic tool for EoAR. For computer to understand a table, for example, in the same way that a human would, requires one or both of

- Extensive restrictions on the way outputs are produced, including the naming of variables and acceptable categories
- The analyst producing a detailed explanation of the statistic in such a way that the computer can unambiguously evaluate the statistic

In other words, at present the time saving from having a computer check all the results is outweighed by the extra time of telling the computer what results it is looking at. For example, using tauArgus or sdcTable to carry out the analysis would require creating the metadata and the table specifications, and then manually reviewing the results.

Whether EoAR is carried out by a human or computer, there remain the problem of assessing final results when the judgement requires a knowledge of the data (e.g. dominance, or the comprehensiveness of categories). This is addressed by either the analyst supplying additional supporting information (e.g. demonstrating p-rule is met) or by reproductive review, below.

3.2.2 Reproductive review (RR)

RR is where the final set of outputs to be released, once confirmed, are run against the original dataset. For example, Statistics Denmark requires users to run tabular outputs through tau-Argus before self-approving for release. CASC in France offers reproductive review as a service to users and journals as a way of ensuring the validity of outputs, rather than checking that the confidentiality of the outputs has been maintained.

RR allows for a much more robust check of the data. Even if the check is carried out by the analyst, the objective of the review process (eg the risk score from sdcTable or tauArgus) balances out the natural tendency of the analyst to favour release. Moreover, if the check is carried out by the analyst, understanding idiosyncratic variable names and categories is no longer a problem. A problem is that analysts need to be familiar with the automatic tools: tauArgus, although menu-driven is not very intuitive and requires complex metadata setup; sdcTable requires good knowledge of R programming as well as SDC.

A second problem with this is that automatic tools currently available carry out a very small amount of checks, mostly frequency and dominance checks with limited functionality to check multiple tables for differencing. Smith *et al.* (2012) demonstrated that tauArgus (and by implication sdcTable) did not provide comprehensive differencing protection (although the open-source releases of tauArgus since 2016 are claimed to have addressed this problem). However, this reflects the historical development of such tools (tauArgus was originally designed for the hierarchical tables typically produced by NSIs, rather than the idiosyncratic tables produced by researchers), and does not preclude the development of more tools; moreover, for researcher outputs, simple SDC checks (ie primary disclosure) are likely to be more than adequate.

The third problem is that this introduces an additional stage into the clearance process, and one which is, at present, quite clunky. This may limit user acceptance and hence compliance.

3.2.3 Seamless within-activity review (SWAR)

Within-activity review takes reproductive review to its logical conclusion: why wait until the processing is completed before providing feedback on the acceptability of outputs? Approving or refusing outputs as soon as the analyst produces them means that the analyst can adjust research in real time to produce only acceptable output. Remote job systems such as microdata.no and Lissy provide immediate feedback to help researchers refine their output requests.

'Seamless' review means replacing the standard user commands with alternative commands that call the necessary disclosure checking subroutines. Thus, for example, the 'table' command in Stata would be replaced by a code file. The user would invoke the 'table' command as usual and see the expected tabulation, but in

addition a disclosure check would be carried out on the tabulation. Ideally the user would be informed whether the tabulation would be an acceptable output or not, allowing the user to adjust his or her output.

The replacement command could have additional options reflecting its additional functionality. For example, one option may allow the user to automatically apply cell suppression, or to request an exception, if the resulting table does not meet SO standards.

At present, there are two implementation of this outside of remote job systems.

One is DataSHIELD (Wallace et al, 2014), which arose from a series of projects researching the safe sharing of biomedical data between sites to facilitate research. A researcher submits a query to a central node, which then sends requests to a set of data centres which reply with statistics results from their sub-populations of data. The central node compiles these to create a 'population-level' aggregate; this is then sent back to the data centres to check if the summary statistics are consistent with the data they hold, a process which iterates until all are satisfied. The authors claim that this produces results that are identical to that which would be obtained if all the data was held centrally. Most relevantly to this project, SDC is carried out automatically when the output is sent to the centre. Simple threshold rules are used (cells of 1 to 3 observations are suppressed). This is seamless to the users.

Of more direct relevance to the project, the Berlin-Brandenburg Federal Statistical Office has developed SDC-friendly replacements for popular output commands in Stata including tabulation, regression test and summarising commands. These work by the user coding as normal, then running a search-and-replace program to replace the original commands with the revised ones. These produce SDC-checked output with minimal need for the user to adjust his or her coding. At present, we do not have more detailed information on this work (which was first developed some five years ago), but it is likely that much can be learned from this, particularly in terms of user acceptance.

3.2.4 Discretionary within-activity review (DWAR)

Discretionary within-activity review is the same as SWAR except that it requires users to select which outputs are sent to the review process. This requires an additional step by the researchers, to choose to run their outputs through the review package. Using this therefore requires some user training. The rationale for making this discretionary is that users typically produce a large quantity of outputs in the course of their research; only a small subset would be requested for release. Running SDC checks just on outputs that are needed reduces the volume and complexity of checks.

3.2.5 Non-production of high-risk statistics in the research environment

One radical solution is simply to ban high risk/high resource outputs from the research environment. For example, if all the variables of interest are available for tabulation on the SOs website, then there is much less need to allow frequency tables to be released. Developments such as the cell-key method allow more complicated tables to be built dynamically whilst still preserving confidentiality.

The difficulty is that researchers often create their own categories, and will select subsets of the data based on their own criteria. At present no SOs offers a tabulation program with sufficient flexibility; and tabulations only make up a part of research outputs. However, we note that this may be an area for future development with substantial benefits for SOs.

4 Developing a proof-of-concept

4.1 Review of characteristics specific to Eurostat

Eurostat has identified the key characteristics of the users of the Eurostat Secure RDC:

- Users are likely to have travelled considerable distances, and so are unwilling to spend time on activities which do not directly enhance their research
- Users are likely to leave preparation of outputs to the last moment, if they prepare at all
- Users usually run a significant number of analysis and produce a big number of outputs which are currently unfiltered (ie Stata log files)
- Users usually produces metadata (key for variable codes and description of new variables; again produced in log files)
- Users usually spend one day in preparing the database (cleaning, merging, recoding etc) and they run analysis remaining days
- The visits are usually well prepared as the users already worked with anonymised microdata and can design the analysis and even prepare the scripts (the same variable names); most visit after analysing these results at home
- Users have a chance to talk to the clearance team and discuss outputs before they leave, but most users do not take this opportunity
- As a result, any system which requires more preparation before leaving the safe centre is unlikely to be welcomed by researchers
- Any system which can extract the relevant outputs and supporting information directly and avoid reading log files should substantially reduce the workload for the output checkers.
- There is comprehensive training online (Eurostat, 2015) but it is not clear how many read this, and it is likely that far fewer apply it
- Researchers are also given some guidance from Eurostat on arrival on the secure facility; again, it is not clear whether they take the lessons in the spirit intended; it appears that they do not absorb the lessons effectively
- Additional user training is therefore unlikely to be effective unless it is memorable and has a clear and direct value to the researchers
- At present, Eurostat is an ad hoc system (in the Alves and Ritchie, 2019, terminology) and does not use the 'safe statistics' concept
- The Eurostat RDC includes both business and personal data

4.2 Outline proposal

This project will develop a solution using Stata. The reasons for this are:

- The developers are familiar with the language
- Stata is popular with Eurostat Safe Centre researchers
- The macro language in Stata makes it easy to replace or augment standard commands with new versions
- The macro language allows users to direct output efficiently.

The use of Stata is not intended to indicate a long-term solution, particularly as Stata is a proprietary product. Longer-term solutions could include, for example, exploiting the programmable nature of the (open source) sdcTable. The value of Stata here is as a rapid development tool, as agreed with Eurostat.

Deliverable 1: Understanding output checking

At this stage, we envisage that the output checking tool will ideally generate Excel output in the form of approved outputs and those requesting an exception. We intend that the production of the output in a re-usable form (Stata does not generate Excel output easily) will increase the take-up by users.

We will consider both discretionary and seamless checking. At this stage we expect the discretionary approach to have greater acceptance by both users and Safe Centre staff as it restricts the volume of output.

The tool will aim to cover the majority of outputs created by a researcher (including analytical outputs), but not all. The tools will carry out primary disclosure, dominance checks, and some secondary disclosure.

The generated output will be editable by the researcher. In other words, the researcher is trusted to not falsify results. One alternative is for generated outputs to be read-only by the researcher; this could require some substantial configuration of the Safe Centre IT system, and it will be difficult to deal with outputs sent through the tool by mistake. Another option would be to use some form of check-digit/hashing verification, which is likely to require additional work on the part of the Eurostat output checker to run the appropriate tests on. We will consider these options as part of the project summary, but we believe trusting users is acceptable for the proof of concept.

We assume that Eurostat will need to see the output before it goes out, as automatic emailing would be unacceptable. The aim will be to present the output in such a way that the Eurostat team can release approved results with minimal intervention.

Finally, as far as possible and with Eurostat's agreement, we would suggest trialling the solution with genuine users in other SOs. We expect that some MSs will be willing to participate, and SOs in Canada UK and Australia have expressed an interest in contributing.

4.2.1 Outstanding questions

We expect that the tool will reject unacceptable output and ask the user to either reformulate it, or ask for an exception. An alternative is to apply SDC rules (for example suppressing or rounding invalid cells). How this is done is likely to require some discussion with users and Eurostat.

The major question will be how far other outputs generated from the dataset should be considered for secondary disclosure. Options include:

- All previous outputs
- All previous outputs cleared by the tool
- Previous outputs from that researcher/project
- Previous outputs from the researcher/project cleared by the tool

And each of these could be considered as simple paired comparisons or complex multi-way comparisons.

Checking only outputs cleared by the tools simplifies the problem as there is no need to create a history from unstructured outputs. However, it means that the secondary disclosure will need to 'learn' over time. This may be something to consider as part of a longer-term goal of allowing AI solutions to clear some outputs.

References

ADSS (2016) Data Access Project: Final Report. Australian Department of Social Services. June. http://eprints.uwe.ac.uk/31874/

Alves K. and Ritchie F. (2019) "Runners, repeaters, strangers and aliens: operationalising efficient output disclosure control". Working papers in Economics, Bristol Centre for Economics and Finance no 20120904 https://ideas.repec.org/p/uwe/wpaper/20191904.html

Brandt M., Franconi L., Guerke C., Hundepool A., Lucarelli M., Mol J., Ritchie F., Seri G. and Welpton R. (2010). Guidelines for the checking of output based on microdata research. Final report of ESSnet sub-group on output SDC. https://www.academia.edu/38191675/ESSNet_SDC_-_Guidelines_for_the_checking.pdf

Bond S., Brandt M., de Wolf P-P (2015) Guidelines for Output Checking. Eurostat. https://ec.europa.eu/eurostat/cros/system/files/dwb_standalone-document_output-checking-guidelines.pdf

Eurostat (2015) Self-study material for Microdata users. Eurostat. https://ec.europa.eu/eurostat/web/microdata/overview/self-study-material-for-microdata-users

Green E. Ritchie F., Newman J. and Parker T. (2017) "Lessons learned in training 'safe users' of confidential data". UNECE worksession on Statistical Data Confidentiality 2017. Eurostat. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2017/7_lessons_learned_training.pdf

Griffiths E., Greci C., Kotrotsios Y., Parker S., Scott J., Welpton R., Wolters A. and Woods C. (2019) *Handbook on Statistical Disclosure Control for Outputs*. Safe Data Access Professionals Working Group. https://doi.org/10.6084/m9.figshare.9958520

Hafner H-P., Lenz R., Ritchie F., and Welpton R. (2015) "Evidence-based, context-sensitive, user-centred, risk-managed SDC planning: designing data access solutions for scientific use", in UNECE/Eurostat Work session on Statistical Data Confidentiality 2015, Helsinki. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/20150/Paper_9_Session_4_-_Various__Hafner_et_al._.pdf

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

ONS (2019) Safe Researcher Training. Office for National Statistics.

Ritchie F. (2007) Statistical disclosure control in a research environment, mimeo, Office for National Statistics. Edited and reprinted as WISERD Data and Methods Working Paper no. 6 (2011). https://uwe-repository.worktribe.com/output/957311/statistical-disclosure-control-in-a-research-environment

Ritchie F. (2008) “Disclosure detection in research environments in practice”, in Work session on statistical data confidentiality 2007; Eurostat; pp399-406 https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2007/12/confidentiality/wp.37.e.pdf

Ritchie F. (2019a) "Analyzing the disclosure risk of regression coefficients". Transactions on Data Privacy 12:2 (2019) 145 – 173 http://www.tdp.cat/issues16/abs.a303a18.php

Ritchie F. (2019b) “10 is the safest number that there’s ever been”, Work session on statistical data confidentiality 2019; Eurostat. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S3_UK_Ritchie_AD.pdf

Ritchie F. and Elliot M. (2015) "Principles- versus rules-based output statistical disclosure control in remote access environments", *IASSIST Quarterly* v39 pp5-13

Ritchie, F., & Smith, J (2018) Confidentiality and linked data. In G. Roarson (Ed.), *Privacy and Data Confidentiality Methods – a National Statistician’s Quality Review*. , (1-34). Newport: Office for National Statistics. https://uwe-repository.worktribe.com/output/856040

Smith J.E., A.R. Clark, A.T. Staggemeier, and M.C. Serpell. (2012) “A Genetic Approach to Statistical Disclosure Control”. IEEE Transactions on Evolutionary Computation, 16(3):431–441.

Wallace, S. E., Gaye, A., Shoush, O., & Burton, P. R. (2014). “Protecting Personal Data in Epidemiological Research: DataSHIELD and UK Law”. Public Health Genomics, 17(3), 149–157. http://doi.org/10.1159/000360255