Setting up output checking processes

A guide for data services

Version 1.0 October 2024

Elizabeth Green, Cara Kendal, Felix Ritchie and Kyle Alves

Bristol Business School, University of the West of England, Bristol, and Data Research Access and Governance Network www.uwedragon.org.uk

Table of contents

List of figures

List of tables

Executive Summary

The purpose of this document is to help data services set up, review and evaluate their output checking (OC) processes. This is intended to be both a living and practical document, so further input and comments on the materials are very welcome. This document is based on; literature, the authors' own experience working with data services and data service staff, and input from seminars and working groups.

While output checking can be seen as cost, it nevertheless is something that all secure research environments should be investing in. If resources are limited, there are good practices to mitigate the cost and improve the efficiency and effectiveness of the process while maintaining confidentiality. Good practice is identified as

- Train researchers to produce good outputs; ideally through active training but as a minimum through well-designed self-study material
- Clear statistics rather than full papers
- Check only for disclosure risk, not statistical quality, and rely upon self-certification to check for scope violation
- If third-party review is required, make sure that the purpose of this review is clear and that response metrics reflect delays outside the control of the data service
- Use a single-pass model of clearance, especially if screen-sharing is allowed
- Operate a principles-based model

These are not intended to be prescriptive, but organisations choosing to follow other procedures should consider the constraints that lead them down alternative paths.

In terms of building processes, there is strong evidence of what works for each stage of the output checking process

Stage	Key success factors
Researcher generates outputs	• Researcher training • Clear processes and allowable formats
Researchers sends outputs	• Ensure that processes reflect the variable nature of research • Minimise unnecessary processing
Service team allocates output	• 'Four eyes' model with sequential checking
Checking takes place	• Different skill sets in staff to reflect demand • Clear guidelines shared with researchers • Outsource rare skills if necessary
Approve or reject	• Train staff in providing constructive feedback • Allow flexibility in how staff respond
Repeat submissions or rejections	• Train staff in developing positive behaviours • Have clear escalation procedures
Third party approval	• Ensure this is advisory as far as possible

	• Offer training to third party assessors

The choice of an IT system can make a substantial difference to both the data service staff and the researchers. There is experience in the data service community of what makes a good IT system.

Data services need to consider what their operating parameters and service-level agreements (SLAs) will be. These should be based on fairly assessed clearance times, and not on 'just-in-case' basis. At present. There is no clear agreement on wider metrics a data service should be collecting, but some potential management information is suggested.

When building the team, it can be useful to identify what stage the data service currently operates at. The guide identifies what might count as 'minimal' (just enough to run a service), 'adequate (enough to run a reasonably effective service) and 'ideal' (what a fully-resourced and efficient service looks like), and suggests ways that an organisation might move up through the stages.

Finally, the guide moves away from operations to consider strategic factors: what are the internal and external drivers and constraints, what are the strengths and weakness of data services in relation to output checking? Data services often face significant resource constraints and external pressures, both coming from the same source: that output checking is a specialist task that few outside data services fully understand. However, this also provides an opportunity for data services to demonstrate their unique value and expertise, and take more control over their destiny.

Acknowledgements

We are grateful for the input from workshop attendees and members of the SDC-REBOOT community for developing this document, particularly Amy Tilbrook and Ting Wang for detailed comments. All errors, omissions and views expressed are the responsibility of the authors.

Introduction

The purpose of this document is to help data services set up, review and evaluate their output checking (OC) processes. This is intended to be both a living and practical document, so further input and comments on the materials are very welcome.

This document is based on; literature, the authors' own experience working with data services and data service staff, and input from seminars and working groups for the SACRO and Future Data Services projects. It also draws heavily on the findings of the March 2024 Output Checking Retreat attended by 20 individuals from 15 organisations, representing their own views rather than those of their organisation.

In this document, we target data services, and we study data checking processes in the context of a 'trusted research environment' (TRE) where output checks are enforced by the TRE processes. However, the general principles and questions raised in here are also applicable to research environments where enforcement is not possible or required, but where the research team wishes to understand and follow good practice.

This document was produced as part of the ESRC Future Data Services programme.

Assumptions and core concepts

We assume that the reader is familiar with the concept of output checking and core statistical concepts such as statistical disclosure control (SDC). For more information, see the SACRO Guide to Output Checking[1]. We also assume a familiarity with the RRSA (runners, repeaters, strangers, aliens) model[2].

We focus on the risk for quantitative outputs from traditional statistical research, as these form the vast bulk of clearance requests at present. We consider the special issues raised by qualitative data, machine learning models, and other unusual outputs in the Appendix.

We define the relevant actors as:

- The *researcher* or analyst who produces the output
- *Output checkers* who review the output and make decisions, and who may work in a *team*
- The *data service* which decides about how output checking operates
- *External organisations* who may have an influence on the process

Some people/organisations may play multiple roles.

We assume that the output checking process looks something similar to this:

[1] https://zenodo.org/records/10054629

[2] Alves, K., & Ritchie, F. (2020). Runners, repeaters, strangers and aliens: Operationalising efficient output disclosure control. Statistical Journal of the IAOS, 36(4), 1281-1293. https://doi.org/10.3233/SJI-200661

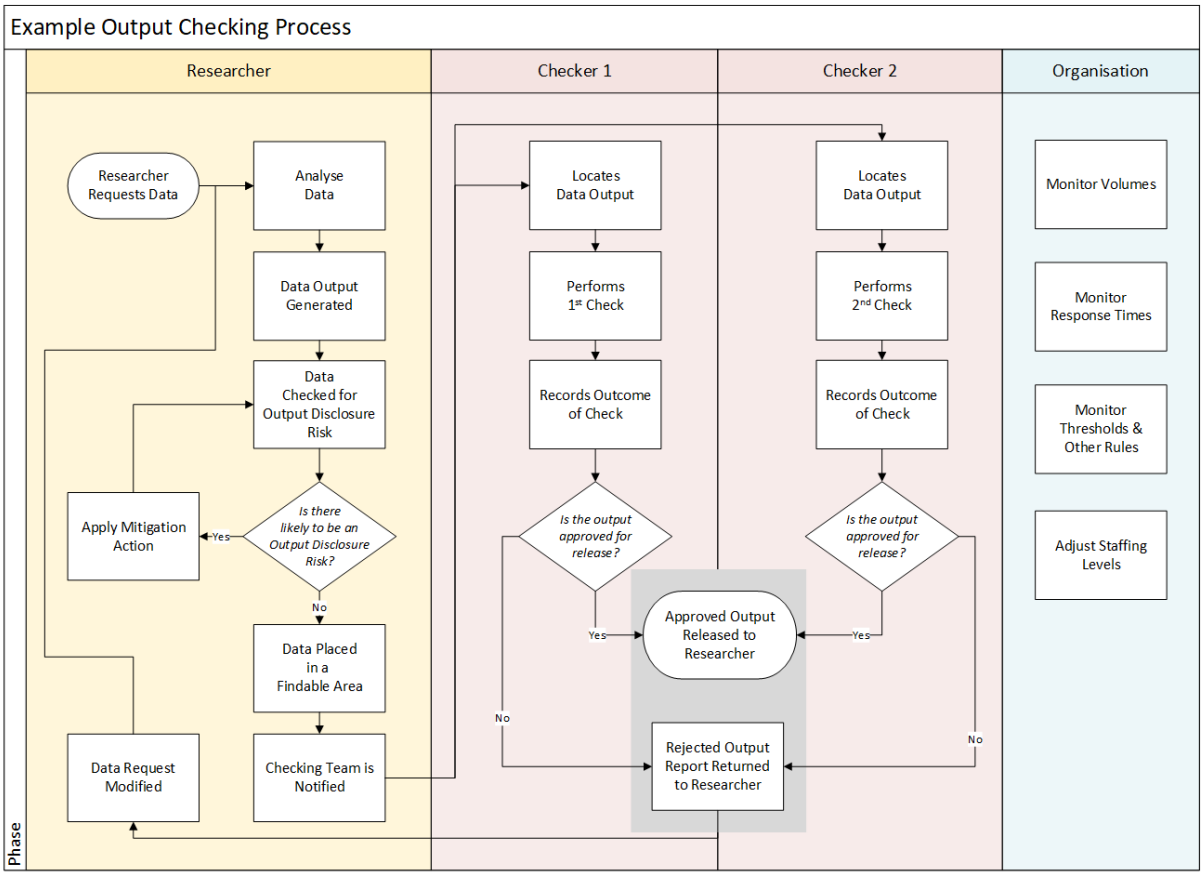

Figure 1 Example OC process

The researcher:

- Analyses data and produces output
- May check for disclosure risk and/or apply mitigation
- Places the output in a findable place and notifies the checking team

An output checking team member

- Locates and evaluates the result
- Approves or rejects the output and notifies the researcher accordingly
- A second checker may evaluate the request, in parallel or in sequence

The data service

- decides the operating model and general risk appetite
- will set response times and allocate human and technical resources

External organisations

- may set thresholds or in other ways express a risk appetite different from the standard model
- may require oversight or right-of-refusal for outputs

This document makes several references to ‘good practice’. These are not intended as mandatory requirements, but the expectation would be that organisations not following these suggestions should be able to justify their choice of a different solution.

Preparation: identifying purpose, context and goals

Why bother checking outputs?

The first consideration for any data service is whether checking outputs is necessary. The case for not doing output checking is:

Reason	Rationale	Valid reason?
Resources	This is a resource intensive process requiring specialist (expensive) staff; fewer resources are therefore available for deployment elsewhere	Yes. All data governance is a risk management process
Negligible risk	In genuine research outputs, risk is very low	No. The question is whether this risk, although low, has been managed effectively
Negligible risk in a risk-managed system	OC is usually part of a suite of checks and controls ('the Five Safes') which provide checks against more significant risks	Yes. The risk may be managed more effectively in other ways, for example by training or data reduction
Limiting benefit	OC may block some useful outputs, leading to public benefits foregone	Perhaps. This is an empirical issue. Evidence so far suggests not.
Legality	There is rarely a legal requirement for OC. The UK Digital Economy Act 2017 does mandate an output clearing process, but does not specify what this is. Official guidance (eg UK Information Comissioner) does not mention OC, or puts it in such general terms that any process is compatible.	No. Ethics is the basis for data governance, not law.
Subjectivity	Good OC for research is best done with a subjective component. This may introduce errors and induce an unwarranted sense of security; checkers may favour or discriminate against certain users	No. Little evidence of this, and the remedy is better staff training, not stopping OC

Table 1 Reasons for not checking outputs

In short, an OC process may not be needed if the data service has very limited resources, and the risk can be effectively managed in other ways: for example, reducing data detail, or training and incentivising researchers to reliably produce non-disclosive outputs.

Reasons for having an OC process are:

Reason	Rationale	Valid reason?
Impact of error	The risk of disclosure from research outputs may be low but the impact of a breach can be very high	Perhaps. Little evidence to date
Poor research behaviours	Researchers can be flawed actors – errors, lack of focus, specific personality flaws – and so can't be relied upon to produce 'safe' outputs.	Yes. Good evidence of this
Reputation	Demonstrating good OC processes improves reputation for competence with funders, data owners and the public.	Yes. Good evidence that this is now expected from secure facilities.
Staff skills	Developing output checking skills can increase general skills in handling confidential data and research outputs	Yes.
Researcher protection	By output checking thoroughly, we protect researchers from consequences of running unsafe research.	Perhaps. Little evidence to date
Ethics	It is the 'right' thing to do.	Yes, if resources allow. Ethics is the basis for data governance, not law.

Table 2 Reasons for having OC processes

These are counterpoints to some of the reasons given in Table 1 for not having an OC process, and they tend to be better evidenced. There are two major lessons from running OC processes in the 21st Century.

The first is that researchers are liable to make mistakes in disclosure protection, if they can be persuaded to undertake this at all; therefore, relying upon researchers as the main protection is problematic. Instead, if no output checking processes are to be put in place, the data service should concentrate on placing more protection in the data. This is the assumption made by data archives that distribute data, such as UK Data Archive or Eurostat. While they may give advice to users on how to carry out disclosure control, the main protection is in the limited data detail in the downloaded file.

The second lesson is that reputation is a significant factor in persuading data holders to make deposits with data services, and for the public to approve of the use of their confidential data. Although very few individuals outside data services are likely to have a good understanding of the risks and mitigations in research outputs, the idea that this is a risk to privacy has now become well-

established. As such, an organisation that can demonstrate how it manages that risk can enhance its reputation for good data handling practices.

Whilst most secure environments check all outputs, a few place more trust in researchers and operate a policy of checking random outputs, or placing limits on the amount of material that can be withdrawn from the secure environment. The effectiveness of this is not known, but where resources are very constrained, this may be sensible.

In summary, it is not necessary to have formal output checking processes, but as the data becomes more sensitive it becomes both good practice and good stakeholder management to have them.

Good practice

- Some form of output checking process unless there is negligible risk in the data
- If resources are very constrained, random/partial checking may be sufficient

For the rest of this document we assume the decision to have a formal OC process has been made and consider how those processes can be made efficient and effective.

What outputs are expected?

A data service needs to decide on

- whether just the statistics (plus supporting information) need to be submitted for review, or to review whole papers/documents
- whether to have a single pass, or to require an intermediate and final clearance stage
- whether to include non-statistical materials in the review
- how much researchers can re-use or re-present the same cleared results

Each of these has operational implications (and the choice may be driven for operational reasons). Data services in the UK have used a mix of these, sometimes as an active choice and sometimes for historical reasons. Practices have also changed over time.

Statistics or papers?

Table 3 enumerates pros and cons of asking researchers only to supply the statistics to be checked, or requiring something closer to full papers.

	Pros	Cons
Clearance of statistics only	• minimal information to check • easier to identify information to check • context not for publication can be added • supports informal outputs such as presentations, blogs, reports • supports early sharing with co-researchers	• Researchers may not provide enough information to make assessment • Researchers may request more output than is strictly necessary • Researchers may add additional contextual information after release
Clearance of completed papers	• Provides extensive detail on outputs • Can confirm that outputs cleared are for the public domain • Should be the minimum necessary for the paper	• Large amount of information to sift • Researchers may still not provide enough information to make assessment • Researchers are using data service resources to write text documents

Table 3 Pros and cons of full versus minimal submission

The last consideration "using data service resources to write text documents" is not relevant for remote access but it remains an issue where physical space or access to the server is limited, such as in SafePods. It may also be relevant where data services do not provide researchers with the facilities to write papers, for example in remote job models, or where results are submitted using a markup framework such as Jupiter Notebooks. Otherwise, it can be ignored.

In theory, the statistics-only model is more efficient as only the necessary information is checked. In practice, researchers often fail to include sufficient information to assess the output, obviating the efficiency gains. The statistics-only model may be also less efficient (from the data service's perspective) because researchers produce more outputs: for example, a set of tables for a presentation, then a revised version of those tables, then a version for a paper; this in theory also presents a differencing risk. The reflection that is part of research is partially taking place outside the data service. This is where training is important. Screen sharing (by researchers on the same project) can also help to reduce the amount of output requests; see box.

Screen sharing

Screen-sharing is controversial. It is enormously valuable for researchers, improving productivity. It also reduces output requests for data services. Some data services formally allow screensharing by researchers working on the same project, some ban it, but many have no clear policy.

Sharing via external software (Teams for example) is considered by some as data leaving the secure environment through an unauthorised channel, and allowing the service provider to 'capture' data. This is based on a misperception of research use. Screen sharing typically involves discussions over code, outputs, or occasionally unusual data points. Data is generally de-identified, and so re-identification is only possible with extended and unrestricted access to the full dataset. Those looking at the shared screen cannot independently explore the dataset, but can only ask questions of the person logged in. For a data breach to occur, the screen would have to linger on a single record where there is sufficient information in the screen image alone to allow a data subject to be re-identified. This is not a reasonable risk.

The argument for having full papers reviewed is that (a) these should contain all the information necessary about the statistics (b) other information, such as general statements about the data or individuals, can also be identified and checked (c) the statistics released should be the minimum necessary to make the researcher's argument. However, published papers may not provide the necessary information to check (for example, only showing weighted statistics, or omitting empty classes). The amount of text to check requires sufficient resources to ensure this is done accurately, and checkers may spend time reviewing irrelevant parts of the papers. What happens when revisions are needed? Does the whole document need to be submitted, or just changed statistics?

'Minimum necessary' will depend on the type of output and stage of the process. Research produces a wide variety of outputs; for example, any article is likely to have been circulated through multiple public conference presentations and working drafts before it is 'final'. Should these be ignored when clearing an output if the statistical findings are unchanged? The data service needs to be clear about what uses an output can be put to, to prevent clearing the same findings twice; but this may create perverse incentives to generate unnecessary intermediary outputs, the activity the full-clearance model is designed to prevent.

Overall, the full-paper model assumes a simpler linear expectation of the research process than is normally the case. It is also incompatible with other methods to make the process more efficient; specifically, semi-automatic tools and researcher-triaged outputs.

Semi-automatic tools such as SACRO only check statistics. Therefore if a data service wants to review full papers it either has to choose not to use automatic tools, or to find some way to link cleared outputs to full papers.

The process in Figure 1 assumes that all outputs are the same. However, the RRSA model shows that efficiency gains come from segmentation of outputs into different types. One option is to encourage researchers to 'self-segment'; that is, to bundle 'runners' together in one request, and 'repeaters' or 'strangers' in a separate request. This triaging allows the data service to focus its checking resources more effectively. For example, runners can be checked by computer or by junior staff with minimal training, leaving more experienced staff to deal with the outputs requiring judgement. At present we are not aware of any organisation that requires self-triaging or trains its user in this way, but this has the potential for efficiency gains for both researchers and data services.

Good practice

- Check statistics rather than papers
- Train researchers to produce good quality outputs
- Encourage screen sharing (within project teams) to reduce the amount of intermediate results

Privacy or data quality?

There can be a perception that OC is also a quality checking process. This is particularly the case for government departments who might be sensitive about results, or comments made about their activities in full papers. Some feel that checking for quality is important: issues in quality may indicate lack of care for data security, or be an aid to junior researchers who are first learning. Others feel that this establishes a dangerous precedent to expect the quality of output to be checked by TREs; it may encourage less care when developing the outputs. There is anecdotal evidence that researchers interpret TRE checks on quality as less needing from the researchers.

Output checking is about preventing (perceived) breaches of confidentiality. While individual checkers may point out errors in statistics to the researchers, the ultimate choice of what to publish is the researchers'. Similarly, if researchers make erroneous statements about datasets, the checker can inform the researcher but the researcher can ignore the comments; enforcement outside the data service is likely to be very difficult. When complete documents are cleared, rather than just statistics, there is a heightened risk of mis-interpretation of OC as a quality review process

Ultimately, if the expectation of the TRE is quality checking outputs, this needs to be communicated and understood by the staff and an acknowledged part of the role. This also needs to be resourced as an additional activity, and the TRE needs to ensure that staff have the appropriate skills to carry out quality checks.

Some data providers may request sight of results before they can be released; this may be a condition of access. Data services indicate that this canhave a major impact on operations, and yet it is tom some extent beyond the data service's control.

In these cases it is important to establish:

- Does the data provider have the right to refuse outputs, or is this advisory only?

- What precisely is the data provider checking for – confidentiality risk, appropriate use of data, concerning findings?

If confidentiality is the reason, then ask where the data provider considers your checks are not up to scratch.

Third party oversight needs to be communicated to the researcher. Any delays caused by requiring oversight by a third party should also be accounted for in the metrics, and separated in the data service's service commitments.

Good practice

- Do not check for quality, unless it has a meaningful confidentiality implication
- If third parties ask to review outputs, confirm whether this is advisory or proscriptive
- Record third party checking times in metrics

Scope

Many TREs claim to check for the scope of the outputs, and require users to confirm that outputs are within the scope of the project. This is seen as a legal requirement, to make sure that projects are carrying out work that has been agreed and not straying into new areas.

It is not clear whether this is a meaningful check. We have acquired evidence of two cases where the outputs were not in scope; in both cases, the requested outputs were in obviously different subject areas to the original proposal. However, it is not clear how a checker would identify out-of-scope work in more nuanced cases. Consider a project studying low pay labour markets. It is hard to envisage an output checker successfully challenging any statistics around wages, employment, training, social security, commuting patterns, or housing as being out of scope, especially if the researcher disputes their judgement.

Apart from researcher error, scope breaches are most likely to occur where the procedures for reviewing and changing the scope of a project (or applying for a new project) are slow and cumbersome. This is not a problem of output checking, but it does illustrate how all parts of a data access system need to work together. If TREs are concerned about unapproved scope creep, it may be more productive to review access procedures – after first having done some random scope reviews to establish whether this is a genuine problem.

The common question about 'Please confirm that these outputs are within the scope of your project' does not seem overly burdensome, and it may cause some researchers to stop and review their outputs (or change them if they are aware that they are not in scope). So we propose that TREs retain this question for psychological reasons, even if the TRE does not actively check it.

Good practice

- Do not check for scope as part of regular outputs
- Ask researchers to confirm this is within scope
- Review outputs periodically if there is evidence that out-of-scope outputs is a problem

One or two passes?

Once cleared, the data services involvement can come to an end. However, some data services release an 'intermediate' output, and the request that a 'final' output be sent for approval before it can be considered fully public. The purpose of this is to balance research support and security. In theory, intermediate outputs can be released more quickly and easily because there is a second check, and that second check need only take place on final results intended for publication.

This was thought a sensible solution when TREs were being set up, and physical restrictions on access meant being flexible on first stage release was crucial for supporting research. There is substantial evidence that, while there were regular mistakes with uncleared final results being presented to a public audience, overall researchers adhered to the two-stage model.

With the move to remote access it is not clear that this is valuable; moreover, data services have tightened up the first stage review so that effectively the review process is doubled. We therefore recommend a single-stage process as the default. If there are to be two stages, then the first stage needs to be genuinely light-touch; that is:

- No differencing checks
- Accepting a broad range of exception requests
- Not making exceptional checks: missing categories in the 'frequency' statbarn, dominance of largest observation for Herfindahl indexes etc
- Giving researchers the benefit of the doubt (for example, allowing just weighted frequencies)

You could also consider only one pair of eyes to look at outputs; although a better solution would be to have two pairs of eyes at the intermediate clearance stage (where public release effectively happens), and just one pair at the final output stage.

Good practice

- Use a single-pass model unless there are very good reasons for a second check
- If using a two-pass model, ensure that the first pass is genuinely 'light-touch'

Principles- versus rules-based

A rules-based system means that disclosure control rules are not subject to negotiation. Rules-based systems generally do not work well for research environments, at least if applying all the good practice SDC rules. Some rules-based environments get round the problem by ignoring rules that do not work for them. Others do apply all the rules but have an implicit model of exceptions. This damages credibility, and potentially can lead to unequal treatment as the basis for differential treatment is not specified.

Under a principles-based system, researchers can ask for exceptions, where the output is not disclosive and important; effectively, it takes a rules-plus-implicit-exceptions model and makes the latter explicit. This is applied in most UK TREs. It does require training of researchers.

It is sometimes argued that rules based is more resource-efficient as rules can be applied with little or no knowledge, and no negotiation. This is likely to be the case if rules are followed exactly, but all rules-based data services allow informal exceptions, meaning that trained staff are still needed. A

principles-based system can be made very much like a rules-based system by tightening up on when exceptions may happen, but the key point is that those exceptions are still accepted.

Good practice

- Operate a principles-based model, training researchers and staff appropriately

Researcher training

The importance of training researchers has been mentioned several times. The evidence shows that this has a significant impact on the operational effectiveness of OC processes, as well as encouraging researchers to understand and appreciate (if not like) the processes.

In terms of helping researchers understand and buy in to processes, highly interactive face-to-face training is seen to be the most effective; there appears to be little difference between physical and virtual engagement for current training. As interactive training is resource-intensive, it may not be feasible for all data services to offer; however, there are common training programmes (in the UK, Safe Researcher Training) which researchers can be sent on.

It is not clear at present whether this is also the most effective for specific SDC training; there is some evidence that passive training materials could be sufficient. This is under review by the DRAGoN team and the SRT Expert Group, with results expected in early 2025.

Good practice

- Plan for, at a minimum, clear passive materials so researchers can self-learn
- If resources allow, develop active researcher training or collaborate with other data services to offer common training

Implementation: building and reviewing processes

Process decision points

The output checking process has multiple stages, at each of which the data service is either making an assumption or making a choice. Figure 2 below reflects the input from process actors consulted during the review process. Individuals professionally engaged with output checking were asked to record ‘choices’ and ‘assumptions’ made at each stage of the building and review process. These were then placed at the relevant process decision points using sticky notes.

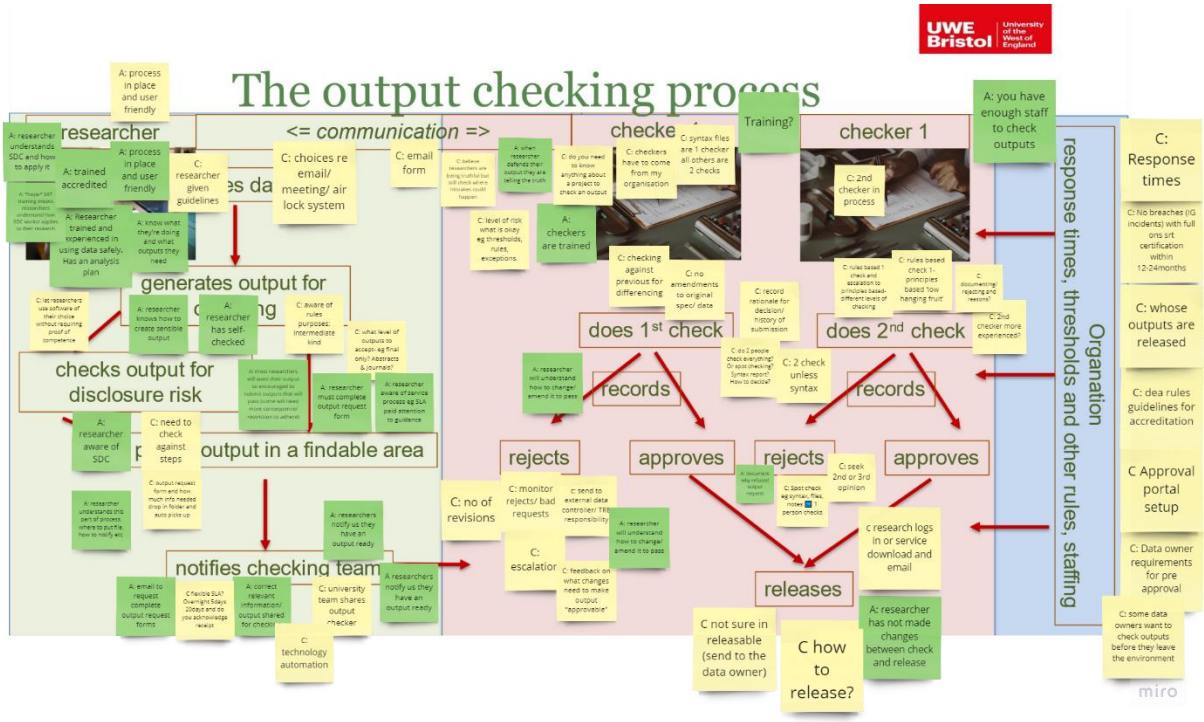

Figure 2 Whiteboard comments on the output checking process

In the following images, we link these comments more clearly to the relevant component parts of the Process Flow diagram presented earlier in *Figure 1*, above. Each of the vertical areas is presented separately. The 'choices' and 'assumptions' provided by workshop participants are linked to the relevant process steps. This approach illustrates which areas of the process carry the greatest amount of focus.

The first image focuses on the part of the process owned and executed by the Researcher. It begins with the researcher requesting data and ends when the Checking Team is notified that the data output is ready for review by the data-holding organisation.

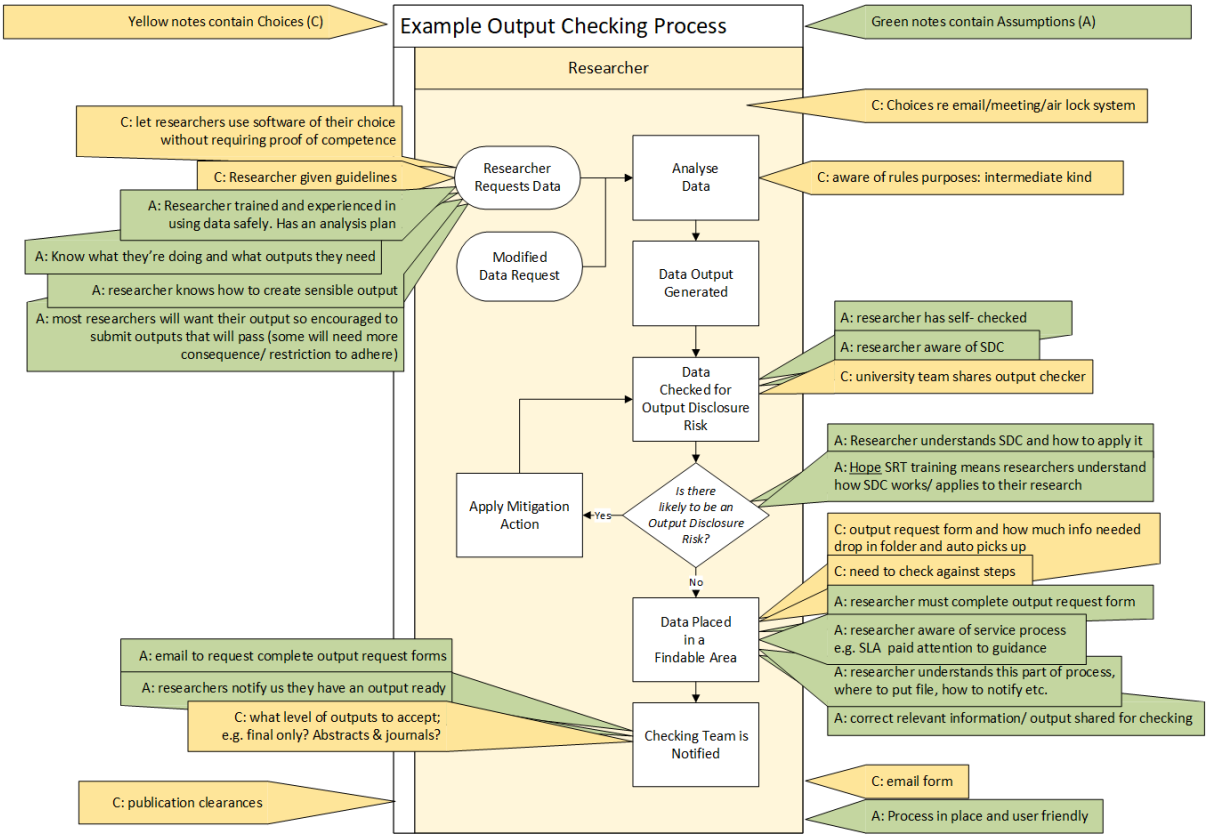

Figure 3 Practitioner comments on Researcher-performed process of Output Checking

This is followed by the part of the process where the data outputs are checked by the data-holding organisation. Note the significant number of comments provided regarding the checking-activity itself.

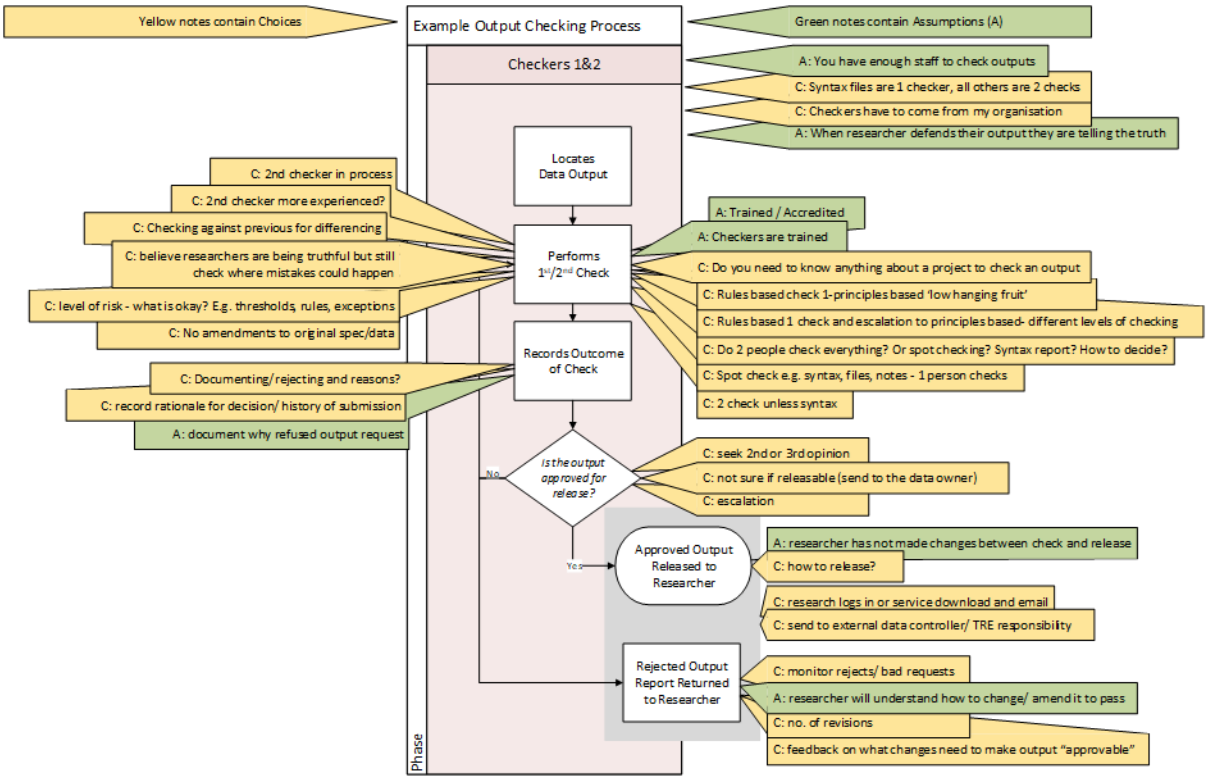

Figure 4 Practitioner comments on the process of the Output Checkers

The final area of process consideration focuses on the requirements of the data-holding organisation as a whole. The process activities illustrated here are considered 'management' processes that provide guidance and regulation. These are not processes that are directly engaged with checking outputs for researchers.

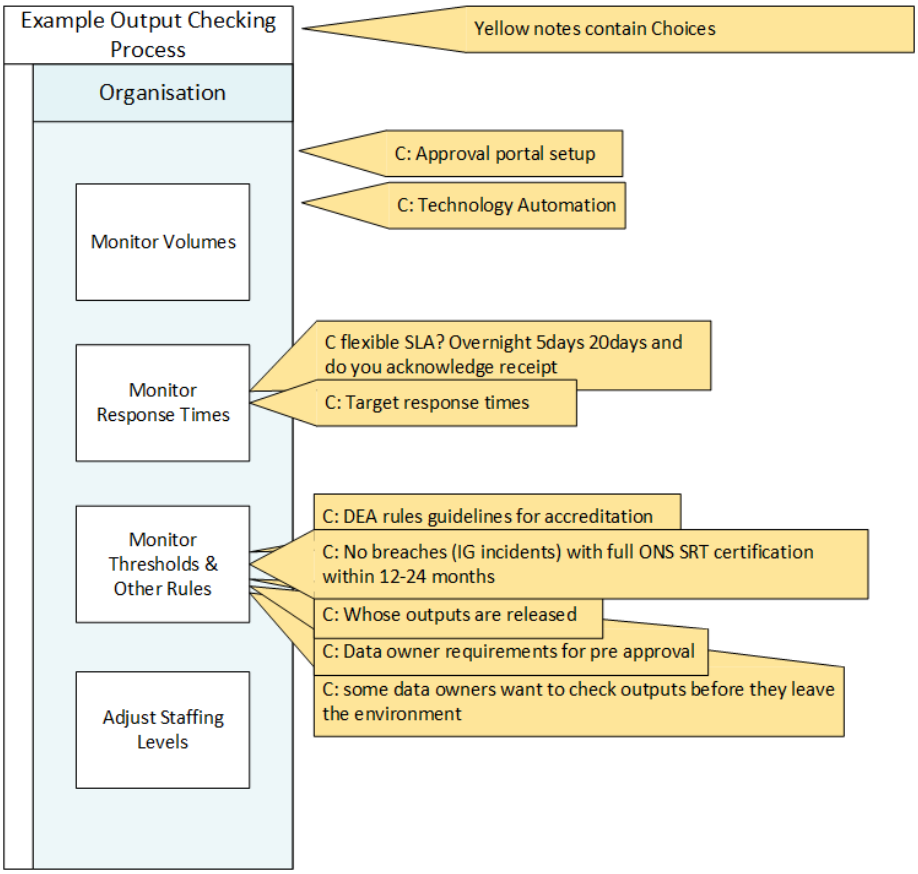

Figure 5 Practitioner comments on the management processes in Output Checking

These comments were used to inform the following analysis.

The tables below go through the process as described in the Introduction. To keep this general, we do not consider how the processes happen (e.g. whether a notification happens by user-generated email, automatic email, some other mechanism). We focus on the questions the data service should have answers to, and the assumptions that the data service acknowledges. We also recognise that not all good practices are feasible in all environments (e.g. in remote job systems, the researcher may have little control over submission of outputs for checking).

At each stage, we consider

- A: assumptions that can/should be made about this stage
- C: choices that the data service needs to make
- G: good practice, as currently understood

Stage 1: researcher generates outputs

	Assumptions being made **Choices to be made**	**Comments**
A	The researcher is not deliberately falsifying outputs to extract data	Reasonable assumption; good evidence
	The research is not hiding information in the output file	Reasonable assumption; good evidence

	Researcher is telling the truth when requesting an exception	Reasonable assumption; good evidence
C	Do we train the researcher to produce safe outputs?	Best introduced at induction; perhaps most effective when revised shortly before producing first outputs
	Do we train the researcher to produce easy-to-check outputs?	Include with statistics training; may need to be ongoing for recidivists
	What formats are allowed?	Unrestricted formats are easiest for researchers. A very limited set helps output checkers. Some services have concerns about information being hidden (see above).
	Should log/markup files be allowed?	These combine both code and results and, if not carefully written, can be messy and hard to read. It may be hard to separate code, wanted results and intermediate results in markup files (e.g. Jupiter Notebooks).
	How do you deal with researchers who refuse (or are unable) to follow good practice guidelines?	There are always some. Data service has a big stick: refusing access until they comply/are retrained, justified on resource grounds. Data service also has a small stick: applying special measures for certain users which will involve delay.
G	Provide researchers with training materials, covering both statistics and producing easy-to-check outputs	
	Have a training response for non-compliant outputs	
	Have an escalation procedure for non-compliant repeat offenders, and use it	
	Agree a set of 'allowable' formats – there should be a relatively large set which covers most research use and is easy for staff to check	
	Do not allow unrestricted log files: ask for separation of code and clean log files (ie required results only)	
	Markup files: guidelines need to be defined (see below)	

Table 4 Process stage: researcher generates outputs

Markup files contain both code and outputs – when the files are run, the outputs are inserted into the code. As a result they can be seen as code files (before being run) or log files (after being run). It is currently not clear how to manage these, as markup files are likely to contain intermediate results not needed for release.

One data service treats them as coding scripts (without results) but requires them to be runnable to generate the requested outputs; however, it also allows for researchers to submit outputs in a separate document ie the usual code file + output file practice. We provisionally suggest this recommended but we welcome views form data services on this.

Stage 2: researcher sends outputs for checking

	Assumptions being made Choices to be made	Comments
A		
C	Are all outputs sent for checking, or does the researcher select?	Research generates unneeded or duplicate intermediate results. Asking the researcher to be selective reduces the volume and increases the importance of checking
	How does the researcher notify the checking team?	Automatic notification reduces work for the researcher, but may increase work for the checking team as unnecessary results may be sent for checking (see previous point)
	Should the researcher split results into easier/harder (runners vs repeaters and aliens)?	Triaging by the researcher can make the process more efficient – more experienced staff resources can be reserved for more complex outputs. Needs some checks to ensure that researchers understand and are following guidelines.
G	Research should make an active decision which outputs are to send for checking, and when they are ready Researchers should be able to override/delete outputs previously selected before they choose the final set for checking Explore self-triaging with researchers, offering this as an option for researchers who already produce good outputs Output request forms should be kept to a minimum: only ask questions that acquire useful and novel information, that can be answered by the researchers, and where the answers are meaningful.	

Table 5 Process stage: researcher submits outputs

Stage 3: data service team is notified of results and sends for checking

	Assumptions being made Choices to be made	Comments
A	The notification process includes where the files are	Minimal requirement for notification
	There exists a pool of people equally proficient at basic checks	May be a good starting point – everyone can make 'first pass' assessment
C	How many people should check this?	Best practice is currently seen as two checkers for an output (the 'four eyes' model). Not clear if this is best done concurrently (both checking at same time) or sequentially (second checks only if first passes it).

		The latter is more efficient (and also the model for automated output checking) but could lead to over-confidence by second checker. Also not clear if code releases also need double-checking.
G	Use a four-eyes model with sequential checking A two-eyes model is appropriate for code checking	

Table 6 Process stage: data service team receives and distributes outputs

Stage 4: checking takes place

	Assumptions being made **Choices to be made**	**Comments**
A	Data service staff have all received basic training to review runners, identify other outputs, and escalate as necessary	Minimum requirement to do output checking if the checking is not outsourced
	A pool of more experienced checkers is available for repeaters, aliens and exception requests	Reasonable assumption – if not, the data service will have to outsource checking
C	Should data service staff do the checking in-house?	There may not be sufficient expertise in-house, or this may not be a good use of skilled resources. Remote access to TREs makes outsourcing of output checking a feasible option. However, there might be difficulties specifying appropriate contracts to avoid perverse incentives (eg processing as many outputs as possible by rejecting all but the best), and this limits the opportunity for the data service to build in-house expertise as a strategic goal.
	Should all disclosure control rules be applied?	The SACRO manual covers all but two statbarns so rules should exist for almost all outputs. Not all processes can handle all rules, and some data providers may not be bothered about some rules (typically, dominance rules are omitted).
	Can all staff deal with all outputs?	Should all staff be trained to deal with all outputs except the most unusual/novel cases? Or should there be differing levels of expertise e.g. junior checker/triage checker vs senior reviewer? Runners/repeaters vs strangers may be an efficient split, as the latter should only be occasional
	What is the escalation procedure (runners=>repeaters	The circumstances under which a junior checker can approve runners or exceptions should be clear, and when

	or exceptions, repeaters and exceptions =>strangers/aliens?	they pass to a more senior staff member, should be clear. Checklists may help.
	What are the rules for agreeing an exception?	This ultimately has to be at the discretion of the checker (determining whether the output is important, and whether the researcher is asking for too many exceptions). Guidelines may help but checker training should emphasise that this is subjective; the data service should make clear that its supports checker decisions.
	How certain must output checkers be of their conclusions?	Checkers may be incentivised to over-protect outputs, which reduces public benefit and irritates researchers. As this is a subjective process, this is difficult to monitor, but the data service should be support output checkers to be confident but not defensive.
G	Train all staff to deal with runners and repeaters, but escalate to senior reviewers for strangers and aliens The rules that are implemented should be communicated to researchers and the public, and ideally contrasted with the best practice guidelines in the SACRO manual to show the level of compliance Develop guidance for allowing exceptions (if principles-based) Make staff aware of the risk of a defensive default position, and encourage or develop self-confidence Outsource only if the need for checking is very rare and does not merit the development of in-house skills in the short or long term (for example, qualitative or machine learning outputs – see Appendix)	

Table 7 Process stage: checking takes place

Stage 5: approving or rejecting results

	Assumptions being made **Choices to be made**	**Comments**
A	The data service provides a way to send out and record decisions	Minimal functionality of an OC process
	The decision mechanism include the option to provide comments to the researcher.	Minimal functionality of an OC process

C	What comments should be sent to the researcher in the case of a rejection?	Output checkers need to send comments to (1) correct problems with the output (2) reduce the chance of the researcher making the same or similar errors. The latter is more important in the long term. Checkers should be trained in having these conversations.
	Can outputs be partially approved?	Yes, this can be efficient. But checkers should not have to edit the output to extract results which can be approved; they may *choose* to do it, but it should not be expected, and checkers retain the freedom to reject the whole suite of outputs. This is so that researchers do not start using the output checkers as editors.
	Do all decisions need to be recorded?	Yes; not just for operational purposes but in case a complaint is received, or if the data service wants to review internal effectiveness. Ideally, recording the decision should happen in the same process as communicating with reviewers (i.e. *not* sending an email to the researcher and then having to upload that email to the recording system).
	Does all discussions need to be recorded?	No. A quick phone call to the researcher may be much more effective than emails (both in terms of time saved, and in terms of building a positive relationship with the researcher). It may be particularly helpful if the checker suspects an output is fine but wants clarity. The formal response should be simple, direct and avoid complex discussion of the output ("figures in Table 7 are weighted but that Table 3 gives the unweighted numbers...")
	What is the response time?	Response times should be measured to evaluate performance. Metrics should distinguish between activity under data service control (i.e. checker's time to respond) and activities beyond their control (waiting for the researcher to reply).
G	Output checkers are trained in providing useful and educational feedback to researchers Partial approval is allowed at the absolute discretion of the checker Invest in an effective IT system (see below) Encourage phone conversations/other informal methods and formalizing the discussion or decision only when this has been agreed.	

Table 8 Process stage: output checkers communicate the result

Recycling: dealing with repeat submissions and repeat rejections

In this table, we consider what happens when an output is re-submitted in response to comments and rejection by the output checker; or when a researcher is observed as repeatedly getting outputs rejected.

Assumptions being made	**Comments**
Choices to be made	

A	The researcher is genuinely trying to correct errors	A weak assumption, but necessary. Some researchers might be positively responding to the rejection, others will just be looking to see what is the minimum necessary to get the output over the line.
C	Has the researcher done the minimum necessary when correcting a repeat submission?	This is fine; hence, it is important that checkers can express their rejection notices in such a way that it makes clear what is to be done.
	How many times can an output be resubmitted?	Not clear that any general guidance can be given; see next comment
	What if the researcher is continually getting rejected outputs?	The researcher may be genuinely confused, or just stroppy. Either way, there is clearly a need for training. Informal one-to-one training may be most effective initially. The data service retains the big stick of refusing access or formal re-training if the researcher does not learn.
G	Checkers should be trained in developing positive researcher behaviours Escalation processes in the case of refusal to learn should be short, and very short where the researcher is uncooperative	

Table 9 Dealing with repeat submissions

Third party approval

	Assumptions being made **Choices to be made**	**Comments**
A	The third party has a service level agreement to make a decision on the output	The data service needs to be comfortable it can require the third party to meet its service obligations
C	Is the third party approving the output (proscriptive), or is it receiving the output for information and only requires a delay for notification (advisory)?	This needs to be clarified before clearance can begin. Ideally the third party's interest should be advisory. If it wants to be proscriptive, explore why; if the answer is that the third party has extra secret knowledge, get trained in it.
	Should the data service offer training to the third party?	Yes. Very few individuals outside data service are trained in disclosure checking of research outputs; disclosure checking of official statistics is not really helpful. Training is an investment to help the third party make better decisions.
	Can the data service override a rejection by the third party?	The third party is likely to be more defensive than the data service. Where the third party's interest is advisory, yes; probably not, where is it proscriptive. But in both cases the data service should engage with the

		third party to try to come to a common view. This should be seen as an investment in knowledge.
	Does the researcher/data service understand the delay in approval?	This stage is out with the data service's control; therefore it should be made clear to the researcher, and in any metrics, where the delay arises. Service Level Arguments need to be enforceable.
G	Agree that oversight is advisory, and agree a response time which will be strictly adhered to. Offer training to the third party if necessary, requested, or likely to reduce friction Ensure that metrics and researcher communications clearly delineate when the checking process is dependent on the timeline of the third party.	

Table 10 Dealing with third party requests for oversight

Using automatic approval tools

Within-activity review tools such as SACRO intercept the researcher's commands and automatically review outputs against a pre-defined set. Review decisions are done for each statistic, and are one of

- Accept: meeting rules
- Accept: meeting rules because automatic mitigation (suppression) has been applied
- Reject: not meeting rules, no mitigation
- Notify that the statistic does not meet rules but an exception is requested (if the tool is being operated in a principles-based system), and a reason for the exception is given

This changes the data service's options significantly. First, only statistics are checked, not complete papers. Second, the tool has all the information necessary to carry out the check, and so researcher providing insufficient information is no longer an issue except in the case of exception requests. Third, the tool carries out tests exactly and verifiably (rules of automatic checking are in the SACRO manual; specific parameters are set by the data service/data providers) so human error in the review of runner is minimised. Finally, the tool should provide greater consistency in checks by removing some of the subjectivity from decision-making.

	Assumptions being made **Choices to be made**	**Comments**
A	The tool assesses the rules for runners accurately and comprehensively	Basic functionality for the tool.
	The researcher cannot edit results of the automatic review except by deleting the output	May be achieved through people training or technical measures
C	Should the tool replace one set of eyes for runners, or should the first review be a check on the tool?	If a new tool, may need to have some oversight to provide reassurance to the data service and third parties that the tool is doing its job correctly.
	How will code releases work?	Is a separate mechanism needed?

	Are exceptions/repeaters still to be seen by four eyes?	The tool can provide some insight but the decision remains with the checkers
G	This is still developing as the tools are not yet in operation, but using the tool to replace one pair of eyes and simplify the job of the second checker is its primary purpose	

Table 11 Using automatic tools for checking

Investing in IT systems

Data service staff repeated highlighted that an effective IT system makes a significant difference to the time needed to manage the output checking process. Characteristics of an effective IT system include:

- One-touch for notifying/recording
- Locking of submitted outputs
- Recording of the decision and communication with the researcher as a single process
- Recording of notes from conversations
- Recording of conversation threads
- Recording of offline conversations (getting on the phone to a researcher is often the most effective communication method)
- Simple identification of the stage an output request is at
- Identification when processes are under the control of the data service and when the data service is waiting for researcher input
- System can seamlessly interact with/be a part of the general customer management, including project and researcher management systems

Good practice

- While at present there is no consensus on recommended IT systems (and such systems have to fit in with wider university systems), there is substantial expertise in UK data services, and views on what works, which should be exploited to identify effective solutions.

Setting operating parameters

The data service needs to set the service level agreement (SLA), the length of time before a researcher can expect to get results cleared. External and internal targets can be different, and should be based upon expected time to clear outputs, not on resources available.

Suppose the data service takes

- One day to process 85% of outputs submitted on that day
- Three days to process 10% of outputs because of some sort of complexity

and the remaining 5% take an indefinite amount of time because of some specific feature of the output that requires substantial new thinking/second opinions.

The 5% need not contribute to the SLA or to regular clearance metrics (see below) because they skew figures towards much higher mean values. However, making a commitment to *respond* to an

output request (even if this is to say “this is a particularly complex output, and we need extra time to review it”) can cover the 5%.

On that basis the **external SLA** can be three days, as all outputs should be dealt with in some way by then. For researchers, a key part of the external metric is predictability. However, the expected time to respond is 1x85% + 3x15% = 1.3 days, and so the **internal SLA** could be 2 days to allow for seasonal variations, staff holidays, unexpected unavailability of checkers and so on. For the internal SLA it does not matter if some responses are longer and some shorter, as this is an internal metric to help monitor performance and allocate resources, not something to be held to.

Setting an SLA much longer than the expected time does not help the service. Evidence suggests that response times follow Parkinson’s Law, and will creep up as the SLA is increased. This is partly because it makes it harder for the data service to ask for resources when it does not meet its expected response times, as it is still meeting external response times. If actual response times are close to the SLA rather than the expected response time, this is an indication of a backlog. Extending the SLA does not deal with the backlog, but extends it. A growing backlog is a clear indicator of insufficient resources, not an unrealistic response time.

For third parties the same argument holds over SLAs. In this case the data service needs to be sure that the third party can keep to its SLA, and also clearly explain to the researcher where any delays are coming from.

Good practice

- Estimate expected response times and generate an internal SLA based on that, with some flexibility for staff absence
- External SLA should be a strong commitment to either assess the output or, in exceptional cases, explain why this is going to take longer than the SLA; it should be close to the maximum expected time for checking all but the expectional outputs
- Do not set a long SLA ‘just in case’; use holding responses to manage exceptional cases
- Monitor performance against the internal SLA as this is an indicator of whether resources are sufficient

Building the team and resources

In this section we consider what should be seen as minimum, acceptable and ideal standards for different parts of the output checking process: the researcher, the output checker, the output checking team, and the resources available to the team. We identify the standards as

- Minimum: the lowest standard necessary to run an output checking process
- Acceptable: expected standards to run a reasonably efficient and effective process
- Ideal: standards that are met in the most efficient services

We break these down into subsets, such as resources or people skills. We then identify measures that might help data services move from minimal to acceptable, and acceptable to ideal. We assume that acceptable and ideal levels inherit standards from the lower levels.

These views are derived from the March 2024 Output Checking Retreat. We are grateful to participants for their expert insight. The sections below largely reflect their collective views, with some editing.

What can be expected from the researcher?

Although the researcher is not under the direct control of the data services team, researcher training (formal and informal) has an impact of the effectiveness of the service. Therefore, it is important to consider what characteristics are desirable in a researcher, and whether training or other resource investments can help develop positive behaviours.

	Minimal	Acceptable	Ideal
Practical skills	• Communicates where to find specific files	• Attempts to minimise outputs and limit exception requests • Arguments for exceptions are clearly stated	• Minimal output requests • Leaves plenty of time before deadlines • Actively raises problematic outputs with OC team before submission
Training and data skills	• Absence of malicious intent in data access and use • Is aware of disclosure control rules and can apply basic measures	• Attempts to apply SDC requirements to output requests • Attempts to follow procedures and leave time for review	• Can apply SDC to high level • Well-presented high quality requests
Personality / people skills	•	• Is polite	• Sees output checkers as being part of the team

Table 12 Progression states for researcher development

Moving up from state to state can be achieved by:

- Researcher training (formal or informal) to develop statistical skills, explain procedures discuss the role of exceptions, repeaters and strangers
- Communication with research to encourage rapport with the clearance team

Researcher training, developing shared understanding and ownership of responsibility, encourages better cooperation with output checking teams. Positive encouragement is thought to be more effective than threats, but moving a researcher from 'minimal' may require (the threat of) punitive measures for those unwilling to learn and collaborate. Discussion of negative consequences for harmful practices, if necessary, should initially be framed around some version of "it's your own time you're wasting".

What should be expected from the output checker?

In this table we consider the characteristics of output checkers.

	Minimal	Acceptable	Ideal

Resources	• Time to review outputs. • Escalation points for issues and contacts if output checker needs help. • Robust and clear output checking processes	• Access to experts where needed • No duplicate processes	• Multiple people who are capable of supporting each other • Maximal automation of processes
Training and data skills	• Basic training in clearing runners, not including statistics or data • Knowledge and training in procedures	• Training in repeaters and dealing with exceptions • Training in statistics and data. • IT skills and software knowledge • Knowledge of data sources	• Training to handle strangers or aliens • IG knowledge • Prioritisation skills • Understanding research process and cycle • Awareness of AI issues
Personality / people skills	•	• Interpersonal skills: team player, able to communicate and persuade • Takes the process of output checking seriously	• Willingness to challenge authority • Patience • Communication skills, including holding difficult conversations • Able and willing to seek help when needed

Table 13 Progression states for output checkers

Moving up from state to state can be achieved by:

- Refresher training to help output checkers keep up to date with their skills
- Diverse skill sets among the staff to help skill sharing within a team
- Workload prioritisation aligned to skills for operational effectiveness (but note this may lead to fewer skills development opportunities)
- Adjusting the process and managing expectations to allow output checkers breathing room, and to provide space to develop skills and reflect
- Collaboration with other TREs to enhance local skills
- Clear progression opportunities tied to skills

The ESRC-funded Odyssey project https://odyssey-project-co-uk is currently developing job specifications and career pathways for data service professionals, reflecting the need for progression options. The latest information can be found on the website.

What should be expected from the output checking team?

In this table we consider how the output checking team as a whole needs to fit together, and support the individual checkers. We assume there is more than one person on the OC team.

	Minimal	**Acceptable**	**Ideal**

Member knowledge and team standards	• Members with basic training to deal with runners • Short escalation process, able to handle frequent requests • Needs some access to expertise for non-runners • Minimise exceptions – run rules-based as far as possible • Consistent guidance	• At least one team member trained to deal with most outputs • Escalation path for exceptional requests	• More than one individual with full training • Capability to handle (almost) all outputs and develop new guidelines • Variety of backgrounds within the team to ensure well rounded nature • Team has access to knowledgeable members and experts, possibly in other organisations • Consistent approach and standard for checking.
Numbers and time investment	• Enough team members to meet demand in most cases	• Enough team members to meet expected demand • Team members have some downtime from output checking	• Enough team members to meet demand sensibly in all reasonable situations • Large enough team for responsibility sharing and discussion. • Team members can commit time easily • Team members have time (and are encouraged) to pursue other research interests
Member temperament	• Team seeks to maintain positive internal relationships.	•	• Members who are engaged and enjoy their work • Team members feel empowered to flag issues, improvements, or changes with decisions makers.

Table 14 Progression states for teams of output checkers

Moving up from state to state can be achieved by:

- Additional resources, particularly time and flexibility
- Appropriate training in building positive relationships

What resources should the output checking team expect?

Finally, we consider what resources should be available to run different stages of output checking service. The table presents similar ideas as in the previous two tables, but organised differently by process, technology and people.

	Minimal	Acceptable	Ideal
Process	• Robust output checking processes • Clear/well described submission procedures • Tracking requests from airlock (audit)	• Defined process to follow with metrics and KPIs. • No duplicate processes • Processes are failsafe (default action is the preferred action)	• Maximal automation of processes/minimal manual processing by staff • Processes that empower the team and researcher without creating pressure to take on uncompensated responsibility or risk • Continued professional development throughout career. • Varied workloads with options to specialise
Technology	• Airlock or equivalent software.	• Tools to support and audit output checking (see above for discussion of basic IT system requirements)	• Researcher passport linked to ORCID or similar • Tool to allow semi-automatic checking (runners and repeaters) • Automatic metrics and auditing
People	• Someone to accept and release output (not necessarily any review). • Minimum of SDC checkers on rules-based judgement call. • Sufficient FTEs to meet external SLA	• Skills are appropriate, acknowledged and renumerated • Sufficient FTEs to meet internal SLA	• Motivated people • Necessary but diverse skills among team • Realistic time allocation allowing for non-OC activity • A supportive team and network • Multidisciplinary aid and skill sets

Table 15 Resource needs at different development states

Moving up from state to state can be achieved by:

- Using the expertise of SDC trainers and professionals to review and revise processes
- Attention to workload balance to allow time to be dedicated to long term sustainability and progression
- Building a self-supportive team
- Building links with the wider community for group learning
- Resourcing a culture of continuous improvement
- Encouraging independent learning
- Seeing both staff and researcher training as an investment
- Using process metrics to measure effectiveness, identify limitations, and make the case for resources

Developing useful metrics

At present there is little evidence or agreement how to develop useful metrics. Not all elements of the process are under the data service's control (e.g. third-party approval or waiting for the researcher to respond to a request for more information). Some obvious metrics (number of clearance requests dealt with) are easily obtained but of limited value. Some metrics (e.g. number of email messages between researchers and checkers) could generate perverse incentives. There is further work needed on this.

Nonetheless, some metrics suggested by the March 2024 Retreat may provide food for thought. These are listed below. Some may be extractable from administrative processes

Researcher Experience and Interaction:

- User experience - problems and complaints that may not be expected
- How many comments? How many of these were issues with output?
- How many refusals?
- How many cancels/retractions?
- Requests in vs requests out
- Reasons for revisions (missing info? SDC issues?)
- Requests for changes after submission
- Use of automated checking (by users or output checkers)

Output Metrics:

- Number of revisions needed for outputs
- Sizes of outputs (files, pages, sheets, cells, lines)
- Number of SDC outputs
- Number of returns to researcher – failures
- Metrics for statbarn class
- Composite complexity score: statbarn + number of outputs + perceived risk + …?

Output Checking Team Operations:

- Extensions to timelines – why were these given/needed? (e.g. illness, researcher availability)
- Number of routine vs exceptions vs number of outright reject
- Number of outputs missed by one checker but picked up by another

Researcher or Project Specific Metrics:

- Who makes requests? Professors? PhD students?
- What is the distribution of requests among projects?
- Breaches of procedure - what training did they receive?
- Number of requests for 'quick' review (i.e. more urgent than external SLA)
- Number of times rules broken
- Number of exception requests made and granted
- How does all this relate to the type of researcher (e.g. PhD, professor, consultant)

Time Metrics:

- Time spent on task
- Time from submission to first response to final release
- Time from submission to time actioned by team
- Time to post request

Some of these metrics have direct operational value (e.g. identifying the number of rejected exception requests and developing new guidelines for researchers). Others are more long-term measures that provide insight into the nature of the process (e.g. are the output requests from professors better than those of their PhD students?), and which could lead to indirectly to changes in processes.

We will update this section in 2025 when we have feedback from the data services.

Strategy: ongoing considerations for operational systems

In this section we assume that the OC processes are operating effectively. We now consider how a data service can ensure that it remains in control of its destiny and use its expertise effectively. We first begin by nothing the drivers, constraints and external influences on the data service's operations. We then bring these together in a SWOT (Strength, Weaknesses, Opportunities and Threats) analysis to help the data service identify its scope for independent and effective decision-making.

There are no good practice recommendations in this section, as this is about identifying items for consideration which are likely to be specific to the data service.

Identifying drivers, constraints and external influences

A number of entities external to the data service can have a significant impact on the service's activities:

External agent	Impact
Commercial entities	Commercial entities can have large scale impacts on implementation processes especially with newer projects by dictating technological solutions
Legal teams	Legal teams can place requirements for working in particular ways that can hinder or slow workflow
Researchers	Researchers can place unreasonable pressure on the teams
Data owners	Data owners can have expectations on how output checking should be done and what specifically needs to be checked
Universities/ data service host	HR processes and other resource models may find it difficult to understand the workload allocation model (including time not doing output checking) to run an effective OC operation

Publishers	Publishers' deadlines may be used as arguments to speed up the clearance process or skip checks
Output checking community	Other members of the output checking community can offer help and advice.

Table 16 External influences on the data service

Note that all but the last are seen as restricting the freedom of the data service to operate. However, these can be turned into positive relationships. For example, the legal team is almost certain to have little or no understanding of output checking, and if asked to express an opinion, is likely to provide a generic, and limiting, response. But the evidence shows that a commitment to communication, expertise and goodwill on both sides can generate a positive working relationship where the freedom of action of the data service is bolstered by the support of the legal team.

We now consider what drives change and improvement in output checking operations and what constrains them. We separate this into internal factors (arising from the output checking team, and under their control) and external factors which the output checking team may only be able to indirectly influence.

What drives change in output checking processes?	
Internal drivers	**External drivers**

• Innate desire to improve operations
• Desire to maintain/improve reputation for good practice
• Frustrations with current processes
• Ambition when compared to other TREs
• Awareness of IG good practices
• Staff attitudes:
  ○ Enthusiasm from new staff
  ○ Motivated old hands
  ○ Push from leadership
	• TRE community – suggesting good practices and new ideas
	• Specialist advisers eg DRAGoN group
	• Researcher deadlines and timelines
	• Funders who support improvement and (particularly) efficiencies
	• Data controllers

What constrains changes in output checking processes?	
Internal constraints	**External constraints**
• Lack of resources to maintain operations (funding, time and technology) • Lack of resources to design or implement change • Inertia • Unhelpful metrics • Bad infrastructure • High turnover of staff • Competing priorities	• Lack of funding • Constraints imposed by funders • Constraints imposed by data providers • Lack of trust in the data service • Researchers' resistance to change • Needing to follow best practice/ common standards/ other data services, even if inappropriate for your data service

• Staff attitudes:
  ○ Lack of autonomy
  ○ Team members opposed to change
  ○ Insufficient appreciation of internal skills
  ○ Lack of change management skills
  ○ Lack of support from leadership	

Table 17 Drivers for and constraints on change in output checking processes

Note that senior leadership, researchers and funders are all seen as both potential drivers of change, and potential sources of resistance to change.

SWOT analysis of output checking

A SWOT analysis (strengths, weaknesses, opportunities, and threats) is a popular tool for helping organisations review their current operations and identify areas for change. A SWOT analysis of the full data service would consider customers, funding, staff resources, strategic goals, and so on. Here we focus just on the particular features of the output checking processes: that the OC team is likely to be an island of expertise in a much larger operation:

Strengths	**Weaknesses**
• The OC team is probably the only source of expertise in the organisation, and can be recognised as such • Other TREs run OC operations which can support the views and practices of the OC team	• The OC team is probably the only source of expertise in the organisation, and so its view might be ignored, and its operations subject ill-informed requirements from others • Likely to be little support within the organisation for • May be very poor understanding of resource need for OC activities
Opportunities	**Threats**
• The OC team can build a reputation in the organisation as value experts • The OC team can work with external organisations to create an expert network, reinforcing good practices, sharing information on resource needs, and contributing to national/formal guidelines • The OC team can enhance the overall reputation of the data service by communicating their expertise to external parties such as the general public	• OC may not be valued as an activity, leading to fall in staff morale • OC may be treating as a functional task, rather than a specialism, leading to ineffective allocation of resources.

Appendix: novel and non-statistical outputs

Machine learning models

Machine learning models present particular challenges for output checking. Two substantial UKRI-funded projects (GRAIMATTER https://arxiv.org/abs/2211.01656 and SACRO[3]) have explored this at length. They have produced guidelines and metrics for data services, as well as tools for researchers to use, such as 'safe wrappers' (see https://arxiv.org/abs/2212.01233). However, these remain specialist analytical areas The key issues are that

- ML models may contain, or allow to be reasonably surmised, record-level information on respondents in the training data
- Metrics for ML models are not well-understood; they require interpretation by specialists, and what counts as an 'acceptable' risk is not well-defined at present ('definitely safe' and 'definitely unsafe' are reasonably well understood and agreed on, but there is a very large area of 'it depends' where most models are likely to fall)
- The risk from ML models arises not just for the model itself, but how it is going to be used; an ML model used to make predictions where requests can be monitored and limited presents fewer risks than an ML model where unlimited predictions can be made

At present the guidance is that researchers should engage with ML specialists in the data service at an early stage of research, to discuss the models to be used, likely outputs, and the way those outputs will be used. As not every data service has the resources to provide this level of support, data services should seek to build relationships with other data services that do, possible with a view to provide formal ML support.

The SACRO-ML team is continuing to explore ways to make checking more accessible to data services; see https://github.com/AI-SDC/SACRO-ML for updates.

Qualitative data and other complex output

For qualitative data there are relatively few guidelines on how to assess outputs. In theory they should follow the same processes as for statistical outputs. However, the lack of guidance and training means that resources to assess such outputs may be very limited. If a data service wishes to provide access to qualitative data, it needs to ensure that staff are adequately trained in assessing such outputs. Given the lack of current guidance, the data service should be prepared to invest in writing and evaluating its own guidance, and circulating this to other data services who are not so far along this road.

Complex outputs, such as anonymised images from MRI scans, can present an additional problem, in that the outputs may be very large. Systems designed to store copies of released outputs may not be able to cope with large binary objects being released.

[3] Paper and presentation by Jim Smith at Eurostat/UNECE workshop (session 'Output checking in RDCs) https://unece.org/statistics/events/SDC2023

At present we are not aware of obvious best practices to share. There are some presentations providing a basic introduction[4], and some references. However, we welcome contributions from data services as to what they currently see as good practice (whether they operate in that way or now), and we will aim to update this guide as such information becomes available.

[4] DRAGoN/NIHR Summer School on data governance in LMICs; slides by Elizabeth Green https://www.saferesearchertraining.org/data_governance/NIHR%20governance%20week%205%20outputs%20part%20b%20quals%20and%20complex%20data.pptx.