11th Biennial IFC Conference on “Post-pandemic landscape for central bank statistics”

BIS Basel, 25-26 August 2022

Sharing researcher-generated code and value-added documentation in a trusted research environment[1]

Louise Corti and Hannah Hodge-Waller,
Office for National Statistics, UK

[1] This presentation was prepared for the conference. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the event.

Sharing researcher-generated code and value-added documentation in a Trusted Research Environment

Prepared by Louise Corti and Hannah Hodge-Waller (Office for National Statistics)[1]

Abstract

This paper sets out work that the ONS Secure Research Service is progressing to enable sharing of documentation and code created by researchers accessing controlled data. We have seen an increasing demand from researchers, funders and some data owners to enable wider access to this value-added material. This is less about reproducibility concerns, and more about recognising and building upon the significant amount of labour that goes into preparing research-ready data, new derived variables, and measures and histories. The paper sets out some of the key issues we have been addressing with researchers and data owners and sets out options for sharing of code in our Trusted Research Environment.

Keywords: data sharing, code sharing, reproducibility, trusted research environment, INEXDA, data access

Content

[1] Louise is Head of Analytical Insights and Impact, and Hannah the Statistical Code Sharing Manager for the Integrated Data Service Strategy Division, at the Office for National Statistics. The authors would like to thank colleagues at the Secure Research Service at ONS who are contributing to our SRS code sharing pilot work: Ian Banda, Beth Hopkins, Lorraine Ireland, Lauren Kerry, Alice McTiernan, Beth Routley and Sabrina Corps.

1. Introduction

This paper provides an overview on pilot work to progress code sharing within the Office for National Statistics' (ONS) Secure Research Service (SRS), a Trusted Research Environment (TRE). The work builds on early ideas discussed at a workshop on researcher code sharing challenges, hosted by the Office for National Statistics and the UK Data Service in February 2022. At the meeting representatives of INEXDA joined colleagues from UK data access organisations and researchers, to consider the pros and cons of reproducible working in Trusted Research Environments (TRE) or RDC (Corti and Engeli, 2020). This paper is intended to share plans and experiences gathered so far regarding sharing of code based on research use of granular data.

It is now generally appreciated that that in addition to transparency around reported analysis, and the increasing use of reproducible code and processes, there is also value of making better use of the research work that goes on behind the scenes when getting data into shape for analysis. Recognising the significant legal and technical challenges involved in providing access to granular data, any wider sharing of code outside the closed researcher environment requires safeguarding statistical confidentiality, validation of its quality, and disclaimers to manage data owner and publishers' reputation.

Working to identify strong use cases, we have begun building polices, protocols, templates, and guidance for writing, submitting, quality assuring and publishing 'code' created by researchers. In the first phase, we have been focussing on 'value-added code' rather than purely analytical code; so, the preparatory work undertaken by researchers to prepare data, such as making their data 'research-ready' for modelling, creating new variables, histories and so on. There is no reason why 'analytic code' cannot be included in a TRE code repository, but typically this is made available outside of the TRE, for example, when a package of code is requested by a journal to confirm that published results can be reproduced on the same data.

The paper introduces the SRS, discusses code sharing benefits and efforts around reproducibility within UK government. It goes on to introduce our use cases for the code sharing pilot in the SRS and the protocols being developed. It finishes with feedback from researchers around sharing code and plans for capacity building in this space.

A final note is that the work is still very much a work in progress. Having recently been successful in recruiting a dedicated *Statistical Code Sharing Manager* role, we feel that the detailed work can now accelerate, and we can look to share completed protocols and case studies over the next few months.

2. About the ONS Secure Research Service

The Office for National Statistics (ONS) Secure Research Service (SRS) is a Trusted Research Environment (TRE). It gives accredited or approved researchers secure access to a wealth of de-identified, unpublished data (microdata) to work on research projects for the public good. The SRS is accredited as a Processor by the UK Statistics Authority (UKSA) under the UK legislation, the *Digital Economy Act (DEA) 2017* for the provision of data for research purposes. (Office for National Statistics, 2022)

To ensure safe use of these data sources, the Secure Research Service (SRS) makes use of the Five Safes Framework; the set of principles adopted by secure labs, which researchers and their organisations must adhere to when undertaking research. The Five Safes protocols provide complete assurance for data owners:

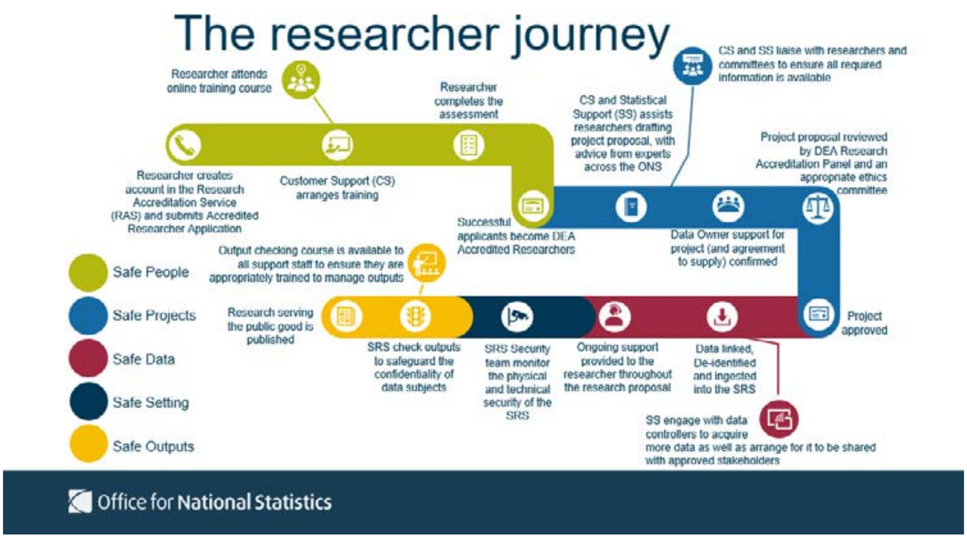
- **Safe People**: trained and accredited researchers trusted to use data appropriately
- **Safe Projects**: data that are only used for valuable, ethical research that delivers clear public benefits
- **Safe Settings**: settings in which access to data is only possible using our secure technology systems
- **Safe Data**: data that have been de-identified
- **Safe Outputs**: all research outputs that are checked to ensure they cannot identify data subjects

Researchers must become an Accredited Researcher and submit a project proposal for it be Accredited, via an online submission system.

Figure 1 sets out the typical journey for a researcher using the service[2]. Platform access is typically via remote access to the Windows-based cloud environment platform, where a unique project space is set up for Accredited Researchers to work on their Project. Data sets that have been approved for the work are mapped to the project space and requests can be made to add additional data. A range of software is available with the most popular being Stata and R.

The SRS researcher journey

Figure 1


The researcher journey
Researcher attends online training course
Researcher completes the assessment
CS and Statistical Support (SS) assists researchers drafting project proposal, with advice from experts across the ONS
CS and SS liaise with researchers and committees to ensure all required information is available
Project proposal reviewed by DEA Research Accreditation Panel and an appropriate ethics committee
Researcher creates account in the Research Accreditation Service (RAS) and submits Accredited Researcher Application
Customer Support (CS) arranges training
Successful applicants become DEA Accredited Researchers
Data Owner support for project (and agreement to supply) confirmed
Output checking course is available to all support staff to ensure they are appropriately trained to manage outputs
Research serving the public good is published
Project approved
SRS check outputs to safeguard the confidentiality of data subjects
SRS Security team monitor the physical and technical security of the SRS
Ongoing support provided to the researcher throughout the research proposal
Data linked, De-identified and ingested into the SRS
SS engage with data controllers to acquire more data as well as arrange for it to be shared with approved stakeholders
Safe People
Safe Projects
Safe Data
Safe Setting
Safe Outputs
Office for National Statistics

[2] Accessing secure research data as an accredited researcher. https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/approvedresearcherscheme#accessing-the-secure-research-service-srs

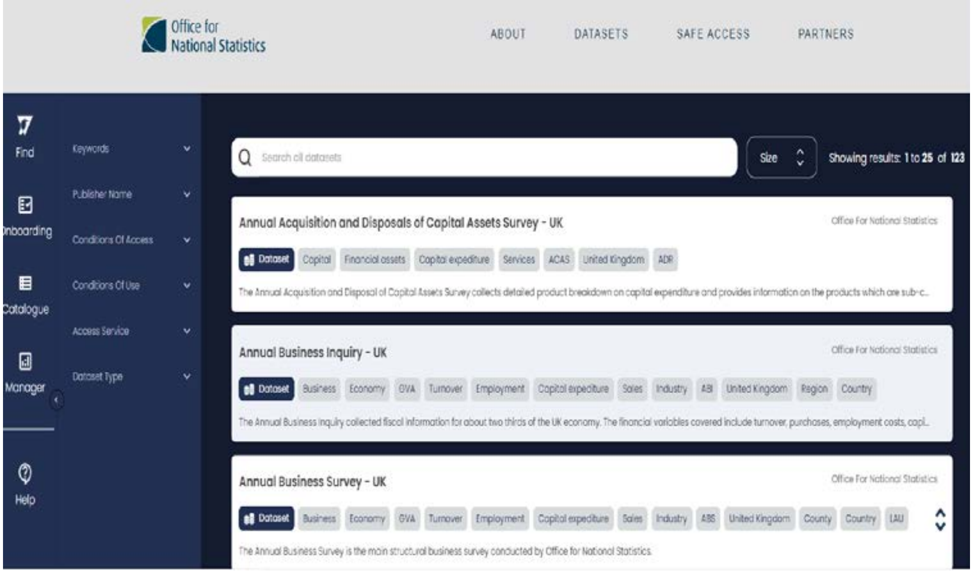
Source: SRS Engagement

Most datasets are available to access through remote access to the SRS, though in very few instances, data can only be accessed from an approved safe setting. The online metadata catalogue lists available datasets and any associated access restrictions[3]. There are around 125 datasets listed that can be requested for use in Accredited projects (Figure 2). These span a number of themes, with business data being amongst the most requested (Figure 3).

SRS data catalogue

Metadata catalogue for SRS data

Figure 2



Showing results: 1 to 25 of 123

Annual Acquisition and Disposals of Capital Assets Survey - UK — Office For National Statistics

Dataset · Capital · Financial assets · Capital expediture · Services · ACAS · United Kingdom · ADR

The Annual Acquisition and Disposal of Capital Assets Survey collects detailed product breakdown on capital expenditure and provides information on the products which are sub-c...

Annual Business Inquiry - UK — Office For National Statistics

Dataset · Business · Economy · GVA · Turnover · Employment · Capital expediture · Sales · Industry · ABI · United Kingdom · Region · Country

The Annual Business Inquiry collected fiscal information for about two thirds of the UK economy. The financial variables covered include turnover, purchases, employment costs, capi...

Annual Business Survey - UK — Office For National Statistics

Dataset · Business · Economy · GVA · Turnover · Employment · Capital expediture · Sales · Industry · ABS · United Kingdom · County · Country · LAU

The Annual Business Survey is the main structural business survey conducted by Office for National Statistics.

Source: SRS metadata catalogue, ONS. https://ons.metadata.works/

In August 2022 the SRS catalogue plans to roll out Digital Object Identifiers (DOI) 2220 for the SRS metadata catalogue, to uniquely identify and cite the datasets. This is the first DOI pilot within UK government digital publishing, and the use case will feed into wider guidelines for implementation in data catalogues and other government published material with research value.

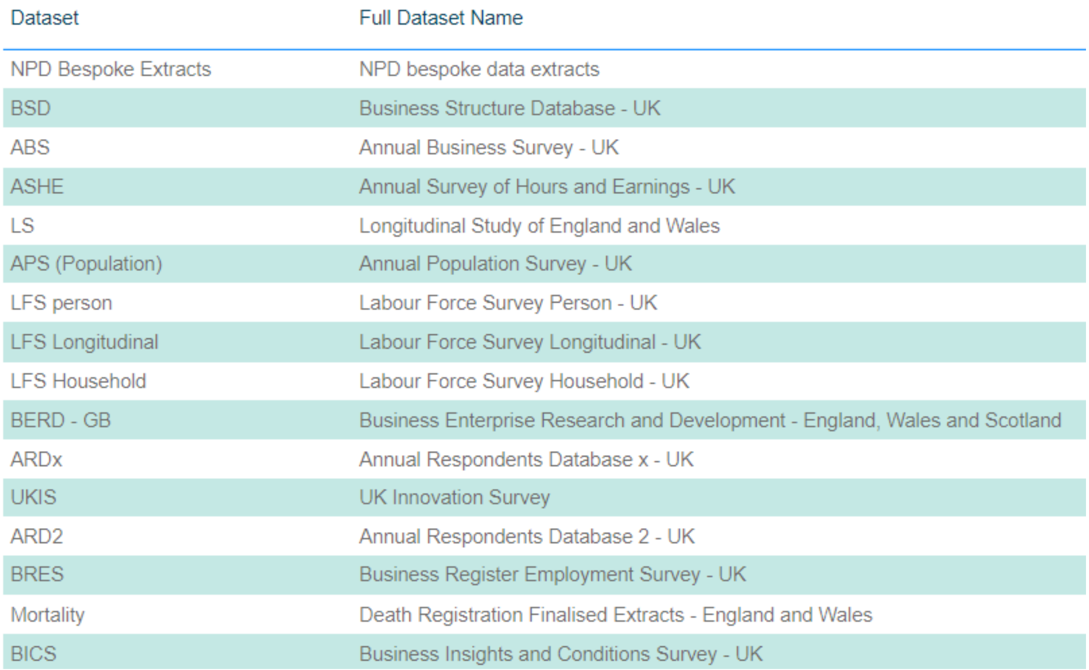
Figure 4 gives a summary profile of the volume of activity in the SRS over the past year. Overall, the service is supporting around 600 live projects, although research projects are not actively worked on all the time. The most popular research themes for this last 12 months have been education and 'health, primarily due some important linked data assets around the theme of education being ingested, and data to support analysis during the pandemic.

[3] Secure Research Service (SRS) metadata catalogue. https://ons.metadata.works/domain/index.html

Top data sources being used requested

May 2022*

Figure 3



Dataset	Full Dataset Name
NPD Bespoke Extracts	NPD bespoke data extracts
BSD	Business Structure Database - UK
ABS	Annual Business Survey - UK
ASHE	Annual Survey of Hours and Earnings - UK
LS	Longitudinal Study of England and Wales
APS (Population)	Annual Population Survey - UK
LFS person	Labour Force Survey Person - UK
LFS Longitudinal	Labour Force Survey Longitudinal - UK
LFS Household	Labour Force Survey Household - UK
BERD - GB	Business Enterprise Research and Development - England, Wales and Scotland
ARDx	Annual Respondents Database x - UK
UKIS	UK Innovation Survey
ARD2	Annual Respondents Database 2 - UK
BRES	Business Register Employment Survey - UK
Mortality	Death Registration Finalised Extracts - England and Wales
BICS	Business Insights and Conditions Survey - UK

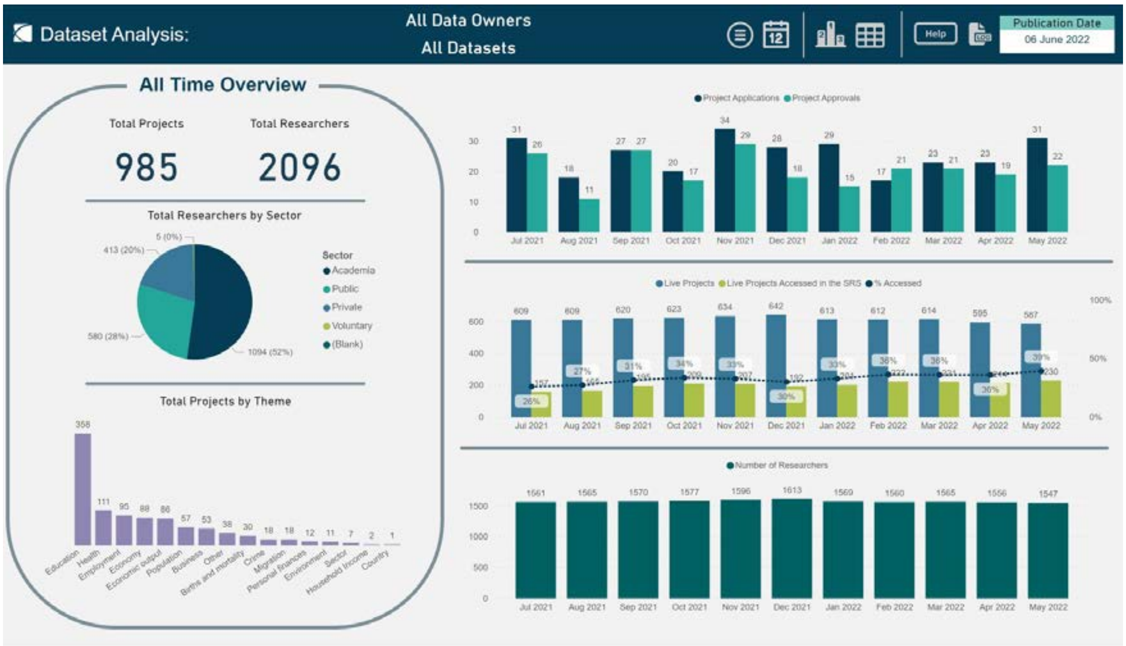
* NPD is National Pupil Database (Education administrative data)

Source: SRS Data Owner Dashboard

Research Activity in the ONS SRS

Cumulative number of accredited projects and researchers in the SRS, since July 2021*

Figure 4



Dataset Analysis:

All Data Owners
All Datasets

Publication Date
06 June 2022

All Time Overview

Total Projects: 985

Total Researchers: 2096

* Metrics are shown for the past 12 months, to provide an example of volume of activity.

Source: SRS Data Owner Dashboard

The Integrated Data Service (IDS) is a new and upcoming UK cross-government service that aims to build on the success of the Secure Research Service (SRS).[4] The IDS has the Office for National Statistics as the lead delivery partner and will enable co-ordinated secure access to a range of high-quality data, critical to informing policy decisions and improving public services.

3. Why share code?

The drive towards transparency and accountability in research has seen community and government-led policies being developed for opening access to research resources. An emerging 'reproducibility crisis' in some disciplines demands that more data and code should be released for substantiating published results. Indeed, code sharing is increasingly being viewed as best practice in empirical scientific research.

As a relatively high bar, research reproducibility for a published research article is where the authors provide all the data, code, and processing instructions necessary to rerun exactly the same analysis and obtain identical results. Contributing one's own code voluntarily is also valuable for the future progress of science. It demonstrates a willingness to follow open science principles, promotes a positive and collaborative approach and ensures that researchers are able to return to their own work later and remember what they have done.

Despite some disciplines leading the way with making underlying code available, sharing code is still not widely done across all in quantitative social and economic science. Economists, psychologists and political scientists have pushed forward with reproducibility mandates, keen to demonstrate research accountability. For example, the American Economic Association journals employ data editors to rerun code and validate results. As a tightknit community, demographers have also moved as a discipline towards R for analysis, and code sharing for complex routines is quite prevalent. Other disciplines, such as sociology or gerontology, appear more wary of sharing code/syntax, worrying more about not having been taught to write 'good code', and exposing this aspect of their work. This is both a cultural and a capability issue but is likely declining due to new research cohorts trending towards using open-source software and better code management practices.

Our interactions with researchers have shown that many feel anxious about exposing code they have written, considering that they may not have the skills to write 'good code'. Others do it routinely, submitting analytic code with their publications and publishing it openly in GitHub.

Gold standard code can be submitted to an external service for validation and receipt of a certificate of reproducibility, for example, the cascad[5] or CODECHECK[6]

[4] Integrated Data Service. https://integrateddataservice.gov.uk/

[5] The cascad service is a certification agency for scientific code and data. The certification allows researchers to signal the reproducibility nature of their research and to their peers. https://www.cascad.tech/

[6] The CODECHECK service supports code checkers with a workflow, guidelines and tools to evaluate computer programs underlying scientific papers. The independent time-stamped runs conducted by code checkers awards a "certificate of executable computation" and increase availability, discovery and reproducibility of crucial artefacts for computational science. https://codecheck.org.uk/

services. And code can be published in an open data repository such as or Zenodo[7], aiding citation and visibility (for example, with a DOI). Intellectual property for code can stay with the researcher but working collaboratively with data owners likely merits joint ownership.

As users of data, researchers can build upon existing code to avoid recreating basic recoding routines or derivation of complex histories. Access to syntax for derived variables created by data owners also supports the derivation of new variables for analysis.

There are challenges of reproducing work that's undertaken in a trusted research environment (TRE). Data access is restricted and unpublished material that has not been disclosure checked and approved cannot be taken out of the TRE. Any code shared outside the secure environment must first be reviewed for disclosure risk. Code tracking and versioning tools can also be used inside a TRE such as R Markdown, Jupyter Notebook and GitLab [8] to manage and document code.

The current 'As is' process for code sharing in the SRS does not include the use of a dedicated code repository but relies on a bespoke request from a researcher to be allowed to use code from another of their own or colleague's projects, using the existing statistical disclosure control (SDC) output' clearance mechanism. This process does not currently check code for good practice, and it almost impossible for researchers to know what useful code might be available, thereby preventing opportunities to experiment with useful data preparation work and derivations. The draft 'As is' and proposed workflows are set out in Section 6.

4. Good practice around coding across UK government

In its National Data Strategy, ONS signals its commitment to good practice and a commitment to transparency. Transparency, in its broadest sense, can be represented in many ways, but the ability to track back from 'fork to field' is important when it comes to analysis and modelling based on public data assets.

The UK response to the pandemic powerfully illustrated the benefits of responsible and effective use and sharing of data to understand COVID –19, to support people, and cooperate across borders. Stemming from this we are seeing a clearer understanding around good practice and how we enable our commitment to transparency. Making code available that underpins both data preparation and analysis promotes openness and provides value-added resources on which new analysis can build. Indeed, the IDS service signals its commitment to code sharing by providing improved toolsets and track-back functionality.

The Government Digital Service (GDS) has been a leader in promoting reproducible ways of working, setting high standards and providing capacity building resources for the government analysis community. Help on learning to code at GDS

[7] Zenodo is an open science repository for European scientists to submit research materials. https://zenodo.org/

[8] R Markdown (https://rmarkdown.rstudio.com/, Jupyter Notebook (https://jupyter.org/) and GitLab (https://about.gitlab.com/)

is included in their training programme.[9] The cross-government Civil Service network, the Government Analysis Function, also helps support good practice, and covers around 17,000 people involved in the generation and dissemination of analysis. It provides a solid learning curriculum covering methods, analysis and reproducible coding.[10]

At the ONS, the data engineering teams have been building capacity and capability by developing skills and reuse around processes and code, especially for statistical production. Reproducible Analytical Pipelines (RAP) experts work alongside with business areas to create sustainable data pipelines.[11] The ONS Data Science Campus routinely use data science methods and provide guidance on creating and sharing high quality sustainable code and pipelines. ONS also offers a host of great learning resources and opportunities, some of which are open to all. The ONS Data Access Platform Capability and Training Support (DAP CATS) team create and promote useful learning materials and coding resources, including a number of excellent Jupyter Executable Books[12] (Turrell, 2022; UK Government Analytical Community, 2020; Data Access Platform capability team, 2020).

The IDS is building on these forward-looking high standards that will meet reproducibility needs and avoid reinvention of wheels along the statistical production journey. Data can be easily updated through systematic engineering, and blocks of code for derived variables/measures and analytic outputs can be constructed and rerun by analysts. Not only is the IDP aiming for maintainable outcomes that can be updated and future proofed, shared methods libraries and code repositories will also be available.

5. Use cases for SRS code sharing pilot

While discussions in the SRS around enabling code sharing have been underway for some time, a number of timely and varied use cases have presented themselves; which we have adopted as exploratory pilots.

[9] Government Digital Service (GDS) https://www.gov.uk/government/organisations/government-digital-service, and its helpful blog https://gds.blog.gov.uk/. Blog on Learning to code at GDS https://gds.blog.gov.uk/2019/07/18/learning-to-code-at-gds/

[10] Government Analysis Function. https://analysisfunction.civilservice.gov.uk/ and the Government Analysis Function learning curriculum. https://analysisfunction.civilservice.gov.uk/learning-development/

[11] Infrastructure for Reproducible Analytical Pipeline (RAP) by the Government statistical Service (GSS):

https://gss.civilservice.gov.uk/reproducible-analytical-pipelines/infrastructure-for-rap/

[12] Jupyter Book (https://jupyterbook.org/intro.html) is an open-source project for building publication-quality books and documents from computational material. It builds on the Jupyter Notebook web-based interactive development environment for notebooks, code, and data. See the 'Executable Books' website: https://executablebooks.org/en/latest/gallery.html

5.1 ADR funded projects and Fellowships

The first opportunity concerns a group of funded Research Fellows working on SRS projects, supported by the UK's Administrative Data Research (UDRUK) Fellowship programme. ADR UK have been supportive of enabling researcher code sharing in the SRS and worked with ONS to (i) add a clause around delivery of code into the Fellow's contracts and (ii) fund a dedicated post within the ONS to help progress the work.[13] The Contract Terms and Conditions state that: *"Enhanced or derived data, code, products or tools for reuse created during this grant will be deposited in the ONS Secure Research Service as set out in 'Intellectual Property' below"*. This requirement certainly sets a bar.

In January 2022, the ONS impact team organised and hosted a workshop to discuss putting this code sharing clause into practice with the Fellows. Following a short introduction on good practice, the workshop used breakout groups to discuss challenges and opportunities. Attendees contributed their experiences and expectations when working with code – either supporting delivery of services or as part of research and intelligence activities, within a trusted research environment (TRE) or as an active contributor to public platforms such as GitHub/Lab.

The exercise gave a useful view on 'readiness' or sharing code and to plan where to start, as well as identifying what level of guidance would be needed to help support capacity building in this area.

Following from that one of the ADR funded projects, the Wealth and Employment Dynamics in Britain (WED)project, agreed to prepare code for wider sharing and work with us on a template and good practice guide for documenting Stata code. [14] The WED project is expected to increase understanding of how people's wages progress through their career, factoring in key demographic characteristics, as well as the particular dynamics of low-pay labour markets. The ONS helped supported the linkage of three key data sources on: employee earnings, a snapshot in time of all firms in the UK registered for tax, the 2011 Census data for England and Wale), and tax data relating to employment spells and earnings, as well as income from occupation pensions and benefits.

This project has the potential to transform understanding of wage and employment issues in Britain, from labour market entry, through job mobility and career progression to retirement decisions. The project has worked to create a new linked dataset and to fully document the Stata code detailing the data manipulation operations to create this new data. The code is currently being reviewed, using our emerging standards. In Section 6 we set out areas for review of code, including QA such as understandability and comprehensiveness of annotation, and sdc.

[13] ADR UK Research_Fellowship_Terms_and_Conditions.

https://www.adruk.org/fileadmin/uploads/adruk/Documents/GRADE_Research_Fellowship_Terms_and_Conditions_Aug_2021.pdf

[14] The 'Wage and employment dynamics in Britain' project is an ADR UK -funded data linkage project aiming to provide important new insights into the dynamics of earnings and employment in Britain. https://www.adruk.org/our-work/browse-all-projects/wage-and-employment-dynamics-in-britain-143/

5.2 Longitudinal Educational Outcomes (LEO) data preparation code

LEO is a de-identified, person level administrative dataset that brings together data on individual's education, employment, earnings data and benefits claims. The asset links data provided by five separate government departments. The dataset is widely recognised as having the potential to provide transformative insight and evidence on the longer-term labour market outcomes and educational pathways of around 38 million English learners, supporting government decision making in order to improve services.[15] The first iteration of this unique linked data source was acquired and delivered by ONS through the Secure Research Service.

There are currently around 50 researchers using, or applying to use, the data in the SRS across nine separate projects. Given the very large size and complexity of the data, researchers have to request (and justify the need for) a number of variables. The permitted variables are accessed via bespoke SQL views. Some projects actively worked with primary data owner, the Department for Education, in the early stages to produce derived variables and so on. Some of the code has been shared across ONS-led projects, and these are a target for wider sharing.

The code is primarily written in R and covers data manipulation activity, rather than pure analysis. One of the ONS projects was the first to externally publish policy relevant findings based on the data (Tolland, Tierney and Bathgate, 2021).

5.3 Covid infection study (CIS) code repository

Following the onset of the Covid-19 outbreak, the UK government agreed it was crucial to understand how Covid-19 was spreading across the population in order to control the pandemic and its effects. To assist the government's response, ONS, in partnership with the University of Oxford, the University of Manchester, and Public Health England, set up the Coronavirus Infection Survey (CIS).[16] Launched in April 2020 the study tested people for the virus whether they have symptoms or not, with data routinely collected from each individual in a household, via nose and throat swabs (measure of infection rate) and blood samples measure of antibodies). Socio-demographic and employment characteristics were also collected, along with symptoms experienced, self-isolating or shielding, and exposure to a suspected carrier of Covid-19.

The CIS was made available for analysis in the SRS. A large Accredited Project with around 80 researchers on it, began to analyse data in the SRS. Given the limitations of the secure platform, and the significant modelling ask, some analyses, using R, moved to the ONS Google Cloud Platform (GCP) where less granular CIS data were available.

GCP already uses GitHub to share code, using source repositories and Jupyter notebooks, and set up around ten separate repositories for each key analytical

[15] Accessing the Longitudinal Education Outcomes (LEO) dataset. https://www.gov.uk/guidance/apply-to-access-the-longitudinal-education-outcomes-leo-dataset

[16] Coronavirus (COVID-19) Infection Survey, UK Statistical bulletins. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveypilot/previousReleases

pipeline. Work in the SRS is more challenging to version control without access to an external location for remote repos, and associated tools for merging branches. In lieu, CISA built version-controlled code using a master folder as a remote location and cloned repos from there. Most pipelines did not use GIT but had only basic change control indicated (a combination of comments within the code and a commented changelog header at the top of the file).

This use case lends itself well to the pilot, as the multiple strands of code, and pipelines could be reviewed, combined and documented into key reference and actionable code for future use. This is very much a typical scenario, and, in hindsight, it is easy to question why analytic code wasn't meticulously managed or curated. However, the urgency of producing analysis for senior officials and policy makers at the time took precedence over code management. The experience of working across sectors also unearthed different approaches to documenting code.

The use case also lend itself well to a 'lessons learned' exercise around having gold standard protocols for the future. While retrospective tidy up can be lot of work, there is an opportunity to take a review. This example also provides a good example to see how the data producer feels about analysts' code; for example, would they wish to add disclaimers on any researcher-generated code, so as to distinguish added-value work from a formally published dataset?

5.4 'Discuss and Collaborate' space for the Integrated Data Programme TRE

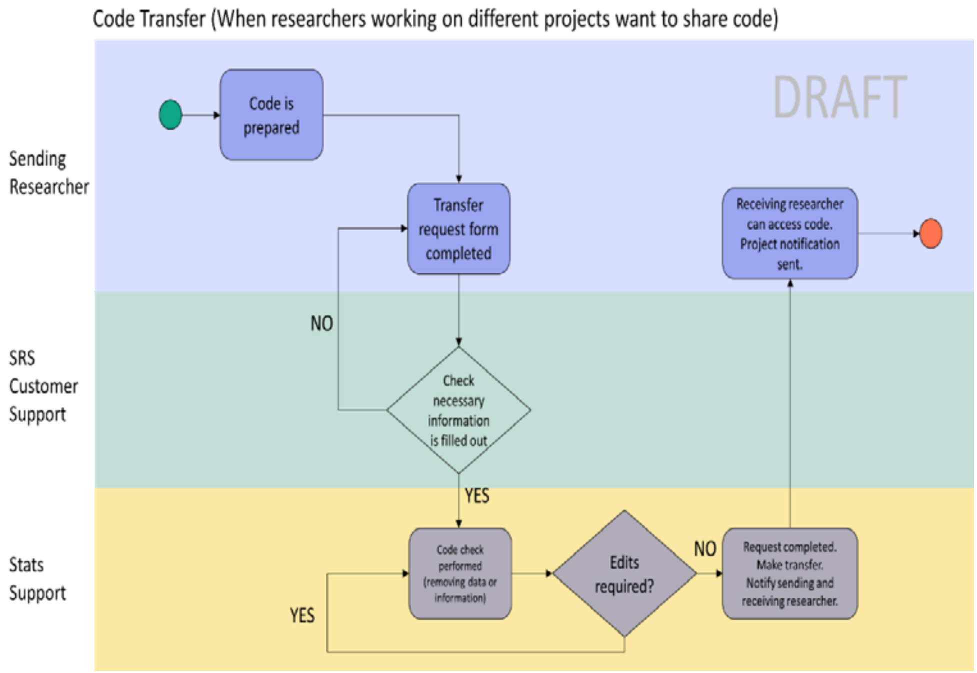
The IDS has aspirations to build on the great work of the ONS SRS. The platform aims to facilitate cross-government working and discussions are underway to plan a collaboration' space, where Accredited Researchers could formulate joint ideas for analysis and share methods, analysis and code outside of their own project spaces. The pilot work being undertaken in our SRS pilots is helping to discover both service, researcher, data owner and future user needs.

6. Workflow and protocols for code sharing in the SRS

The following diagrams show the process maps for the current and draft proposed journeys for code sharing in the TRE. Figure 5 shows the As-is process.

Current ‘As is’ for sharing of researcher code in the SRS

Figure 5


Code Transfer (When researchers working on different projects want to share code)
DRAFT
Sending Researcher
SRS Customer Support
Stats Support
Code is prepared
Transfer request form completed
Receiving researcher can access code. Project notification sent.
NO
Check necessary information is filled out
YES
Code check performed (removing data or information)
Edits required?
NO
Request completed. Make transfer. Notify sending and receiving researcher.
YES

Source: IDS Analytical Insights Team, Office for National Statistics

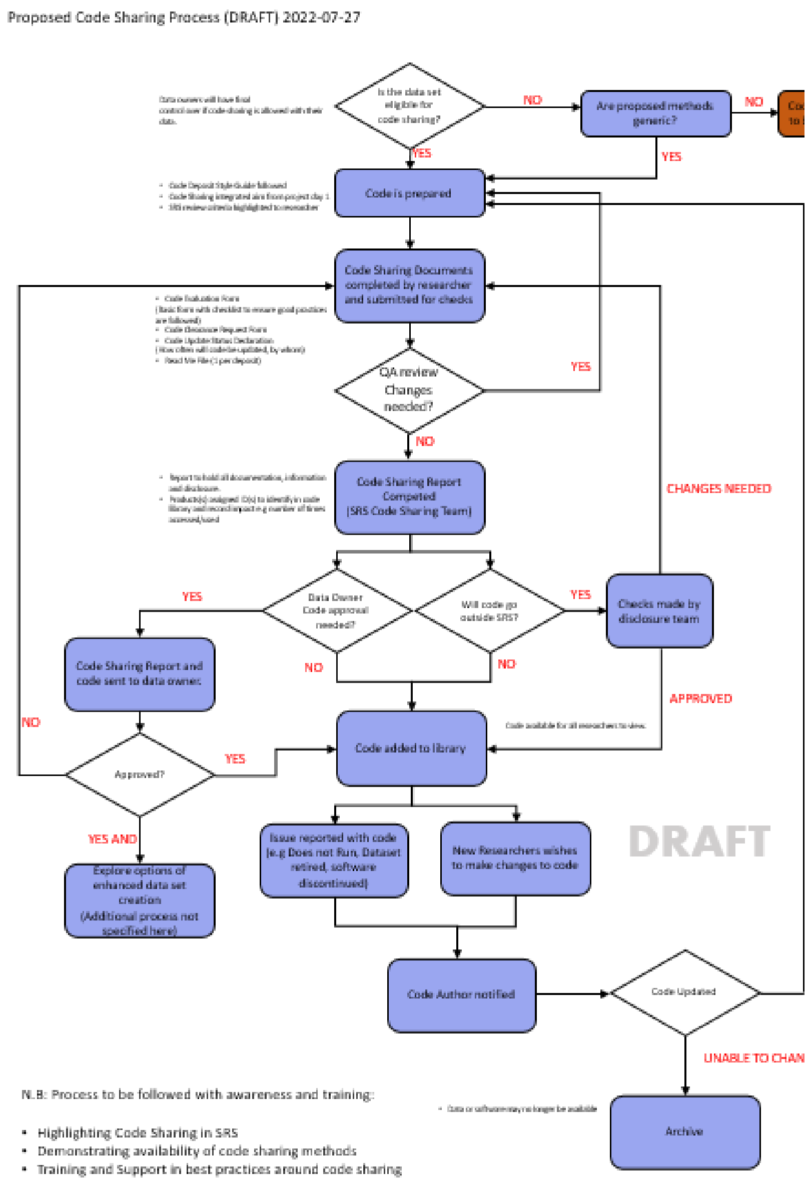
Wider code sharing models beyond 1-1 transfer between research projects has been scoped out. The **SRS Code Repository** or **Code library** must meet the skills of two sets of researchers, demarked by their existing code management practices.

These are:

- Group 1. Those used to managing and versioning code using a Git type environment; primarily R and Python users;
- Group 2: Those used to storing and managing code or syntax files using traditional folder structures, possibly (and hopefully) with file naming conventions (primarily Stata and SPSS users).

The plan is for the **SRS Code Repository** to be an SRS fully managed repository that is accessible to all those entering the environment. Thus, code should be findable, searchable and well documented, and have a workflow for submission, review and publishing. All code that is offered for deposit must be sdc reviewed and quality assessed. Figure 6 sets out the possible planned workflow.

Proposed workflow for sharing of researcher code in the SRS   Figure 6


Proposed Code Sharing Process (DRAFT) 2022-07-27
Data owners will have final control over if code-sharing is allowed with their data.
Is the data set eligible for code sharing?
NO
Are proposed methods generic?
NO
Co to
YES
YES
Code Deposit Style Guide followed
Code Sharing integrated aim from project day 1
SRS review criteria highlighted to researcher
Code is prepared
Code Sharing Documents completed by researcher and submitted for checks
Code Evaluation Form (Basic form with checklist to ensure good practices are followed)
Code Disclosure Request Form
Code Update Status Declaration (How often will code be updated, by whom)
Read Me File (1 per deposit)
QA review Changes needed?
YES
NO
Report to hold all documentation, information and disclosure.
Product(s) assigned ID(s) to identify in code library, and record impact e.g number of times accessed/used
Code Sharing Report Competed (SRS Code Sharing Team)
CHANGES NEEDED
YES
Data Owner Code approval needed?
Will code go outside SRS?
YES
Checks made by disclosure team
Code Sharing Report and code sent to data owner.
NO
NO
APPROVED
NO
Code added to library
Code available for all researchers to view.
Approved?
YES
YES AND
Explore options of enhanced data set creation (Additional process not specified here)
Issue reported with code (e.g Does not Run, Dataset retired, software discontinued)
New Researchers wishes to make changes to code
DRAFT
Code Author notified
Code Updated
UNABLE TO CHAN
N.B: Process to be followed with awareness and training:
Highlighting Code Sharing in SRS
Demonstrating availability of code sharing methods
Training and Support in best practices around code sharing
Data or software may no longer be available
Archive

Source: IDS Analytical Insights Team, Office for National Statistics

In addition to assessing the most suitable structure for either Git or folder structure libraries, Figure 7 sets out the mandatory documentation and user files that will be needed.

Code submission documentation

Figure 7

Stage	SRS Code Sharing Policy	Internal policies
Administration/ Internal	Data owner agreement	Include data owner options for review or disclaimer. Option to add to Data Sharing Agreement)
	Admin Check list	Check list to track governance, submission and code publishing steps
	Code Repository conventions	Set out repository conventions and is an Annex to the Code Deposit Guide, Includes: file naming and labelling. User Guide, Disclaimers, Citation statement
Before Project Starts	Code Templates	Sets out recommended styles, headers, types of annotation. For different software languages
	Code Deposit Guide	Guide to Good Coding and requirements (SRS oriented)
	Training Documents/Sessions	Optional; but intended to help assist new researchers or those new to code sharing with promoting best practices
During Project	Code Advice/Drop in Sessions	Focussed on best practice. Embedding sharing and reproducibility
Code submission point	SRS Code Sharing Policy	External policy
	SRS review criteria	Criteria used to review code that is submitted. Covers sdc and QA and required documentations e.g., ReadMe
	Code clearance request form	Simple form for submitting code with basic details
	Code clearance confirmation	Email to confirm clearance
	Code Update Status	Agreed with researcher, how often is code reviewed; process map if changes are needed to code
	Read Me file, per deposit	Describe the code file(s) contents and any key user information. Includes the citation for the material

Source: IDS Analytical Insights Team, Office for National Statistics

7. SRS user engagements on code sharing

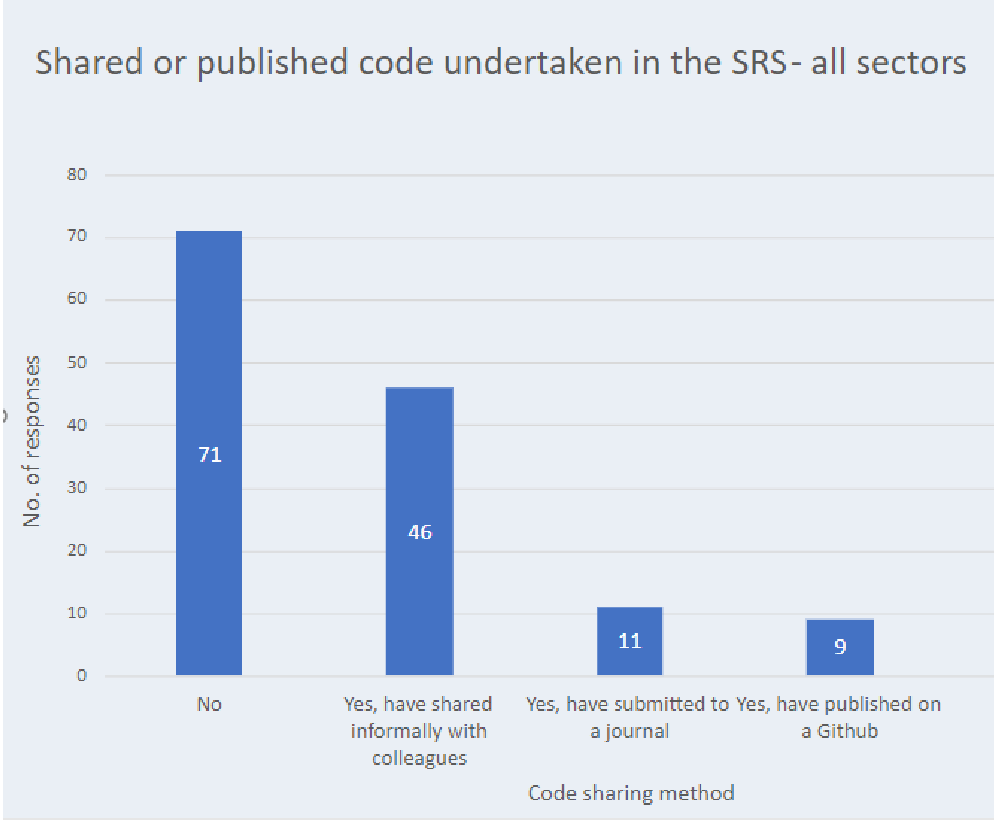
Over the past year, the Insights team have been seeking the views of SRS users on code sharing, in tandem with the pilot work to progress the sharing of researcher-generated code and additional user-focussed documentation. Overall, the engagement exercises showed that that both analysts and data owners see the value of sharing code.

Every year the SRS runs an Annual User Experience Survey in early spring. The survey asks researchers about their satisfaction with various aspects of the service, about data needs and on suggested improvements.

A question on code sharing was specifically included this year. Figure 8 suggests that code sharing is indeed happening, via various means inside and outside of the TRE including via the SRS internal libraries and externally via other organisations and websites. More government departments and private sector researchers had shared code with colleagues compared to within academia.

Code sharing through the SRS

Shared or published code undertaken in the ONS SRS* Figure 8



* Annual User Experience Survey (Secure Research Service). Delivered online February-March 2022. N = 57

Source: ONS Analysis Insights, Integrated Data Service, Office for National Statistics

When asked more generally, in an open question, about how ONS might better support analytical needs, Figure 9 shows code sharing as the second most volunteered suggestion. Comments volunteered by survey respondents around code **sharing ability and software packages** included:

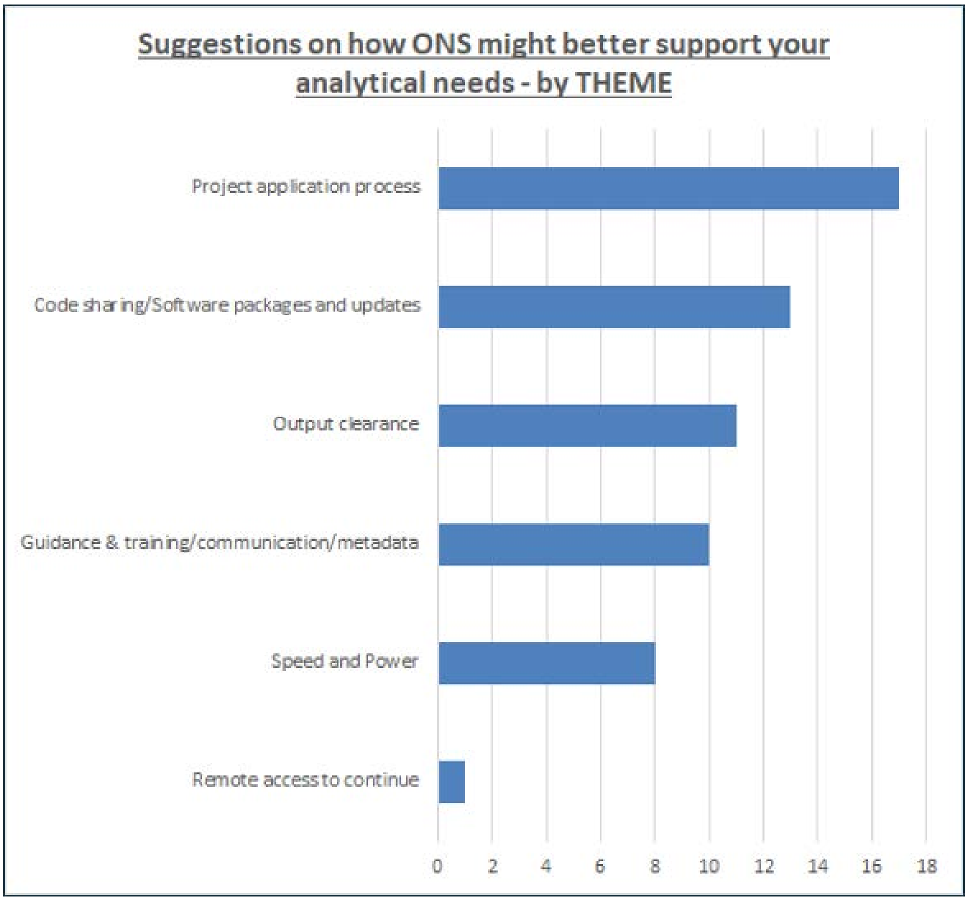
- Improve functions when adding to the environment
- Enable shared working areas
- Ability to share code
- Greater flexibility in the R libraries installed
- Enable Reproducible Analytical Pipelines in SRS, for instance RMarkdown
- Promote standard cleaning code across the government

Users were also invited to note whether using the SRS had enhanced their capability. Of the 95 responding, 40 mentioned coding skills. Of interest, 65 mentioned sdc awareness, and 45 noted data management.

How might ONS better support analytical needs?

Open responses coded by theme to the question, “How might ONS better support analytical needs”? in the Office for National Statistics (SRS)*

Figure 9



[*] Annual User Experience Survey (Secure Research Service). N = 56. Ran online February-March 2022

Source: ONS Analysis Insights, Integrated Data Service, Office for National Statistics

Following the survey, in a similar vein, in Spring 2022, the IDS Analytical Insights team undertook a series of focus groups with SRS users from government departments. While the main focus was on seeking a better understanding of the types of analysis being undertaken in the environment, the kinds of public *research outputs* being created and to hear about *actual or expected outcomes***,** we used the opportunity to ask about software needs and code sharing. Specifically, questions asked about **analytic and software needs** were:

- What **software** are you using to perform the majority of your analyses at present?
- Does it **meet your needs**?
- Does your team you have any **skills gaps** for the analyses?
- What tools do you use to **manage your code** within a secure environment? e.g., for collaboration or QA.

Comments volunteered on code sharing included:

"It would be great if there was a standard cleaning code across the government"

"Would be good to get better resources to enable Reproducible Analytical Pipelines in SRS, for instance RMarkdown."

We found out that many researchers did now know that there was a Git GUI already available to them for managing and versioning code for their projects. It is part of a standard set of software that users can access but has never been actively promoted.

8. Conclusion and future plans

This paper has set out an update on code sharing pilot work being undertaken for the ONS Secure Research Service. The timing is excellent, given that the challenges about how to go about demonstrating transparency and reduce repetition and redundancy in creating and running analysis pipelines come to the fore. Further, the topic is being raised in many forums, by researchers, funders and sponsors and engineers. The SRS Code Sharing group is continuing to connect in with other areas with ONS and promote its work, and will start to publish case studies, protocols, guidance and templates once they are ready.

Going forward, we have two strands of engagement activities being planned. The first is a call through our researcher networks for early adopters, so that we get a pipeline of keen researchers who want to have their code assessed and shared. The second is a series of Drop-In sessions and webinars on Writing and Documenting Good Code. There is still the need to highlight benefits and openly examine barriers and show how they can be overcome. Sharing tips and templates for how newcomers can get started is good first step.

References

Corti, L and Engeli, A (2020). #LoveYourCode2020. Blog, UK Data Service Impact Blog, 23 December 2020.

Corti, L., Ritchie, F. and O'Reilly, M. (2021). To share or not to share: code sharing in social science, National Centre for Research Methods (NRCM) News, 22 February 2022.

Office for National Statistics (2022). Accessing secure research data as an accredited researcher, ONS website.
https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice

Tolland, A., Tierney, J. and Bathgate, H. (2021). Education, social mobility and outcomes for students receiving free school meals in England: initial findings on earnings outcomes by demographic and regional factors. Office for National statistics.
https://www.ons.gov.uk/peoplepopulationandcommunity/educationandchildcare/ar

ticles/educationsocialmobilityandoutcomesforstudentsreceivingfreeschoolmealsinengland/initialfindingsonearningsoutcomesbydemographicandregionalfactors

Turrell, A. (2022). Coding for Economists, Data Science Campus, ONS. https://aeturrell.github.io/coding-for-economists/intro.html

UK Government Analytical Community. (2020). Quality assurance of code for analysis and research (version 2021.9). Office for National Statistics, Quality and Improvement division: https://best-practice-and-impact.github.io/qa-of-code-guidance/

Sharing researcher-generated code and value-added documentation in a Trusted Research Environment

August 2022

Louise Corti, Head Analytical Insights and Impact, ONS

louise.corti@ons.gov.uk

Overview

- Introduce researcher code sharing in the UK Secure Research Service (SRS)
- Highlight proposed processes, policies, documentation and support & training activities
- Highlight use cases

Who has shared code with analysts other than with close colleagues?

Why share code?

✓Viewed as best practice for demonstrating transparency and accountability in empirical scientific research

✓Enables building upon existing code to support the derivation of new variables, avoid recreating complex recoding routines

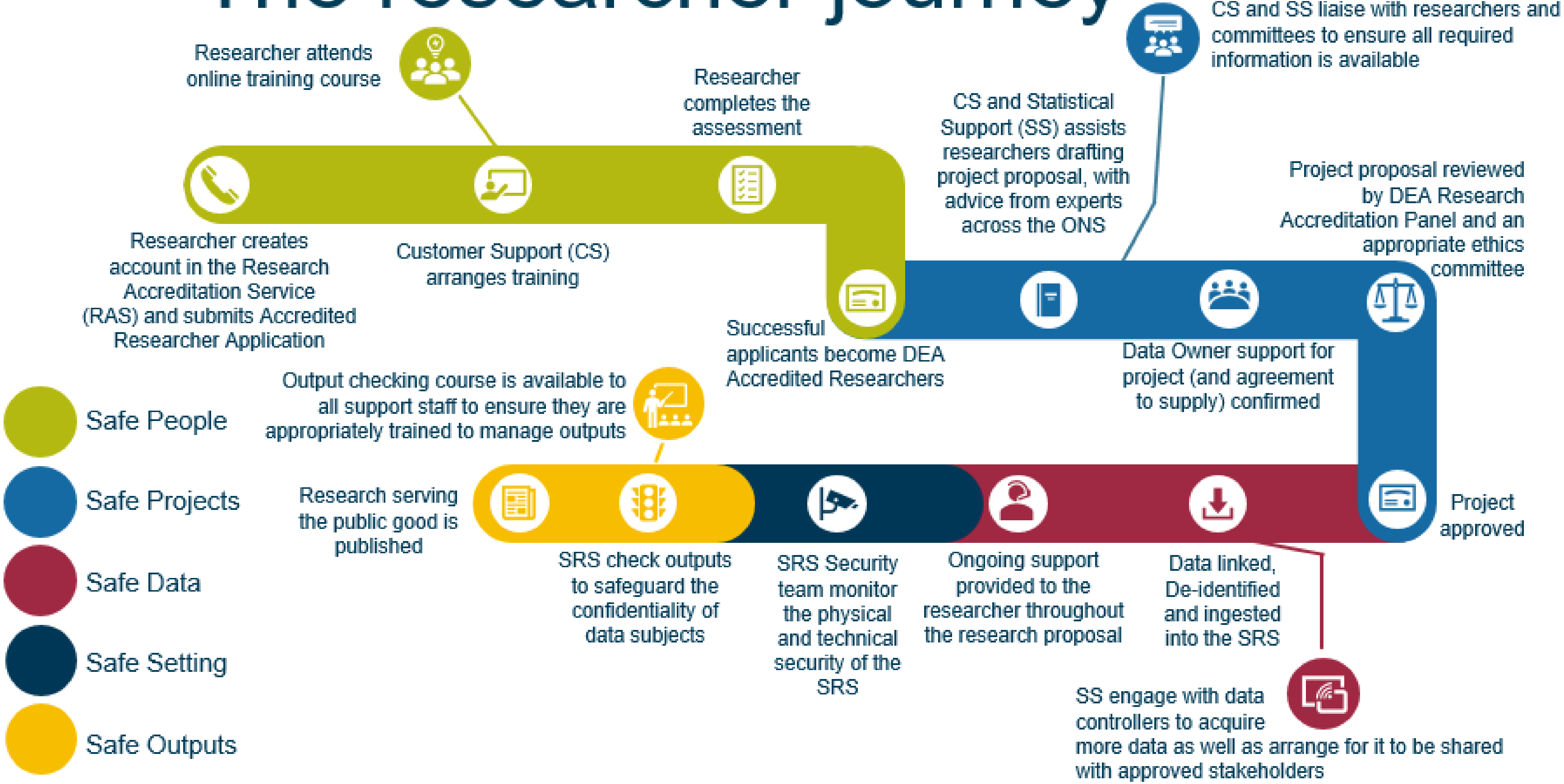
✓Exposes code for peer review /validation/ promotion

✓Some journals require underlying code submission

❖*Users can feel anxious about exposing their code; consider they don’t have skills or time to write ‘good code’*

❖*Users ask if data owners can supply code for derived variables*

The researcher journey



Researcher creates account in the Research Accreditation Service (RAS) and submits Accredited Researcher Application

Customer Support (CS) arranges training

Researcher attends online training course

Researcher completes the assessment

Successful applicants become DEA Accredited Researchers

CS and Statistical Support (SS) assists researchers drafting project proposal, with advice from experts across the ONS

CS and SS liaise with researchers and committees to ensure all required information is available

Data Owner support for project (and agreement to supply) confirmed

Project proposal reviewed by DEA Research Accreditation Panel and an appropriate ethics committee

Project approved

SS engage with data controllers to acquire more data as well as arrange for it to be shared with approved stakeholders

Data linked, De-identified and ingested into the SRS

Ongoing support provided to the researcher throughout the research proposal

SRS Security team monitor the physical and technical security of the SRS

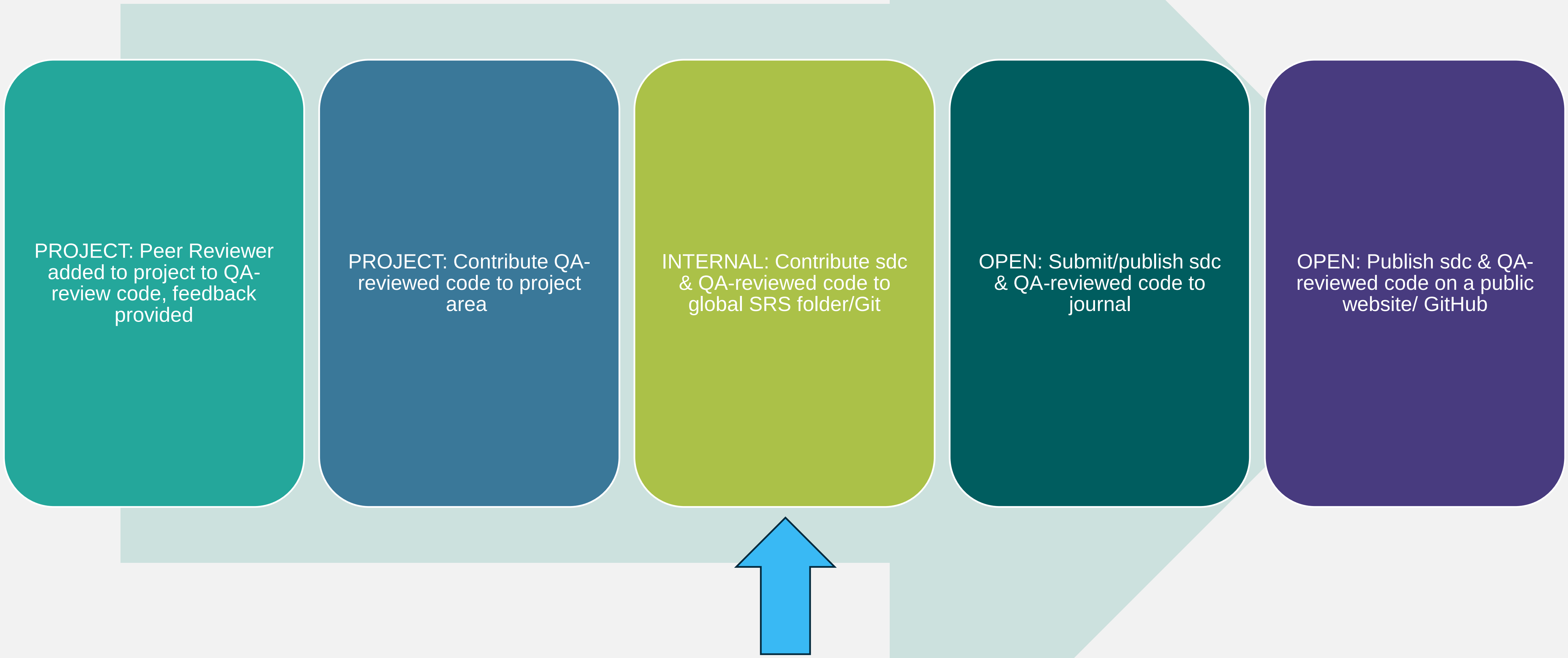
SRS check outputs to safeguard the confidentiality of data subjects

Output checking course is available to all support staff to ensure they are appropriately trained to manage outputs

Research serving the public good is published

- ● Safe People
- ● Safe Projects
- ● Safe Data
- ● Safe Setting
- ● Safe Outputs

Options for code sharing



PROJECT: Peer Reviewer added to project to QA-review code, feedback provided

PROJECT: Contribute QA-reviewed code to project area

INTERNAL: Contribute sdc & QA-reviewed code to global SRS folder/Git

OPEN: Submit/publish sdc & QA-reviewed code to journal

OPEN: Publish sdc & QA-reviewed code on a public website/ GitHub

Aim of pilot work

- Facilitate planning - code sharing group set up and manager role recruited
- Locate suitable use cases and start investigations/solutions

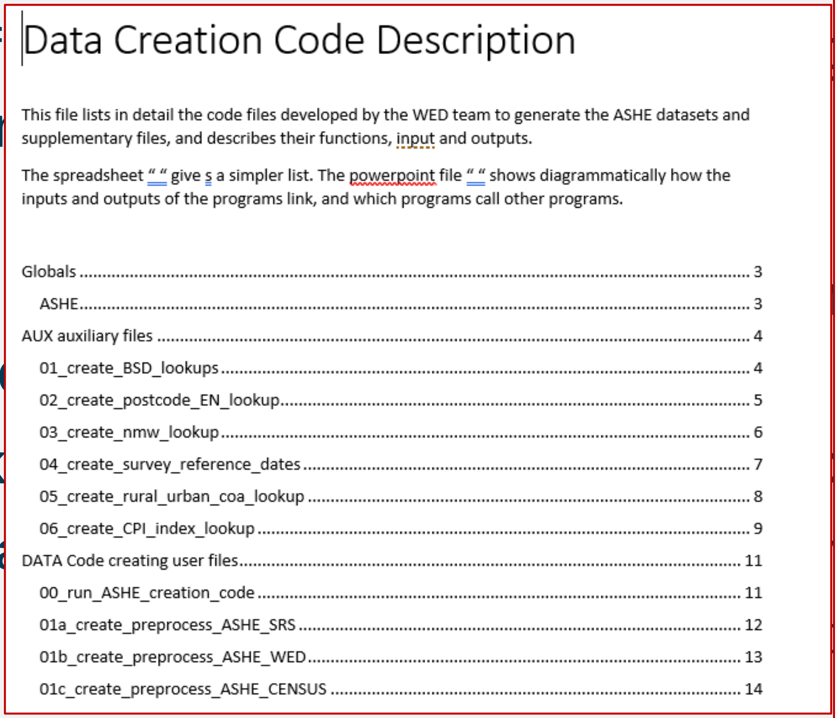
Explore and develop workable processes and protocols:

- ➢ 'As is' and proposed workflows
- ➢ Governance and administration, resourcing
- ➢ SRS and user policies
- ➢ User guidance and templates
- ➢ Capability building activity: webinars, 1-1 drop in sessions, blogs and case studies

- Roll out early adopter call and training sessions

Use case 1: Wealth and Employment Dynamics in Britain (WED) project

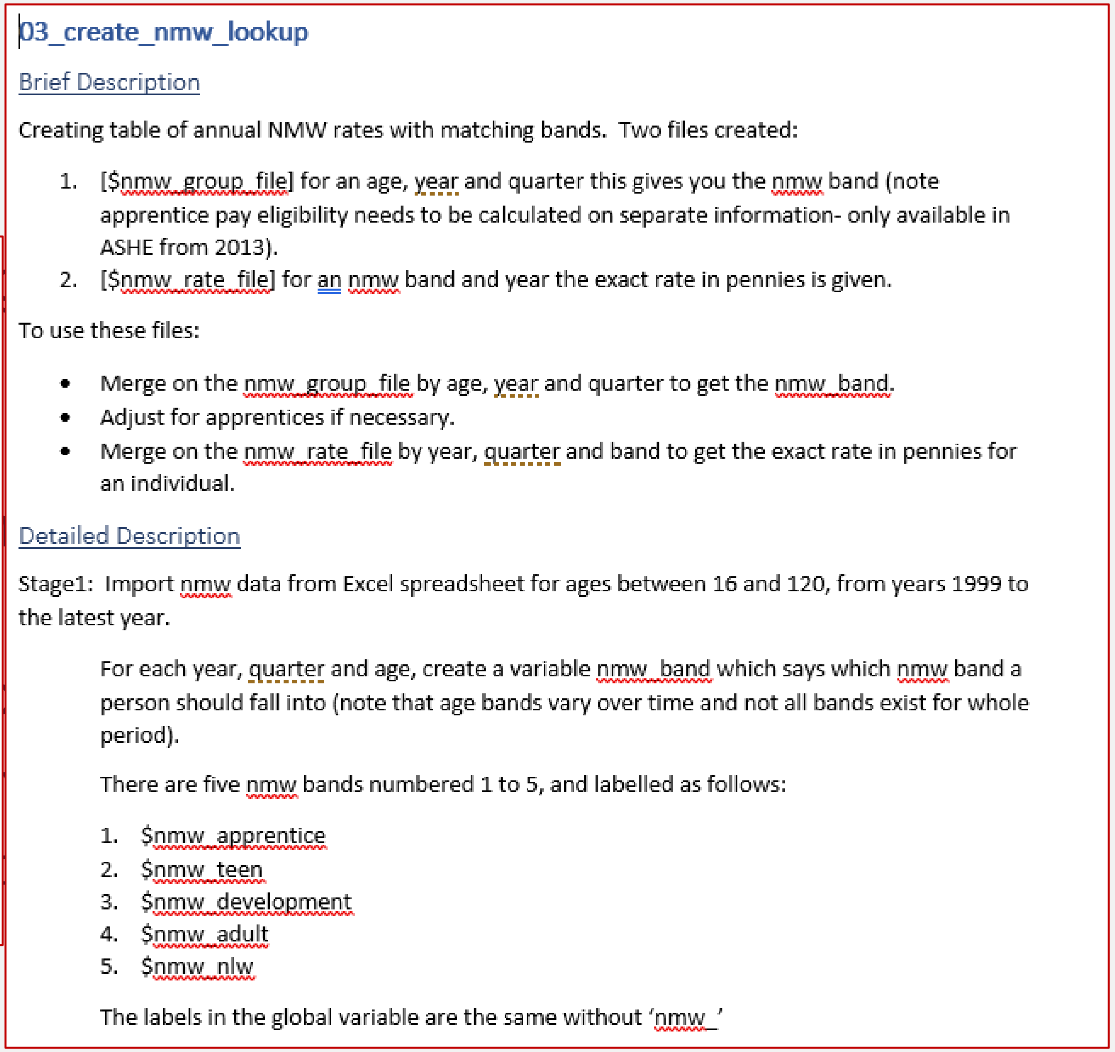
- Project aims to transform understandi
  Britain, f
  retiremer
- Involves
  (ASHE),
  occupati
- Project k
  - Data ma
  - Testing



Data Creation Code Description

This file lists in detail the code files developed by the WED team to generate the ASHE datasets and supplementary files, and describes their functions, input and outputs.

The spreadsheet " " give s a simpler list. The powerpoint file " " shows diagrammatically how the inputs and outputs of the programs link, and which programs call other programs.



03_create_nmw_lookup

Brief Description

Creating table of annual NMW rates with matching bands. Two files created:

1. [$nmw_group_file] for an age, year and quarter this gives you the nmw band (note apprentice pay eligibility needs to be calculated on separate information- only available in ASHE from 2013).
2. [$nmw_rate_file] for an nmw band and year the exact rate in pennies is given.

To use these files:

- Merge on the nmw_group_file by age, year and quarter to get the nmw_band.
- Adjust for apprentices if necessary.
- Merge on the nmw_rate_file by year, quarter and band to get the exact rate in pennies for an individual.

Detailed Description

Stage1: Import nmw data from Excel spreadsheet for ages between 16 and 120, from years 1999 to the latest year.

For each year, quarter and age, create a variable nmw_band which says which nmw band a person should fall into (note that age bands vary over time and not all bands exist for whole period).

There are five nmw bands numbered 1 to 5, and labelled as follows:

1. $nmw_apprentice
2. $nmw_teen
3. $nmw_development
4. $nmw_adult
5. $nmw_nlw

The labels in the global variable are the same without 'nmw_'

Use case 2: Longitudinal Educational Outcomes dataset

- LEO is a de-identified, person level administrative dataset that brings together data on individual’s **education, employment, earnings data and benefits claims**
- Asset links data provided by **five separate government departments** via the SRS
- Dataset has the potential to provide transformative insight and evidence on the longer-term labour market outcomes /educational pathways of @**38 million English learners**
- 9 projects /50 researchers using data, including government users
- Early data manipulation work to create '**research-ready 'datasets/new variables**
- Some R code shared across ONS-led projects; target wider sharing in the SRS

Use case 3: Large scale Covid survey analysis

- April 2020 new survey launched from ONS, Universities, Public Health England: the Coronavirus Infection Survey (CIS), available in the SRS
- Interviews with each individual in a household, including nose and throat swabs (infection rate) and blood samples (antibodies)
- Large project with 80 researchers with **urgency and significant modelling asks**. Directly **informed government decision-making**
- Varied software use: R users preferred ONS Google Cloud Platform with less granular data
  - GCP uses GitHub to share code - ten repositories set up for each key analytical pipeline
  - SRS - initial poor management of code; later built basic version-controlled code using a master folder
- Review the code repositories for **lessons learned** for large multi-sector projects
- Work with data owner to review **publishing of analysts' code** in the SRS; distinguish added-value work from formal data documentation

Useful ONS resources

- Blog: https://www.gov.uk/government/news/coding-from-zero and https://intranet.ons.statistics.gov.uk/blog/coding-from-zero/
- Quality Assurance of Code for Analysis and Research
- Reproducible Analytical Pipelines (RAP) Champions
- Data Accelerator programme
- Reproducible Analysis — Coding for Economists (aeturrell.github.io)
- Tips for Better Coding — Coding for Economists (aeturrell.github.io)