**The role of Statistical Computing in delivering quality**

**Amy Large**

**Gareth James**

**Office for National Statistics**

**(ONS)**

**Abstract**

The UK Office for National Statistics recently published its on-going strategic aims. Several relate to improving output and process quality. One way to achieve these aims is through a more innovative and coherent approach to Statistical Computing.

The ONS Statistical Computing Branch (SCB) was established to co-ordinate projects and act as a central focus for development in the field of statistical computing. The main objective was to help bridge the gap between statisticians and the IT teams responsible for coding and maintaining statistical production systems by drawing on methodological expertise and giving statistical direction to projects.

In addition, the team provide a resource for small-scale redevelopment projects using standardised software and more flexible approaches whilst maintaining standards. This is vital in reducing risks inherent in legacy systems for which large scale re-engineering is not immediately appropriate.

This paper will explain how the work of SCB contributes directly to the strategic aims of improving quality; minimising the risk of errors; having flexible and efficient processes; all underpinned by sound methodology. This will include discussion on the operational aspects of resourcing the work, developing standards, and opportunities for re-use of solutions, highlighting recent successes, which will be of interest to other NSIs.

1. **The context of Statistical Computing in ONS**

In April 2012, the Office for National Statistics underwent significant structural change. One of these changes was to bring together the IT Applications and Development and Methodology Divisions into what would later become the Research, Development and Infrastructure Directorate. The motivating factor for the change was a clear understanding that, within a Statistics Agency, to be innovative, responsive and to produce outputs of the best quality, we need to ensure that both our Methodology and our IT capability is robust and efficient, and closer working across these two areas would greatly benefit the wider process of statistical production.

In order to facilitate this closer working, the Statistical Computing Branch (SCB) was formed. As a small team not tied to a single part of the Statistical Business Process Model (SBPM), SCB were free to offer assistance not just to the teams within their Directorate, but across the Office to bridge the knowledge gap between IT, Methodologists and areas involved in the production of statistics. The goal of the Branch is to help identify systems and processes throughout the Office which are deemed to be high-risk (often due to inherited IT systems, the use of unsuitable software or the lack of suitable automation, allowing for human error), and to help reduce that risk.

* 1. ***Team structure and resourcing***

The team consists of one team leader and (at this time) four senior researchers with expertise in a variety of different business areas across social and business statistics. All have an interest in continuous improvement, a proven track record of good practice in programming and are confident in carrying out consultancy activities with project owners and other stakeholders.

In terms of work, projects can be allocated to the Branch through several routes:

* Direct contact from a project owner,
* Branch involvement in strategic projects,
* Projects being requested through the IT team, which are then deemed to be better suited to an SCB project,
* Identification of risk (often as a result of a strategic review) and a subsequent project being proposed to the business areas.

Some funding is allocated to the Branch to secure time for strategic projects (such as National Accounts and the Beyond 2011 population estimates research programme). However, outside of those commitments the Branch is free to engage across the Office based on what work is deemed to be of highest priority without the need for additional funding from any business area at this time. This means that, if a business area identifies a particular need but does not have funding available for a full IT project and solution, they can still look at addressing the problem through the Branch.

Due to the small size of the team, the Branch also works in a ‘hub and node’ approach. The Branch acts as a central resource (the hub) and the members can take on project work themselves. Throughout the Office there are also individuals who have been identified as having particular skills in coding, specific methodological areas or business processes. After a process of accreditation, and with agreement from their managers, these individuals can then be called on to act as resource to carry out project work. The projects may be within their own business area (and is usually the case). However, projects often arise from areas focused on the collection of data or the production of an output that requires specific methodological knowledge. In this case, nodes may find themselves working on projects for other business areas too. If project work is being carried out by a node, the hub acts as the project manager and offers guidance and quality assurance throughout the project.

* 1. ***Project types***

The types of project that the Branch is involved with fall into two categories:

1. Type A: These are large strategic projects. The Branch normally sits on the associated Boards, and can help raise awareness of technical and methodological issues throughout the project. These include National Accounts, Measuring Well-Being and Beyond 2011 (looking at the future of population estimate production). Other Branch services to these projects can be to provide guidance and training on best practice to the researchers involved in the work directly, and assist with delivery of small scale elements of the project (similar to Type B projects).
2. Type B: Small scale redevelopment work. This can be a system or area that has been identified as strategically high risk by the Office, a piece of work that the business area have identified as a local high risk or an area that the Branch can see large benefits to the Office from being redeveloped. Redevelopment work can be carried out by the branch or by nodes with hub managing the project. Any approach requires substantial communication with business areas. Solutions are often built in short sprints (the Agile approach), giving benefit as soon as possible and allowing for changes to develop as requirements change or are refined.
3. **ONS Strategic Aims**

The ONS Strategy [1] was launched in March 2013. It details the mission, vision, values and strategic aims for the Office for the next subsequent ten years. The nine strategic aims outlined were:

* Inform debate and have greater impact on decision making.
* Dramatically improve the communication of our statistics and analyses.
* Be highly regarded by our customers for producing trustworthy statistics and analyses that anticipate their needs.
* Be at the forefront of integrating and exploiting data from multiple sources.
* Have flexible and efficient processes and systems for statistical production, underpinned by sound methodology.
* Improve quality and minimise the risk of errors.
* Keep the data we hold secure.
* Be a statistical powerhouse at the heart of the Government Statistical Service and the European Statistical System.
* Have skilled and motivated people who are enthusiastic for change.

The Branch directly contributes to two of these Strategic Aims, as detailed in the following sections.

* 1. ***Flexible and efficient processes***

The Branch focuses on helping build flexibility and efficiency into processes by encouraging business areas to make better use of the following techniques:

* Modular code structures – By building systems in modules (usually one module per specific element of the process, or per function required), this means that, later in a systems life, blocks can be easily switched out and replaced or modified without the need for a complete redevelopment of the system. It can also make a complex statistical process easier to understand (if appropriate naming conventions are utilised), and can encourage the sharing of code across processes with common elements.
* Parameterisation – By encouraging coders to consider what will change over time (such as datasets, reference years or particular values that drive the system) and making these parameters, as opposed to hard coding values in, the system becomes more flexible and will require less updating and maintenance.
* Sharing code – By developing robust code in a modular fashion, it is possible for elements of common functionality to be shared across the wider Office. This will enable us to share best practice, potentially test new and innovative methodology across multiple processes and continuously review functions with an aim to making them more efficient and robust.
* Improvements in documentation and making code more transferrable – Putting suitable support in place will ensure that systems are understood, and the time taken to learn how to operate each system will be greatly reduced.
* Right software / platform / method – By making coders more aware of the software and platforms available, and how these fit with proposed methodologies, we can pick the right combinations to ensure that systems are efficient and flexible.
* Programming good practice – Encouraging a culture of good programming practice across the Office, providing suitable training and leading by example and through collaborative working ensures that there is greater awareness of the issues around efficiency and flexibility, and a greater acceptance of the time and steps required to make these a reality.
  1. ***Sound methodology***

When undergoing any kind of system redevelopment, the Branch strongly advises that any new system is tested and compared against any pre-existing system. This is to ensure that any methodology implemented has been consistently applied, ensuring that any new version has the same sound methodology as the old. This also acts as quality assurance that the new system is performing as expected, and can provide a useful performance indicator by comparing system running times to check for efficiency.

As mentioned in section 2.1., utilising a modular coding approach allows new methodology to be tested and implemented easily, ensuring continuous improvement in methodology. The facility to share these blocks of code also allow the innovative sharing of common and methodological processes.

* 1. ***Improving quality***

The dimensions of Output Quality (as defined by the ESS) are relevance, accuracy, timeliness and punctuality, accessibility and clarity, comparability and coherence. Several of these are directly affected if we are using systems that are poorly developed, irregularly reviewed and maintained, badly defined and inefficient. It is these factors that the Branch is actively seeking to raise awareness of and improve.

In addition to Output Quality, the ESS also defines Process Quality. The dimensions are efficiency, flexibility, transparency, robustness, effectiveness and integration. By encouraging coders to take more time over their system builds, and to carry out development in a more robust way, taking into consideration on-going support and documentation and robustness over time, the Branch is helping to improve Process Quality across the Office.

* 1. ***Minimising the risk of errors***

The main problem that has been described during IT software development is that of a lack of common understanding between the developer and the business area. When business areas develop their own code, there is often a lack of understanding of the tools used to minimise the occurrence of error that are implemented in an IT build. By getting a better understanding of how errors can occur in a process, appropriate steps for avoidance and monitoring can be introduced. The work of the Branch is often dictated by a review process that has identified an error, which may include:

* Human error – several high profile system or output errors have occurred due to copy and paste errors in spreadsheets, or providing incorrect files to a process.
* Data error – the data is in an unexpected format, contains inappropriate values or is for the wrong time period / industry etc. If these errors are not identified they can filter through to later processes
* Process error – where an incorrect method has been applied. This may also occur if an inappropriate version of a system is used.
* Lack of QA – all of the above errors and many others can often be identified within a process if suitable QA mechanisms have been introduced in the process. These include error trapping and handling in systems, and consistency checks on data and parameters.

1. **Branch objectives**

As has been detailed in the previous sections, the Branch is heavily involved with project work. This tends to be either of a long-term and strategic nature, or responsive to an identified risk or issue. In addition to these types of work, the Branch has several objectives to make process improvement and good practice in programming an integral part of methodological development.

* 1. ***Standard setting***

Part of the process of continuous improvement is to identify those areas that need improvement. Through collaboration with colleagues in IT and engagement with areas involved in statistical production around the Office, the Branch identified that the actual process of developing code or software outside of IT was not robustly developed. This leads to systems that gradually become unsupportable to due poor coding, lack of documentation and lack of understanding of what the code can and will do. As a response to this, the Branch has developed a one-day training course in good practice in programming. The course is aimed at researchers, statisticians and analysts involved in coding, and in any language – although the majority of the examples are given in SAS, as this is the current corporate tool for the Office. They do not need to have had formal programming training.

The syllabus covers:

* Program structure and design
* Version control
* Documentation
* Performance and efficiency of code
* Testing and error checking
* Programming conventions
* Context / Awareness of the IT software development lifecycle

The course has been running for a year and has had over 100 attendees across a range of areas. Feedback has been consistently positive with the majority of attendees agreeing that these are elements of development that they do not always think about. The provision of this course has also helped develop relationships, often resulting in project work, and raised the profile of the Branch.

* 1. ***Reduction of risk***

By initially utilising the outcome of strategic review work across the Office to indicate where project work would give most benefit, the Branch has quickly developed a reputation for helping to reduce risk. Again, this work has resulted in business areas being more aware of the processes in their areas which they would class as being high risk, and the steps they need to take to help mitigate that. The existence of the Branch has ensured that those looking to reduce risk have a team they can approach with whom they share a common methodological vocabulary, but with an overview of the IT process within the Office too, which enables them to question their processes in a much more structured fashion.

* 1. ***Solution re-use***

As well as helping to build new systems, the Branch is aware that there exists a substantial amount of well-written, well-documented code (predominantly in SAS). The Branch is actively collecting this code and seeking a way to share it more effectively across the Office, enabling codes to utilise a modular approach when developing a system, and utilising code that has already been tested and deemed of a high standard rather than write their own.

This code currently exists as a suite of macros that cover functions such as:

* Error trapping and graceful system shut down
* Sorting and merging
* Checking for existence of files, variables and values
* Copying and deleting files (file management)
* Processes for handling duplicate and missing values.
  1. ***Appropriate solution selection***

Systems that are developed and maintained by business areas are often built using software that the developer is most familiar with. The standard tools usually include:

* Standard packages like Microsoft Excel – highly flexible and often what most people are familiar with. They can be prone to human error, and may make complex statistical processes and structures difficult to view, understand and QA.
* Specific coding languages like Visual Basic – these require substantial training to learn and understand, often resulting in systems that are difficult to maintain unless time is spent on unilateral training.
* Statistical programming languages, like SAS – again, these require training, but if a common language is chosen across the Office, the burden of this training is reduced, and knowledge sharing can flourish
* Common software – for processes with similar elements, it may be possible to integrate them into a custom-built processing environment which can be shared across multiple statistical outputs. This is often an IT built solution, and, whilst the functionality is very much hard-coded, the processes are usually very robust.

The objective of the Branch is to ensure that, whenever a new system is being developed or an existing one is redeveloped, the project owner makes a conscious and informed decision about what type of software they will use. This decision should incorporate the functionality they require and the current corporate strategic software choices as well as the availability of resource and the skills of their immediate team.

1. **Recent successes**

Some of the recent successes in the Branch have included:

* Project work:
  + Processing of the Zero-Hours contract data. This output is still in the experimental stages, undergoing substantial Methodological redevelopment and discussion. The output was produced within a very tight timescale and generated a substantial amount of media interest.
  + Continued process requirements when faced with decommissioned software. A decision was taken not to renew a tabulation production package that has been historically used for producing commissioned tables for census data. In light of that, the Branch worked closely with IT colleagues to shape a solution that would allow data to be held securely, accessed efficiently and produce comparable outputs within the same timescale as the previously used solution.
  + Replacement of several systems due to software requirements. These have included spreadsheet based systems (prone to human error or inefficient in terms of design and processing requirements) and those in software that is not in-line with the current software strategy for the Office. This has included areas such as National Accounts, Well-Being and in providing management information on the learning and development activities of the Office.
* Strategic work:
  + Pilot of an alternative computing environment. The Office is currently investigating alternatives in terms of processing environment. The Branch provided, and sourced additional, resource and scenarios to carry out the testing.
* Training:
  + Good Practice in Programming. This course, developed by the Branch, has had over 100 attendees in under a year. The course is not software specific, but encourages those involved in programming to think more carefully about how they go about developing their code. The course content includes topics like planning and documentation, version control, error trapping and quality assurance, efficiency and flexibility, and how to build software in line with our IT colleagues – usually in a modular fashion.

1. **Conclusions**

The Branch is working consistently towards helping the Office achieve its Strategic Aims, especially with respect to process improvement and the reduction of risk. This new and innovative way of working is already providing benefits, with high risk systems undergoing redevelopment. The Branch is also working to educate those involved in Methodological development and programming on how to ensure that their code is more efficient, flexible, supportable and robust.

As the Branch is growing, both in terms of resource and reputation, the opportunity to provide this kind of service across the Office is increasing. There is substantial support for the work and no shortage of areas making requests for projects. It is the hope of the Branch that the focus on process improvement continues, and that we can continue to contribute in ways that provide rapid benefit and help foster enthusiasm for this kind of work.

1. **References**

[1] ONS Strategy. Available at [www.ons.gov.uk/ons/dcp14298\_323384.xml](http://www.ons.gov.uk/ons/dcp14298_323384.xml) (Last accessed 15/05/2014)