Frameworks, principles and accreditation in modern data management

Felix Ritchie and Elizabeth Green, UWE Bristol

Abstract

The Five Safes framework is increasingly widely used for data governance. Since its conception in 2003, it has influenced data management in many ways, particularly in the public sector. As it has become established, both the advantages and limitations have come to the fore, along with an understanding of modern data management principles.

This paper explores the history, strengths and limitations in the Five Safes, as well as recent suggestions for deepening or extending the framework. It places the Five Safes in the context of contemporary developments in principles-based design, user-centred planning, and the evidence-based decision-making. It discusses the different variations on the framework over time, and questions whether there is a need for a fundamental rethink.

The paper argues that the framework works best when aligned simultaneously to an accreditation process, a principles-based design ethos, an evidence-base, and a user-centred decision-making process. It examines two countries, the UK and Australia, which are moving in this direction but through very different routes.

Key words

Data governance, data management, Five Safes, confidentiality, privacy, principles-based

1. Introduction

At the beginning of 2003, the board of the UK Office for National Statistics (ONS) was presented with a proposal to set up a virtual research data centre (RDC) to allow academic researchers the opportunity to analyse sensitive microdata under controlled conditions. Accompanying this proposal was a framework that divided confidentiality management into four dimensions: safe projects, safe people, safe settings and safe output. Later on, with the addition of 'safe data', this framework became known as the Five Safes. The Five Safes is widely adopted by organisations across the world, particularly National Statistical Institutes (NSIs) in countries as diverse as Australia, Mexico, Norway and New Zealand.

Over time, both the advantages and limitations of the Five Safes framework have become clearer. The advantage of the framework is its generalisability: it can be applied to any data management problem, such as; designing internal administrative systems, setting up secure onsite facilities or releasing information on the internet. However, this advantage is also its main disadvantage: it is a framework for thinking, but does not explicitly state a solution. Furthermore the Five Safes has also been challenged as being too restrictive due to only having 5 components. Some researchers (eg Ritchie and Tava, 2020) have suggested that some dimensions should be prioritised over the rest; other researchers have suggested as many as ten dimensions are now necessary (Oppermann, 2018). There is even confusion over scope: is it a framework for data governance, or a checklist for practical data management? This ambiguity has resulted in a variation in both practice and implementation of the frameworks.

The last twenty years has been witness to a number of other developments in confidential data management. Some are technical developments, such as differential privacy or synthetic data; some are conceptual developments such as the attitudinal model encapsulated in the 'EDRU' (evidence-based, default-open, risk-managed, user-centred) approach first formalised in DSS (2016). However, perhaps the most important change has been the growth in the principles-based approach to data management. A principles based approach suits the ever changing context of the digital world, and provides a basis for policy and legislation, for example in the UK Digital Economy Act 2017 (UKDEA) and the European General Data Protection Regulation (GDPR). This allows research to continue and advance alongside digital advancements; there is less need to revise legislation in the light of changing circumstances as implementations can evolve without breaching core principles.

The alliance between the Five Safes and the principles-based approach has been taken furthest in Australia, where the federal government is implementing the 'Five Data Sharing Principles' which combines both concepts. This ambitious proposal for the whole of the commonwealth government has been accompanied with an extensive consultation with data users, data producers and the public, and has generated significant debate in Australia and abroad.

At the same time, some earlier approaches to data management have reappeared; in particular, the data-centric approach and a quest for spurious objectivity. Now therefore seems an apposite time to review the history and development of the Five Safes and related topics.

The next section introduces the Five Safes and related concepts. Section 3 considers the limitations of the concepts and whether new developments enhance or disrupt the framework. Section 4 considers the principles-based approach in detail and the role of accreditation. Section 5 details examples of the synthetic approach in Australia, the UK and Canada. Section 6 concludes.

For clarity, we shall use 'data access' to cover the whole range of ways that data is being used, and 'the access system' to refer to particular implementations. Examples of access systems could be:

- Allowing one's own analysts or third party researchers to produce statistical results from confidential data
- Generating and releasing an anonymised data file for unrestricted use

- Creating files to be shared under licence
- Government departments sharing administrative data for policy analysis
- A private company sharing customer data with suppliers for operations or marketing

Whilst some authors (eg Arbuckle, 2020) have applied the Fives Safes to private sector contracting and data management, we focus here on the public sector. We will also focus on research use such as producing statistical analyses, as operational use (eg designing HR systems) presents some different challenges.

Given differences in terminology across different countries, in this paper we use these terms:

- Microdata: individual records
- Aggregate statistics: statistical results which cannot be traced to a source respondent
- Identified data: microdata with direct identifiers such as name, full address, social security numbers, health service numbers, and so on
- De-identified data: microdata with direct identifiers removed but which would reasonably allow sources to be identified if data use were unregulated
- Anonymised data: microdata with no reasonable risk of source re-identification even in unrestricted use
- Data subject: the individual about whom data was collected, whether from surveys or operational/administrative sources

2. The Five Safes

2.1 Concept and history

Data access is a complex issue, involving question of law, ethics, IT, statistics, contracts, HR, accreditation, and so on. The Five Safes splits data access questions into five elements or dimensions of control:

Element	Typical question	Example of problems being addressed
Safe projects	Is this appropriate use and management of the data?	What is the purpose of the access request? Is this an ethical and lawful use of the data? What is the benefit to society or to the organisations sharing data? Is there a data management plan in place? What happens to the data at the end of the project?
Safe people	How much can I trust the data users to use it appropriately?	Does the users have the necessary technical skills? Do the users need training in handling confidential data? Are users likely to follow procedures?
Safe settings	How much protection does the physical environment afford to the data?	How is data stored? Are there physical restrictions on the users? Does the IT prevent unauthorised use? Are mistakes by authorised users likely to be detected?

Safe outputs	How much risk is there in the outputs of the access breaching confidentiality?	If the aim of access is to produce statistics, is there any residual risk by for example, highlighting outliers? If the aim of the access is to produce data for onward transmission, how do we make sure that the data is appropriate?
Safe data	Is the level of detail in the data appropriate?	Is there sufficient detail to allow the project to go ahead? Is this excessive detail which is not necessary for the project?

These are separate but joint elements of any solution; see Ritchie (2017). In each element, there is a wealth of knowledge about how to achieve the goals (eg how to run ethics committees, or how to anonymise data, how to design IT systems). Therefore each element can be considered separately, taking the others as solved problem. For example, a system designer can make decisions on the project element on the assumption that user training, IT, data detail and output checks can be implemented to any relevant standard. The security and efficiency of the overall solution depends on how the components work together but this joint-but-several approach greatly simplifies decision-making.

This joint-but-several approach makes clear that the overall objective of 'safe use' might have multiple solutions – and also that the effectiveness of control efforts needed to be assessed against the risk being managed. For example, Ritchie (2019) argues that restricting the type of output allowable in a remote job system, because of a fear of researcher misbehaviour, is aiming at the wrong target. The output is not the problem; the researcher training is.

In general, the Five Safes is focused on risk-reduction: what risk factors have been identified, and will such an intervention/control meaningfully reduce that risk? However, the framework can also help to identify the need for administrative processes, and the coherence (or not) of those. This indirectly can reduce risk, as poorly designed procedures are themselves a risk factor.

There is a question of terminology. Describing these as five 'elements' tends to work well for designers and implementers, as it implies a self-contained component which can be tackled as part of a project plan. However, in discussions with data holders and the general public, who tend to be more concerned that risk is being managed, 'dimensions of control' seems to be a more meaningful description; this carries the sense of "what can't you control, and what can you?".

Other writers have described the elements as themes, security controls, components, or standards. The Mexican Statistical office (Volkow, 2019) uses five "secure elements", as does OECD (2012). Some organisations (eg the NORC Data Enclave in the US, the German federal statistical office) describe their operations in terms of a 'portfolio approach'.

These five dimensions of control are scales, not targets; different solutions will have different levels of control in different dimensions. Treating controls as scales demonstrate one of the most useful aspects of the Framework – the ability to dial up or down controls based on context and how the safes are considered jointly. Each control should be effective and contribute to reducing risk, therefore limiting unnecessary risk control efforts. In addition, there may be non-existent controls in some dimensions, so long as the collective solution is appropriate. For example, contrast the controls used when releasing open data on the internet, when downloading end-user licence (EUL) files from the UK Data Archive, or when using a secure research data centre (RDC) such as those run by most OECD statistical agencies:

Situation	Controls				
	Project	People	Settings	Outputs	Data
Open data	None	None	None	None	Very high (full anonymization)
EUL download	Some (online application)	Some (licence)	Some (online guidance)	None	High (eg little geographical detail)
RDC use	High (application with human review)	High (compulsory training)	High (isolated environment)	High (all outputs manually reviewed)	Minimal (de-identification only)

Originally, there were just four 'safes' (project, people, setting and output) as this was designed specifically for the ONS RDC, and then applied in similar facilities in other countries. This four-safes model was adopted in 2005 by the NORC/NIST Data Enclave, which is why US models often only have these four. The fifth safe, data, was added by 2008 to allow the framework to describe the whole range of ONS statistical outputs, including open data.

The framework was initially known as the 'ONS security model' or 'VML security model' after ONS' secure research facility. Around 2012, the name was replaced with the more generic 'Five Safes' following a suggestion from Statistics New Zealand. However, the use of 'safe' has proved problematic. Particularly in the early 2010s, organisations began interpreting the framework to mean that a 'safe' solution must be 'safe' in every dimension, in the ordinary sense of the word 'safe'. Desai et al (2015) try to correct these errors, and in general the framework is now appropriately referenced by data professionals. However, the language is still problematic for non-specialist audiences, which is one reason why the Australian government adopted the phrase "Five data sharing principles" instead of "Five Safes" in its strategy.

There is no inherent precedence in the Five Safes. Ritchie (2020) argued that for system design, safe projects takes precedence over the other four. In modern data system design using the Five Safes, safe data is usually treated as the residual: when you know who needs the data, for what purpose, and how they will access it, you can adjust the level of the detail in the data to the access arrangements and user need (Ritchie, 2017). However, this is not a rule. There may be cases where, for example, the detail in the data is fixed in advance, and the other dimensions must adjust to it.

Finally, it needs to be recognised the use of the framework itself is no guarantee of good practice. A governance plan of "There's no need for ethics as we know what we're doing; we'll only give the data to people we trust, who say they'll look after it safely; we don't need to reduce the data detail, as we couldn't find anyone we know in it, so there's certainly no risk in outputs…" is certainly using the Five Safes, but not necessarily to good effect.

2.2 Use

The Five Safes was created to describe data management systems, particularly RDCs. It is currently used to describe the governance arrangements for all the general-purpose UK government and academic RDCs (ONS, HM Revenue and Customs, Northern Ireland Statistics and Research Agency, Scottish Centre for Administrative Data Research, and the UK Data Archive). Eurostat, the German central bank, the Dutch research infrastructure ODISSEI, and the NSIs of Canada, Australia, New Zealand, Mexico, Norway describe their RDCs in this way.

Over time, its application has moved away from RDCs to encompass all aspects of data governance. For example, it has been used to cover primary data collection in Nepal (nirc.org.np), anonymization of health data for sharing (HIPAA, 2017), design of scientific-use files (Hafner et al, 2019) and interoperability of standards between organisations (OECD, 2012; Ritchie, 2013). As NSIs often have a strong influence over the data strategies of other parts of the public sector, the adoption of the Five Safes by NSIs has had significant spillover effects, particularly in Anglo-Saxon countries; although as found by OSR (2019), organisational commitments to the concept can vary.

The Five Safes framework has three main uses: description, including pedagogy; design and evaluation; and regulation.

In its early days, the Five Safes was applied to existing systems as a way of describing them (eg Ritchie, 2008; Corti, 2015; Bujnowska, 2018), even by organisations that do not formally use it in their internal models. The growing familiarity with the model provides a handy short-cut for conference presenters, and provide a common frame of reference. It has also been used in confidential data management training since 2004 (Eurostat, 2016; Green et al, 2017; ONS, 2020); the ready-made structure appeals to the trainees, and the framework provide a context for the training itself as part of the 'safe people' element.

In the last ten years awareness of the Five Safes has preceded planning, and so it has become more common to use it for designing data strategies (eg OECD, 2012; DSS, 2016: ICON, 2016; OSR, 2018; Cranswick et al, 2019; DfE, 2020). Private sector organisations advising on data strategies have also begun to use the five safes (eg Security Brief, 2019; Arbuckle, 2020). The predefined structure can simplify evaluation, particularly in cases where the system being evaluated was designed using the Five Safes (eg the risk assessment of the UK Data Archive RDC; ONS, 2011) but not exclusively (eg ICON, 2016).

The Five Safes is used in formal legislation, such as the UK Digital Economy Act 2017 or the state legislation in New South Wales, Victoria and South Australia (DPMC, 2019). With the growth in principles-based regulation (see below), the Five Safes has become a useful standard for organisations to demonstrate compliance. In the UK, for example, the ONS Regulation bases its guidance on the Five Safes (OSR, 2018), while key academic funders (Administrative data Research UK, Health Data Research UK, The Innovation Hub), all require bidders to 'address' the Five Safes in their data management plans.

2.3 Related developments

Two important developments in data access have been the EDRU model and the principles-based approach.

Evidence-based, default open, risk-managed, user-centred (EDRU) data planning

As Ritchie (2014) notes there are strong incentives in the public sector, to take a highly risk-averse approach to decisions around confidential data. This can encourage defensive planning across the five safes, most obviously when considering 'safe data'.

Statistical disclosure control (SDC) can be applied to microdata to reduce the identifiability of the data itself, including the production of anonymised datasets ('input SDC'). SDC is also applied to statistical aggregate to reduce the risk that a statistical finding may inadvertently reveal confidential information ('output SDC').

Input SDC is a well established field of research stretching back to the 1970s. It has a coherent methodological framework, a large range of statistical techniques, open source software to support data holders, and general agreement on the advantages and disadvantages of various solutions.

However, Hafner et al (2015) and Hafner et al (2019) argue that the application of input SDC in real situations over-protects the data and undermines public benefit due to

- A data-centred perspective which assesses inherent data risk in isolation from its use
- The use of hypotheticals and mathematically tractable 'worst case' models rather than evidence in the assessment of risk

- A ‘default-closed’ position, placing the onus on the data holder to consider whether all possible risks have been addressed
- Use of arbitrary statistical models to generate spurious objectivity

Similar concerns can be raised about historical attitudes to the other five safes. Often stemming from an assumption that data users are inherently untrustworthy and require active policing. these reflected the incentives for defensive decision-making in the public sector (Bhatta, 2003; Lofstedt, 2004; Yang and Holzer, 2011; Ritchie 2014).

The EDRU model (Hafner et al, 2015; DSS, 2016) proposes an alternative conceptual framework, reversing many of the implicit assumptions used in decision-making. The elements are:

- Evidence-based: use of empirical (rather than modelled) evidence and plausible scenarios for risk assessment, and acknowledging uncertainty
- Default-open: intending to share data unless negative factors cannot be overcome
- Risk managed: considering risk as a spectrum; including benefits foregone by not sharing; acknowledging trade-offs made
- User-centred: identifying user needs as the primary objective and working backwards

This approach emphasises decision-making focused on achieving the goal of data access: “how do we make X happen?” This contrasts with the historical, defensive approach which is perhaps characterised by “Should we make X happen?”

The EDRU approach is a relatively recent attempt to conceptualise and integrate coherently a number of underlying themes in this area, such as ‘safe statistics’ (Ritchie, 2008), ‘active researcher management’ (Desai and Ritchie, 2009), ‘circles of trust’ (OECD, 2013) or user-centred training (Brandt et al, 2010). As such, it is difficult to say whether organisations have adopted this approach or are merely selecting useful elements from it. DSS (2016) is an exception where the organisational data strategy was explicitly designed using EDRU motifs.

In private sector organisations, it is not clear whether the ‘defensive’ approach dominates, or whether organisations are active and focused on the ‘how’. This is an area for further research.

Principles-based models of data planning

The principles-based approach to data management derived from discontent at rules-based regulation which was the dominant approach until recently.

The rules-based model of regulation aims to specify in a binary manner what is allowed or not allowed. Under this model, the primary source of direction is the regulation itself. The principles-based approach instead focuses on what any system is trying to achieve, and then questions whether the system actually achieves those objectives. In a principles-based system, implementation decisions are primarily under the control of the implementor; regulation is there to specify the goals, and to identify what evidence should be presented that the goals have been achieved.

Consider the example of tax regulation. A rules-based system would try to identify and specify all the allowable tax breaks that can be charged against income. This is clear, but can encourage firms to spend unproductive time searching for loopholes, and requires the regulator to spend unproductive time seeking to close holes; and it assumes that the tax regulator can adequately specify the exceptions. A principles-based tax system could have a general rule along the lines of “only expenses relating to the trading activity are chargeable”, and then evaluate using expert opinion when tax returns are submitted. Clearly the second case is much more complicated; there is negligible value to finding loopholes in the law, but more incentive to misrepresent activities. The success of it depends very much on how the regulator communicates with firms, and the reputation of the regulator.

In the context of the Five Safes, the two may be contrasted using the following examples, which all exist somewhere:

Safe…	Rules-based	Principles-based
Projects	Identify list of valid uses	Specify benefits that must be demonstrated, and risks to be considered
People	Require specific accreditation eg meet Civil Service appointments criteria	Require 'appropriate' training
Setting	Follow government IT standard	Follow ISO27001 practices (ie choose system and be able to demonstrate integrity of it)
Output	Apply threshold rule tabular statistics	Apply threshold rules but allow appeals if important and demonstrably non-disclosive
Data	Clear boundary between anonymised and other data	Data must be 'appropriate' to the environment

The Five Safes therefore can accommodate either approach, or a mixture of the two. Older data management solutions tend to be rules-based and may or may not use the Five Safes. The moves towards principles-based models are more likely to be done in the context of the Five Safes.

3. Developing the Five Safes

3.1 Should there be more safes, or more sub-safes?

As noted above, the Five Safes is a generic framework. Some authors have tried to give more meaning to the dimensions. For example, Ritchie (2013), when considering how international data-sharing standards might be developed, suggested breaking down 'safe people' into 'knowledge' and 'incentives', and 'safe setting' into 'access' and 'networks'. Statistics Canada breaks the 'safe people' component into organisational and personal criteria, so that this component is itself multi-dimensional.

The element that has grown most in scope since 2003 is 'safe projects'. Initially there to cover the process of getting approval to ONS' secure facility, it now is seen to cover all the project planning aspects: ethical approval, data management plans, benefit/cost assessment and user identification. As Ritchie (2020) notes, when designing data management from scratch this element takes centre stage. This has also been the focus of much of the Australian consultation, particularly the onward use of data from data access requests.

'Safe outputs' has also changed. Originally concerned with the production of statistical outputs by researchers, different applications of the Five Safes has means that this has described user queries to an HR system, datasets for onward management, or products to support compliance strategies.

The Five Safes website lists a number of sub-fields within each of the dimensions as a way of breaking complex data management/evaluation problems down. It also provides a list of questions or challenges for designers.

However, there have been some proposals to increase the number of dimensions. For example, ACS (2019) suggests:

- Safe organisations – ensuring the organisation has in place and follows agreed procedures
- Safe lifecycle – time-relevant issues such as archiving and the sensitivity over time
- Safe outcomes – the ultimate uses of the project output
- Safe use – the impact of using the project output, apparently in a moral/ethical context
- Safe response – dealing with accidents

While these are all important subjects, it is not at all clear that the extra 'safes' help data planning. The safe projects/outputs/outcomes/use overlap underlies the difficulties – as more dimensions are added, the distinctions between the dimensions become weaker, and it is not clear what question is being addressed and how each dimension can be considered separately.

Ironically, it can also create unhelpful distinctions. For example, most organisations consider the safe organisations/safe people argument as part of the same question – what skills, training or assurances are needed to ensure that the users act appropriately? For bodies which explicitly consider organisations as part of the solution (such as Statistics Canada or Eurostat), separating organisation from person makes no sense. Excessive subdivision may also lead to micromanagement and turf wars, outcomes the Five Safes was originally designed to avoid.

This is not to say that raising these issues is unhelpful in terms of thinking about data access. For example, 'safe response' emphasises that a well-run organisations would have processes in place to identify, manage and communicate breaches of data security. It is not clear why it is helpful to think of this separately from the other dimensions. People and IT planning normally includes a breach policy, and part of a cost-benefit or privacy impact assessments should include the risk of failure and mitigation. However, it could be argued that this is necessary to make organisations explicitly confront the issue. Addressing these issues is one reason why the Australian Government included advice on what to do before and after applying the Five Data Sharing Principles in its Guidance.

Perhaps the reason why these struggle to make the case for new 'safes' is that the additional measure are about implementation rather than concepts or themes. Claiming that a designed system is 'safe' because safe projects, people and setting have been considered assumes that project planning, people training and robust system design are done competently. There is of course nothing in the Five Safes or the extended scheme to guarantee this. In contrast, the extensions to 'safe people' being trialled in Canada, for example, does not add new concepts, but refines them.

Nevertheless, the ACS's detailed critique of the Five Safes should be welcomed for challenging assumptions, precisely because they can be accommodated in the current framework. One particular problem they highlight is of artificial intelligence (AI) systems. Should a new 'safe' be created for them? Perhaps in the future there will be a new dimension for autonomous systems, but the current debate, whether AI fits into the 'people', 'settings' or 'output' category, is stimulating much debate precisely because trying to fit AI into an existing category forces us to confront our understanding.

3.2 Should the five safes be more objective?

The Five Safes is an explicitly subjective framework. It does not attempt to quantify 'safe' in any dimension, let alone try to balance 'safety' in one dimension against 'safety' in another dimension. There are five reasons for this.

First, there is no meaningful metric in any dimension, let alone a common metric. Consider defining a metric for 'safe projects'. There could be a case for a cost-benefit analysis of the value to the public: lives saved as a result of a data linking project versus implementation cost and expected cost arising from an unknown breach. Most of these costs and benefits are unknowable and not measurable. This is why cost-efficiency analysis is more often used to assess the worth of a project, but, by design, they do not question whether the project itself brings an acceptable risk or not.

It could be argued that the common measure across all dimensions is "what is the risk of re-identification?" This is often the proposed measure (assuming it could be measured) but is not the only one. Re-identification risk can almost always be substantially reduced by expenditure in one or more dimensions, and so this risk alone is not a measure.

The second issue is that discussion of 'trade-offs', 'risk-utility maps' or similar concepts implicitly assumes the independence of protection measures. Assume, for expository purposes, that people and setting are the only relevant dimensions. The trade-off argument assumes these are independent and linked by some function *f* to create safe use:

$$\text{safe use} = f(\text{people, setting})$$

However, what if in reality the risk of re-identification for each depends on the value of the other:

$$\text{safe use} = f(\text{people}_{|\text{setting}}, \text{setting}_{|\text{people}})$$

For example, the likelihood of an individual making a mistake might be dependent on the IT system; the effectiveness of IT controls might depend on training for the user. This is no longer a linear trade-off, but a complex non-linear model – again, with unknown values in every place.

Third, trade off models do not allow for discontinuities. A secure RDC presents qualitatively different risk to an end user licence model, and both will be differently affected by an ethical approval process. This topic is characterised by discontinuities across all the different potential activities.

Fourth, any data that does come out of modelling is subjective. Consider the most objective measure in this whole field: assessing the disclosure risk in a given dataset. There is a large literature on this, going back to the 1970s. Techniques have been developed for many types of data, statistics and attack models. There are software tools, such as µArgus/τArgus or sdcTable/sdcMicro, to calculate risk probabilities to many decimal points, and which are invaluable for exploring the *relative* risk in data and tabular outputs. Overall, there is a large and stable body of literature on the relative pros and cons of all the different statistical disclosure control (SDC) procedures.

None of these have any external validity. The disclosure modeller faces choices at every stage of the process for which he or she has to make a subjective choice:

- the attack scenario
- the attacker's motivation
- how much information the attacker already has
- the time and resources available to the attacker
- alternative data sources available to the attacker
- alternative published statistics available to the attacker
- what similar results will be published in future
- whether repeated attacks are possible

And so on. As noted in section 2 above, SDC analysts often discuss 'worst case scenarios'. The SDC literature has developed this concept to a common standard so that the results from analytical research articles can be usefully compared and lessons learned about the pros and cons of different measures. However, this does not mean that models used for methodological development should be applied in practical cases, as the 'worst case' may no longer be relevant. Moreover, these models are typically the *mathematically tractable* worst-case scenarios. They cannot represent events which cannot be modelled (spontaneous recognition, for example; Ritchie, 2017) or unknowable events (the likelihood of an authorised user selling market-sensitive information). In short, claims to objectivity are spurious.

Consider differential privacy (DP), which has become popular among private sector consultancies selling data protection services. DP is claimed to mathematically guarantee a level of uncertainty around a statistic, which it does; but only in a very constrained set of circumstances. It is easy to demonstrate that it is little better than any other noise addition tool in protecting confidentiality; it can also produce nonsensical results except on a very well-defined use case (basically, frequency tables with no rare events). It is noticeable the statistical agencies, with the exception of the US Census Bureau (Dajani et al, 2017), have largely rejected DP in favour of more user-centred methods such as cell-key perturbation.

Finally, there is the quantum problem: the design of the system affects the interaction of the participants in the system. Most obviously, user training and IT systems are designed to affect behaviour; but in practice most systems evolve over time in unexpected ways.

Until ten to fifteen years ago, data access was still dominated by the statistical approach, where quantitative risk assessments were given great prominence. This was most noticeable in the statistical agencies; as these had positions of authority in national systems, this was the prevailing viewpoint in the public sector. However, in the last decade this has increasingly been abandoned in favour of the multi-dimensional, explicitly subjective, EDRU approach.

The choice of whether to quantify/not quantify risk also affects the public debate. Numerous studies have shown that public perceptions of whether data sharing is a public benefit are highly sensitive to the questions asked (eg Hallinan et al, 2012; Wellcome Trust, 2013; Understanding Patient Data, 2018). Studies (eg Jenkins et al, 2017) also show that numbers are more likely to be perceived as reliable than verbal description, irrespective of their actual credibility. For example, during the Covid-19 pandemic, there was substantial public debate about the value of 'R0', the base infection rate. Public engagement centred on whether this value was greater or less than one. There was almost no debate outside scientific circles of confidence intervals, sample bias, distributions or other statistical factors which make the simple 'what is R0?' question meaningless.

In this light, the lack of quantitative measures for the Five Safes is not a gap in the literature, but an essential strength of the model. It focuses decision-makers' minds on the need to collect, evaluate and use subjective evidence, and to build consensus for decisions. Quantitative models, such as those used by SDC professionals, can provide useful supporting evidence, particularly in terms of relative risk, but only supporting evidence.

4. The Five Safes, principles-based planning and accreditation

In recent years, the Five Safes model has become increasingly associated with the principles-based approach to regulation. There is an affinity between the two concepts. The Five Safes provides a framework for planning; the principles-based model provides a way of suggesting how goals should be specified. Neither is specific on the actual implementation; but both provide a way that the effectiveness of any implementation can be measured.

As noted above, there is no inherent reason for organisations to apply both, but in practice principles-based regulation is increasing framed in a Five Safes framework. This is explicit in legislation in the UK, and in the forthcoming legislation in Australia. The European GDPR is not explicitly principles-based and does not cite the Five Safes. Nevertheless, its balancing of data detail against 'procedural and technical measures' and the avoidance of specific technical or statistical standards in favour of solutions 'having regard to' outcomes places it in the same camp.

The popularity of principles-based regulation is that it seems to address some of the flaws of older legislation which struggled to provide adequate guidance. Rules-based regulation works well when the terms can be unambiguously defined. For example, the UK Data Protection Act 2018 makes (attempted) re-identification of de-identified or anonymous data a criminal act except in very specific, unambiguous, easily understood cases. However in general specification of rules in this field is difficult.

Consider, once again, the issue of anonymization. Much of the older legislation distinguishes between 'anonymous' and identifiable (identified or de-identified) data. While this is helpful in setting the limits of legislation (which typically allows data to be unregulated once anonymised), it provides little guidance as to how identifiable data should be managed. There are definitional problems: 'anonymised' is never defined unambiguously, and some laws distinguish between 'personal' and 'identifiable' (so that data can be anonymous yet still personal information). In the case of data legislation, it is almost certain that technological developments will lead to changes in the standard of what is 'anonymous' (NSQR, 2019). Most importantly, defining exactly what the legislation means in any specific case requires guidelines and, sometimes, legal judgments.

The principles-based approach aims to sidestep this by being explicit about the need for ambiguity at the legislative level, as long as the appropriate checks and balances are available at the implementation stage. For example, a principle could be “data which has been fully anonymised does not count as ‘personal information’ in the context of this legislation”. This is clear in its purpose, and the legislation does not need to define ‘fully anonymised’; additional regulation fills the legislative gaps by design.

A key element of making the principles-based approach work is accreditation. Rather than specifying actions, actors, systems and procedures undergo ex ante validation to ensure that they are fit for the purpose. An analogy is with driving: while there are rules of the road, the primary mechanisms for compliance with the rules is to ensure that drivers are fit; drivers infringing good practice (either by breaking explicit rules, or by infringing subjective measures such as ‘dangerous driving’) have their accreditation revoked. In many financial markets, good conduct is hard to specify, and so regulators rely on the accreditation of practitioners (Keenan, 2020).

The advantages of the principles-plus-accreditation approach are

- efficiency, as solutions can be adjusted to circumstance rather than a legislative context
- flexibility, as multiple accreditation pathways can be set up
- adaptability to circumstances, rather than requiring legislative/formal change
- adaptability to collective learning as processes develop
- cultural change through positive reinforcement and engagement with goals
- engagement with stakeholders

The last is a consequence of the problems of the principles-plus-accreditation approach, compared to a rules-based regulation model:

- It may be less clear, at least at first sight
- There may be concerns about unfair or uneven treatment, and so transparency in processes is essential
- Practitioners might want clear guidelines in advance to implement solutions, rather than post-hoc approval

An effective principles-based system therefore requires a much greater level of engagement.

We consider how this works in practice by studying two national systems, one (UK) which evolved over two decades, and one (Australia) currently being created from, relatively speaking, a standing start.

Microdata access at the UK Office for National Statistics

In 2002, when the Five Safes was conceived, access to the business data at ONS was governed by the Statistics of Trade Act 1947. This Act had no conception of data being collected for any other purpose than producing aggregate tabulations for measuring the economy. As a result, finding an efficient access path which met both the spirit and letter of the Act, required a complex legal structure which blocked a number of important potential users, such as PhD students (Ritchie, 2014).

In 2007 the Statistics and Registration Services Act created a simple access gateway, the ‘approved researcher’, and the basis for a flexible data acquisition and access path. These provisions addressed the limitations of the 60-year-old legislation. It did not build a new conceptual model, but this new legislation and the growing impact of the Five Safes on ONS data planning led to a culture change about how ONS should approach data access issues.

By the time of the Digital Economy Act 2017 (DEA), the pieces had fallen into place, and the legislation served to embody this new model into law. In respect of research data access, the DEA used the Five Safes for its structure, set out the principles on which access would be granted, and created an accreditation process for each of projects, people, settings and outputs (data is effectively the residual, as Ritchie, 2017, recommends). The UK Statistics Authority oversees all the accreditation processes.

The advantages of this approach are best seen in the 'people' and setting' accreditation. There are five general-purpose RDCs in the UK holding detailed ONS and other government data. These all operate slightly differently (some for example, are only accessible on the premises, some allow home remote access in special circumstance). However, the same principles of safe setting apply to all. In contrast, there is one accredited researcher training programme ('Safe Researcher Training, or SRT; Green et al , 2017) which all the RDCs use, as well as a number of other services. There is nothing to stop another provider setting up an accredited research training programme. However as the SRT was designed to cover good practice in a range of settings, there has been no substantial demand for local variations.

In the UK therefore, legislation has largely followed practice. That practice has evolved over some twenty years, with many opportunities for learning and correction (and input from other countries), and the move to a principles-plus-accreditation system has been uncontroversial.

Data access in the Australian Federal Government

Until recently, the Australian federal government took a data-centred default-closed perspective on data sharing and access. In 2014, after a critical review by the public auditors, the Australian Bureau of Statistics (ABS) took the decision to move towards a default-open environment for research access, and adopted the Five Safes as the governance framework. Building on the ABS experience, other departments followed suit, including the Australian Institute of Health and Welfare and the Australian Government Department of Social Services.

In 2016, the Productivity Commission was asked to carry out a root-and-branch review of the options for improving the sharing of data across all government. The subsequent report (Productivity Commission, 2017) exhaustively researched best practice and new developments around the world, and argued for a systemic shift in emphasis towards an EDRU-style principles-based model using the Five Safes as the structure. In 2018 the Office of the National Data Commissioner was set up to implement the recommendations, and began a two-year programme of public consultation and legislative drafting for a user-centred principles-plus-accreditation model to be applied across the federal government. As the drafting period has progressed, the consultation has moved from exploring problems and concerns to developing guidelines to help practitioners implement the principles.

The Australian experience therefore represents a very different development trajectory compared to the UK. Notwithstanding the early changes implemented by individual departments such as ABS, this is a clear break with past traditions. This has necessitated the extensive consultation period, as an entire culture change is being proposed. This is a substantial leap beyond current practices in other countries, and so presents more of a risk than a piecemeal approach. However, none of the individual elements being introduced is unknown in the rest of the world, and the whole-of-government approach provides a unique opportunity to build in consistency and shared understanding of principles from the beginning. At the time of writing (June 2020) this process is only partially complete, but it seems likely that the Australian experience will provide valuable lessons for many other countries.

5. Conclusion

The Five Safes model has become well-established, not through diktat but because it offers a useful structure for addressing data access, management and governance. It is increasingly the design structure for statistical organisations and RDCs, and has prompted the development of useful practices such as output statistical disclosure control and active researcher management.

It remains an empty structure, by design. The usefulness and applicability of the concept is because it helps to order discussions, design, evaluation, regulation, but makes no specific requirement on any of these. Using the Five Safes to design a data governance plan is no guarantee of competent delivery, any more than ignoring the Five Safes is a certain recipe for failure.

The Five Safes is explicitly subjective and qualitative, again by design. While some recent proposals to develop quantitative measures of risk are no doubt well-intentioned, the lesson of the last twenty years is that spurious metrics limit genuine risk assessment, lower credibility, and reduce the opportunity to extract public value from data sharing and use. Subjective decision-making, properly implemented, provides consistency and transparency but with the capability of responding to individual circumstances.

There have also been proposals to extend the range of the Five Safes; Statistics Canada, for example, has spent much time rethinking how 'safe people' can be expressed as the interaction of the personal and the organisational responsibility. Others have suggested having more 'safes' but at the moment it is not clear that the gain in coverage outweighs the increase in complexity.

What has become apparent in the last decade, as the number of users of the Five Safes increases, is that there is a need for more detail. Many organisations have developed their own guidelines or interpretations; perhaps the next stage in the evolution of the Five Safes will be common understanding the ten or twenty or however-many 'Sub-Safes'. The scale of the Australian federal government project is likely to become a benchmark in this regard.

Finally, we note that the Five Safes did not develop in a vacuum, and continues to interact with other concepts. The evidence-based, default-open, risk-managed, user-centred attitudinal (EDRU) approach aligns well with the Five Safes as it has the same perspective on the importance of subjective, evidence-based decision–making. But the biggest change in the next decade is likely to be the growth in principles-based data regulation, where the Five Safes provides a natural structure for the accreditation processes essential for effective principles-based operation.

References

DSS (2016) Data Access Project: Final Report. Australian Government Department of Social Services. June.

Arbuckle L. and El Emam K. (2020) Building an anonymization pipeline. O'Reilly Publishers

Bhatta G. (2003) "Don't just do something, stand there! Revisiting the issue of risks in innovation in the public sector". The Innovation Journal

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

Bujnowska A. (2018) Access to European Microdata for Statistical Purposes. https://ec.europa.eu/eurostat/cros/system/files/04.access_to_microdata.pdf

CESSDA (2017) CESSDA Strengthening and Widening. Confedaration of European Social Science Data Archives, Deliverable D3.5. October.

Corti L., van den Eyden V, Bishop L. and Wollard M. (2020) Managing and Sharing Research Data: A Guide to Good Practice, 2nd edition. Sage.

Cranswick K., Tumpane S. and Stobert S (2019) Virtual data labs - A more flexible approach to access Statistics Canada microdata. UNECE/Eurostat Work Session on Statistical Data Confidentiality, The Hague, October

Dajani A. Lauger A., Singer P., Kifer D., Reiter J., Machanavajjhala A., Garfinkel S., Dahl S., Graham M., Karwa V., Kim H., Leclerc P., Schmutte I., Sexton W., Vilhuber L., and Abowd J. (2017) The modernization of statistical disclosure limitation at the U.S. Census Bureau. UNECE/Eurostats Worksession on Statistical Data Confidentiality 2017, Skopje.

Desai T. and Ritchie F. (2010) "Effective researcher management", in Work session on statistical data confidentiality 2009; Eurostat; forthcoming

Desai T., Ritchie F., and Welpton R. (2016) The Five Safes: designing data access for research. Working papers in Economics no. 1601, University of the West of England, Bristol. January

DfE (2020) How we share pupil and workforce data. Department for Education. https://www.gov.uk/guidance/data-protection-how-we-collect-and-share-research-data

DPMC (2019) Data Sharing and Release Legislative Reforms Discussion Paper. Commonwealth of Australia, Department of the Prime Minister and Cabinet.

Eurostat (2016) Self-study material for the users of Eurostat microdata sets

Green E., Ritchie F., Newman J. and Parker T. (2017) "Lessons learned in training 'safe users' of confidential data". UNECE worksession on Statistical Data Confidentiality 2017. Eurostat.

Hafner H-P., Lenz R., Ritchie F., and Welpton R. (2015) "Evidence-based, context-sensitive, user-centred, risk-managed SDC planning: designing data access solutions for scientific use", in UNECE/Eurostat Worksession on Statistical Data Confidentiality 2015, Helsinki.

Hafner, H., Lenz, R. & Ritchie F. (2019). User-focused threat identification for anonymised microdata. Statistical Journal of the IAOS, 35(4), 703-713. https://doi.org/10.3233/SJI-190506.

Hallinan D., Friedewald M. and McCarthy P. (2012) Citizens' perceptions of data protection and privacy in Europe. Computer Law & Security Review v28:3 pp263-272 https://doi.org/10.1016/j.clsr.2012.03.005

Groves R. and Harris-Kojetin B. (Eds) (2017) Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps. National Academies of Sciences Engineering and Medicine.

ICON (2016) Hellenic Statistical Authority Mission on microdata access: final report. Eurostat

Jenkins S., Harris A., and Lark R. (2017) Maintaining credibility when communicating uncertainty: The role of communication format. In: Gunzelmann G., Howes A., Tenbrink T. and Davelaar E. (eds.) CogSci 2017: Proceedings of the 39th Annual Meeting of the Cognitive Science Society. (pp. pp. 582-587). Cognitive Science Society: London, UK. Lofstedt R.E. (2004) “The swing of the regulatory pendulum i Europe: from precautionary principle to (regulatory) impact analysis”. J. Risk and Uncertainty v28:3 pp237-260

Keenan P. (2020) “Dictum meum pactum”: UK regulation: Rules or Principles. Keenan Regulatory Consulting.

OECD (2014) OECD Expert Group for International Collaboration on microdata Access: Final report. Organisation for Ecoomic Co-operation and Development.July.

ONS (2020) Safe Researcher Training 2017 onwards. Last reviewed June 2020

Opperman I. (ed.) (2018) Privacy in Data Sharing: a guide for business and government. Australian Computer Society. November.

OSR (2018a) Joining up data for better statsitics. Office for Statistics Regulation Systematic Review Programme Report. September

OSR (2018b) Regulatory guidance – Building confidence in the handling and use of data. Office for Statsitcs Regulation. Updated October.

OSR (2019) Joining up data for better statsitics. Office for Statistics Regulation Systematic Review Programme Report. October

Productivity Commission (2017). Data Availability and Use – Inquiry Report. Australian Productivity Commission.

Ritchie F. (2008) "Secure access to confidential microdata: four years of the Virtual Microdata Laboratory" in Economic and Labour Market Review; Office for National Statistics; May, pp 29-34

Ritchie F. (2008) "Secure access to confidential microdata: four years of the Virtual Microdata Laboratory" in Economic and Labour Market Review; Office for National Statistics; May, pp 29-34

Ritchie F. (2013) "International access to restricted data: A principles-based standards approach". Statistical Journal of the IAOS v29:4 pp289-300. DOI 10.3233/SJI-130780

Ritchie F. (2014) Resistance to change in governments: risk inertia and incentives. UWE Departmetn of Economics working paper no.1412. December

Ritchie F. (2017) The 'Five Safes': a framework for planning, designing and evaluating data access solutions. Data For Policy Conference 2017. September.

Ritchie F. (2017) Spontaneous recognition: an unnecessary control on data access? ECB Statistical Papers no.24. European Central Bank. August.

Ritchie F. and Tava F. (2020) Five Safes or One Plus Four Safes? Musing on project purpose. Bristol Centre for Economics and finance blog

Security Brief (2019) IXUP embeds Five Safes framework in platform. https://securitybrief.com.au/story/ixup-embeds-five-safes-framework-in-platform

Understanding Patient Data (2018) Public attitudes to patient data use: A summary of existing research. Understanding Patient Data. September

Volkow N. (2019) Harnessing the potentiality of microdata access risk management model. UNECE/Eurostat Work Session on Statistical Data Confidentiality, The Hague, October

Wellcome Trust (2013) Qualitative Research into Public Attitudes to Personal Data and Linking Personal Data. Wellcome Trust, London. July

Yang K. and Holzer M. (2006) "The Performance–Trust Link: Implications for Performance Measurement" Public Administration Review. v66:1 pp114–126