emerald PUBLISHING

Records Management J

Risk identification and management for the research use of government administrative data

Journal:	*Records Management Journal*
Manuscript ID	RMJ-03-2019-0016.R1
Manuscript Type:	Research Article
Keywords:	Government administrative data, Information governance, Risk Management, Sensitive data, Open data

Risk identification and management for the research use of government administrative data

1. Introduction

Government administrative data have enormous potential for public and individual benefit, through improved educational and health services to citizens, medical research, environmental and climate interventions and exploitation of scarce energy resources. Administrative data is 'collected primarily for administrative (not research) purposes by government departments and other organisations for the purposes of registration, transaction and record keeping, usually during the delivery of a service', such as health care, vehicle licensing, tax and social security systems (https://esrc.ukri.org/funding/guidance-for-applicants/research-ethics/useful-resources/key-terms-glossary/). Administrative data are usually distinguished from data collected for statistical use, such as the census. Unlike administrative records, they do not provide evidence of activities and generally lack metadata and context relating to provenance. Administrative data, unlike open data, are not routinely made open or accessible, but only on request to named researchers for specified research projects, through research access protocols which often take months to negotiate and are subject to significant constraints around re-use, such as the use of safe havens. Researchers seldom make use of freedom of information or access to information protocols to access such data, as they need specific datasets and particular levels of granularity and an ability to re-process data, which are not made generally available. This article draws on research undertaken by the authors as part of the Administrative Data Research Centre in England (ADRC-E). The research examined perspectives on the sharing, linking and re-use (secondary use) of administrative data in England, viewed through three analytical themes: trust, consent and risk. The article presents the analysis of the identification and management of risk in the research use of government administrative data and presents a risk framework.

Risk management (that is, coordinated activities which allow organisations to control risks, Lemieux, 2010) enables us to think about the balance between risk and benefit for the public good and for other stakeholders. Mitigating activities or management mechanisms employed to control the identified risks depend on the resources available to implement the options, on the risk appetite or tolerance of the community and on the cost and likely effectiveness of the mitigation. Mitigation and risk do not work in isolation and should be viewed holistically, keeping the whole information infrastructure in balance, across the administrative data system and between many different stakeholders.

This article seeks to establish a clearer picture of risk with regard to government administrative data in England. It identifies and categorises the risks arising from the research use of government administrative data. It identifies mitigating risk management activities, linked to five key stakeholder communities, and discusses the locus of responsibility for risk management actions. The identification of the risks and of mitigation strategies are derived from the viewpoints of the interviewees and associated documentation and therefore reflect their lived experience. The five stakeholder groups identified from the data are: individual researchers; employers of researchers; wider research community; data creators and providers; and data subjects and the broader public. The main sections of the article, following the methodology and research context, set out the seven identified types of risk events in the research use of administrative data, present a stakeholder mapping of the communities in this research affected by the risks, and discuss the findings related to managing and mitigating the risks identified. The conclusion presents the elements of a new risk framework to inform future actions by the government data community and enable researchers to exploit the power of administrative data for the public good.

2. Methodology and research context

2.1 Methodology

Between 2014 and 2017, the researchers conducted four case studies on government administrative data for education, transport, energy and health. The purpose of the research was to examine stakeholder perspectives about the sharing, linking and re-use (secondary use) of administrative data. Following a scoping study, the qualitative research undertook 44 semi-structured interviews, plus one focus group, supported by documentary analysis and a literature review. The secondary use of government administrative data by academic researchers was the core of each case study. The research data was enriched by interviews with government bodies as data providers, regulatory bodies, research funders and lobby groups. In the education case study, we also interviewed data subjects. In spite of extensive discussions with transport data providers, none were willing to be interviewed for the study. The analysis is limited by the stakeholder views and may not present a complete picture: for example, we did not interview many records and information managers in the research, and we did focus on academic researchers. The choice of the four case studies allowed for some exploration of different disciplinary perspectives (health, education, transport, energy) on the same topic (government administrative data). The two years of the study did not allow for a complete study of all classes of administrative data from all government providers, but by considering four disciplinary perspectives, we were able to explore differences and avoid mono-disciplinary approaches and assumptions. All interviewees were anonymised and extracts are referenced by a code (eg. A10). Table 1 provides a summary of the research data collection. The case studies, the key datasets accessed by academic researchers and the interview protocol are explained in [anon] 2018, pp4-7. As explained there, interview transcripts were thematically coded line-by-line, assigning a coding label to each component and refining the codes into themes derived inductively from the data, in an iterative process of analysis and assigning meaning. Most of the coding was undertaken by two of the authors, reviewed by a third, and the analysis was discussed by the whole group. Three themes, trust,

risk and consent, framed the data analysis. The findings presented here arise from the coding analysis of the research data, concentrating on our analysis of the identification and management of risk in the research use of government administrative data: articles published elsewhere report on trust (anon, 2017) and consent (anon, 2018).

Case study	Number of interviewees	Types of interviewees	Coding range	Dates of data collection
Scoping study	5	Academic researchers	A1-A5	06/2014-07/2014
Education data	12, plus 4 in a Focus Group	Students (data subjects), HEI student data manager, academic researchers, data providers	FG1, A6-A17	12/2014-04/2015
Transport data	9	Academic researchers	A22-A25, A28, A29, A31-A33	09/2015-11/2015
Energy data	5	Academic researchers, data providers	A26, A27, A30, A34, A35	10/2015-01/2016
Health data	18	Academic researchers, data providers, policy advisors, information & data managers, lobby groups	A36-A53	05/2016-10/2016

Table 1: summary of interviewees

2.2 Research context

The identification and management of risk is discussed in the information compliance, governance and records management literature. ISO 31000 (2009b, revision 2017) defines risk as the 'effect of uncertainty on objectives', that is a deviation (positive or negative) from the expected outcome. ISO 31000 notes that risk is 'a combination of the consequences of an event … and the associated likelihood of occurrence'. In an information security context, AIC (risks which compromise the Availability, Integrity and Confidentiality of information) measures risk impact. ISO PD Guide 73 (2009a) describes a process of risk identification

comprising four elements (risk sources, events, causes and consequences, stakeholder needs). Risk analysis includes calculations of the likelihood and potential consequences of a risk and its severity or impact. Risk management and mitigation options can then be identified, responsibility for action delegated, and costs and effects estimated. Routine hazard or operational risks may be accepted and not do great damage. Control or uncertainty risks must be constrained within acceptable limits. Opportunity or strategic risks may involve failure or loss, but may lead to better outcomes.

Doty (2015) adopts Beck's characterisation of risk as a lens through which we seek to make sense of an increasingly unpredictable society. Doty asserts the close relationship between risk assessment and information, concluding that while risk assessment depends on information, it is never possible to have a complete set of information and therefore, it is impossible to eliminate risk altogether. Risk management is not always seen as a positive activity: Sprehe (2005) warns against using management of risk as an argument in favour of records management, describing it as 'an inadequate rationale for enterprise-wide records management because it is essentially a defensive strategy'. We would argue for the positive benefits of risk management as a proactive, not reactive, strategy, which enables an organisation to assess risks, against an agreed risk appetite, and take action. Risk is also related to benefit: a risk might be worth taking because of the benefit it will bring.

As Lemieux (2010) sets out, risks can be categorised in a variety of ways, and different disciplines take different views on the best grouping. ISACA (2010) groups IT risks into strategic risk; environmental risk; market risk; credit risk; operational risk; and compliance risk; whereas ARMA International (2010) divides records and information risks into administrative risks, records control risks, legal/regulatory risks, and technology risks. The discussion of risk is often in the context of digital systems and cyber-security, which are perceived to increase some information risks and therefore warrant further consideration,

such as that given by Quigley et al (2015) in their study of rhetorical analysis in cyber-security communications.

Our research is situated in the records management literature and adopts a risk management frame which broadly reflects information governance approaches. The *Records Management Journal* 'Special Issue: Information governance and ethics: information opportunities and challenges in a shifting world' (vol 29: 1-2, 2019) emphasised the interest in information governance in the records field. Several papers illustrate the relationships between information governance and the identification of risks to information security, which can then be addressed by a variety of rules and controls (eg Daneshmandnia, 2019, Xie, S L. 2019). Lemieux (2010) suggests that 'in contrast to varied definitions of risk management across distinct disciplines, the processes associated with its practice have evolved to become relatively standardized': thus a study of risk derived from administrative data management processes might have applicability across records and information management disciplines as well.

3. Results and Discussion

3.1 Identifying risks in the research use of administrative data

The thematic coding of the interview data identified seven different types of risk events in secondary use of administrative data, which can be further analysed for their sources, causes, impact, likelihood and potential consequences. Most of these are caused by unmanaged disclosure, whether accidental or deliberate, although research participants also distinguished risks caused by linkage and misinterpretation which may not entail disclosure. The seven risk events which were identified in the research data are summarised in Table 2 and in Figure 2.

Risk event #	Risk event description
1	*Identification of anonymised individuals*: leads to data misuse, harm (4, 5). Risk increases with data volume, making data more 'disclosive'. Some research requires identification. Risk appetite varies depending on sensitivity of data (eg health data), granularity of data (eg transport).

2	*Data linkage affordances*: increases other risks eg identification of subjects (1) or locations. Linkage errors, missing data, mismatches in data granularity and increased complexity may magnify risk and affect data reliability. Linked data can benefit individuals and groups, but low risk appetite may lead to opportunity risk (7).
3	*Misinterpretation of data*: lack of metadata, poor data documentation, linkage error, and researcher inexperience may lead to misinterpretation of data. Users misunderstand findings or misuse them for other purposes.
4	*Malicious misuse of data*: deliberate misuse leads to risk of harm (5).
5	*Harm to individuals and groups identified in the data*: results from deliberate misuse (4) or accidental re-identification, leads to potential psychological, physical, emotional, financial, reputational & other harm.
6	*Risk to commercial confidentiality*: privatisation of public functions, complexities of data production & ownership leads to data breaches, commercial risks, conflict between public policy benefit and commercial-in-confidence.
7	*Opportunity risk of not using data for research*: risk appetite varies between data providers, individual data subjects, and researchers, resulting in no agreed risk appetite. Risk aversion (eg legal) results in data not available for research in the public interest.

Table 2: Risk event identification

Risk event 1: identification of anonymised individuals

The most prominent risk is that of identification of individuals represented in the data, including re-identification from de-identified data. 'De-anonymisation is a risk primarily to the person who is de-anonymised' (A7) with consequences of other risks occurring, including malicious misuses (eg. identity theft) and harm to the individual. 'Any use of data has disclosure risk, it's impossible to reduce that to a zero chance.' (A12).

Identification risk has long been a serious concern (and constraint) in the research use of educational administrative data, indicating a risk averse appetite by data providers. 'Respondents were concerned that the level of anonymisation used in the National Pupil Database was not sufficient to prevent personal data about individuals held within the database from being discovered.' (D1). One interviewee argued that the likelihood of identification varied with the data collected, 'the risks in administrative data are often lower than in the risks in survey data, just because of the numbers'. (A12)

In several case studies, discussions of the risk of identification referenced the normative

legislative framework of UK data protection, with its content-based definitions of 'personal' and 'sensitive' data. Some participants sought to articulate how these might be understood, for instance in an education data context, where date of birth information can link to other personal information. However, A10 suggested that educational 'attainment data, I would say it's not as disclosive as health data … that's very, highly sensitive.' A16 compared accidental release of 'your GCSE [school exam] French result' with 'releasing all of your earnings data', describing 'a different degree of privacy'.

Some suggested that the risk of identification may be re-assessed against shifting attitudes in society at large, whereby privacy may no longer be considered an absolute right (and consequently the researcher's obligations may diminish). A15 remarked that 'the younger people are, the less worried they are about privacy, about data privacy… So, you know, lots is changing here about public attitudes towards privacy'.

Others pointed out that some uses of administrative data in fact *require* identification of individuals in the data. Not doing the research would risk loss of opportunity. A9 reported 'We're doing a lot of randomised controlled trials. So for that, with the identifying, the whole thing is we can identify children so we can track them over time.'

Interviewees perceived health data as highly sensitive and confidential, therefore the risks to the data subjects (identification, misuse, misinterpretation) were significant. A51 articulated this, alongside an acknowledgement that public perception of the uniquely sensitive nature of health data is still a relative unknown.

> Certainly where health data is concerned people are particularly sensitive. … I think there are a lot of members of the public who have got things in their health record that they don't want other people to know. Now you may say that is true of people's financial information, and for some people that may be the case, but I would have thought there is a more general point around health and care data and people's

medical histories and so on being more sensitive. (A51)

Perhaps as a result of the access which transport researchers have enjoyed to detailed statistical datasets over many years, interviewees expressed the view that use of administrative data in transport research carried fewer risks relating to disclosure than in other domains. As A24 noted, 'with the transport data it's just oh, you know, oh you usually cycle or you usually go by car'. Transport researchers recognised that sensitivity also relates to increases in granularity or detail in the data. A25 considered that 'there's so much sensitivity about the geo-location because when you get down to that unit of analysis it can be easier to identify people', although A28 remarked on the lack of consistency, 'I couldn't understand why something would be classified disclosive when something else isn't'.

The risk of disclosure, inadvertent or deliberate, increases in line with the volume of data about each individual. As A32 said, 'the dataset is anonymised but at the same time there is a lot of information in them, so you can narrow down what people are in the sample.' This was reflected by data subjects in the focus group: 'I feel more comfortable with the idea of my data being used for research at an aggregate level, whereas when you get down to a kind of individual level, it kind of makes me feel that I'm losing control of how it's being used'(FG1)

Risk event 2: data linkage affordances

First, the likelihood of the risk of re-identification is increased by data linkage. Data subjects recognise this: 'Yes, there's the idea of, not just one piece, but once you link all the pieces together, you could be re-identified I think.' (FG1). As do researchers: 'As data become more and more detailed through linkage, whether it's longitudinal linkage or linkage to other datasets, then they become more disclosive' (A15). FG1 suggested that risks to individuals increased by data linkage affordances might outweigh the public and private benefits.

Estimates of the increased risks through data linkage varied with different types of data, and among different groups of interviewees. A12, using educational data, said that adding further

information to 'datasets which are used by researchers in situ with people's educational achievement, anonymised, … wouldn't be any additional risk.' Whereas linkage of transport data increased the likelihood of identifiability of individuals and specific geographic locations or areas, as A23 reported, through air pollution data.

Issues unique to linkage of health data emerged, including conflation in the minds of data providers and policy makers between different types of secondary uses. A39 identified significant differences in the risks of creating linked population data to support individual patient management including 'the use of population based linkages to identify individuals who can be targeted [for preventive treatment]', with linkages to support secondary uses for 'understanding patterns of service use, risk factors, outcomes in groups. … public health monitoring, service evaluation, performance management and research.' (A39). This warrants further exploration, particularly given the proposed models of consent and opt out that frame the issues of privacy, data sharing and secondary use (Caldicott, 2013).

Secondly, interviewees identified the risk that the largely unknown degree of linkage error or missing data affects the reliability of data and its analysis. Examples included identifying the wrong individual and 'missed links' especially in 'highly mobile populations' (A39). A14 felt that 'if we link more datasets that the risk of something going wrong increases'. A12 suggested that academics might assume that data processed by a third party in a safe setting was itself reliable, and not look out for missing data.

Opinions varied considerably however on the impact of linkage error as a risk to research integrity. A14 remarked that 'the way data is being linked and the methods and the reliability is a black box' and researchers need to maximise linkage rates and account for linkage error better. A32 suggested that linking transport data, 'because they're in different forms and in different scales, merging them into a dataset that we can use is a very complex task. … it is very easy to do the wrong type of merge' (A32). A35 used energy data where datasets are

available at different levels of granularity, so the resolution of the linked data is low.

A third issue was increased complexity, leading to greater likelihood of disclosure and of mismatches in data linkage. With energy data, where many different bodies provide, analyse and process data for the NEED (National Energy Efficiency Data-Framework) dataset, which holds 'records for approximately 26 million households, and close to 30 million properties in the data-framework. The quantity of data available gives rise to privacy risks.' (D39). A30 noted that energy data was collected monthly, daily, even half hourly, and 'the data privacy concerns and risks do grow with the granularity of the data'.

Risk event 3: misinterpretation of data

Researchers identified several risks which might contribute to an increased risk of misinterpretation of the data. Linkage based on population assumptions, described by A25, leads to misinterpretation through unfounded assumptions or inappropriately generalised data. Cavoli et al. (2015) give examples of assumptions made from a national dataset that cannot be applied to a local population or *vice versa* and from datasets which are insufficiently detailed to facilitate investigation of specific issues, such as unusual injuries or effects upon particular minority communities.

Misunderstanding of characteristics in administrative data leads to misinterpretation. Cavoli et al. (2015) give the example of cycling casualty rates growing in line with an increasing popularity of cycling. Similarly, A29 noted differences between datasets: 'if you have car licence … administrative sources always overstate the number … if you die, you usually don't get round to telling the DVLA [Driver and Vehicle Licensing Agency] that you no longer have a driving licence, so you're counted in their set ... But you've dropped off the population set.'

Misinterpretation can arise through variations in data capture, which may not be evident. A11 reported, 'the UCAS [Universities and Colleges Admissions Service] data that I looked at on

schools, you find that certain schools code pupils in a certain way, other schools don't'. A34 reported the risk of the researcher not understanding the data creation process: 'The problem is if you expect the data to be perfectly clean and suitable.' Misinterpretation (or misunderstanding) also poses challenges to data providers. A13 commented, 'Sometimes customers don't necessarily know what they're asking for. And then when they get it, they don't quite understand what it is that they've been sent.'

A second aspect of the risk of misinterpretation is where the researcher's data analysis itself is misunderstood or is misused, posing a risk to the researcher's professional reputation. A7, A9 and A11 all gave examples of this type of risk. A9 worried that while his data analysis could properly be used to evaluate educational outcomes over time, it could be misused through ignorance, for instance in teacher performance reviews. While A11 reported not publishing some findings for fear of misuse by parents in the school selection process.

Risk event 4: malicious misuse of data

The risk of misuse describes the possibility of researchers or third parties acting with malicious intent towards an individual or group represented in the data. It results from re-identification risk, but whereas the risk of disclosure is inadvertent or accidental, misuse is a deliberate act. A16 gave a lurid example, 'if you are Hungarian and you're an abusive father, you can then look up the name of every kid who took Hungarian GCSE [school exam], and … with a bit of work, you'd be able to find your ex-wife.'

Risk event 5: harm to individuals and groups identified in the data

Both deliberate misuse and accidental re-identification may result in harm to an individual. Similar to 'damage and distress' in the UK Data Protection Act (s. 10), interviewees recognised that the risk of causing harm to an individual extends beyond tangible damage such as physical injury or financial loss to emotional or reputational harm. A15 commented 'by harm we don't just mean physical harm, because that's usually unlikely, it's reputational

harm or it's psychological harm.'

Groups or individuals within a particular group potentially risk harm. In education data, 'this was a particular concern for respondents from bodies representing groups where the unique characteristics of their representatives made them easily identifiable. Several respondents were of the opinion that access to the National Pupil Database should remain restricted and/or there should be an opt-out mechanism.' (D1). A17 suggested the risk was most likely to occur in a population with similar characteristics where you could 'identify those people who were different.'

The risk of prejudicing members of a particular group was noted by A15, 'let's say you come to the conclusion that all people of the Muslim faith in Wolverhampton are likely to be jihadist … if you did something like that you would stigmatise a whole group inappropriately.' This risk was also recognised by data subjects, such as A17 who said 'categorising people by their different nationalities or different religions. It … give[s] minorities a disadvantage, the way they're treated.'

Risk event 6: risk to commercial confidentiality

Interviews with energy researchers and data providers raised concerns for the largely privatised energy industry about risks to commercial confidentiality. A government data provider (A26) commented 'I just think their main worry is the data gets lost, it gets leaked … there are sort of two consequences, it would be embarrassing … and also some of the information might be considered to be commercially sensitive. So their competitors might use it to their advantage.' Energy researchers, however, tended to be sceptical about the genuine risks to commercial confidentiality. A35 commented that 'the reasons cited by the suppliers… is commercial confidentiality, or … the commercial worth of the data. …in reality all the suppliers know exactly what all the other suppliers are doing.'

Commercial risks were also raised by individual businesses being identifiable even in aggregated data owing to geographic location or other distinguishing attributes, as noted by A35, or data variables which might be commercially sensitive, such as property tenure, 'whether you own or merely occupy the building is a key piece of business information.'

Risk event 7: opportunity risk of not using data for research

Data providers varied in their risk appetite. A24 said that transport data was more easily available with 'less hoops to jump through' than health data. A28 justified quite an aggressive attitude towards the use of administrative data in transport research, the disclosure risks notwithstanding, saying 'I think we should always be pushing to use more. I think we should always be testing, testing the boundary.' A30 identified a conflict of interest between the commercial interests of energy suppliers, who must also comply with government policy aimed at better public outcomes (fuel poverty, energy efficiency, climate change).

Assessment of risk amongst data providers is, at best, inconsistent, evidently a source of much frustration amongst researchers, including A28, 'I don't think they're consistently applied. … I think a lot of it just comes down to the person on the day. … certain people are more pragmatic than others.' A12 suggested that data reuse 'is further up the list of priorities in some Departments than others, and that reflects … the extent to which the senior politicians or policy makers feel that it's an important issue.'

A balance between risks of using the data and the opportunity risk of not using it needs to be struck. A31 reported meeting the data provider requirements, 'the training and then I could use the dataset only on a safe environment, a secure environment, to access the web, their server. But… I abandoned this dataset.' A24 warned that 'being risk averse makes research less useful.'

Risk aversion in broader public opinion is also an opportunity risk. A28 suggested that 'the biggest risk of course is this sort of public perception issue and, you know, the extent to

which the public understand the benefits of the data and also the safeguards that have been put in place to ensure that there's nothing particularly disclosive.' Researcher A27 outlined the opportunity risk if access to key administrative data sources on energy demand and use was withheld, or if data are not provided in appropriate detail. Risks that future policy development and industry-level evaluation may be hindered by restrictions on data access to NEED is also recognised by DECC (Department of Energy and Climate Change) which states 'the development of UK energy policy has required more detailed data to help deliver and monitor reductions in energy use and emissions.' (D38)

One of the strongest emergent themes in the interviews with health researchers revolved around access to data from NHS Digital following the Partridge Review (2014) and the care.data programme (Digital Health Intelligence, 2014), which for a period of time resulted in a lock down on the processing of data requests. The implementation of new procedures with a stronger scrutiny and control led to prolonged delays to data access, and more access request refusals. Researchers faced an increase in risk of lost opportunity. The interviews provide evidence of how funded projects had to be significantly modified in quality and scope to get around the lack of access to data, such as A40, forced to abandon a project because of delays getting data, we 'simply haven't been able to answer the research questions we set out with'. Coping with multiple data providers also risks access approvals and delays, as A45 reported. Opportunity risk is increased by uncertainty and risk aversion.

Figure 1: Risk events summary [insert]

Research is put in jeopardy if the interdependencies between stakeholders, types of risk, and risk management approaches are not balanced. Without a holistic understanding of this interdependency, risk mitigation designed to protect the interests of one group may open up more risk for others. Risk management must be viewed at a collective, societal level where these interdependencies are balanced in order to best mitigate the risks for the benefit of all.

3.2 Stakeholder mapping

The research helped to identify stakeholder communities affected by the different risks who have a role in mitigation and risk management. As Figure 1 illustrates, we identified in the research data five entities with a stake in the beneficial use of government administrative data.

Figure 2: Stakeholder mapping [insert]

Individual researchers, usually working in an academic setting, seek as much unfettered access to relevant data for research, but want to avoid reputational risks for themselves, their research group, discipline and employer. Secondly, employers, mainly of academic researchers, have legal and ethical interests in information security and privacy, mitigating reputational risk, while enabling researchers to carry out high quality research. Thirdly, the wider research community, of individual researchers, research groups, their employers and representative bodies (eg subject associations and networks like Administrative Data Research Network), seeks to manage the risks of data re-use. The research community influences data provider and researcher behaviour and can act as a broker. The fourth stakeholder group are data creators and providers, mainly central government agencies but also trusted third party processors, commercial providers of data (especially in the energy data field) and regulatory oversight bodies, such as NHS Digital's IGARD (Independent Group Advising on the Release of Data). Data providers should give transparent and fair access through their protocols, to balance the risks associated with providing data to researchers and the opportunity risk if data was not provided. Data subjects and the broader public in whose interest research is carried out (including individuals, households, and businesses) provide the fifth stakeholder community. Once data subjects have provided their data, they cede control over its use. But they express views on governance mechanisms and on the risk appetite of public bodies. Promoting better public understanding of research use of

data and the risks and benefits attached to it, among individual citizens and collectively, will strengthen the voice of this stakeholder group.

Risk appetite varies within the five stakeholder communities and between the different stakeholders. It changes over time as part of a chain of interrelated consequences. One interviewee (A30) compared 'the risks from the perspective of an energy consumer, a householder' with 'the perspectives of government departments and commercial organisations, energy suppliers, National Grid'. Commenting on risk appetite, A30 said 'I think there is over-estimation of these risks. You know, government partners are extremely risk-averse, and they fear, more than anything, the headline in the *Daily Mail*.' A10 noted the interdependent responsibilities for managing data risks dispersed across their research team, the university employer's ICT infrastructure and wider research community. Interviewees A24 and A28 also, however, expressed frustration where risk management was excessively burdensome given the data risks. 'I've done safeguarding, …ADRN [Administrative Data Research Network] training, safe researcher training… lab induction, and …background checks.'(A28). While A24 noted that whatever was done, 'the risk is not zero…'. Table 3 summarises the five stakeholder groups.

Stakeholder group	**Risk focuses**	**Risk mitigation**
Individual researchers	Access to relevant data for research; reputational risk; linked data	Researcher training; credentials; documenting data processing; data immersion; linkage protocols; statistical disclosure controls; compliance with ethical and provider frameworks.
Employers of researchers	Legal and ethical risks; organisational reputation risks; information security	Implement information security (eg safe harbours) & research data management practices; ethical research approval; legal compliance.
Wider research community	Reputational risks; legal and ethical risks.	Peer review and ethical frameworks; sanctions for researcher misbehaviour;

		influencing sector attitudes to data; developing linkage protocols and techniques; supporting researcher training and credentials; good data management practices.
Data creators & providers	High control over access and access conditions to data; balancing access risks v. opportunity risks.	Data supply agreements; data security; detailed data documentation; auditing and monitoring requirements (credentials, training, trusted third parties, information security); fair costs and charges for data access; publishing open data; commissioning research; project specific or time limited access to data.
Data subjects and the broader public	Public benefit v. private harm; anonymization risk;	Governance mechanisms; risk appetite; public engagement.

Table 3: Stakeholder groups, their risk focus and mitigations

3.3 Managing and mitigating the risks in the research use of administrative data

The analysis and coding of the research data revealed 22 risk mitigation actions considered or taken by the interviewees, which we aggregated into four clusters of risk mitigations. These are sectoral expectations of behaviour and professional reputation; data supply and data security protocols; controls of access to data and metadata; and changing attitudes to open government data. They are summarised in Table 4.

Risk mitigation cluster	Risk mitigation actions
1: sectoral expectations of behaviour and professional reputation	1. Ethics and research integrity protocols 2. Sanctions for un-ethical behaviours 3. Researcher credentials 4. Researcher training 5. Peer review 6. Documenting the research process 7. Data immersion
2: data supply and data security protocols	1. Data supply agreements 2. Data documentation 3. Data governance auditing and monitoring 4. Data security protocols

	5. Good data practice 6. Linked data affordances
3: controls of access to data and metadata	1. Limited time, limited data, project specific access 2. Statistical disclosure controls 3. Anonymization and aggregation of data 4. Granularity of accessible data 5. Data sharing 6. Commissioned research 7. Privatisation
4: opening up data	1. Promoting a dialogue of public engagement and public understanding 2. Proactive publication of open data

Table 4: risk mitigation clusters and actions

Risk management 1: sectoral expectations of behaviour and professional reputation

The first group of mitigation actions focuses on ways of ensuring that researcher behaviour complies with the proper expectations of the wider community, as expressed in ethics and research integrity protocols, credentials and training, and peer review. A7 identified a conflict, saying 'It's a slightly tricky position of being my own policeman but you know I've got professional standards'. Education researcher A16 commented that individuals clearly understood and accepted their obligations. Risk management engendered less discussion in the transport case study, possibly because such data was less sensitive. A24 remarked, 'It's crazy to think that anyone who does it will be doing anything malicious with the data.' While A28 said 'I thoroughly agree that people will be responsible in the way they approach the analysis of the data and that kind of thing'.

Ethics and research integrity

Some risk mitigation responsibilities rest with the wider research community: local institutional infrastructure and across academia. For example, A15 stated, 'primarily of course it's the research community itself which has to satisfy any ethical considerations and requirements, which is done routinely through research ethics committees, or institutional review boards'. Some interviewees saw a separation between the practical management of data by, for example, the university IT infrastructure compliance with data security standards,

from the individual researcher's data use. For instance, while there might be an agreement to destroy the data, the practical execution involved several parties, 'You can't leave it to the researchers' (A14). Ethics processes do not appear to have evolved fully to accommodate the use of unconsented personal data in research, to reflect the changing legal environment, or distinguish between the risks of using aggregate and individual data. A7 noted 'my institution… take[s] the view that aggregate data aren't particularly important, that it's all about individual level data when ethics comes into play'.

Sanctions

The consequences of 'deliberately breaching confidentiality' are clear, including criminal prosecution, or action by the data provider against both the researcher and their employer (eg D10). A15 set this out in detail: 'most researchers work in a research institute or a university, if they're getting access to administrative data … their employer is bound in to a set of conditions about how they must behave with respect to those data, and the penalties for disobeying that or abusing that set of rules and guidelines is fairly extreme. …a lifetime ban on their receipt of [Research Council] funding,… a ban on funding to their organisation … the penalties which are now in place are quite draconian, so … the institution has a very strong incentive to ensure that their employees do not do anything inappropriate.' (A15)

Researcher credentials

However, sanctions do not adequately address risks in advance. A12 stated, 'it's better to focus on credibility in the quality of the people who are accessing the data, … checking who is getting access to data, as much as what data they see.' A credentials-based approach is also recognised by government data providers, for example, in handling of repeat requests for educational data, which receive priority if 'The request is made by an accredited requester who has already demonstrated their credentials' (D10 – National Pupil Database User Guide).

Researcher training

Researcher training helps to ensure users properly understand the data and avoid mistakes. A12 explained, 'if you were training new researchers, it's in the descriptive statistics, the table at the beginning of your paper, that you're most likely to accidentally make a mistake.' Training is a requirement of data supply agreements, for example access to educational data, 'shall be restricted to only those persons who have received appropriate training regarding data protection and security' (D2). Sometimes, the data provider delegated training to the researcher or their employer. For example, 'The Requester [for NPD] shall ensure that each proposed Permitted User receives appropriate training regarding data protection and security to enable the Requester to comply with principle 7 of the Data Protection Act' (D6). However, A8 described the increased requirement for researcher training and more stringent data security protocols as 'a real balancing act'. 'Thinking, oh, we should put more processes in place, … I think that always seems sensible when you're foreseeing all these problems. But then when it's practically implemented … that will discourage people from using [data]'. Training requirements demanded of researchers seeking access to data seem out of line with the requirements for data capture. For example, a data manager responsible for capturing student data (A6), in response to a question about training or background checks, responded, 'No. I've been working on this kind of data for about fifteen years I suppose, but I don't remember anything.'

Peer review

The traditional apparatus of community-based quality control in academia, peer review, helps to manage quality risks and, as A14, remarked, 'I think that's what the peer review process is for. I think there'd be a danger if they [government data provider] were taking an interest that it would be politically driven.' However, peer review is compromised if administrative datasets are not available to reviewers or journal readers. A14 reported that 'a lot of journals insist on making all the data you've used in your papers publicly available. So far I've got

away with saying that we could arrange for any independent third party to have permission to access the data to check our results but we can't make it publicly available. … But that, that is a problem'. A1 identified conflicting priorities, 'the data provider's stipulations trump the funder's stipulations. I think it's legitimate as a researcher to say that the work in this paper was done on data with restricted access.' A related issue concerns the abilities of the reviewers themselves, who, given the narrowness of access channels and novelty of research using administrative data, may lack expertise to judge the data analysis. This may be a systemic issue for quantitative social science research, as A8 reported 'I'm not sure the peer review process would necessarily pick up the appropriate usage of any data, whether it be administrative, large scale survey, census data or whatever, for sure.'

Documenting the research process

A34 argued that solutions lay in documenting the process of research analysis, as much as in documenting the data. 'People need to be recording clearly how they created variables, for instance, so what was the cleaning process, in our field, the energy field, it's just terrible in terms of the lack of documentation, lack of clarity…'. But she noted the paucity of guidance available to researchers about documenting data analysis, 'You can have a whole library just on how to perform…various statistical computations etc., and then I've managed to find one book, one book … on how to actually document the steps as you go.' Transport and energy researchers suggested that quantitative researchers could learn from qualitative research methods of analysing secondary data sources, 'upping …the amount of time that's just spent on, you know, quality assurance of data, rather than just only run the stats' (A34).

Data immersion

A thorough understanding of administrative data is vital to mitigating the risk of misinterpretation. As A22 asserted, 'it's your job as a researcher to understand how that [data collection] may influence your results'. He gave an example of traffic accident data

(STATS19) which extended well beyond documentation, training, or desk-based processing. 'My boss actually rode along with one of them [the police] as sort of like an experience, you know, this is how data is collected.' (A22). He also emphasised the need for the researcher to understand data variables, such as standard response categories, in detail. Data providers need to make available sufficiently detailed documentation, including changes in data collection mechanisms over time, to facilitate data immersion by researchers.

Risk management 2: data supply and data security protocols

The second group of mitigating activities focuses on protocols which govern and regulate the supply and re-use of data.

Data supply agreements

Data providers have varied and detailed requirements embedded in their data supply agreements. Some, such as the Department for Education, depend upon the researcher's individual declaration of compliance with the terms. A13 explained: 'When they received the data they signed up to the terms and conditions, and how they were going to be making the data available, of outputs that were going to be arising out of that data, and …they can let us see what it is they're going to be publishing, just so that we're happy.'

Data documentation

Opinions varied as to the quality of the documentation supplied with energy data (A34, A35), although A27 suggested certain advantages to raw state data. 'So they don't have a big book of standards that are necessarily always coming with these datasets. So it does take quite an effort, for the most part, to come to grips with all that could be in there, … we want to know how things are being recorded, and we want to be able to see the messiness of it, in order to evaluate it'. It may help to understand how data anomalies arise, as A35 reported 'where we do have problems is they step outside their standard codings, and their standard descriptions. … they use that as a means of making the description more specific, but it does make our life

a bit more difficult because we have to clean it all up'.

Data governance auditing and monitoring

Data governance requirements have increased in recent years, as A15 suggested, 'they're engaged in much more auditing and policing of what happens to data.' Data providers have established regulatory oversight bodies to advise, such as NHS Digital's IGARD. Yet potential loopholes remain over the auditing of conformity. A12 noted 'they don't check it directly in the sense of coming down to check your computer, I guess they're relying on another external process to ensure that it's safe.' While A11 asked, 'Do they ever audit? Er no, not to my knowledge…they might ask for proof … but they haven't checked.' Data providers often rely on researcher good practice, thus reducing the compliance burden, as A14 reported, 'With the NPD you just use an ordinary computer, NPD-HESA [National Pupil Database linked with Higher Education Statistics Agency] you just use … on a secure computer that only you can access, but the fact that there's no controls over… what you take away and nobody examines your outputs'. A24 suggested that researchers in turn rely on institutional systems, 'we sort of delegate it to our IT department. All the data are held on computers, and … I haven't heard of any university datasets being hacked into.' Delegation of governance might result in an increase in risk.

Data security protocols

Data providers contribute to risk management by taking a balanced approach to data security. For example, with education data, A12 suggested, 'the kinds of information that the government department is interested in, tends to be, can this data be inadvertently released and reveal identity? That's the overriding concern.' Department for Education seeks to 'protect the privacy and confidentiality of individuals' in education data (D10 4.1). They warn that 'NPD data users need to consider the risk of identifying individuals in their analyses prior to publication / release' and that researchers 'should be on the alert for any rare

and justified and unintentional breach (D10 4.3). A12 put emphasis on 'the security questionnaire... We need to know that the organisation has got the relevant data security base.'

Researchers accepted data security protocols, as A15 remarked, 'I think data security is an absolute given. You have to have data security if you're going to persuade people that what you are doing is not open to any kind of abuse of their personal information.' 'So gone are the days of people just leaving data sitting on laptops or leaving laptops on trains, or leaving diskettes, you know, disks and USB sticks and so on around' (A15). Researchers felt that they had not been the culprits in data breaches, 'it's always the government department that's lost the data, not the academics' (A1), while A9 remarked 'education researchers… we're all now suffering because of some mistakes made by other people.'

Some expressed concern that the balance between data security and access had tipped too far in favour of security. A12 reported that, 'When they decided that there would be individuals in each government department responsible for the data, there was sea change in people's attitudes to releasing the data, and government departments became much, much more careful… But the downside of that is … never to release the data, that way you take no risk.' A9 commented that data breaches 'make the system harder for everybody.'

Good data practice

Even where cross-departmental best practice does exist, such as HM Treasury (2015) Aqua Book, which resulted from a review of quality assurance of government analytical models, departmental boundaries (and academic disciplines) limit adoption. Tools, guidance and templates (GB. Department for Business, 2015) targeted at researchers internal to government would also useful to external researchers using government administrative data, but seem not to be widely adopted: energy researcher A34 was the only interviewee to mention the Aqua Book.

Linked data affordances

Linked data affordances can be a source of risk, but also a risk mitigation. For example, linking government datasets to existing cohort studies might expose flaws and biases in the data collection processes. A15 suggested that 'one of the benefits that arises from giving researchers better access to administrative data is that they start to uncover all sorts of problems in the data, and also the documentation for data is often, you know, severely limited, lacking, or of poor quality. And it causes organisations to think about the way in which they document what they are doing. It also reveals what can happen when administrative processes change … what data are generated, how they're generated, how they are preserved, how they are linked through time' (A15). A14 added, 'having access to survey data, you can sort of unpick some of the biases in these administrative [data].'

Linking datasets from other providers allows beneficial sharing of research ethics best practice. A12 said, 'they have research ethics committees, they have a very formal structure, they have a group collectively considering the ethics of the linkage… individuals who have expertise and competence, including lay members of these ethics committees to make judgements about whether linkage is a good idea.'

Too many different parties may, however, lead to unclear lines of accountability, lack of agreed procedures, confusion and delay, 'if two datasets are linked, there has to be an explicit decision I guess as to who is responsible for the data' (A12). A27 also identified confusion over responsibility, 'I've been applying to join two different datasets together for a while now, and again the interest is there, everything's there but actually there is not a clear recognition as to who says yes.'

Risk management 3: controls of access to data and metadata

A third group of risk mitigation actions focuses on controlling access to data and to metadata. Data providers and third party data processors manage many of these mitigations, which

sometimes restrict or prevent beneficial research access to data.

Limited time /limited data/ project specific access

A common means of controlling researcher access to administrative data is to make it available for a time-limited period only, for limited data types, and for a specific, pre-defined research project. Data subjects would welcome an explanation of the limits around their data, as a student asked, 'In what way have they accessed it and [who] is it given to…' (FG1). Researchers were largely accepting of the consequent need to destroy data at the end of a project, although they expressed frustration at the impact this has upon their ability to carry out follow-up research. A9 said, 'In theory, I've got the data to do it, for another project… I'm constrained from doing it. Now I could re-apply, but I don't know how long it would take.' An alternative approach is to replace project-specific permissions with a licence for accredited persons to use the data. A9 explained, 'you have a licence to have this data, right, and within a set of guidelines about anonymity and protection and security, and not asking for data which is more sensitive than you need, but then you have a licence to basically play with the data then and come up with research questions and research ideas… to find out why this happened or what you could do about it'.

Some data providers manage risks by providing only 'a sample of data would be used for a number of reasons, including the reduction in the risk of data disclosure or loss' (D39). Similarly, educational research does not usually need access to a full set of attributes held in a particular dataset. Sensitive data could be withheld, as reported by A16 and A9: 'it's got nothing to do with individuals. I am not interested. … In fact, when I ran the analysis the first time, I deleted every column except the local authority, the school, the free school meal and the ever-free school meal...I chop out everything that I'm not going to need'.

Statistical disclosure controls

Administrative and other data is sometimes subject to statistical disclosure controls to ensure

that anonymised individuals are not identifiable. Linking data which includes unusual characteristics or clusters can allow anonymised people or places to be re-identified. Data supply agreements often require the mechanism of statistical disclosure control or rounding mechanisms 'to ensure that all statistics published are at a level of anonymisation and aggregation which will ensure that no Personal Data or Sensitive Personal Data are published, and thereby ensure the confidentiality of individuals' (D2, D6). A15 explained, 'As you create a longitudinal record, in other words you link the information that people have given you through time, whether per person or the family, then you're creating a … richer source of information but you're also creating something which is more and more disclosive through time, even if the identity of the individual is removed before those data are made available to the research community. And therefore you've got to engage in disclosure control mechanisms, so that …you can say that, you know, these data have been used for research but you were not identified and your identity cannot possibly have been revealed in the process of conducting that research' (A15). A11 followed a similar protocol, 'my models are all aggregate, results are displayed in the aggregate, and typically sort of cloaked in a sort of geo-demographic classifier which, I mean you could argue that's another level of sort of abstraction from the individual.'

Anonymisation and aggregation of data

Anonymisation can go beyond individuals to entities, such as businesses or schools. A14 suggested a way to 'anonymise the school data, so, so you could just, you know, make sure the schools all had the same identifier but anonymise it, effectively that would be sufficient to completely anonymise the data and you know, make it publicly available. …I've always said to journals we can't make this data available ... but I can see that changing' (A14). Researchers were keen to stress that no identifying data about pupils would be published: 'We're not even identifying schools, let alone individuals' (A9).

Energy data from the linked data sources in NEED is anonymised. A26 reported, 'The whole idea of anonymising the dataset was to prevent individuals from being identified and individual households from being identified, so we had to take out all of the data that would allow it to be linked.' Multiple layers of aggregation also make re-identification difficult.

Granularity of accessible data

Decreasing granularity of data makes it less disclosive, but as energy researcher A30 suggested, less useful for some research. 'So there is a fairly reasonable quality and quantity of aggregated energy statistics available… at the national level or kind of sub-national level, the sort of going down to local authority level and maybe a little bit lower, you can get reasonable energy data. Going beyond that is extremely difficult'. In contrast, the Valuation Office Agency data on business rates was cited as a world-leading example of the research potential of raw administrative data. 'It's completely disaggregate, which is one of the huge values of it. … As a resource, it's essentially unparalleled in the world' (A35).

Data sharing

The research identified a need for government agencies to collaborate on data access requirements and criteria, and apply lessons learned from data sharing initiatives to requests for access and linkage of data for re-use. Interdepartmental collaboration within government (or the absence thereof) was raised in the transport case study and came through strongly in interviews with energy researchers, such as A35, 'There's no joined up thinking between government departments in terms of data. It just isn't there.' Different approaches to data sharing between government bodies leads to inconsistent risk assessment about similar types of data, and consequently their release to researchers. A27 reported 'the argument of not allowing detailed individual access … I've requested it before and have been turned down, 'cause they treat everything like an FoI, and they say, well, it's to do with privacy.'

In spite of the work of ADRN, negotiations for data access generally proceed on a case-by-case basis and in isolation from each other, such that decisions reached do not provide precedents when similar requests arise in other parts of government. Access is re-negotiated from first principles, resulting in protracted negotiations, such as researcher, A30's 3 year negotiations with energy suppliers over access to data on energy efficiency interventions under government-funded schemes. Whilst DVLA has been at the forefront of transport data sharing initiatives for administrative purposes within government and with industry, there is no evidence that this has smoothed the path for researchers wishing to use the same data for secondary analysis.

Commissioned research

Delays and inconsistencies in the approval process for obtaining administrative data led some researchers to seek other routes to data access. One is via government-sponsored research, directly or through third party organisations. A10 hoped that saying, 'we can access NPD data from other funders, we'll say we can get access' would facilitate her education research. However, the need to specify the purposes for which the data will be used makes this a frustrating experience, particularly where linked data is concerned. A9 reflected on 'a tactical error' running three projects together in an application for pupil data resulting in confusion, questions and delays, 'three different things here and we need different things for different reasons, and then there were three parties involved'.

Privatisation

Many researchers were concerned about government data going behind paywalls, where publicly funded researchers had to pay government departments for access to data to undertake research in the public interest, and the effect of privatisation of government functions curbing access to data. A35 gave an example of privatisation of government Executive Agencies, such as Ordnance Survey Ltd, fearing another 'Building Research

Establishment … a government research department on buildings which got privatised and then they said, oh well, we're keeping all of our data then, despite the fact that it was collected with public money, you can't get it.' Privatisation is a rather extreme example of control of access to data which at least has the benefit of reducing the risks associated with data sharing and re-use.

Risk management 4: opening up data

The final group of mitigating activities relate to changing the risk attitudes about data access, partly through encouraging more proactive publication of open data, and by promoting a dialogue of public engagement and public understanding of research access to government data.

Public engagement

The researcher's individual responsibility extends well beyond managing the risks in data and academic outputs into proactive public engagement activities to demonstrate the balance of risks and the public benefits of secondary research using administrative data. A14 felt that 'researchers have a responsibility to actually show the public worth, this involves, you know, getting support from the groups, and making a case.' Public engagement would also help to mitigate risks such as misinterpretation, allowing the researcher to explain 'why these data may or may not show what they appear to show and so on' (A7) and to demonstrate the public value of the research.

Open Data

Making data as widely available as possible is one radical approach to managing risk. A12 exclaimed, 'You know, the best way to counter that is to have data accessible to as many people as possible …. in education that's exactly what's happened, so making this data available, particularly to other disciplines has meant that some of the debate around certain controversial issues that in the past have been very dominated by individuals or small

datasets that are not replicable, now it's much harder because the data is publicly available.' A16 said, 'some people think we should just put the National Pupil Database on the web so anyone who wants to use it… and obviously in a world of open policy-making and open data, there's much to be said for that.' A compromise between confidentiality and open access to a dataset might be to look for new ways to make parts of the dataset publicly available. For instance, A16 suggested issuing a 1% or 2% open sample of NPD data.

Open data helps researchers to understand what kinds of administrative data might be useful for their research work, before applying for data access. A1 gave an example from HESA, the Higher Education Statistics Agency, 'they publish some open tables on their website and they also have non-open data, but it's fairly clear from the open data what sort of things they have, and you can imagine what research work you might be able to do using that.'

Transport researchers appear to be enthusiastic advocates of open data as a means of validating data and reducing the risks of misinterpretation. A24, suggested that 'If data are publicly available … somebody who's … queuing to cross check and validate every dataset, and they might say, I've found this dataset, I can prove it is all inaccurate. … So if everything's publicly available, it means they can be scrutinised more easily.'

4. Conclusions

The discussion has analysed elements of the risk framework associated with research use of government administrative data, brought together in Figure 3. Table 2 and Figure 1 identifies and describes seven risk events, including re-identification, data linkage, misinterpretation, misuse, harm, commercial and opportunity risks. Table 4 summarises 22 risk mitigations in four clusters which can be adopted by the five groups of stakeholders (researchers, employers, research community, data providers, data subjects) identified in Figure 2 and Table 3.

Figure 3 Risk framework elements [insert]

The analysis derives from the experiences of those interviewed for the study, captured in the interview transcripts and association documentation. It is therefore reflective of that experience and does not attempt to present all possible risks or all possible mitigations. By interviewing stakeholders in four case studies, we sought to avoid mono-disciplinary views and assumptions on risk. We believe that the analysis presented here will be of value to those working in other disciplinary areas and to those such as records and information managers who may find it helpful to better understand the perspectives presented here. It was surprising that records managers were not more visible either in data provider or research user organisations as they have relevant skills in data management, however they were seldom directly involved in administrative data in our case studies. We do not claim that the risk framework will be applicable in all data settings, rather that it adds richness to our understanding of data risks and how best to manage them, especially by highlighting the views of academic researchers who use administrative data.

Further discussion is needed between stakeholders to develop an agreed framework for managing risks collaboratively, which can then be implemented, monitored, reviewed and continually improved. At present the data community lacks mechanisms to agree and implement risk mitigation and risk management frameworks. As a result, risks are experienced locally, lessons are not learned, and there is no opportunity for the research community to better manage risks. No forum or community of practice exists in which discussion of best practice in administrative data management, and specifically risks and risk mitigation, can take place. A balance is required to create a risk management environment that benefits all stakeholders, since the exercise of risk mitigation by one group while decreasing the risk for that group, may in fact shift the burden and increase the risks for another. The approach to risk must reflect the nature of the data (sensitivity), the conditions of its creation (confidentiality, legal frameworks), its depth (granularity) and its breadth

(degree of linkage) which combine to influence the weighting of risk factors. An understanding of risk weighting, not discussed in this article, should inform the approach to risk. Further research is needed into cultures of risk management or risk appetite: perceptions of what is too risky vary across the different stakeholder groups. Risk mitigation cannot simply focus inwards on effective systems and processes but must also look outwards and include public awareness of data sharing, and the public perception of risk. It is in the interests of all those responsible for the creation, management and re-use of government administrative data to make progress on a common mechanism, in order to ensure that the power of administrative data is fully exploited for the public good.

References

ARMA International (2010). *Evaluating and Mitigating Records and Information Risk*, ARMA, Overland Park, KS.

Caldicott, F. (2013). *Information: To Share Or Not To Share? The Information Governance Review.* London, Department of Health.

Cavoli, C., Christie, N., Mindell, J., Titheridge, H. (2015). Linking transport, health and sustainability: Better datasets for better policy-making. *Journal of Transport & Health*, 2:2, 111-119.

Daneshmandnia, A. (2019). The influence of organizational culture on information governance effectiveness. *Records Management Journal*, 29: 1-2, 18-41.

Digital Health Intelligence. (2014). *Care.data: a row waiting to happen.* https://www.digitalhealth.net/2014/01/care-data-a-row-waiting-to-happen/ Accessed 31.01.18.

Doty, P. (2015). U.S. homeland security and risk assessment. *Government Information Quarterly*. 32: 3, 342-352.

GB. Department for Business, Energy & Industrial Strategy (2015). Aqua Book resources. https://www.gov.uk/government/collections/aqua-book-resources. Accessed 31.01.18.

GB. HM Treasury. (2015). *Aqua Book: guidance on producing quality analysis for government*. https://www.gov.uk/government/publications/the-aqua-book-guidance-on-producing-quality-analysis-for-government Accessed 31.01.18.

International Standards Organisation. (2009a). *PD ISO Guide 73:2009 Risk management — Vocabulary.*

International Standards Organisation. (2009b). *BS ISO 31000:2009 Risk management — Principles and guidelines (revision consultation ISO/DIS 31000 :2017(E).*

ISACA (2010). *The Risk IT Framework*, ISACA, Rolling Meadows, IL.

Lemieux, V L. (2010). The records-risk nexus: exploring the relationship between records and risk. *Records Management Journal*, 20: 2, 199-216.

Partridge, N. (2014). *Data Release Review*. Leeds: Health and Social Care Information Centre.

Quigley, K., Burns, C., Stallard, K. (2015) 'Cyber Gurus': A rhetorical analysis of the language of cybersecurity specialists and the implications for security policy and critical infrastructure protection. *Government Information Quarterly*. 32: 2, p108-117.

Sexton, A., Shepherd, E., Duke-Williams, O., Eveleigh, A. (2017). A balance of trust in the use of government administrative data. *Archival Science.* 17:4, 305-330.

Sexton, A., Shepherd, E., Duke-Williams, O., & Eveleigh, A. (2018). The role and nature of consent in government administrative data. *Big Data & Society*. 5:2, 1-17.

Sprehe, J.T. (2005). The positive benefits of electronic records management in the context of enterprise content management. *Government Information Quarterly*. 22: 2, 297-303.

Xie, S L. (2019). A must for agencies or a candidate for deletion: A grounded theory investigation of the relationships between records management and information security. *Records Management Journal*. 29: 1-2, 57-85.

Figure 1: Risk events summary

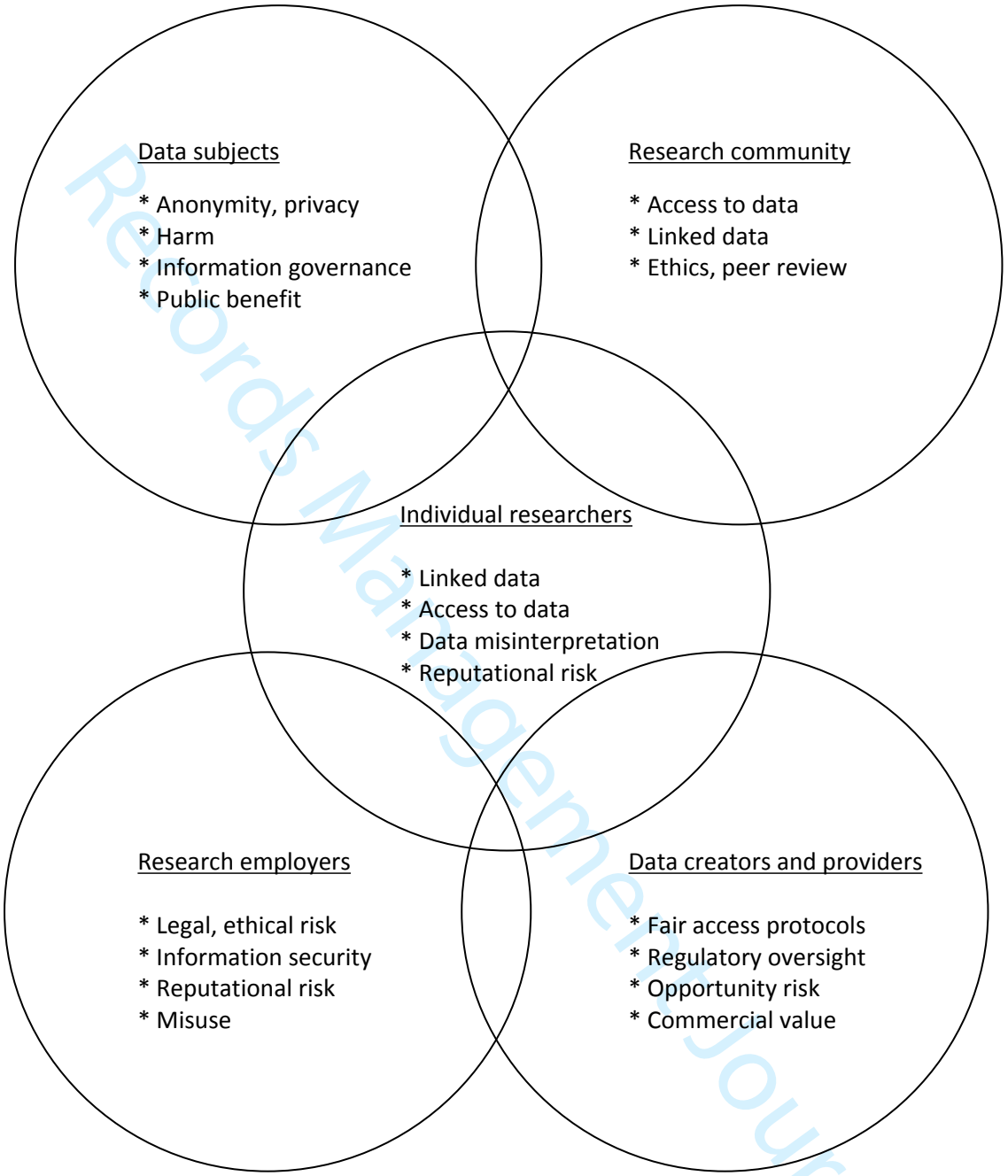

Figure 2: Stakeholder mapping

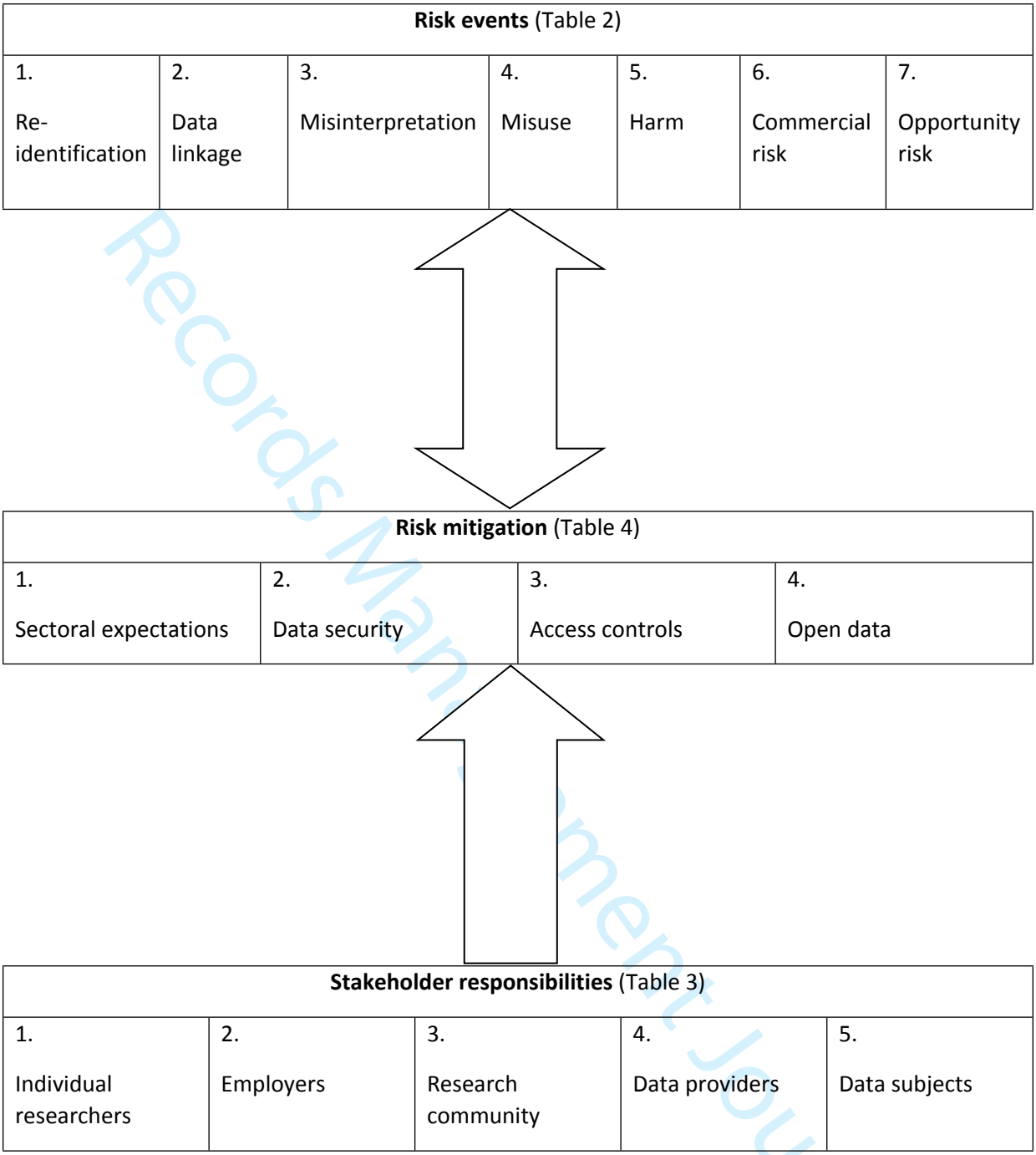

Figure 3: Risk framework elements