Principles- Versus Rules-Based Output Statistical Disclosure Control

Bridging the Business Data Divide

Improving the Quality of Digital Preservation Using Metrics

IN THIS ISSUE
Privacy, Disclosure and Surveillance

ON PAGE 5
Editor Karsten Boye Rasmussen

ON PAGE 33
Membership Information
In this issue

4 Editor's notes
Karsten Boye Rasmussen

5 Principles- Versus Rules-Based Output
Statistical Disclosure Control In Remote Access Environments
Felix Ritchie and Mark Elliot

14 Bridging the Business Data Divide: Insights into Primary and Secondary Data Use by Business Researchers
Linda D. Lowry

26 Improving the Quality of Digital Preservation Using Metrics
Mari Kleemola
Editor’s notes

Avoiding disclosure, using data, and improving digital preservation

Welcome! The three papers of this second issue of Volume 39 of the IASSIST Quarterly (IQ 39:2, 2015) are about data. Matters of privacy, disclosure and surveillance have been, and will continue to be, hot topics in society. Within science we have obvious reasons to ensure that research cannot be accused of breaches, and therefore researchers need principles and guidance on how to avoid statistical disclosure. Staff of research data centers and universities demonstrates in the first paper their investigation of this sensitive area in ways that we all can learn from. The second paper brings insights in the use of data collections and research design in business master’s theses at a Canadian university. Better insight of the use means better guidance for users. The third paper shows how modern data archives are continually improving their data procedures by applying standards and using measurement. All three papers will give you insight into these areas and they are equally well furnished with references so you can explore the areas further.

The first paper ‘Principles- versus rules-based output statistical disclosure control in remote access environments’ is authored by Felix Ritchie from Bristol Business School, University of the West of England, and Mark Elliot, School of Social Sciences and Data Research Institute, University of Manchester. Delivering data for research requires measures against disclosure risk. The authors demonstrate the differences between the two approaches where ‘rules-based’ has been the traditional model. This is where a set of formulated rules are applied ahead of dissemination to researchers. A common example is ‘A table may only be released if there are at least three observations for each cell’. The set of rules can be comprehensive and each rule has a trade-off between confidentiality and efficiency. The authors argue that disclosure control based on principles is to be preferred, but demands more training which will in turn build a culture of confidentiality expertise.

Linda D. Lowry is the liaison librarian at Brock University in St. Catharines, Ontario. Her paper Bridging the business data divide: insights into primary and secondary data use by business researchers’ is based on analysis of 32 master’s theses within Management at Brock University. The applied content analysis is demonstrated and the coding of the theses is explained in informative appendices. The analysis describes the research designs and data collection methods found as well as the distribution of theses in business subfields: accounting, finance, operations and information systems management (O & ISM), marketing, and organization studies. It turns out that the distribution of master’s theses shows nearly half are within the finance subfield. Furthermore, all of the finance theses are ‘archival - quantitative’ while nearly all marketing theses are ‘questionnaire’-based. The conclusion is a recommendation of content analysis for examining data practices within social science disciplines.

Mari Kleemola is Information Services Manager at the Finnish Social Science Data Archive (FSD), and she outlines FSD’s venture into digital preservation standards and assessments in the contribution Improving the Quality of Digital Preservation Using Metrics. The paper demonstrates the journey through standards, checklists and assessments, such as: Open Archival Information System (OAIS), the Audit and Certification of Trustworthy Digital Repositories (TDR) Checklist, and finally the CESSDA Trust Process. The complete process is described with the key issues of measurements and requirements of each step in the process. What started as an internal handbook and a formation plan in 2003 has - especially during the last four years - been developed until the FSD received the Data Seal of Approval certification in September 2014.

Articles for the IASSIST Quarterly are always very welcome. They can be papers from IASSIST conferences or other conferences and workshops, from local presentations or papers especially written for the IQ. When you are preparing a presentation, give a thought to turning your one-time presentation into a lasting contribution to continuing development. As an author you are permitted ‘deep links’ where you link directly to your paper published in the IQ. Chairing a conference session with the purpose of aggregating and integrating papers for a special issue IQ is also much appreciated as the information reaches many more people than the session participants, and will be readily available on the IASSIST website at http://www.iassistdata.org.

Authors are very welcome to take a look at the instructions and layout:
http://iassistdata.org/iq/instructions-authors

Authors can also contact me via e-mail: kbr@sam.sdu.dk. Should you be interested in compiling a special issue for the IQ as guest editor(s) I will also be delighted to hear from you.

Karsten Boye Rasmussen
September 2015
Editor
Principles- Versus Rules-Based Output Statistical Disclosure Control In Remote Access Environments

by Felix Ritchie¹ and Mark Elliot²

Abstract
In recent years, the level of detail in confidential data made available to social scientists has increased dramatically. One particularly important growth area has been ensuring that research outputs do not present any residual disclosure risk. Traditionally this has been managed by specifying rules for researchers (the ‘rules-based’ model), but it is increasingly recognized that a ‘principles-based’ approach is both more secure and more cost-effective. The principles-based approach requires a higher level of expertise from those managing access to data, and places the subjective assessment of risk at the forefront of decision-making; these two factors make data managers uncomfortable. In addition, knowledge of this approach is concentrated amongst a relatively small community, whereas the rules-based approach has dominated for half a century; data managers may not be aware of an alternative perspective. This paper reviews the arguments for the two different approaches. They are not mutually exclusive: both take simple rules as a starting point, but the rules-based approach also finishes there. This has advantages in some circumstances, but the value of the principles-based approach increases with the sensitivity of the data and the scope of researchers to innovate. The paper considers how the two approaches can be implemented. Although the principles-based model requires greater initial investment by both researchers and those managing access to data, this can bring substantial auxiliary benefits to the latter. The paper therefore concludes that a principles-based approach is generally preferable, and it is essential for the remote research data centres which dominate access solutions for the most sensitive data.

Keywords
Data access, data security, statistical disclosure control, principles-based, output SDC, researcher management

Acknowledgements
We are grateful for comments by members of the Administrative Data Research Network; Don Webber and Richard Welpton; and the anonymous referees, who suggested the inclusion of the summary table in section 4.

The main advantages of a rules-based model is the certainty and lack of ambiguity

Introduction
Since the early 2000s, social scientists have seen an explosion in data availability. The most important has been the increasing research access to highly confidential but high utility data (see for example Trewin et al., (2007) or Elliot and Purdam, (2015) for a discussion of this trend). This has been made possible by the development of secure data access solutions.
These allow the managers of such facilities to control the data access process with a great deal of security, but researchers are no longer necessarily restricted to physically visiting the facility manager’s site; instead, they increasingly do so through remote access systems.

Some organisations have invested in secure remote job submission, where the researcher submits code to a server holding the data and receives back statistical results; for some data this is the most appropriate model, but for many types of data the dominant model is the remote research data centre (RRDC), where researchers access data through ‘thin clients’ (that is, where all the processing is done by the server holding the data, and the researcher controls the research through a web browser, for example). These allow researchers almost complete freedom to work with the data (other than removing them from the facility) and so are popular with researchers; they are also popular with data access managers, as such systems still allow a high level of control to be applied without the need for continual surveillance.

A key element of that control is checking to ensure that statistical outputs do not present any residual disclosure risk: a researcher will use the data to produce statistical outputs, and it is possible that those outputs could inadvertently breach the confidentiality of the underlying data (for example, by revealing that only one person in a small area has a particular illness). A critical point here is that when a researcher publishes output they are effectively moving data (for output is still data) from a highly secure setting to a completely insecure one. The change in data environment, means that the status of the data can, in principle, change from non-personal to personal (see Mackey and Elliot, 2013, for a review). Hence, disclosure checking of output is a vital part of the governance of secure data access systems.

Traditionally, ‘output statistical disclosure control’ (OSDC) has been managed by specifying rules for researchers to follow; for example, a requirement for all table cells to have at least three observations contributing to that cell. If the table meets the rules it can be released; if not, not. This ethos is reinforced by a half-century of statistical research focused on making tabular outputs safe (Hundepool et al., 2012).

However, in the last ten years it has become clear that simple models for tables have limited value in modern research environments, and increasingly common to discuss ‘output SDC’ as a separate research field1. The complexity of outputs led some data managers (e.g. Ritchie, 2007) to argue that the rules-based approach was both unsafe and inefficient, instead, a ‘principles-based’ approach could be both more secure and more cost-effective.

The principles-based approach uses rules to first-approximate a decision to release or not; but all preliminary decisions are subject to review and change if the researcher or the person responsible for approving the release of output can make the case. The key is that both parties agree on the aims of the SDC process — and one of those aims can be to use the resources of the facility efficiently. The principles-based approach does not accept that outputs can be definitively classified as safe or not, only that the balance of probability says so. Finally, the principles-based approach acknowledges the value of research output in any decision. These differences may seem subtle, but they have profound implications for the way the facility and the researchers are managed.

The principles-based approach requires a higher level of expertise from the managers of research facilities, who must have both technical knowledge and an understanding of the research environment and researcher. In addition, it places the subjective assessment of risk at the forefront of decision-making. These two factors often make facility managers uncomfortable, as such organisations are typically risk-averse (Ritchie, 2014a). In addition, knowledge of the principles-based approach is concentrated amongst a relatively small community, whereas the rules-based model has been the dominant approach for half a century. Hence, facility managers may not be aware that there is an alternative perspective; if they are, they may not appreciate the subtleties of the principles-based approach, preferring instead a simpler model of data security (Ritchie and Welpton, 2014).

This paper aims to help facility managers make decisions about SDC, using the current best understanding of the pros and cons of each of the two process methodologies. For explanatory purposes, it uses the example of deciding on an approach to SDC for a remote RDC. This is because this is the case in which the difference between the two is starkest, and this paper demonstrates that the value of the principles-based approach increases with the sensitivity of the data and with the degree of freedom that researchers have to innovate.

Our conclusion is that for remote RDCs the advantages of principles-based SDC are clear. Beyond this, there are also lessons for other environments. The two approaches are not mutually exclusive: both take simple rules as a starting point, but the rules-based approach also finishes there. This has advantages in some circumstances (for example, rules-based is more appropriate for European Statistical System outputs; Eurostat, 2014), but as the principles-based approach is the generalisation of the rules-based approach, facility managers would do well to consider both.

The paper notes that, although the principles-based model requires greater initial investment by both the facility managers and the researchers, the necessary training can bring substantial auxiliary benefits to the facility manager. Put simply, the necessity to train researchers gives the facility manager an opportunity to encourage other positive behaviours, leading to increased legitimacy and improved researcher behaviour (Ritchie and Welpton, 2014). Again, this is not always feasible, which is why rules-based modelling is sometimes more appropriate. The aim of this paper is to show when these benefits can be realised.

The next section considers the definition, benefits and costs of the rules-based approach, whilst section three evaluates the principles-based approach. Both consider the evidence for claims made. This is particularly important for the principles-based model, which brings risk-assessment to the fore. Section four summarises the discussion, and concludes that the principles-based approach is essential for RDCs, whether remote or not. Section Five reviews implementation issues; Section Six concludes4.

Some definitions are necessary for the paper:

An ‘output’ in the context of this paper is any statistical product arising from use of the data, intended for distribution beyond the technical confines of the research facility. An output may be a table, graph, regression model, frequency count, survival function etc., or it may be a paper containing multiple statistical outputs.

A ‘commentary’ in the context of this paper refers to any discussion about the analysis produced by the researcher. This includes written and verbal discussion.
‘Statistical disclosure control’ (SDC) means techniques to ensure that a statistical product or data does not breach confidentiality guidelines. ‘Input SDC’ (often just referred to as SDC) is concerned with protection of data before researchers have access to it, and is the subject of a different paper. ‘Output-based SDC’ (OSDC) is only concerned with statistics for distribution, and is the focus of this paper.

‘Disclosive’ is used in this paper as a short-hand for output which should not be made generally available as it retains a non-negligible residual risk that an individual population unit could be identified within them; there might be variations in outputs between what is strictly unlawful and what is unwise and undesirable. For the purposes of exposition, we assume that disclosure equals a breach of confidentiality, with legal consequences and/or ethical implications.

The discussion below sits within the ‘five safes’ framework (Desai et al., 2014; see Camden, 2014, or Sullivan, 2011, for examples of use), which is a way of identifying sources of risk in data access:

- **Safe projects** – whether the data use is lawful
- **Safe people** – whether the researchers can be trusted to hold and use the data appropriately
- **Safe settings** – whether the manner of accessing the data offers protection
- **Safe data** – whether there is any inherent protection in the data
- **Safe outputs** – whether the outputs from the research pose a disclosure risk

The final criterion recognises that, however well-intentioned and competent the researcher is, accidents can happen – either through ignorance or through complexity.

This paper only considers the ‘safe outputs’; that is, it is based on the assumption that data access is lawful and that researchers are not deliberately trying to misuse breach data confidentiality. How the researchers access the data is not relevant to this discussion.

Inherent protection in the data is mostly irrelevant to this discussion of general principles and procedures, (although it does have some bearing on training issues and rules-based approaches, to be discussed later). Therefore, this paper does not place any limits on the data used to generate these outputs.

**Rules-based OSDC**

*How it works*

Rules-based OSDC consist of the application of a set of rules to determine whether an output should be released or not. Example rules might be:

- “A table may only be released if there are at least three observations for each cell”
- “A regression may be released if not based entirely on categorical data”
- “A Herfindahl index of over 0.3 should only be released as ‘over 0.3’”
- “Variance-covariance matrices XX may not be released”

These are hard rules; they are expected to be applied consistently. This generates certainty in what output is acceptable, and allows machine-based SDC to be applied.

**Rules-based OSDC**

Rules-based OSDC is simple and transparent. It is popular with data owners as it reflects the guidelines used to create official statistics, which are largely tabular. It is also necessary for the increasing number of automated systems which allow researchers to produce tables and analysis on the fly.

The main advantages of a rules-based model is the certainty and lack of ambiguity. This allows untrained (or partially trained) staff to clear outputs, and requires no training on the part of the researcher. Researchers can be given all the information they need in the form of hand-outs. This, for example, is how researchers at the Eurostat Safe Centre have been advised.

**Criticisms of Rules based OSDC**

There are three main disadvantages with the rules-based approach. All three are a consequence of the inevitable trade-off between confidentiality and efficiency problems. Consider devising a rule for the number of observations that have to be in a cell for it to be released:

- **The Confidentiality Problem**: a low limit increases the probability of disclosive cells being published
- **The Efficiency Problem**: a high limit increases the probability of non-disclosive findings not being published

First, no rule can guarantee non-disclosure, and the application of strict rules can provide a false sense of security. For example, in some cases no amount of units in a cell prevents that cell breaching confidentiality in some way. Hence, the rules based approach may under-protect the data in some cases.

Second, a rules-based approach tends to over-protect the data. Protection will tend to dominate user value: the rule is the only protection against disclosure, and hence has to be much stricter than if other factors (such as the specific data being tabulated) are taken into account. As well as being inefficient, this can also create credibility problems when expert users are asked to follow rules which do not make sense to them.

Third, rules cannot cover all conditions; new rules need to be devised and agreed as new possibilities occur. A proliferation of rules is possible and this can in turn lead to contradictions. Consider defining dominance rule for table cells of N units (that is, determining whether one or two units contribute so much to a cell total that the cell can be considered, for all practical purposes, to only include those units). Rules that have been put forward include:

- the top unit does not account for more than w% of the total of the bottom N-2 records
- the top unit does not account for more than x% of the cell total
- the top two units do not account for more than y% of the cell total
- the Herfindahl index does not exceed z%

Each rule identifies a potentially problematic distribution of data, but the rules will not necessarily agree. Requiring all four rules to be met is over restrictive. Finally, unless the person clearing the output knows how the output was created, it is not possible solely from the output to determine whether the rules have been met or not.
In summary, the simplicity of rules-based OSDC is also its main limitation, particularly when dealing with an expert user base where rules-based OSDC can suffer from credibility problems. The blunt instrument of rules-based OSDC can under-protect in some cases, but is more likely to over-protect. This can cause frustration in researchers, which in turn is one of the factors associated with confidentiality breaches (Desai and Ritchie, 2010).

**Principles-based OSDC (PBOSDC)**

**How it works**

Principles-based OSDC is characterised by:

- researchers and output checkers both trained in SDC
- rules-of-thumb rather than hard rules
- freedom to approve any output in principle
- no duty to release any output
- responsibility for producing good output resting with the researcher
- output checkers considering the value of the output
- output checkers considering resource constraints

PBOSDC starts from the same perspective as a rules-based model: a set of rules exist to guide output approval. The difference is that these now become rules-of-thumb, rather than hard rules. They are there to guide the output-checker, but do not necessarily need to be followed. The output checker has complete freedom to exercise discretion, both to release an output and to decide not to release it. Clearance for release thus becomes a negotiation between researcher and approver. But an unrestricted negotiation is inefficient and so the rules-of-thumb provide the starting point.

It is important that both parties recognise the costs and the benefits of checking an output for clearance. Both parties want outputs to be processed quickly. The researchers want their outputs to be cleared; approvers want to be satisfied that the outputs are non-disclosive. A rejected output imposes costs on both parties; both parties therefore want to avoid this. The key to PBOSDC is that the person best able to assess whether an output should be released is the person who created it. The researcher knows whether the data was sampled, whether there are any dominance issues, how many observations there are in a cell, and so on. Most importantly, the researcher knows the value of the output to his or her research.

If the researcher knows the broad criteria on which output is checked he/she can ensure that (1) the output meets those criteria (2) the output checker has the information necessary to come to the same conclusion quickly and easily. If the output does not meet the prima facie conditions but is of high importance to the researcher, the researcher can try to persuade the output checker that this case is a valid exception to the rules-of-thumb. As this is likely to involve effort on both parties (and the output checker is under no obligation to concede the arguments), the researcher will have to consider whether the output is worth it.

The final part of the system is that the outputs can be rejected not just on the basis of disclosures, but on the basis of whether they are a good use of the output checker's time. It is irrelevant how much time the output checker actually has; the point of this rule is to allow the output checker to reward compliant behaviour and punish the malcontents by setting up appropriate incentives (Welpton and Ritchie, 2011, Ritchie and Welpton, 2012).

Consider a (male) researcher wanting to produce a number of outputs which are (to his mind) non-disclosive, but nevertheless breach the rules-of-thumb. He has three alternatives:

a. Send the outputs through and hope for the best
b. Not send the outputs through
c. Raise the issue with the output checker and try to identify a solution which works for both parties

Option (a) is often a good strategy as a one-off; output checkers should be tolerant of researchers periodically sending through output which requires more work to check. However, as a repeated strategy it risks aggravating the output checker who can delay or stop checking future outputs, particularly if the researcher shows a failure to understand the principles of clearance.

Option (b) is common in practice; researchers working in restricted environment typically produce a large amount of outputs, not all of which is wanted. As a general rule, PBOSDC aligns with good statistical practice; for example, in discouraging very low cell counts.

Option (c) is the most interesting one, and frequently used in restricted facilities using PBOSDC. Researchers learn that a ‘no surprises’ policy appeals to output checkers, who also want an efficient clearance process. This can lead to inventive solutions which work for both parties; for example, validating program code rather than outputs.

Hence, researchers are incentivised to produce good output, and are encouraged to talk to output checkers before problems arise. Output checkers are encouraged to promote good practice amongst researchers, and to listen to user perspectives on the value of certain outputs.

To be most efficient, a PBOSDC process should adopt the ‘safe’/‘unsafe’ statistics dichotomy (Ritchie, 2008; Brandt et al., 2010; Ritchie, 2014b). A ‘safe’ statistic is something which has very little or no inherent disclosure risk, such as regression coefficients. An ‘unsafe’ statistic is one which presumed to present a disclosure risk unless proved otherwise; for example, a simple tabulation. ‘Unsafe’ statistics, being ex ante disclosive, are much more time consuming to check.

Under PBOSDC, researchers can expect that ‘safe’ statistics will be cleared unless the output checker can demonstrate that it is disclose; by definition, there will be almost no instances where this is the case. In contrast, ‘unsafe statistics will not be cleared unless the researcher can demonstrate that there is no disclosure risk. ‘Unsafe’ statistics passed for clearance impose a cost on the researcher related to the output checker’s cost of clearance. The researcher should therefore be incentivised to concentrate on producing outputs made up of ‘safe’ statistics, and because the researcher is aware of how the value of specific outputs, there is an incentive to focus on important results and not large quantities of ‘nice to have’ information. This dynamic needs to be emphasised in the service access training.

Finally, PBOSDC ideally attempts to raise awareness of speculating about the identity of records when commenting on results. For example, although a researcher may produce good statistical outputs, he or she might draw attention to the presence of a particular outlier which affects results. This cannot be dealt with by
a rules-based approach, but only by ensuring that researchers are aware of the risks posed by incautious discussions of data quality. In summary, the aim of PBOSDC is to create an atmosphere where clearance is seen as the joint responsibility of researchers and output checkers. The common interest encourages the production of safe and useful outputs cleared by an efficient process.

However, for this to happen effectively, both need to understand the incentives of the other party, as well as the principles of SDC. Hence, there is a need for training of researchers and of output checkers. This is a major difference compared to the rules-based approach.

Advantages of Principles based OSDC
The direct advantage of PBOSDC is that it should allow both more and safer outputs than the rules-based approach. Consider again the potential errors noted above arising from a threshold rule, the confidentiality problem (too low a limit —→ disclosive outputs) and the efficiency problem (too high a limit —→ safe outputs rejected). Under PBOSDC, there is no conflict: a high limit, likely to remove almost all disclosure risk, is set as the initial rule of thumb. If a researcher feels that this is too high in a particular case, he or she can make that argument, confident that the arbitrary rule of thumb will be replaced by a review of the circumstances in the specific case.

This works because most researchers requiring access to detailed data want it for analytical purposes, which are more likely to be safe statistics. When unsafe statistics are the main focus of output (for example, the OECD commissioned work on high-growth companies which required very many tabulations from the VML), output checkers and researchers have an incentive to work together to agree in advance the range of permissible outputs. The indirect advantage of PBOSDC is the ability to develop a culture of confidentiality awareness amongst researchers. This has positive feedback effects: a demonstrably educated and trustworthy researcher base can have systems designed to reflect that knowledge and trust, and a more appropriate working environment encourages positive behaviour from researchers. Desai and Ritchie, 2010; Ritchie and Welpton, 2014. This has been the pattern of development in UK Research Data Centres over the past decade.

Criticisms of Principles based OSDC
The three main criticisms of PBOSDC are uncertainty, inconsistency, and resource requirements.

PBOSDC, by design, introduces uncertainty into the process, as the whole ethos of PBOSDC is that decisions are taken in specific contexts. Training and ongoing engagement are therefore required to build trust.

If several output checkers are reviewing outputs, they may make different decisions; it is also possible that the same output checker will make apparently inconsistent decisions in different circumstances. This is because the checkers are making decisions responding to the particular circumstance of output, data and researcher and not just a rule about an output. To ameliorate this good training (of both researchers and output checkers) ensures that there will be agreement on the principles. An important part of the process is that is that there is no rules-based yes/no answer; therefore output checkers should be aware that any decision necessarily has an element of subjective judgement and may need to be justified.

It has been argued that the need for output checkers rather than automatic processes (or checking by staff with limited expertise) increase the costs of a facility, as does the explicit allowance for researchers to challenge decisions. While this has not been the case to date in facilities where PBOSDC is fully implemented, it is clear that some expenditure on training for staff and researchers is necessary; when one facility failed to train its new staff, there was an immediate impact on clearance rates and quality.

PBOSDC in practice
In practice, none of the criticisms suggested in the previous section have proved significant. The evidence for this comes from the eight years of PBOSDC at the Virtual Microdata Laboratory (VML) at the UK Office for National Statistics, where the process was developed, as well as more recent experience at SDS and HMRC Data Lab, also in the UK.

Uncertainty does exist. However, there should be no uncertainty about the process or the criteria for deciding whether to release an output or not. The purpose of researcher training is to help researchers understand the uncertainty and manage it. Researchers have made mistakes, and while these are mostly oversights on the part of the researchers, a small number are due to researchers not understanding the principles of SDC. In general these were dealt with by discussion with the researchers, explaining the error. Recidivism rates were negligible, but a very small number of VML researchers were asked to re-attend training (less than five people in seven years, out of over eight hundred trained researchers).

Although there are no figures, the number of requests for output refused at the VML was believed to be around 5%.

Inconsistency exists but there is little evidence to date of it being significant. Over seven years around twenty output checkers were employed at ONS, periodic checks were carried out, and although output checkers showed some slight variance, there was no practical difference in outcomes. It was discovered that one output checker was seen as ‘softer’ by some researchers, but even those outputs were well within the safety margin. In one case, a researcher’s output was seen by five output checkers (including the ‘softer’ checker), all of whom independently gave the same opinion. So whilst differences do exist, in practice these have no notable impact.

If the research facility is not centrally run, or if several facilities are trying to co-ordinate SDC policies, there is another level at which variance may happen: between centres. There may be some cultural variability and the potential for local ethos to develop should be acknowledged and monitored. A common training framework where the principles of SDC are gone into in some depth, practice sharing between centres, “test” submissions and cross-centre case study reviews can help to mitigate this.

The reason inconsistency and uncertainty have not been major problems to date is because of the built-in safety margins. As noted above, ignoring the Efficiency Problem means that the rules of thumb to address the Confidentiality Problem can be made much stricter; a confidentiality breach therefore requires a considerable error by both researchers and output checkers.
This margin of error is also at the heart of the efficiency of the process. Output checkers can clear large amounts of output because they have confidence in the extra-safe rules-of-thumb for “unsafe statistics”, and because “safe statistics” require little scrutiny. At its peak the VML was dealing with 2500 clearance requests a year, or roughly ten every working day; one person was allocated to be output checker for that day, and the target was that this should take no more than half an hour, a target generally achieved. As a ‘clearance request’ typically consisted of several regressions, a couple of descriptive tables (and in one case over fifty graphs), and/or a log file, the actual number of statistical outputs being checked was much more than ten per day. This work rate was maintained by emphasising the researcher’s role in clearance. For example, the VML output checkers had an informal policy of clearing the easiest outputs first; this was communicated to researchers at training sessions. This gave researchers an incentive to build up a reputation of producing ‘good’ (i.e. easy to check) outputs. Finally, there is evidence that the training sessions (and the relationship built up between researchers and output checkers) do foster a culture of confidentiality awareness, with examples of researchers being self-policing. It is unlikely that academics (Brandt et al., 2010), and why researcher training should be seen as an investment rather than an expense. **Rules-based versus principles-based OSDC** The various aspects of the two approaches can be summarised as follows. Although described above as alternatives, there is a relationship between rules-based and principles-based output SDC. The rules act as the starting point for PBOSDC output checking (see Brandt et al., 2010). However, there are two crucial differences: • In PBOSDC, the rules are ‘rules-of-thumb’ – explicitly ad hoc, and amenable to adjustment, up or down, depending on circumstances. • these rules of thumb can then be more restrictive as prima facie efficiency has a low priority in the setting of the default values. For ‘safe statistics’ the ‘hard rules’ and ‘rules-of-thumb’ are the same; by definition, ‘safe statistics’ are those which are amenable to the identification of simple yes/no cases (Ritchie, 2008). Ultimately, PBOSDC takes rules-based SDC as a good first-order approximation, but gives expertise and experience the final decision. For this reason it is the recommended approach for the RDCs. In situations where facility managers have less opportunity to manage researchers’ activities and outputs, PBOSDC may be harder to implement, as the active engagement of researchers is essential. **Implementing PBOSDC** **Conceptual development** The most detailed expression of PBOSDC is the Eurostat-approved guidelines of Brandt et al. (2010). These were derived almost entirely from the VML rules, with the exception of the dominance rules. Since the publication of the Eurostat guidelines there have been a number of minor developments. For example, on regression, several authors (e.g. Reznek and Riggs, 2005; Ronning, 2011; Bleninger et al., 2011) have investigated options for deliberately creating misleading regression results, and US and Australian works have studied regression models in remote-execution systems. Ritchie (2012, 2014b) incorporates most of these results, but new queries appear (for example, on the tabulation of binary variables, and how single-observation categories are treated in regressions) In addition, not all UK work (for example, on variance-covariance matrices) was adopted as it was felt to be too obscure for a general document. 

<table>
<thead>
<tr>
<th>Complex</th>
<th>Rules-based</th>
<th>Principles-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flexible</td>
<td>No</td>
<td>Generally not – rules-of-thumb usually sufficient</td>
</tr>
<tr>
<td>Transparent</td>
<td>Yes</td>
<td>Yes, if output checkers record factors leading to exceptional judgment</td>
</tr>
<tr>
<td>Consistent</td>
<td>Yes</td>
<td>Yes, if output checkers record factors leading to exceptional judgment</td>
</tr>
<tr>
<td>Secure</td>
<td>Limited by efficiency</td>
<td>Yes - able to handle lower and higher risk cases</td>
</tr>
<tr>
<td>Efficiency</td>
<td>Limited by security</td>
<td>Yes - tailored to circumstances</td>
</tr>
<tr>
<td>Risk management</td>
<td>Yes/no model</td>
<td>Explicit ‘balance-of-risks’ model</td>
</tr>
<tr>
<td>Sensitive to context</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Sensitive to user needs</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Suitable for automation</td>
<td>Yes</td>
<td>Only as initial gatekeeper – humans are final arbiters</td>
</tr>
<tr>
<td>Requires researcher training</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Requires researcher engagement</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Requires co-ordination of output checkers</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Cost</td>
<td>Low</td>
<td>High initial cost, low or negative ongoing costs</td>
</tr>
</tbody>
</table>

using the VML, SDS and HMRC Data Lab would see themselves as being particularly SDC aware, as this is their only experience of it’. Nevertheless, a comparison with users of facilities in other countries shows that the UK academia is much better informed. This is why the UK model of researcher training was adopted largely unchanged by Eurostat as recommended best practice. None of these are major challenges but they highlight the need for updating the current state of knowledge and communicating this to facility managers and researchers. When the VML guide to SDC was the only source and was owned by the VML team, updating was straightforward. One of the negative consequences of the wider use of the PBOSDC approach is that there is no clear mechanism for maintaining and disseminating knowledge.
Facilities wishing to implement PBOSDC may therefore need to collaborate with other facilities to ensure that a consistent approach can be developed and applied. It has been suggested that the theoretical basis for PBOSDC and the safe/unsafe statistics model does not provide sufficient reassurance to potential data suppliers. In addition, theoretical models may miss some plausible outcomes that (is, those likely to occur in practice) arising from, for example, naive researchers.

One solution is to set up an 'ethical hacking' mechanism to probe both genuine and fake outputs to test what information could be acquired. A second would be to have periodic audits of outputs from the facility by a competent body, perhaps another facility. This has the added advantage of encouraging consistency across facilities.

**Need for training**
Researcher training is essential for PBOSDC. Untrained researchers cannot be expected to know the basis of SDC, and the training itself should be used to develop an affinity between the researchers and the facility. There are several PBOSDC training manuals/presentations which can be drawn upon and updated. This training could be part of any user accreditation process.

There is already significant training experience in the UK and other countries, and therefore any facility would not need to start from scratch. However, as one of the key elements is to encourage researcher engagement, training needs to be sensitive to the needs and indeed interests of the researcher community; what may appeal to Mexican researchers may be different from the expectation of Norwegians. Training also needs to be sensitive to the data. Much of the extant training focuses on social and business survey data, which are of more limited risk, but administrative data brings additional problems (data may be a census; health data is typically more prone to outliers and retains its sensitivity over time). This is particularly important if tables are likely to comprise a lot of the outputs.

Finally, there is the need to train and update the knowledge of output checkers. Periodic peer review has proved useful in the past. Discussion forums allow knowledge to be shared, discussed and updated across facilities. If made available to researchers, these would also have value for researchers, although not necessarily in the same detail and with a full range of views expressed. One option would be to set up a discussion forum for facility staff, with a summary/FAQ page for researchers.

**Trusted researchers**
One decision to be made is whether some researchers could have different rules applied. Either particular types of people (e.g. full professors) or those who have built up reputations with the output checkers could have fewer checks imposed on them. The argument is that the burden for output checking is unnecessary, and creates ill-will amongst senior researchers who have proven their expertise. Some facilities do use such differentiated models. In practice, these arguments do not hold. Seniority is no guarantee of good practice. Indeed the opposite can be true; experience shows that junior staff are more enthusiastic adopters of safe practices. Similarly, familiarity may make mistakes less likely but does not eliminate them. As mistakes are far and away the most likely reason for failed clearances, long-term usage does not seem sufficient cause on its own to remove checks.

These arguments are also based on the assumption that all outputs are equal. As should be clear from the discussion above, over time researchers should develop a sense of what is and is not allowed, and tailor their outputs accordingly. In other words, the PBOSDC encourages the development of expertise in output assessment by all parties, and thus efficient exchanges. There is no need to create an artificial group of 'good' researchers; besides, creating 'classes' of users could discourage knowledge and experience sharing.

**Malevolent researchers**
Any output disclosure control system will have additional difficulties with a user who deliberately sets out to breach the system. PBOSDC assumes that researchers are well-intentioned and interested in generating good statistical outputs. It may be possible to spot unauthorised outputs, but in practice a user set on breaching procedures would be able to disguise inappropriate outputs (for example by burying discovered data in complex model output). It should be noted that all the data used in current training programmes for controlled environments is wholly invented, yet plausible.

The success of PBOSDC therefore depends upon the 'safe people/safe project' dimensions of the Five Safes model. That said, at present, there are no known examples of academic researchers maliciously breaching confidentiality rules. There are numerous examples of well-intentioned researchers making mistakes, and a smaller number of cases of researchers deliberately ignoring procedures to make life easier for themselves (but again, without intending to disclose confidential data).

Setting up a system which would stand a chance of picking up malicious attacks therefore has no realistic prospect of success, and would increase greatly costs and clearance time. If a facility believed that malicious attack was a significant risk, then examining user logs and codes would be a more useful place to look for malpractice – and would also be necessary for forensic evidence in the event of a disciplinary event.

In summary, PBOSDC cannot stop ill-intentioned researchers; it is designed to deal with cases of accidental rule breaking and errors. If malicious attack is felt to be a problem, then this should be tackled at the 'people' level.

**Multi-stage clearance**
The UK VML operated a two stage-clearance procedure: 'intermediate' outputs could be released to researchers who could work on the further analysis at their home institution. 'Final clearance' was given when papers were ready for general distribution. The UK Secure Data Service only allowed 'final clearance': papers are fully prepared within the SDS virtual facility. These two choices reflect physical differences. The VML involved travel to a specific location; it was not thought to be a good use of restricted facilities to have researchers editing papers, and the opportunity to discuss results with co-researchers was limited. In the SDS these two issues are less important.

Note however that VML applied full PBOSDC at the intermediate stage; the assumption was that once an output had left the VML's control it could end up in the wild however well-intentioned the researcher. This assumption turned out to be correct, even if unsanctioned releases were rare. The VML's 'final clearance' stage imposed a level of super-checking on outputs – asking researchers to limit outputs to the minimum necessary (compared to the
intermediate stage were multiple variants might be tried). As this reflected the research stages (exploratory work, produce many outputs, refine, and then publish) the model worked. The two-stage model did introduce an extra layer of administration, as now both intermediate and final clearances had to be checked and recorded. However, it did speed up the intermediate level clearance, as VML staff knew that if they made a mistake they had a 'second chance' to address it. This increased the already-wide margin for error in the VML PBOsdc procedures.

Given that researchers were only bound by VML procedures (and not required by law) not to publish intermediate outputs, the 100% co-operation (allowing for mistakes) can be seen as a positive reflection of how researchers respond to appropriate training. However, the VML did make researchers aware that failure to follow procedures would affect the way that future access to data was viewed. This may have had more of an effect, and may be of relevance to a facility choosing to adopt two-stage clearance. Allowing intermediate output out implies increasing risk unless adequate mitigation is in place. Relying 100% on trust of researchers without any bounds implies allowing unmeasurable variations in risk and therefore is not sufficient to provide that mitigation. Appropriate risk mitigation implies processes such as:

- Secondary licensing agreements.
- Minimum required security arrangements for intermediate output.
- Specified lists of persons who will have access to the intermediate outputs.
- Occasional random audits.

Summary

This paper recommends PBOsdc be adopted for use in RDCs, whether physical or remote. The reasons for this are twofold: (i) principles based SDC can produce output that is of higher quality at the same or lower level of risk; and (ii) the opportunity of building a relationship with researchers can generate multiple benefits. In short, PBOsdc is both safer and more efficient than rules-based approaches and it encourages the development of a culture of expertise in confidentiality.

There are already guides and training programmes boasting several years’ experience of PBOsdc. A facility wanting to adopt PBOsdc can build upon these, perhaps tailoring them more to its particular researcher group and data. PBOsdc needs to be integrated into a training programme; it assumes that researchers are well-intentioned (if liable to make occasional mistakes). There also needs to be a mechanism to ensure consistency across checkers (and possibility sites). The facility may also want to invest some resources in ethical hacking to provide extra reassurance to data owners. Finally, the facility needs to determine whether it wants a one- or two-stage clearance process. While PBOsdc at the point the output leaves the restricted facility can be the same, the perceived and actual security differs.

For non-RDC environments, the case for PBOsdc is less clear. For example, if researchers’ only sensible interaction with the facility manager is receiving a partially anonymised file on CD plus guidelines for publishing safe statistics, then a rules-based approach may be simpler. However we would recommend that even a discussion of rules should be placed in the context of the principles of SDC: in general, the support officer is more welcomed than the policeman.

References


Ritchie F. and Welpoton R. (2014) “Understanding the human factor in data access”, Working papers in economics no. 1413. University of


Notes
1. Corresponding author: Felix Ritchie, Bristol Business School, University of the West of England Bristol, Coldharbour Lane, Bristol BS16 1QY. Email: felix.ritchie@uwe.ac.uk.
2. Mark Elliot, School of Social Sciences and Data Research Institute, University of Manchester, Oxford Road, Manchester M13 9PL. Email: mark.elliot@manchester.ac.uk
3. There is a large literature on statistical disclosure risk assessment and control methods. We do not discuss this in detail as here we are talking about two top-level process methodologies rather than the specifics of individual technical methods. We would direct the reader who is interested in the detail to recent comprehensive field reviews (Duncan et al. 2011 and Hundepool et al. 2012)
4. A more extended discussion of this argument is available in Ritchie (2007).
5. It is important to stress that “safe” is used here not in its absolute postpositive sense (free from danger or risk) but in its relative sense (the degree to which a solution affords security or protection from risk); see Ritchie (2014b, section 5).
6. PBOSDC is also used at other restricted facilities in Mexico, Germany and the Netherlands, as well as informally in other countries.
7. An exception is medical sciences where data protection has had a much higher profile, and researchers in all countries tend to have a greater awareness of confidentiality issues.
8. In January 2015, all the UK RRDCs initiated a working group on SDC; one of the group’s functions is to determine how guidelines can be effectively maintained, distributed and updated.
9. Note that rules-based OSDC is even more susceptible to malicious attacks, as the yes/no approval process means that anything that looks acceptable will be approved without further checking.
Bridging the Business Data Divide: Insights into Primary and Secondary Data Use by Business Researchers

by Linda D. Lowry

Abstract
Academic librarians and data specialists use a variety of approaches to gain insight into how researcher data needs and practices vary by discipline, including surveys, focus groups, and interviews. Some published studies included small numbers of business school faculty and graduate students in their samples, but provided little, if any, insight into variations within the business discipline. Business researchers employ a variety of research designs and data collection methods and engage in quantitative and qualitative data analysis. The purpose of this paper is to provide deeper insight into primary and secondary data use by business graduate students at one Canadian university based on a content analysis of a corpus of 32 Master of Science in Management theses. This paper explores variations in research designs and data collection methods between and within business subfields (e.g., accounting, finance, operations and information systems, marketing, or organization studies) in order to better understand the extent to which these researchers collect and analyze primary data or secondary data sources, including commercial or open data sources. The results of this analysis will inform the work of data specialists and liaison librarians who provide research data management services for business school researchers.

Keywords: business, primary data, secondary data, graduate students, research data management.

Introduction
A bridge is an apt metaphor for the work of an academic liaison librarian, who acts as a boundary spanner between faculty, students, and the Library. Much of this boundary spanning activity is driven by traditional liaison responsibilities including reference service, information literacy instruction, and collection development. As Canadian academic libraries begin to develop new research data management (RDM) services, liaison librarians have been identified as crucial intermediaries between the library’s services and its researcher community… [who] often have domain-specific expertise and a network of department-specific relationships’ (Steeleworthy, 2014, p.7). Like many of its Canadian peers, Brock University Library has articulated a desire, through its most recent strategic planning exercise, to explore opportunities to support research data management and curation (Brock University Library, 2012). As the liaison librarian to the Goodman School of Business at Brock University, I was quite familiar with the challenges of working with complex, and often expensive, commercial sources of numeric business data such as Compustat and CRSP (Hong & Lowry, 2007), but less familiar with the data practices of business scholars who generated primary data as part of the research life cycle. In order to bridge the business data divide, I needed to acquire evidence-based...
insight into business researchers in their dual roles as data producers and data consumers.

A key phase in the development of RDM services is the discovery phase, which documents and analyses current researcher data practices that may be shaped by a variety of factors such as discipline, funding source requirements, research team composition, and career stage (Whyte, 2014). An independent assessment of the management, business, and finance (MBF) research landscape in Canada, commissioned by the Social Sciences and Humanities Research Council (SSHRC), provides some insight into these factors (Council of Canadian Academies [CCA], 2009a). The number of business faculty in Canada was estimated at just over 2,900 individuals working at 58 different academic institutions (CCA, 2009a, p. 14). Business is diverse discipline comprised of many subfields, some of which are more research-intensive than others. A bibliometric analysis of Canadian MBF research output published between 1997 and 2006 found that the accounting subfield represented 14% of business school faculty but produced only 2% of the research output, while the organizational studies and human resources subfield represented 5% of total business school faculty but produced 11% of the research output (CCA, 2009a, p.22). An analysis of research grants administered by SSHRC between 2005 and 2008 calculated that just 1.7% of these grants went to MBF research (CCA, 2009a, p. 18). The Council of Canadian Academies also examined the level of collaborative activity among MBF researchers and found that: (a) 40% of all papers published between 1996 and 2007 were collaborative; and (b), among the top 25 Canadian universities, 45% of collaborative papers had an international co-author. Data management plans are not currently required for SSHRC-funded research, but researchers who collaborate internationally may find themselves subject to data management and sharing policies required by funding agencies in other countries (Corti et al., 2014).

This study employs content analysis to investigate the research designs and data collection methods found in one form of academic business research output, the master’s thesis, in order to discover to what extent graduate student business researchers collect primary data, or rely on access to secondary data sources for their analysis, and to explore variations within and between business subfields. In order to distinguish between the terms research strategy (which was not considered in this study), research design, and research method, the following definitions were considered:

1 A research strategy refers to a general orientation to the conduct of social research (Bryman et al., 2011, p.579). Commonly cited strategies are qualitative, quantitative, and mixed methods, while other terms used to describe research strategies include strategies of inquiry, traditions of inquiry, or methodologies (Creswell, 2003, p. 13).

2 A research design refers to a framework for the collection and analysis of data’ (Bryman et al., 2011, p. 579). Examples of research designs described in standard accounting, business, and social science research textbooks include experimental, cross-sectional (survey), fieldwork, case study, and archival (secondary analysis) designs (Bryman et al., 2011; Neuman, 2003, and Smith, 2011).

3 A research method can be defined as ‘simply a technique for collecting data’ (Bryman et al., 2011, p. 77) such as self-completion questionnaires, structured interviewing, focus groups, structured observation, ethnography and participant observation, content analysis, and secondary analysis. For consistency’s sake, the term data collection method will be used in this study when discussing research methods.

Reliance on primary data collection has implications for the development of research data management services, while reliance on secondary data has implications for data reference support and collection development planning, particularly due to the high cost, proprietary nature, and complex interfaces of many business data sets. This study sets a baseline measurement for data practices at the master’s level of business research, and can be used in future studies to compare current data practices at other career stages, or at other institutions, at the disciplinary or sub disciplinary level of analysis. This paper is structured as follows: section 2 reviews the literature related to methods of discovering researcher data practices; section 3 describes the purpose of the study and the research questions I will be exploring, section 4 describes the study’s procedures including the setting, and methods of data collection; section 5 presents the findings of the content analysis; section 6 discusses the implications of the findings for research data management, reference support, and collection development; section 7 discusses the limitations of the study; and the final section presents suggestions for future research.

Literature Review

Surveys and Interviews

Academic librarians and data specialists have used a variety of approaches to gain insight into current research data management practices such as case studies (e.g., Key Perspectives, 2010), campus-wide questionnaires (e.g., Parham, Bodnar & Fuchs, 2012), interviews (e.g., Carlson, 2012), and focus groups (e.g., McClure et al., 2014). One study revealed statistically significant differences in research data management practices and attitudes across four research domains (but did not consider discipline-specific distinctions), leading the authors to recommend tailoring data management services using discipline-specific approaches (Akers & Doty, 2013). Another study attempted to examine differences in research data practices by discipline and by methodology but the findings were of limited generalizability due to the low response rate and a survey instrument which conflated research strategies (e.g., qualitative, quantitative, and mixed methods) with research designs (e.g., experimental, survey, field work), and data collection methods (e.g., oral history, textual analysis) (Weller & Monroe-Gulick, 2014).

Motivated by research funding agency requirements for data management plans, most studies have focused on the data curation behaviors and attitudes (such as data preservation and data sharing) of science researchers (e.g., Scaramozzino, Ramirez, & McGaughey, 2012), or if institution-wide, have grouped business scholars with social science or professional schools (e.g., Akers & Doty, 2013), thus providing little, if any, insight into variations within the business discipline. In the next section, I discuss how deeper insight into a researcher’s choice of data collection methods and patterns of secondary data use within a discipline can be acquired by conducting a content analysis of scholarly research publications such as journal articles, theses, and dissertations.

Content Analysis

Content analysis, which is a nonreactive or unobtrusive data collection method, enables researchers to overcome some of the weaknesses of survey research, such as low response rates, sampling errors, or unclear question wording (Neuman, 2003). Several studies provided insight into business research designs.
at the disciplinary level of analysis. Researchers investigating the prevalence of mixed methods research designs in business and management dissertations conducted a systematic content analysis of 186 Doctor of Business Administration theses and confirmed the use of a diverse set research designs and data collection methods (Miller & Cameron, 2011). In a similar study, McLennan, Moyle, and Weiler (2013) explored the role of economics in tourism postgraduate research by conducting a content analysis of 118 doctoral dissertations completed in the United States, Canada, Australia, and New Zealand between 2000 and 2010. Their examination of the frequency of use of specific research approaches methodologies found that 60% of tourism economics theses used quantitative approaches, 21% used qualitative approaches, and 9% used mixed methods approaches, while their analysis of the data collection methods employed identified a diverse range of techniques including interviews, surveys, case studies, econometric forecasting, observation, and econometric modeling (McLennan, Moyle, & Weiler, 2013, p. 186).

Other content analysis studies explored research trends and practices within specific business subfields (e.g., accounting, logistics and supply chain management), thus providing insight into primary and secondary data use at the sub disciplinary level of analysis. An examination of trends in accounting research over 50 year period found that archival research, defined as ‘papers using data from historical market information [such as] stock prices’ (Oler, Oler, & Skousen, 2010, p. 668), has been the dominant research methodology in published accounting papers since the 1980s and comprised more than 60% of all papers published between 2000 and 2007. A review of articles published over a two year span in the Journal of Business Logistics found that 62% of empirically-based studies used primary data from surveys or case studies, while 21% of studies used secondary data methods (Rabinovich & Cheon, 2011). While logistics and supply chain researchers appear to rely less heavily on secondary sources than do accounting researchers, a broader review of recent research in the logistics and supply chain field identified extensive use of secondary data sources for archival data collection, simulation, content analysis, event studies, and meta-analysis, leading Rabinovich and Cheon to advocate for the extension of traditional secondary data methods to include logistics research.

Several studies conducted by librarians also illustrate the value of the content analysis method in uncovering discipline-specific data practices. Nicholson and Bennett (2009) explored the nature of primary and secondary data use and availability within business ethics research through a content analysis of 48 doctoral dissertations. Their analysis revealed that 51% of the dissertations contained only primary data, 12% relied exclusively on secondary data, and 32% collected both primary and secondary data. A review of primary data collection methods identified four main categories: observations, surveys, experiments, and structured interviews, while a review of the secondary data collected identified a range of data types including numeric datasets, corporate annual reports, government filings and regulatory cases (Nicholson & Bennett, 2009). More recently, Williams (2013) analysed the content of 124 journal articles published by 64 faculty members in crop sciences for evidence of data usage and data sharing, in order to identify faculty candidates for data services. An advantage of the bibliographic study (sic) approach was that it revealed a diversity of discipline-specific data practices, but it was time consuming to conduct, because if data sets were used, they were typically not cited in the bibliography, but within the text of the article (Williams, 2013, p. 207).

In summary, content analysis is an unobtrusive discovery method which can provide insight into the prevalence of various research designs and data collection methods in order to determine patterns of primary and secondary data use within specific disciplines, but few studies have examined variations between or within subfields of business. This study attempts to fill that gap by reporting on the findings of an exploratory content analysis study of business master’s theses.

**Purpose**

The purpose of this study was to investigate the research designs and data collection methods of students in a research-based Master of Science in Management (MSCM) program in order to better understand the extent to which these researchers collected and analyzed primary or secondary data. A content analysis of a corpus of 32 master’s theses explored differences between and within subfields of business with respect to research designs and data collection methods. In cases where a thesis used secondary data, attempts were made to identify whether the data sources could be considered open data, or commercial data. This study explored the following research questions:

1. What is the distribution of theses by area of specialization and how does it compare to the distribution of core (supervisory) faculty?
2. What is the overall distribution of theses by research design and by data collection method? What are the patterns of data collection method use within each type of research design?
3. What is the distribution of research designs and data collection methods by area of specialization?
4. What is the overall nature of primary and secondary data collection and use (across all specializations)?
5. What types of secondary sources are used in business research?
   Do these researchers use open data sources, proprietary/commercial data sources, or both?

**Procedures**

**Setting**

Brock University is a large comprehensive university located in Canada which offers a wide variety of undergraduate and graduate programs across seven faculties. Brock University’s Goodman School of Business (GSB) is accredited by AACSB International and has undergraduate and graduate degree programs in accounting and business administration, an enrollment of 2500 FTE students, and a faculty complement of 95 (Brock University Institutional Analysis & Planning, 2014). The GSB launched a research-based Master of Science in Management program during the 2007/2008 academic year with two goals in mind: first, to prepare students to conduct research in industry and government settings, and second, to prepare students for doctoral level studies in business (Brock University, 2007). The MSCM is a two year program which culminates in a thesis based on independent and original research, and currently offers specializations in accounting, finance, operations and information systems management (O & ISM), formerly known as management science), marketing, and organization studies. The organization studies stream was first offered during the 2010/2011 academic year (Brock University, 2010). Although Brock University does not currently offer a doctoral level degree in Business, at one point in time the GSB’s medium to long term plan included the development of a
research-based doctoral degree in Business, perhaps jointly with another university (Brock University Faculty of Business, 2005).

**Method**
In order to better understand the extent to which business student researchers collect and analyze primary or secondary data, I conducted a systematic content analysis of a corpus of 32 Master of Science in Management theses which were deposited in Brock University Library’s Digital Repository. Master’s theses must be published in the digital repository as a graduation requirement, so this sample represented 100% of the MSCM degrees awarded since the inception of the program. Each thesis was hand coded using a hybrid approach of manifest and latent coding, similar to the approach taken by Nicholson and Bennett (2009), in order to identify the business subfield, research design, and data collection method employed, and the extent and nature of secondary data use (see Appendix A). The full text of each thesis was reviewed, with particular attention paid to the title page, acknowledgements, abstract, table of contents, methods, and data sections.

Brock University’s MSCM program offers five subfields (referred to as areas of specialization): which are (a) accounting, (b) finance, (c) O & ISM, (d) marketing, and (e) organization studies. If the area of specialization was not specifically stated on the title page, a code was assigned based on the topic of the theses, and the home subject area of the student’s thesis advisor (who was often cited in the acknowledgements).

Each thesis was coded according to the choice of research design and data collection method and the coding form allowed for the possible use of more than one research design and data collection method, as might be the case in a mixed method research strategy. A core list of research designs and data collection methods was compiled after a review of accounting, business, and social science research methods textbooks (see Appendices B and C).

Finally, each thesis was analyzed for evidence of secondary data use. Each secondary data source was identified by name and by type (i.e., open or commercial or proprietary). Further investigation was required in some cases to determine if a secondary source was a commercial or an open source.

**Findings**

**Distribution of Theses by Area of Specialization**

Table 1 presents a comparison of the distribution of theses and core faculty by area of specialization. The largest proportion of theses came from the finance area, followed by the marketing area. The proportions for these two areas were larger than one might expect, based on the distribution of core faculty by area of specialization as currently listed on the program’s website (Brock University Goodman School of Business, 2015). Three of the five subject area specializations were under-represented (when compared to the distribution of core faculty) including: accounting, O & ISM, and organization studies. The differences in proportions might be a result of several factors such as the relative newness of the MSCM program, the growth of the program over time, and variations in student interest in each of the specialized streams. According to the Appraisal Brief for the MSCM program (Brock University Faculty of Business, 2005), at the time the degree program was proposed there were 21 core faculty distributed across four areas of specialization: (a) eight faculty in accounting, (b) five faculty in finance, (c); five faculty in management science, and (d) three faculty in marketing. Given the two year length of the program, and the fact that the organization studies specialization was not added until the 2010-2011 academic year, it is not as surprising to have just two theses completed in organization studies. The 2014-2015 Graduate Calendar notes that the specialized streams may not be offered every year if there is insufficient student interest (Brock University, 2014).

**Distribution of Theses by Research Design and Data Collection Method**

Three types of research designs were employed in MSCM theses: archival/secondary analysis, survey, and experimental (see Figure 1). There were no examples of case study designs, and none of the theses employed more than one research design. The analysis of data collection method use, as shown in Figure 2, noted three different types of data gathering methods: archival-empirical/quantitative, questionnaires, and archival-content analysis. Patterns of data collection method use within each type of research design appear in Table 2. Both examples of theses with experimental designs used questionnaires for data collection, as did all seven of the theses with survey research designs. Of the 23 theses which employed the archival/secondary analysis research design, only one engaged in a qualitative content analysis, while the other 22 engaged in the empirical analysis of quantitative data. None of the theses employed more than one type of method for the collection of data.

| Table 1 Distribution of Theses and Core Faculty by Area of Specialization |
|-----------------------------|-----------------|-----------------|------------------|
| Area of Specialization      | Theses (%)      | Core Faculty (%)| Over or Under-Represented |
| Accounting                  | 5 (16%)         | 13 (25%)        | Under            |
| Finance                     | 15 (47%)        | 8 (15%)         | Over             |
| O & ISM                     | 3 (9%)          | 8 (15%)         | Under            |
| Marketing                   | 7 (22%)         | 7 (13%)         | Over             |
| Organization Studies        | 2 (6%)          | 16 (30%)        | Under            |
| Total                       | 32 (100%)       | 52 (100%)       |                  |

Figure 1 Overall patterns of research design use in MSCM theses (N=32).
Distributions of Research Designs and Data Collection Methods by Area of Specialization

The distribution of research designs and data collection methods by area of specialization are presented in Table 3 and Table 4. The archival/secondary analysis research design was employed at least once within each area of specialization, but was most heavily used within the finance and accounting specializations. Survey designs were employed in four of the five areas of specialization, while experimental designs were used in two of the marketing theses. The marketing area exhibited the widest variety of research designs, while finance used only one type of research design. An analysis of data collection method use by area of specialization revealed widespread use of the archival – empirical/quantitative method, with evidence of use within four of the five areas of specialization.

In order to make sense of the patterns of research design and data collection method use by area of specialization, I also examined the course descriptions for each area of specialization in the MSCM program. Students in all specializations except finance take a two course research methodology sequence which covers topics such as: multivariate statistical techniques, advanced regression analysis, measurement and scaling, survey research and questionnaire design, sampling methods, qualitative research, and structural equation modeling (Brock University, 2014). Students in the finance specialization take a two course sequence in empirical finance which covers empirical research methods and econometric techniques in investment finance (Brock University, 2014). Students in the accounting stream also take additional courses which cover accounting theory and research methods in behavioural accounting research and market-based research, while the O & ISM specialization includes courses on modeling, data mining, mathematical programming, simulation, and forecasting. Looking again at Table 3 and Table 4, patterns of use begin to emerge, with finance, accounting, and O & ISM theses favouring archival designs and quantitative analysis methods, while marketing and organization studies theses used a variety of research designs and data collection methods. Insights from a case study of an MSc program in finance in the United Kingdom confirmed that students were exposed to secondary data and regression analysis as the model to follow in their own research (Belghitar & Belghitar, 2010, p.578).

### Primary and secondary data collection

This study also explored the nature of primary and secondary data collection and use across all areas of specialization, and within each specialization. Table 5 presents the patterns of primary and secondary data collection across all areas of specialization. MSCM theses showed a greater reliance on secondary data sources, less reliance on primary data collection, and no evidence of combining primary and secondary data collection, when compared to the Nicholson and Bennett (2009) analysis of business ethics dissertations. Primary data collection methods were used in four of the five areas of specialization, and secondary data sources were used in all five areas of specialization (see Table 6). The Finance area relied exclusively on secondary data sources, as did the majority of

---

**Table 2**

Patterns of Data Collection Method Use by Research Design

<table>
<thead>
<tr>
<th>Research Design</th>
<th>Questionnaire</th>
<th>Archival – Content Analysis</th>
<th>Archival – Quantitative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental</td>
<td>2 (100%)</td>
<td>0 (0%)</td>
<td>Under</td>
<td>2 (100%)</td>
</tr>
<tr>
<td>Survey</td>
<td>7 (100%)</td>
<td>0 (0%)</td>
<td>Over</td>
<td>7 (100%)</td>
</tr>
<tr>
<td>Archival</td>
<td>0 (0%)</td>
<td>1 (4.3%)</td>
<td>Under</td>
<td>23 (100%)</td>
</tr>
<tr>
<td>Total</td>
<td>9 (28%)</td>
<td>1 (3.15)</td>
<td>22 (68.7%)</td>
<td>32 (100%)</td>
</tr>
</tbody>
</table>

**Table 3**

Research Design Use by Area of Specialization, Percentage of Row Totals

<table>
<thead>
<tr>
<th>Specialization</th>
<th>Experimental</th>
<th>Survey</th>
<th>Case</th>
<th>Archival</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accounting</td>
<td>0 (0%)</td>
<td>1 (20%)</td>
<td>0 (0%)</td>
<td>4 (80%)</td>
<td>5 (100%)</td>
</tr>
<tr>
<td>Finance</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>15 (100%)</td>
<td>15 (100%)</td>
</tr>
<tr>
<td>O &amp; ISM</td>
<td>0 (0%)</td>
<td>1 (33.3%)</td>
<td>0 (0%)</td>
<td>2 (66.6%)</td>
<td>3 (100%)</td>
</tr>
<tr>
<td>Marketing</td>
<td>2 (28.5%)</td>
<td>4 (57.1%)</td>
<td>0 (0%)</td>
<td>1 (14.2%)</td>
<td>7 (100%)</td>
</tr>
<tr>
<td>Organ. Studies</td>
<td>0 (0%)</td>
<td>1 (50%)</td>
<td>0 (0%)</td>
<td>1 (50%)</td>
<td>2 (100%)</td>
</tr>
<tr>
<td>Total</td>
<td>2 (6.2%)</td>
<td>7 (21.8%)</td>
<td>0 (0%)</td>
<td>23 (71.8%)</td>
<td>32 (100%)</td>
</tr>
</tbody>
</table>

**Table 4**

Data Collection Method Use by Area of Specialization, Percentage of Row Totals

<table>
<thead>
<tr>
<th>Specialization</th>
<th>Questionnaire</th>
<th>Archival – Content Analysis</th>
<th>Archival – Quantitative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accounting</td>
<td>1(20%)</td>
<td>0 (0%)</td>
<td>4 (80%)</td>
<td>5 (100%)</td>
</tr>
<tr>
<td>Finance</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>15 (100%)</td>
<td>15 (100%)</td>
</tr>
<tr>
<td>O &amp; ISM</td>
<td>1 (33.3%)</td>
<td>0 (0%)</td>
<td>2 (66.6%)</td>
<td>3 (100%)</td>
</tr>
<tr>
<td>Marketing</td>
<td>6 (85.7%)</td>
<td>0 (0%)</td>
<td>1 (14.2%)</td>
<td>7 (100%)</td>
</tr>
<tr>
<td>Organ. Studies</td>
<td>1 (50%)</td>
<td>1 (50%)</td>
<td>0 (0%)</td>
<td>2 (100%)</td>
</tr>
<tr>
<td>Total</td>
<td>9 (28%)</td>
<td>1 (3.1%)</td>
<td>22 (68.7%)</td>
<td>32 (100%)</td>
</tr>
</tbody>
</table>
This study investigated both the nature and types of secondary data sources used and revealed that very few theses relied exclusively on open data sources. 28% of theses used commercial data sources exclusively, and 37% of theses used a combination of open and commercial data sources (see Table 5). Many theses used multiple secondary data sources, including four theses which used five or more secondary data sources (Table 5). A detailed listing of these open and commercial secondary data sources, including source type, example names, and frequency of use, is presented in Appendix D. The secondary sources included many of the data types identified in the Nicholson and Bennett (2009) study, but given this study’s Canadian location, it was not surprising to find some Canadian equivalents, such as SEDAR (for corporate financial reports and filings) and CANSIM (for socioeconomic data). Some, but not all, of the theses that used secondary sources relied on commercial numeric datasets hosted by the Library or Business School (e.g., Bloomberg Professional, CFMRC, Compustat, CRSP, Datastream). Other theses used datasets such as the Internet Retailer’s Top 500 Guide, which may have been purchased directly the student or provided by the student’s faculty advisor.

**Table 5**

<table>
<thead>
<tr>
<th>Description</th>
<th>Number ( %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theses with primary data collection</td>
<td>9 (28%)</td>
</tr>
<tr>
<td>Theses with secondary data collection</td>
<td>23 (72%)</td>
</tr>
<tr>
<td>• Collected only open data sources</td>
<td>2 (6%)</td>
</tr>
<tr>
<td>• Collected only commercial data sources</td>
<td>9 (28%)</td>
</tr>
<tr>
<td>• Collected both open and commercial sources</td>
<td>12 (37%)</td>
</tr>
<tr>
<td>Theses with both primary and secondary data collection</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>

**Table 6**

<table>
<thead>
<tr>
<th>Number of secondary data sources used</th>
<th>Number ( %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Source</td>
<td>3 (13%)</td>
</tr>
<tr>
<td>2 Sources</td>
<td>6 (26%)</td>
</tr>
<tr>
<td>3 Sources</td>
<td>3 (13%)</td>
</tr>
<tr>
<td>4 Sources</td>
<td>4 (17%)</td>
</tr>
<tr>
<td>5 Sources or more</td>
<td>4 (17%)</td>
</tr>
<tr>
<td>Total</td>
<td>23 (100%)</td>
</tr>
</tbody>
</table>

Secondary sources types used in business research

This study investigated both the nature and types of secondary sources used and revealed that very few theses relied exclusively on open data sources. 28% of theses used commercial data sources exclusively, and 37% of theses used a combination of open and commercial data sources (see Table 5). Many theses used multiple secondary data sources, including four theses which used five or more secondary data sources (Table 5). A detailed listing of these open and commercial secondary data sources, including source type, example names, and frequency of use, is presented in Appendix D. The secondary sources included many of the data types identified in the Nicholson and Bennett (2009) study, but given this study’s Canadian location, it was not surprising to find some Canadian equivalents, such as SEDAR (for corporate financial reports and filings) and CANSIM (for socioeconomic data). Some, but not all, of the theses that used secondary sources relied on commercial numeric datasets hosted by the Library or Business School (e.g., Bloomberg Professional, CFMRC, Compustat, CRSP, Datastream). Other theses used datasets such as the Internet Retailer’s Top 500 Guide, which may have been purchased directly the student or provided by the student’s faculty advisor.

**Discussion**

Implications for research data management

Libraries planning research data services may use a life cycle model to describe the real-world activities of their researchers (Carlson, 2014). The research data lifecycle consists of the following data related activities: discovery and planning, data collection, data processing and analysis, publishing and sharing, long term management, and reusing data (Corti et al., 2014 p. 17). Viewed through the lens of the research data lifecycle, this study’s analysis of business master’s theses identified variations between and within business subfields with respect to research design and data collection method use which have implications for the development of research data management services for business researchers.

Over 70% of graduate student business researchers could be categorized as data consumers, while less than 30% could be considered data producers. Archival / secondary analysis research designs were employed at least once within each business subfield, and constituted the majority of theses in the accounting, finance, and operations and information systems management subfields. The discovery and acquisition of secondary data sources are crucial real-world activities for these business data consumers. A minority of theses employed survey or experimental research designs and collected research data via questionnaires. Due to the small sample size, clear sub disciplinary patterns could not be found, but the researchers collecting primary data in this study were more likely to be from the marketing subfield. The real-world data activities of these business data producers, such as planning data collection protocols and obtaining informed consent, closely align with social science researchers employing survey or experimental research designs. This segment of researchers could be targeted for future research data management services.

Unlike our counterparts in the United States and the United Kingdom, there is less urgency with respect to developing research data management plans in Canada due to a lack of public funding agency data sharing mandates. The Tri-Agency Open Access Policy on Publication, which applies to peer-reviewed journal publications, was announced on February 27, 2015, but the sharing of publication-related research data is only required by Canadian Institutes of Health Research funding recipients (Tri-Agency, 2015). Barriers to publishing and sharing business data exist due to the
proprietary nature of many of the data sets used in the accounting and financial research which rely on econometric methods. While some top economic journals have mandatory data availability policies which require authors to submit the datasets used in their research, most journals do allow exemptions for research based on proprietary or confidential sources such as Thomson Reuters Datastream (Vlaeminck, 2012, 2013).

**Implications for reference support and collection development**
Liaison librarians and data specialists often provide data reference support to novice researchers such as undergraduate students (Partio, 2010). According to the ‘Seven Ages of Research’ model, master’s degree students (who are in the first age of the researcher’s lifecycle) are engaged in research for a limited period of time and may not be seeking a career in academia (Bent, Gannon-Leary, & Webb, 2007, p. 85). Researchers in this early career stage often turn to thesis supervisors for guidance on the research process, academic writing, and information retrieval, to varying degrees of satisfaction (Bent, Gannon-Leary, & Webb, 2007, p. 89). Librarians serving business schools with research-based master’s degree programs with an accounting or finance emphasis, where students are expected to engaged in archival quantitative research, may want to direct their energy toward providing higher levels of support for secondary data discovery and extraction, and then promoting this service to faculty and graduate students (e.g., Kellam, 2011).

Liaison librarians and data specialists who develop data collections to support academic programs in their institutions will need to work closely with disciplinary faculty in accounting, finance, or other subfields that rely heavily on secondary data, to identify key data sources. Thesis and dissertation content analysis can provide much needed evidence of student usage, and strengthen the argument that these databases support the curriculum (e.g., when writing the library or data services support portions of program appraisal briefs or reaccreditation reviews). A content analysis of faculty journal publications can uncover hidden data sources (e.g., owned or licensed by faculty members for use by their own research groups, but not listed on the Library website or a Departmental resources web page), which may be missed by traditional database benchmarking efforts that examine such listings. One must also keep in mind that traditional citation analysis studies which only look at bibliographies (for a review, see Hoffman & Doucette, 2012), will drastically underestimate the use of secondary data sources which are often only cited within the methods sections of dissertations and journal articles. In times of Library materials budget cuts, expensive subscription-based financial databases with high cost per use may be prime targets for cancellation, but are fundamental to the research process for financial scholars.

**Limitations**
There were three aspects of this study which limited the generalization of the findings to the broader population of academic business researchers. The first limitation was due to the small sample size which, while representative of 100% of the population of Brock University’s MSCM students, did not reflect the sub disciplinary distribution of core faculty supervisors in the MSCM program. The second limitation was due to the focus of the study on only one career stage of researchers. While a graduate student’s choice of research designs and data collection methods may mirror the standard protocols within a discipline, the short duration of master’s degree programs may limit his or her choices (see Belghitar & Belghitar, 2010, p. 579), and therefore may not be representative of research designs employed within doctoral theses or studies published by faculty in peer-reviewed journal articles. The third limitation was due to the use of a single coder, rather than multiple coders. However, the results were strengthened by the nature of the coding scheme, which relied primarily on manifest coding, and focused on measuring the frequency of occurrence of each variable (Neuman, 2003). This exploratory study served as a pilot project to test the thesis coding scheme, which could easily be adapted for broader use at other institutions by using a broader list of business subfields, such as the list of 16 subfields employed by the Council of Canadian Academies in their bibliometric analysis (CCA, 2009b, p. 11).

**Conclusion**
This study used the content analysis method to provide insight into primary and secondary data use by master’s level business students at one institution. The content analysis found variations in the choice of research designs and data collection methods use across the business discipline, as well as variations within business subfields. The results of this exploratory study can serve as a benchmark for future discipline-specific content analysis studies, and the content analysis method can be used to examine variations in data practices within other social science disciplines. The study could be extended to examine business research from multiple academic institutions, multiple career stages (e.g., doctoral students, faculty, and postdoctoral researchers), or a variety of types of research output including doctoral dissertations and peer-reviewed journal articles. The use of multiple coders would allow for the examination of a larger and more representative sample of the broader business research landscape. Given the Canadian Association of Research Libraries’ interest in developing a collaborative, networked approach to building a research data management infrastructure (for a discussion see Canadian Association of Research Libraries, 2013), a collaborative approach to conducting future discipline-specific content analysis studies is recommended.

**References**


Notes
1 Linda D. Lowry is the business and economics liaison librarian at Brock University in St. Catharines, Ontario, Canada. She can be reached by email: llowry@brocku.ca

2 https://dr.library.brocku.ca/
## Appendix A

### Content Analysis Coding Form

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area of Specialization (circle one)</td>
<td>Accounting (1); Finance (2); Operations and Information Systems Management (3); Marketing (4); Organization Studies (5)</td>
</tr>
<tr>
<td>Research Design (circle all that apply)</td>
<td>Experimental (1); Survey (2); Case study/Field Research (3); Archival / Secondary Analysis (5); Other (specify) (6)</td>
</tr>
<tr>
<td>Data Collection Method (circle all that apply)</td>
<td>Questionnaire (1); Structured interview (2); Focus group (3); Ethnography/observation (4); Archival-content analysis (5); Archival-empirical/quantitative (6); Other (specify) (7)</td>
</tr>
<tr>
<td>Type of Secondary Data Collected</td>
<td>Open only (specify sources or names of data sets) (1); Proprietary only (specify sources or names of data sets) (2); Both (3) (specify sources or names of data sets); Not applicable (primary data only) (4)</td>
</tr>
</tbody>
</table>
## Appendix B

### Descriptions of Research Designs

<table>
<thead>
<tr>
<th>Research Design</th>
<th>Description</th>
<th>Examples from textbooks</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experimental</strong></td>
<td>Includes classical experimental designs (random assignment, pretest, post-test, experimental group, control group) or quasi-experimental designs (Neuman, 2003, p. 247).</td>
<td>Experimental (Bryman et al., 2011; Smith, 2011; Neuman, 2003).</td>
</tr>
<tr>
<td><strong>Survey</strong></td>
<td>“Quantitative social research in which one systematically asks many people the same questions, then records and analyzes their answers” (Neuman, 2003, p. 546).</td>
<td>Cross-sectional or Social Survey (Bryman et al., 2011); Survey (Smith, 2011; Neuman, 2003).</td>
</tr>
<tr>
<td><strong>Case Study / Field Research</strong></td>
<td>Case study “entails the detailed and intensive analysis of a single case” (Bryman et al., 2011, p. 571); Field research is “a type of qualitative research in which a researcher directly observes the people being studied in a natural setting for an extended period” (Neuman, 2003, p. 535).</td>
<td>Case Study (Bryman et al., 2011); Fieldwork (Smith, 2011); Field Research (Neuman, 2003).</td>
</tr>
<tr>
<td><strong>Archival / Secondary Analysis</strong></td>
<td>Archival: research using secondary sources such as historical documents, texts, journal articles, corporate annual reports, and company disclosures to conduct time-series, cross-section data analysis, content analysis or critical analysis (Smith, 2011); Secondary Analysis: “research in which one does not gather data oneself, but re-examines data previously gathered by someone else and asks new questions” (Neuman, 2003, p. 544).</td>
<td>Archival (Smith, 2011); Secondary Analysis (Neuman, 2003).</td>
</tr>
</tbody>
</table>
Appendix C

### Descriptions of Data Collection Methods

<table>
<thead>
<tr>
<th>Data Collection Method</th>
<th>Description</th>
<th>Examples from textbooks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Questionnaire</td>
<td>“A collection of questions administered to respondents” (Bryman et al., 2011, p. 579).</td>
<td>Self-completion questionnaires (Bryman et al., 2011); Mail and online surveys (Smith 2011); Mail and self-administered questionnaires (Neuman, 2003).</td>
</tr>
<tr>
<td>Structured Interview</td>
<td>“A research interview in which all respondents are asked exactly the same questions in the same order with the aid of a formal interview schedule” (Bryman et al., 2011, p. 581).</td>
<td>Structured interviewing (Bryman et al., 2011); Interviews (Smith, 2011); Telephone and face-to-face interviews (Neuman, 2003)</td>
</tr>
<tr>
<td>Focus Group</td>
<td>“A form of group interview in which: there are several participants; there is an emphasis in the questioning on a particular fairly tightly defined topic; and the emphasis is upon interaction with the group and the joint construction of meaning” (Bryman et al., 2011, p. 575).</td>
<td>Focus groups (Bryman et al., 2011; Neuman, 2003).</td>
</tr>
<tr>
<td>Ethnography/Observation</td>
<td>A composite category which includes ethnography (immersion in a social setting) and structured observation (observing and recording behaviour) (Bryman, et al., p. 574, 581).</td>
<td>Structured observation, ethnography &amp; participant observation (Bryman et al., 2011); Complete participant, complete observer, participant-observer (Smith, 2011); Nonreactive / unobtrusive observation; Ethnography (Neuman, 2003).</td>
</tr>
<tr>
<td>Archival – Content Analysis</td>
<td>“a systematic analysis of texts (which may be printed or visual) to determine the presence, association, and meaning of images, words, phrases, concepts, and/or themes (Bryman et al., p. 375)</td>
<td>Content analysis (Bryman et al., 2011; Smith, 2011; Neuman, 2003)</td>
</tr>
<tr>
<td>Archival – Empirical / Quantitative</td>
<td>The empirical analysis of secondary data sources (such as cross-sectional or time-series data).</td>
<td>Secondary analysis (Bryman et al., 2011; Neuman, 2003); Archival (econometric analysis of cross-sectional or time series data) (Smith, 2011).</td>
</tr>
</tbody>
</table>
### Appendix D

**Open and Commercial Secondary Data Sources Used in MSCM Theses**

<table>
<thead>
<tr>
<th>Classification</th>
<th>Data Type</th>
<th>Examples</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Open Data Sources</strong></td>
<td>Canadian Public Company &amp; Mutual Fund filings</td>
<td>SEDAR</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Socioeconomic Data</td>
<td>CANSIM; FRED</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Stock Exchange websites</td>
<td>FTSE; ME; NASDAQ</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Canadian regulatory filings</td>
<td>IIROC</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>US Government websites</td>
<td>EPA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Bankruptcy Cases</td>
<td>UCLA-LoPucki Bankruptcy Research Database</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Other academic datasets</td>
<td>Hasbrouck’s Liquidity Estimates</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Organizational websites</td>
<td>Ontario Winery websites</td>
<td>1</td>
</tr>
<tr>
<td><strong>Commercial / Proprietary Sources</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Library subscriptions</td>
<td>Numeric Databases</td>
<td>Compustat,</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Datastream,</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CFMRC,</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bloomberg,</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CRSP,</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Execucomp,</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fundata,</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Risk Metrics,</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TRACE,</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IBES,</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Carbon Disclosure Project</td>
<td>1</td>
</tr>
<tr>
<td>Library subscriptions</td>
<td>Bibliographic Databases and Full Text Publications</td>
<td>Lexis-Nexus</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TSX E-Review</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CBCA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hoover’s,</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Law Source,</td>
<td>1</td>
</tr>
<tr>
<td>Other (not Library-hosted)</td>
<td>Ecommerce data; web search analytics, financial trading and ownership data</td>
<td>Internet Retailer Top 500 Guide</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SpyFu</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Keyword Spy</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Econoday,</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Espeed,</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Global Hysales,</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13f Spectrum</td>
<td>1</td>
</tr>
</tbody>
</table>
Improving the Quality of Digital Preservation Using Metrics

by Mari Kleemola

Abstract
The Finnish Social Science Data Archive (FSD) is dedicated to preserving digital research data in the social sciences and to providing good quality services to the research community. Since the research data landscape is perpetually changing, digital repositories need to stay alert and be ready to review their policies and procedures, continuously. In FSD’s case, it is critical that our operations and procedures are up-to-date and consistent with relevant standards and best practices, and that our stakeholders trust us.

In this paper we outline FSD’s venture into the world of digital preservation standards and assessments. We started by exploring the key standard for long term preservation of digital data, the Open Archival Information System (OAIS). We then proceeded to conduct a self-assessment within the framework of the Audit and Certification of Trustworthy Digital Repositories (TDR) Checklist, followed by the CESSDA Trust Process which resulted in FSD’s successful application for the Data Seal of Approval certification.

The process has been fruitful and we can safely say that in the case of using metrics to improve quality, it is the journey that matters as much, if not more, than the destination.

Keywords: digital preservation, certification, OAIS, trusted digital repositories, assessment, DSA

The Finnish Social Science Data Archive FSD
The Finnish Social Science Data Archive (FSD) is a national resource centre that promotes open access to research data as well as transparency, accumulation and efficient reuse of scientific research data. FSD was established in 1999 to archive and disseminate digital, quantitative data for social science research, teaching and learning. Over the years, it has broadened its services to include qualitative dataarchiving as well as provision of guidance on research ethics and data management.

In many ways the turn of the century proved to be an excellent time to set up a national data archive. In the late 1990s the Internet was already a daily working tool for social scientists, and standards like the Data Documentation Initiative were emerging. Most importantly, the international social science data archiving community was well established and networked, having become a part of the research scene in the 1950s and 1960s. The various social science data archives were actually the very first institutions to handle and preserve digital material (Doorn and Tjalsma 2007). Our new archive was privileged to learn from the practices and experiences of the pioneers of digital data archiving.

Early on we realised that preservation, management and dissemination of datasets as well as building high-quality knowledge-based services needed to be done in a transparent way, at the most appropriate time possible and without over-burdening researchers. It was also clear that for the outcome to be successful, many organisational issues and pieces needed to be in place, including policies, procedures and sustainable resources. Consequently, we paid attention to documenting our processes and procedures from the
very beginning: our first Internal Handbook was created during the first two years of operation and our first Archives Formation Plan, or AWS, emerged in 2003. The AWS is FSD’s highest-ranking document concerning provision of data services. It is based on guidelines set by the National Archives of Finland and it describes tasks and processes, selection criteria, preservation periods and forms, confidentiality and legislative issues, responsibilities, and data systems and security.

Nowadays, FSD is an acknowledged and active national centre of expertise in the areas of preserving and providing access to digital research data. FSD’s core user community consists of researchers, teachers and students from Finland and abroad. Most FSD’s services, such as the Data Management Guide, are openly and freely available via our website and in 2014 the number of successful web page requests reached 1.2 million. Access to FSD’s data holdings is provided via the Aila Data Portal. In June 2015, Aila contained 1200 datasets and had 1300 registered users. FSD is funded by the Ministry of Education and Culture and operates as a separate unit at the University of Tampere. FSD is Finland’s Service Provider for the pan-European Research Infrastructure, CESSDA⁷.

Measuring Trustworthiness of Digital Preservation

Hedstrom (1998) defines digital preservation as ‘the planning, resource allocation, and application of preservation methods and technologies necessary to ensure that digital information of continuing value remains accessible and usable’. Digital preservation is therefore an active, and in the optimal case, even a pro-active process that does not include extended periods of inactivity. Digital preservation is also something that needs to be done presently for the future; we need to think about a time period long enough to be concerned with the impacts of changing technologies or a changing user community (CCSDS 2009, Giaretta 2011).

The demands these definitions make are powerful and illustrate well the difficulties in providing long-term access to digital data. Preserving digital data is a challenge. For the outcome to be successful, many organisational and practical issues need to be in place. Key aspects in demonstrating trustworthiness include: transparency, documentation, adequacy and information security. In addition, one has to keep in mind that one needs to evaluate trust into the future. (Dobratz et al. 2010; Giaretta 2011.)

For quite some time it has been evident that methods are needed to assess the trustworthiness of a digital repository. In 1996, the Commission on Preservation and Access and the Research Libraries Group called for a certification programme for repositories claiming to serve an archival function, and since then several stakeholders have explored certification issues (see Dobratz et al. 2010). The need to be able to test a repository’s claims about digital preservation was also one of the key drivers of the Open Archival Information System (OAIS) Reference model (Giaretta 2011, 461).

In the European data archive world assessment and trust issues came up roughly ten years ago. In 2006, The Council of European Social Science Data Archives (CESSDA) Research Infrastructure was identified as an existing pan-European RI recommended for a major upgrade by the European Strategy Forum on Research Infrastructures Roadmap. As a direct result of this, CESSDA launched the Preparatory Phase Project (PPP) in 2008. The project recommended, amongst other things, that adherence to standards should be a part of CESSDA ERIC’s membership criteria since common use of standards is necessary for compatibility. As possible tools, the project brought forward the OAIS model in particular, and the DSA criteria. (Dusa et al. 2010.)

The OAIS model (ISO14721:2003) is the international key standard for archival systems and for organisations engaged in long-term preservation of digital data (see, for example, Spence 2006; Lavoie 2004; Giaretta 2011). Prior to the CESSDA PPP project, the OAIS model had been explored by the UK Data Archive (Beedham et al. 2005) and the ICPSR (Vardigan and Whiteman 2007). A Finnish version of the standard was published in March 2010 by the Finnish Standards Association SFS⁸.

The OAIS model introduced important concepts and paved the way to a certification standard for digital repositories. OAIS conformance is necessary for trustworthiness but not sufficient since the OAIS model is very general and does not cover, for example, financial aspects. Metrics were therefore needed. The Trustworthy Repositories Audit and Certification Checklist (TRAC) was released in 2007, and its revised version, entitled the Trusted Digital Repository (TDR) Checklist, in 2011. The Checklist derives from the OAIS, and the ISO 16363:2012 standard⁹ is based on it. The first edition of the Data Seal of Approval, DSA, emerged in 2008. It was initially developed for use in the Netherlands, but already in 2009 an international DSA Board was established. The DSA contains 16 guidelines, and it is the first step in the three-level European Framework for Audit and Certification of Digital Repositories⁰ that was developed in 2010. Other certification initiatives include, for example, the Nestor Catalogue of Criteria (DIN 31644) and the DRAMBORA toolkit (for an overview of key guidelines and frameworks, see Kvalheim et al. 2013).

All these developments prompted us to take a critical look at our functions at FSD. We wanted to find out if they were adequate and up-to-date and hoped to find ways to rationalise our processes. We also wished to demonstrate that we can be trusted by our key stakeholders: data depositors, users and funders. Furthermore, an important goal was to ascertain that FSD would able to fulfill the CESSDA membership criteria. To achieve all this, we decided to use suitable standards and metrics to assess our systems, policies and procedures.

Step One: Start by OAIS

Our first step was to familiarise ourselves with the OAIS model in 2010. An OAIS compliant organisation has to support the information model of the standard and fulfil the six minimum requirements. Thus, the organisation must:

- negotiate for and accept appropriate information from information producers,
- obtain sufficient control of the information provided to the level needed to ensure long-term preservation,
- determine, either by itself or in conjunction with other parties, which communities should become the designated community and, therefore, should be able to understand the information provided,
- ensure that the information to be preserved is independently understandable to the designated community,
- follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, and which enable the information to be disseminated as
The OAIS compliance was expected for many reasons. First, the results of our OAIS exercise were encouraging. However, the OAIS standard also contains a functional model consisting of six main entities: ingest, archival storage, data management, administration, preservation planning, and access. Additionally, an organisation has to provide common services, such as various technical support services. Exploring these functional entities in detail provided us with a new perspective about our processes. All the entities were identifiable in FSD’s operations and luckily we were not able to find any serious defects. This in turn resulted in a strengthened conception that we had been doing the right things in the right way. However, it became clear that our processes could be further clarified and improved and that we should, for example, collect and store more extensive preservation information and to better structure it.

The OAIS compliance was expected for many reasons. First, FSD has been established to function as a repository for digital data, and also FSD’s operational model has been based on the examples of well-established social science data archives. All in all, FSD emphasises functions slightly differently in comparison to the OAIS model. While the OAIS describes ingest only briefly and practically omits acquisition, they are central processes in the FSD. On the other hand, the OAIS presentation of administration and preservation planning is more complicated than FSD’s procedures, mainly because FSD is still a relatively small archive. As FSD grows, it must be prepared to adopt more sophisticated administrative processes.

Our experiences about OAIS were very similar to those reported by other data archives, for example by the UK Data Archive (Beedham et al. 2005) and the ICPSR (Vardigan and Whiteman 2007). Also GESIS has carried out a mapping of the OAIS functional model to GESIS’s operations (Schumann and Recker 2012).

**Step Two: TDR Checklist**

The results of our OAIS exercise were encouraging. However, the OAIS model is very high-level and thus did not provide enough concrete details to support the assessment of our day-to-day practices, which we thought would help us ensure the quality of our practices and processes and consequently the quality of our data services. Therefore, in spring 2012, we continued our assessment exercise with the help of the TDR Checklist. This decision was influenced by the UK Data Archive’s draft audit against the emerging ISO 16363 standard (see Woollard, 2011). At the time of our self-assessment, the TDR Checklist was at its final stages of approval as ISO 16363 and was available as CCSDS 652.0-M-1 Magenta Book (CCSDS 2011). We used both the Magenta Book and the preliminary version of the TDR Checklist in Excel format provided by the Center for Research Libraries. The Checklist is designed to cover all the aspects necessary to demonstrate that an organisation can be trusted by its stakeholders. The main sections include organisational infrastructure, digital object management and infrastructure, and security risk management.

For each of the 100+ TDR clauses, we identified and described the written evidence and assessed our compliance using the scale of 0 (not compliant) to 3 (fully compliant). We found FSD to be well compliant (3) or almost compliant (2) with 66% of the clauses and not compliant (0) or only somewhat compliant (1) with 10% of the clauses. Six clauses were thought to be out of scope for FSD. In 18% of the clauses we could not make a straightforward assessment because we did not understand them completely. This is probably at least partly due to the fact that in many cases FSD relies on a manual process and a set of systems whereas the TDR Checklist describes a larger and more automated system. Our estimation is that FSD would be not compliant (0) or only somewhat compliant (1) with most of the “unclear” criteria. However, not all criteria have the same importance or level of risk.

Examples of FSD’s full compliance:
- mission statement exists
- collection policy exists
- preservation policies exists
- short and long term business planning in place
- appropriate deposit agreements exist
- minimum information requirements specified
- minimum descriptive information captured

Examples of FSD’s non-compliance:
- no formal succession plan
- no commitment to regular self-assessment or external certification
- no documented process for testing for an understanding of the AIP Content Information
- no systematic analysis of security risk factors

The good news was that we found no major failures or risks. However, the analysis revealed several weak points. For example, early on in the process it became evident that FSD needed to clarify several processes and to especially improve the documentation that describes the technical infrastructure and security risk management. As a result of the self-audit, FSD made an action plan for various changes and improvements. Some minor adjustments were made immediately and major changes became subject to discussions and have been – or will be – implemented gradually.

Our motivation for the self-assessment was to improve our practices and to plan for the development of our processes. For this, the TDR Checklist worked very well, although the learning curve was rather steep and understanding what constitutes adequate evidence was sometimes difficult. The self-assessment was also rather time-consuming and therefore we decided not to go too deep into the clauses and evidence. Since the goal of this TDR self-assessment was not a certification we did not analyze in detail which of the non-conformities would be acceptable and which were not. Our Checklist assessment is also in draft state (for example, some texts are in English, some in Finnish) and while it is sufficient for internal use, it has not been published. So even though the self-assessment provided us with a lot of insight into
our processes and helped to recognise and correct deficiencies, due to the lack of transparency it is not very useful for building trust with our key stakeholders.

**Step Three: CESSDA Trust Process**

In 2013, CESSDA archives carried out a two-phase self-assessment exercise also known as the ‘CESSDA Trust Process’ that aimed to progress the CESSDA archives in the area of Trusted Digital Repository status. As CESSDA is on its way to becoming an European Research Infrastructure (ERIC), all its Service Providers need to meet the membership criteria.

In the CESSDA Trust Process, the Data Seal of Approval (DSA) was selected as a reference point because it is the base-level in the European Framework for Audit and Certification of Digital Repositories. The 16 guidelines of the DSA allow for verifying quality aspects concerning the creation, storage, use and reuse of digital data, and they can be seen as a minimum set distilled from proposals like Nestor, Drambora and TDR. The guidelines focus on three stakeholder groups: the data producers, the data repositories, and the data consumers. A repository is designated a Trusted Digital Repository if it complies to the ten guidelines for repositories and if it enables data producers and data consumers to comply with their three guidelines (Data Seal of Approval 2014).

At the beginning of the CESSDA Trust Process, the DSA criteria were mapped to the CESSDA member obligations. After the mapping, all the European data archives undertook a DSA self-audit. The process was coordinated by an Expert Panel consisting of four people, representing the British UK Data Archive, the Dutch DANS, the German GESIS and the FSD. The Expert Panel counselled the archives during the self-audit, reviewed all self-assessments and provided feedback and a gap analysis. All in all, European data archives appeared to have good practices in terms of, for instance, data reusability. Long-term preservation of data is also well managed, although the documentation of practices was inadequate across many data archives.

During the CESSDA Trust Process, FSD gained valuable knowledge about the DSA as well as of certification in general. Discussions with other data archives clarified the meaning of the guidelines and helped to understand and identify the relevant evidence ie. the documentation that was needed to demonstrate that we meet the DSA criteria. Feedback from the self-assessment and reviews led us to further improve our documentation.

**Step four: DSA Application**

The CESSDA Trust Process confirmed that FSD was in a good position to apply for the DSA. In May 2014 we implemented major changes in our data services that took us from paper-based data ordering to an advanced online data download system, so we decided to postpone our DSA application until August 2014. This ensured that we had time to revise both our internal and external documentation to reflect the changed service model. Before we sent our application, we also translated some key documents, like our Archives Formation Plan (AMS), into English.

Our actual DSA application process was very straightforward and not time-consuming. It took only a day. This somewhat contrasts with the GESIS experience (Schumann 2012) and reflects the fact that we already had many pieces in place following the OAIS, TDR and CESSDA Trust exercises. For example, we had all the necessary documentation readily available to support our assertions, most of it in both Finnish and English. It is also worth mentioning that FSD had perceived good documentation as a cornerstone of its operations from the beginning so our documentation is the result of years of collective, continuous, and innovative work by FSD experts.

Our documentation consists of three main components: our website, the Archives Formation Plan (AMS) and the Internal Manual. FSD’s website contains detailed information on archived data, archiving and ordering data, and on managing research data. The Archives Formation Plan includes, for example, our preservation plan, ingest criteria, and data protection practices, as well as the Internal Manual contains very detailed practical guidelines for all processes. The extensive and well-maintained documentation ensures, for example, that if any two data managers would process the same data according to the instructions, the resulting Archival Information Packages would be substantially similar.

On September 23rd, 2014 we were awarded the 2013 DSA Certificate, and are planning to renew it regularly since we view it as a tool that can be used in preparing for the changes and challenges we will inevitably face. FSD was the first Finnish organisation to acquire the DSA and we have received several inquiries about our certification. We have been very pleased that, for its own part, FSD’s trust process has raised awareness about trust and long-term preservation of digital data in Finland.

**Conclusion**

Over the last few years, FSD has successfully used metrics to improve the quality of its processes and operations. As was to be expected, FSD conforms to the OAIS model. The TDR exercise confirmed that FSD is doing things in the right way, the CESSDA Trust Process allowed us to compare FSD with other archives, and the DSA certification increased trust with stakeholders.

The use of models and metrics to assess our procedures and policies have raised our awareness about the challenges of digital preservation, revealed existing and possible problems and weaknesses as well as strengths, steered and initiated minor and major changes in our operations, and resulted in improved documentation. As a consequence, many of our processes are now better and more efficient or, they will be better – some of the bigger changes will take time to implement. We are also able to better manage risks, provide more trustworthy services for the research community, and demonstrate FSD’s trustworthiness to our stakeholders. In addition, we are in a good position to meet the CESSDA membership criteria.

Our journey into the world of standards and measurement has so far taken four years, and we see metrics such as the TDR and DSA as tools to be used continually for improving processes and services. The actual intensive work related to OAIS, TDR and DSA has required in total about six person months’ worth of effort which we see as resources well spent. We are also considering taking the next step in the European Framework for Audit and Certification, which is a structured, externally reviewed and publicly available self-audit based on ISO 16363 or DIN 31644. The research landscape and thus the research data landscape is changing rapidly. As a digital repository we need to stay alert, be ready to review our policies and procedures methodically and critically, and build and share our competence continuously.
**Funding Note**

In order to be successful in any project, commitment from management as well as sufficient resources are needed. The work described in this paper has been part of the FSD Upgrade and VERIC projects, both funded by the Academy of Finland and both aimed at strengthening the Finnish national service provision as part of the CESSDA ERIC process.

**References**


**Notes**

1. Mari Kleemola is Information Services Manager at the Finnish Social Science Data Archive. She can be reached by email: mari.kleemola@uta.fi.
4. The Finnish version is called SFS 5972 Viitetallit pitkäaikaissyöryksiköille.
5. CCSDS 652.0-M-1 -- Audit and Certification of Trustworthy Digital Repositories (September 2011) contains the final draft standard submitted to ISO for review and approval and is freely available from the CCSDS website as a Recommend Practice document: http://public.ccsds.org/publications/archive/652x0m1.pdf [3.6.2015]
Embracing the ‘Data Revolution’: Opportunities and Challenges for Research

The 42nd annual conference of the International Association for Social Science Information Services and Technology (IASSIST) will be hosted by the Norwegian Social Science Data Services in Bergen, Norway from May 31-June 3, 2016.
The International Association for Social Science Information Service and Technology (IASSIST) is an international association of individuals who are engaged in the acquisition, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, database librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompasses hard copy as well as machine readable data.

Paid-up members enjoy voting rights benefit from reduced fees for attendance at regional and international conferences sponsored by IASSIST. Join today by filling in our online application:

http://www.iaassistdata.info/