Privacy and Confidentiality in Service Science and Big Data Analytics

Christine M. O'keefe

Privacy and Confidentiality in Service Science and Big Data Analytics

Christine M O'Keefe

CSIRO Digital Productivity Flagship,
GPO Box 664, Canberra ACT 2601, AUSTRALIA
Christine.OKeefe@csiro.au
http://www.csiro.au/people/Christine.OKeefe

Abstract. Vast amounts of data are now being collected from census and surveys, scientific research, instruments, observation of consumer and internet activities, and sensors of many kinds. These data hold a wealth of information, however there is a risk that personal privacy will not be protected when they are accessed and used.
This paper provides an overview of current and emerging approaches to balancing use and analysis of data with confidentiality protection in the research use of data, where the need for privacy protection is widely-recognised. These approaches were generally developed in the context of national statistical agencies and other data custodians releasing social and survey data for research, but are increasingly being adapted in the context of the globalisation of our information society. As examples, the paper contributes to a discussion of some of the issues regarding confidentiality in the service science and big data analytics contexts.

Keywords: Privacy, Statistical Disclosure Control, De-identification

1 Introduction

The future internet in the age of globalisation is turning the so-called *information super-highway* into an *information super-mountain* of data. The Internet of Things continues to grow and touch every aspect of our lives, and every interaction generates a digital record, leading to vast data archives accumulating in repositories everywhere. Put together, these data repositories reveal more and more details about ourselves, our behaviours, and our preferences. On the one hand, these detailed data hold a wealth of information vital to informed decision making, research, services personalisation, and debate within governments and the community. On the other hand, there is a risk that personal privacy will not be protected, where privacy is understood as the interest an individual has in controlling the dissemination of information about themselves.

In this paper, we focus on the use of data archives, irrespective of how they have been established, populated and maintained, and on methods for assuring confidentiality of the people or organisations represented in the data. Such methods are called *statistical disclosure control* methods, since they seek to reduce or

control the risk of disclosure from statistical analysis. To provide a full solution, they must be implemented with an appropriate governance framework and with appropriate information security processes. The methods we will present and discuss were developed in the context of a national statistical agency making census and survey data available for research. Confidentiality remains a major concern for national statistical agencies [11, 19], as well as for a broader range of agencies and organisations which now find themselves holding significant data archives and receiving access requests from researchers.

Thus, this paper aims to contribute to the investigation of what technologies, frameworks, and tools we might need to gain, regain and maintain informational self-determination and lifelong privacy while still extracting useful information from our growing data archives. This would be in addition to the minimum standard required by applicable privacy, data protection, and related legislation.

2 Preliminaries

In this section we describe preliminary notions regarding confidentiality and privacy, where confidentiality is *a status accorded to information about a person* [11, Section 1.1], and: *A disclosure occurs when a person or organisation recognises or learns something that they did not know already about another person or organisation, via released data* [19].

2.1 Types of data

Microdata refers to datasets in which each record is contributed by an individual in the population, so that the record typically comprises values of a number of variables for that individual. A variable can be either continuous or categorical, where a continuous variable value is numeric and a categorical variable value is a category label.

Tabular data result when microdata are summarised and presented as a table with axes corresponding to explanatory variables and cells corresponding to a response variable. Table cells can contain counts, where each data record contributes 1 to its tabulation cells and 0 to all other cells, in which case the data is called *tabular count data* and the table is called a *contingency table.* Table cells can also contain aggregates of one response variable, for example the total or average value of that variable for individuals contributing to that cell, in which case the data is called *magnitude data.*

2.2 Types of disclosure

There are two basic types of disclosure, namely identity and attribute disclosure [10], resulting from a data release. An *Identity Disclosure* occurs if an individual is identifiable from the data release. An *Attribute Disclosure* occurs when the released data make it possible to infer the characteristics of an individual more accurately than would have otherwise been possible.

The main ways that an identity disclosure can occur are:

- Release of identifying information
- Spontaneous recognition - where an individual is sufficiently unusual in a data collection, or the data user knows sufficiently many attributes of an individual, so that individual can be recognised from normally non-identifying attributes. This may occur if the attributes have extreme values such as extreme old age or an unusual combination of attributes. For example, it is generally accepted that households have distinctive patterns of inhabitants and other features that make them vulnerable to spontaneous recognition.
- Matching to another data base - where combinations of so-called key variables in the data occur in other databases sufficiently rarely that data matching reveals identity.

Attribute disclosure is usually achieved through identity disclosure; an individual is first identified through some combination of variables and then disclosure of values of other variables included in the released data follows.

2.3 Balancing disclosure risk with data utility

The balance between protecting confidentiality and allowing the use of data for research has been represented as a trade off between disclosure risk and data utility [10], where *disclosure risk* attempts to capture the probability of a disclosure of sensitive information, while *data utility* attempts to capture some measure of the usefulness of the released data. Confidentiality methods are technical approaches designed to reduce disclosure risk, and are applied in addition to governance and information security measures. Unfortunately, any confidentiality method will also reduce data utility.

The idea of balancing risk and utility advanced by Duncan et al. [10] is that in a specific situation, the data custodian creates a *Risk-Utility (or R-U) Map* as a two-dimensional plot of disclosure risk versus data utility for various parameter instances of a range of confidentiality methods, and chooses the method and parameter instance with the maximum utility given a maximum tolerable risk.

3 Approaches for protecting confidentiality in data

In this section we provide a structured overview of a broad range of approaches for protecting confidentiality in data archives. Many of these appear in the literature, and are described in [11, 19]. The structure of our overview depends at a high level on the system design, the type of method, and the type of data [30].

3.1 Types of methods

In the remainder of this Section we provide a structured overview of approaches to reducing disclosure risk when making data available for research. Importantly, each approach only addresses the disclosure risk inherent in the data, and so each must be implemented within an appropriate legislative and policy environment

and governance structure, and with user community management and IT security, including user authentication, access control, system audit and follow-up. The approaches have different strengths and weaknesses, and so none dominates the others in all data access scenarios. In fact, because there is a range of scenarios, it is desirable to have a range of disclosure risk reduction approaches, so the most appropriate one can be chosen to meet the requirements of a particular situation involving a particular dataset, data custodian, analyst and so on.

Traditionally, there have been two different general approaches with regard to enabling the use of data while protecting confidentiality [12]:

- *restricted or limited access*, wherein the access to the information is restricted; and
- *restricted or limited information*, wherein the amount or format of the information released is restricted.

Often these two approaches are used in combination, such as when access to data is restricted to approved analysts and the data themselves have had identifying information removed and/or dates aggregated to months or years. The relationship between the degree of access restriction and the degree of information restriction required is perhaps best represented in the framework of Marsh et al [26], who noted that a successful disclosure involves first an attempt at disclosure, then success of that attempt. In probabilistic terms, this is $\Pr(\text{disclosure}) = \Pr(\text{attempt}) \cdot \Pr(\text{disclosure} \mid \text{attempt})$. Restricted access seeks to reduce $\Pr(\text{attempt})$ while restricted data seeks to reduce $\Pr(\text{success} \mid \text{attempt})$.

3.2 Restricted access methods

In this Section we discuss various data access strategies used to restrict access to information, noting that they are predominantly implemented as system designs. We present these in generally increasing order of restriction, so that the degree of requirement for information restriction generally decreases correspondingly. At opposite ends of this spectrum are the familiar data access strategies of providing no information to non-authorised users, and full information to fully authorised users.

User agreements for offsite use Under this approach, sometimes called *Licensing*, users are required to register with a custodian agency, and sign a user agreement, before receiving data to be analysed offsite. Typically such agreements specify restrictions on the user, such as, restrictions on the manner of storage and further dissemination of the data, as well as prohibiting attempts to re-identify data records. Such agreements also typically specify sanctions for breaches, and are legally binding. The user community is managed by the custodian, including, possibly, the use of external audits to verify compliance with the restrictions in the agreement.

Examples of this approach include the many Public Use Files disseminated by organisations and agencies, including national statistical and health agencies, see [3, 8, 27].

Remote analysis systems In remote analysis, the analyst submits statistical queries through an interface, analyses are carried out on the original data in a secure environment, and the user then receives the (confidentialised) results of the analyses [16, 44]. In particular, the analyst does not receive any data at all, but only analysis results. Since analysis results can reveal information about the underlying data, the output needs to be confidentialised.

The Australian Bureau of Statistics Remote Access Data Laboratory (RADL) is a secure online data query service that clients can access via the Australian Bureau of Statistics web site [2]. The Australian Bureau of Statistics has recently developed the TableBuilder and DataAnalyser remote analysis systems with automated confidentiality routines that allow uses to build their own custom tables or undertake regression analyses on secured ABS microdata [46]. The Microdata Analysis System under development by the U.S. Census Bureau will allow users to receive certain statistical analyses of Census Bureau data, including regression analyses, without ever having access to the data themselves [23].

We remark that remote analysis systems need to be protected against attacks including massively repeated queries, subsetting to create very small datasets and never-ending loops. Recently-developed systems do not allow user-submitted code but rather implement a menu-driven interface to prevent these and other types of attack.

Virtual data centres Virtual data centres are similar to remote analysis systems, except that the user has full access to the data [31], and are similar to on-site data centres except that access is over a secure link on the internet from the researcher's institution.

An example of a virtual data centre is the US NORC Data Enclave, that provides a confidential, protected environment within which authorised social science researchers can access sensitive microdata remotely [49]. Another interesting example is the Australian Population Health Research Network Secure Unified Research Environment [41], see [29]. Similar systems include the United Kingdom Office For National Statistics (ONS) Virtual Microdata Laboratory [28], and the UK Secure Data Service, that provides secure remote access to data operated by the Economic and Social Data Service [47].

Secure on-site data centres Many national statistical agencies allow researchers access to confidential data in secure, on-site research data centres. Usually the data have undergone a confidentialisation process such as de-identification and some light statistical disclosure control, but have more detail than datasets confidentialised for release to researchers. Analysts are generally not restricted in the analyses they can perform and the intermediate results they can generate

and view. However, only results which have been checked to ensure low disclosure risk, or which have been confidentialised if necessary to reduce disclosure risk, can be removed from the laboratory. Currently this output checking is done manually, as in the guidelines in [19].

Examples of on-site data centres include the U.S. Census Bureau Research Data Centers (RDC) [48] and the Australian Bureau of Statistics (ABS) On-site Data Laboratory [5].

3.3 Restricted information methods - microdata

Restricted information methods normally comprise the application of some statistical disclosure limitation techniques, see [11, 19]. Statistical disclosure control techniques can be perturbative or non-perturbative. Perturbative methods operate by modifying the data values, whereas non-perturbative methods do not modify the data values. Perhaps the most well-known perturbative method is the addition of random "noise" to a dataset, and perhaps the most well-known non-perturbative method is suppression of sensitive values.

In this section we describe the main techniques developed for microdata. The methods are presented in the order of generally increasing restriction on released information, so decreasing disclosure risk, from removal of identifying information to synthetic data. The amount of trust in the analyst therefore generally decreases across the methods, and so access restrictions may also be able to be relaxed across the methods.

3.3.1 Removal of identifying information

Probably the most common method of reducing disclosure risk in data sets is to remove identifying information such as name, address, date of birth, and unique identifiers such as social security number or healthcare identifier. This is often called *de-identification*.

As examples, the Population Health Research Network [33] will enable existing Australian health data to be brought together and made available for health and health related research purposes under protocols that use linkage keys to replace personal information in health records. Similarly, the University of British Columbia Centre for Health Services and Policy Research [6] is the central access point for researchers wishing to obtain and use health data in de-identified format for research in the public interest.

3.3.2 Non-perturbative methods for microdata

Suppression of variables or variable values Entire sensitive variables, such as name of surgeon in clinical data, can be suppressed. It is also possible to suppress certain values of categorical individual variables, where such a value is sufficiently unusual that it leads to unacceptable risk of disclosure via matching.

Variable Recoding A widely-used method for reducing disclosure risk is *variable recoding,* or *coarsening,* that can be either part of the data collection design phase or applied to the resulting dataset. The method can be applied to either tabular data or microdata, and can be applied to any number of the variables.

Variable recoding usually involves reporting the values of the variable with less than full detail. For example, geographic information such as address can be recoded to suburb or postal code area, age can be recoded to 5-year or 10-year intervals, and age over a certain threshold can be recoded to simply being over that threshold.

Sampling Disclosure risk can depend on the existence of microdata records that are both unique in the sample and in the population on a set of potentially identifying cross-classified key variables, since such records can be matched to external datasets with high confidence [42].

3.3.3 Perturbative methods for microdata

Rounding Original variable values are replaced by rounded values rounded to multiples of a given number such as 3 or 5.

Data swapping Data swapping transforms a database by interchanging values of sensitive variables between records in a microdata file.

Additive or multiplicative noise addition Randomly-distributed noise values can be added to the data, or the data can be multiplied by randomly-distributed values. Additive noise can be uncorrelated or correlated, and can be augmented with a linear or non-linear transformation.

Micro-aggregation Micro-aggregation is applied by clustering records into small groups of similar records and replacing individual record values by the cluster average values.

Post-randomisation method (PRAM) The Post-Randomisation Method technique is applied to categorical data and involves a form of intended misclassification using a known and pre-set probabilistic mechanism. Under PRAM, for each record in a microdata file, the value of one or more categorical variables is changed with a certain probability.

Synthetic data Rubin [39] suggested the approach of generating and releasing (fully) *synthetic data*, see also [22] and [35]. In the generalisation to partially synthetic data, the data custodian releases a dataset comprising the original records with some observed values replaced with multiple imputations drawn from distributions designed to preserve important relationships in the confidential data, or from models generated by amachine learning technique.

3.3.4 Examples of restricted information approaches on microdata

Internationally, IPUMS-International is a project dedicated to collecting and distributing census data from around the world [27]. IPUMS-International works

with each country's statistical office to minimise the risk of disclosure of respondent information. The details of the confidentiality protections vary across countries, but in all cases, names and detailed geographic information are suppressed and top-codes are imposed on variables such as income that might identify specific persons. In addition, IPUMS-International uses a variety of technical procedures to enhance confidentiality protection, including:

- Swapping an undisclosed fraction of records from one administrative district to another to make positive identification of individuals impossible.
- Randomizing the placement of households within districts to disguise the order in which individuals were enumerated or the data processed.
- Aggregating codes of sensitive characteristics (e.g., grouping together very small ethnic categories)
- Top- and bottom-coding continuous variables to prevent identification of extreme cases.

There are several examples of partially synthetic datasets already released for research. For example, the US Bureau of the Census has released a partially synthetic, public use file for the Survey of Income and Program Participation including imputed value of Social Security benefits information and dozens of other highly sensitive variables [1]. More recently, a synthetic public use file for the U.S. Longitudinal Business Database, an annual economic census of U.S. establishments, has been approved for release by the U.S. Bureau of the Census and the Internal Revenue Service [21].

3.4 Restricted information methods - tabular data

Although tabular data are aggregated, there still may be unacceptable disclosure risk. Perhaps the most common disclosure risk is associated with a cell of a table that relates to only one individual, where an identity disclosure may occur by data matching from the characteristics in the table.

Restricted information methods normally comprise the application of some statistical disclosure limitation techniques for tabular data, see [11, 19]. As in the case of microdata, tabular statistical disclosure control techniques can be perturbative or non-perturbative. There are two main classes of confidentiality methods for tabular data, namely, pre-tabular and post-tabular. *Pre-tabular* methods modify microdata before aggregation into a table, while *post-tabular* methods modify a table directly.

Pre-tabular methods Perhaps the most widely-used pre-tabular method is table redesign, including collapsing of categories along any axis. In fact, any of the methods presented in Section 3.3 could be used as a pre-tabular confidentiality method.

Post-tabular methods In post-tabular Statistical Disclosure Control for tabular data, the first task is to determine whether any of the cells are sensitive, where a sensitive cell is one for which the release of the data in the cell could lead to a disclosure. The most commonly-used cell sensitivity tests are:

- *Threshold rule* - a cell is sensitive if less than n individuals contribute to its value (frequency and magnitude tables)
- (n, k) *-rule* - a cell is sensitive if less than n individuals contribute at least $k\%$ of its value (magnitude tables)[9]

For a discussion of the shortcomings of these techniques, see [37].

After the sensitive cells in a table have been identified, the second task is to take steps to address the disclosure risk. The most commonly-used techniques are:

- *Deletion of variables* - removing table axes corresponding to sensitive variables and/or variables that lead to sensitive cells.
- *Variable recoding* - adjusting the level of aggregation of variables to reduce the number of sensitive cells. This method aggregates all cells involving the recoded variable, whether sensitive or not.
- *Cell collapsing* - merging pairs of cells until no sensitive cell remains. This method only aggregates the sensitive cells, but can make analysis more difficult.
- *Cell suppression* - suppression of the entry in each sensitive cell, then suppression of entries in non-sensitive cells sufficient to prevent reconstruction of any sensitive value.
- *Rounding* - rounding all cells to a multiple of a chosen positive integer, for example, 3 or 5.
- *Addition of noise* - altering sensitive cell values (and usually also non-sensitive cell values) by the addition of noise sampled from some distribution. Examples of this method include the *Post-Randomization Method* [17] and the key-based method in [25].

The Australian Bureau of Statistics Census TableBuilder [4] is an online tool that allows users to create confidentialised, custom tables of Census data from variables including age, education, housing, income, transport, religion, ethnicity, occupation, family composition and more for all ABS geographic areas [46].

3.5 Analysis output confidentialisation methods

Currently virtual data centres rely on manual checking for confidentiality protection, such as those outlined in [19]. This solution may not be feasible in the long term given the trend of rising user demands for data access. Although it is acknowledged that developing valid output checking processes that are automated is an open research question [11], there have been some recent advances in such methods for remote analysis systems, see [16, 23, 34, 36, 43, 44, 46].

Remote analysis systems now in development in the US Census Bureau and the Australian Bureau of Statistics do not rely on restricted output methods

alone, but also make use of a combination of protective measures from the restricted access, restricted data, restricted analysis and restricted output groups of methods.

4 Confidentiality in Service Science and Big Data Analytics

In this paper, our aim is to contribute to the investigation of what technologies, frameworks, and tools we might need to gain, regain and maintain informational self-determination and lifelong privacy while still extracting useful information from our growing data archives. We do this by means of two examples of growing importance due to the rise of the future internet in the age of globalisation, namely, service science and big data.

4.1 Confidentiality in Service Science

The world is dominated by service-based economies. In developed countries, the sector accounts for over 70% of economic activity, and in a significant number of developing countries it accounts for over 50% [14]. A *service* can be defined as: the application of competencies for the benefit of another [24], see [45]. Further, "Service is performed in close contact with a client; the more knowledge-intensive and customized the service, the more the service process depends critically on client participation and input, whether by providing labor, property or information" [40] see [45]. According to the Journal *Service Science,* "Leading and competitive services enabled by service systems are all remarkably delineated with information-driven, people-centric, e-oriented, and satisfaction/success focussed characteristics". Market and consumer trends in the service economy include: demand for personalisation, customisation of services, and improvement of the customer service experience.

It is clear that a successful service economy relies fundamentally on customers providing information to service providers. This information is needed in the service provision, for example, a delivery address is needed for goods delivery, and service requests often include choice of options. The service provider may need to share the information internally or externally, for example with a courier service.

It is highly likely that service providers also store client information, for future use in service improvement and innovation, including personalisation, customisation and customer experience improvement. Issues of privacy and commercial sensitivity arise, since client information can be complex, personal and sensitive. For example, the information may include direct data such as health status, employment status, or financial status, or may include indirect information revealing behaviours, movements, and preferences. In some cases, such as government services, the client may not have a choice whether to interact or not, so must accept that the information needs to be provided. In addition, given the trends of the future internet in the age of globalisation, service providers would

be increasingly collecting as much transactional and auxiliary information as possible. The service provider may be motivated to address these issues, since assurances of confidentiality protection during service provision may increase the information the client is willing to give, so improving the service and the level of personalisation possible, and hence improving the overall service experience for the client.

Interestingly, we have been unable to locate any articles in the academic press relating primarily to privacy or confidentiality technologies in service science (via a title search on a popular publications database). There has been some discussion of policy aspects, see for example [32], who note that "technology has the potential to regulate behaviour by enabling or disabling it, in contrast with law, which regulates mainly be imposing sanctions. ... Therefore, it is necessary to consider these approaches simultaneously ..."

It is therefore worthwhile to analyse the similarity between the confidentiality protection scenario in service science and that in the research use of data scenario as discussed in this paper. We consider the applicability of the approaches described in Section 3, in order to better understand the need for new methods of protecting confidentiality in service science. This is consistent with the approach advocated by service science experts, including: "Services science is an emerging field that seeks to tap into these and other relevant bodies of knowledge, integrate them, and advance three goals - aiming ultimately to understand service systems, how they improve and how they scale", "The study of service systems is an integrative, multidisciplinary undertaking and many disciplines have knowledge and methods to contribute" [45], and "Synthesis of partial knowledge from individual disciplines is vital for future of service science" [14].

The table in Figure 1 gives a summary of the main similarities and differences between the scenarios of service science and research use of data archives.

We see from Figure 1 that the data providers and data custodians in the two scenarios are broadly similar. The differences in motivation for providing information are probably not sufficient to impact on the data providers' expectations of confidentiality of their data. For the implications of big data on confidentiality protection, see Section 4.2.

One of the main differences between the two scenarios is in the area of data sharing - specifically, whether the dataset is held by the collecting agency or is held in trust for a different collecting agency. In the research use of data, this issue arises in data linkage centres which bring together data from various sources, see the Manitoba Centre for Health Policy and Evaluation, Canada [38], the Oxford Medical Record Linkage System [15], the Scottish Medical Record Linkage System [20], Western Australian Data Linkage Branch [18] and the Welsh Secure Anonymised Information Linkage (SAIL) system [13]. In this situation the issue is best addressed in the governance layer, supported by technological implementations and information security measures.

Finally, two areas of broad similarity between the two scenarios are the range of authorised users and the cohort against which confidentiality protection is

	Service Science	Research Use
Data providers	Clients	Census and survey participants
Motivation for providing information	In exchange for a service, but may be compulsory	Voluntary or compulsory
Data custodians	Service provider agencies and companies	National statistical agencies and other service provider agencies
Dataset size	Moderate trending to big data	Moderate trending to big data
Data sharing	Shared amongst a community of service provider agencies and companies	Usually held by custodian agency, some initiatives in data linkage
Authorised data users	Staff in service provider agencies and companies; increasingly outsourced to contractors	Staff in service provider agencies. Researchers, policy analysts, increasingly the general public
Confidentiality protection needed against:	Unauthorised users – most privacy policies include use for service provision, improvement and personalisation	Unauthorised users and some classes of authorised users

Fig. 1. Main similarities and differences between the scenarios of service science and research use of data archives

needed, if we accept that there is broad analogy between contractors and researchers, policy analysts, and the general public.

Given the similarities between the two scenarios of service science and research use of data, we believe that the approaches for protecting confidentiality outlined in Section 3 are broadly applicable in service science. We note that within the broad service science context, there is still likely to be a range of more detailed scenarios involving a particular dataset, data custodian, analyst and so on. We reiterate that it is important to choose the appropriate confidentiality protection method to address the particular scenario in question.

4.2 Confidentiality in Big Data

As mentioned in the introduction, technological advances and the increasing connectivity of a growing number of computers, devices, and sensors are resulting in massive amounts of data being generated and stored. The term "Big Data" was coined in response to the realisation that traditional methods of storing, processing, and analysing data were breaking down in the face of the so-called "3 V's of big data", namely, volume (amount of data), variety (range of data sources and data types), and velocity (speed of collection and dissemination).

In the big data scenario, it is worthwhile to think about whether the 3 V's of big data pose additional privacy or commercial sensitivity risks. Again, we analyse the similarity between the confidentiality protection scenario in big data and that in the research use of data scenario as discussed in this paper. We further consider the applicability of the approaches described in 3, in order to understand the need for new methods of protecting confidentiality in big data.

Big data involves massively increasing volumes of data, often leading to a situation in which values for more and more characteristics of an individual are being stored. This in turn increases disclosure risk, since the more information we have about an individual, the more likely it is to be able to identify an individual in the data and learn something that we did not know already about that individual. Massively increasing volume of data also places stress on storage and computational infrastructure. These challenges are being addressed through new infrastructure and computational approaches. A important additional privacy or commercial sensitivity risks would therefore arise in the context of information security - the question is whether the new infrastructure and computational approaches for big data are still adequately protecting data from unauthorised access and use.

Massively increasing variety of data has the potential to increase again the likelihood that an individual is identified, with subsequent increased disclosure risk.

Massively increasing velocity of data bring challenges in terms of ensuring that data processing is fast enough to keep up with the rate of data arriving. For example, if it is necessary to remove direct identifiers from data in order to reduce disclosure risk, then how fast does this need to be done in order to ensure that only de-identified data are stored or accessed? Again, these challenges seem to be best addressed through the integration of confidentiality protection routines with the new infrastructure and computational approaches under development to cope with big data velocity.

There have been a number of books and articles published in the academic press relating primarily to privacy or confidentiality technologies in big data. Many of these address the socio-legal or information security perspectives. We believe it is still worthwhile to analyse the similarity between the confidentiality protection scenario in big data and that in the research use of data scenario as discussed in this paper. We consider the applicability of the approaches described in Section 3, in order to better understand the need for new methods of protecting confidentiality in big data.

The table in Figure 2 gives a summary of the main similarities and differences between the scenarios of big data and research use of data archives.

We see from Figure 2 that there are significant differences between the nature of data providers and the motivation for providing information between the two scenarios of big data and the research use of data. In the situation of big data, data providers may not even know they are providing data, or may have little or no choice in the data provision. Given these observations, there is a higher moral/ethical responsibility on data custodians to protect confidentiality in big data, which would be addressed in the design of information management systems, see [7].

There are also significant differences between the nature of data custodians in the two scenarios. While in research use, the relatively few data custodians are subject to enabling legislation containing specific confidentiality protection requirements, in big data almost any entity can be a data custodian subject to

	Big Data	Research Use
Data providers	Any individual interacting with the internet or other electronic device – knowingly or unknowingly	Census and survey participants
Motivation for providing information	Generally trading information for goods and services, but increasingly data collected while individuals go about their daily lives	Voluntary or compulsory
Data custodians	Anyone with a website, electronic equipment, sensors, etc	National statistical agencies and other service provider agencies
Dataset size	Big data	Moderate trending to big data
Data sharing	Anecdotally could be huge and range from controlled to uncontrolled	Usually held by custodian agency, some initiatives in data linkage
Authorised data users	Authorised users for a range of applications	Staff in service provider agencies. Researchers, policy analysts, increasingly the general public
Confidentiality protection needed against:	Unauthorised users and some classes of authorised users	Unauthorised users and some classes of authorised users

Fig. 2. Main similarities and differences between the scenarios of big data and research use of data archives

more general privacy or data protection laws. Similarly, while data sharing is quite controlled in research use, in big data it could be trending to uncontrolled. The degree of community awareness of data collection would also impact community expectations of confidentiality protection from data custodians during collection, storage, sharing and use. There would appear to be an imperative to ensure that legislative frameworks are robust and widely applicable enough to cover the activities of all data custodians, not just the traditionally-recognised ones.

There is broad similarity between the two scenarios with respect to the range of authorised users and the cohort against which confidentiality protection is needed, though these populations may be vastly different sizes. The main difference is in the size of the datasets, though this may disappear in time as research data sets also become larger and larger. In the big data scenario, it will be generally infeasible to transfer datasets to users, implying that there will be a preference for remote analysis systems, virtual data centres and secure on-site data centres. These are likely to rely on a combination of microdata confidentialisation methods with techniques to confidentialise analysis outputs. The growing user numbers are likely to cause a greater reliance on automated methods.

We note that within the broad big data scenario, there is still likely to be a range of more detailed scenarios involving a particular dataset, data custodian, analyst and so on. We reiterate that it is important to choose the appropriate confidentiality protection method to address the particular scenario in question.

5 Conclusion

In this paper our aim was to contribute to a discussion of some of the issues regarding confidentiality in the service science and big data analytics contexts. We believe that these two areas are growing in importance as the future internet in the age of globalisation is transforming our economy into a service economy and is turning the so-called *information super-highway* into an *information super-mountain* of data.

We provided an introduction to a consideration of what technologies, frameworks, and tools we might need to gain, regain and maintain informational self-determination and lifelong privacy while still extracting useful information from our growing data archives. In particular, we gave an overview of methods for protecting confidentiality in the use of data for research, as developed in the context of a national statistical agency making census and survey data available for research and policy analysis. We then discussed the general applicability of these methods to the new scenarios, in order to help pinpoint where existing methods might be applicable and where new methods might be in demand.

In the service science scenario, we found that the approaches for protecting confidentiality in the research use of data are broadly applicable, with adaptations as needed to particular situations. In the case of big data on the other hand, we found that only certain of the approaches were applicable, namely, remote analysis servers, virtual data centres, and secure on-site data centres with automated output confidentialisation routines. We remark that this is a trend underway for enabling the research use of data, and we echo the following: *... recent events in the development of remote analysis servers herald the dawn of a new era in automated confidentiality protection for analysis and we look forward to invigorated research collaborations among NST's and academic institutions to further this research ...* [46].

Acknowledgments. I warmly thank the organisers of the International Federation for Information Processing (IFIP) 9th Summer School on *Privacy and Identity Management for the Future Internet in the Age of Globalisation,* for their invitation to participate. I acknowledge the financial support of the *Authentication and Authorization for Entrusted Unions (AU2EU)* project funded by the European Commission Seventh Framework Programme for Research and Technological Development.

References

1. Abowd, J.M., Stinson, M., Benedetto, G.: Final report to the social security administration on the sipp/ssa/irs public use file project. Tech. rep. (2006)
2. Australian Bureau of Statistics: Remote Access Data Laboratory (RADL), `http://www.abs.gov.au` Accessed 23 October 2014
3. Australian Bureau of Statistics: About CURF Microdata. website (nd), `http://www.abs.gov.au/websitedbs/D3310114.nsf/home/About+CURF+Microdata` Accessed 23 October 2014

4. Australian Bureau of Statistics: Census TableBuilder (nd), http://www.abs.gov.au Accessed 23 October 2014
5. Australian Bureau of Statistics: (website), http://www.abs.gov.au Accessed 23 October 2014
6. British Columbia Linked Health Database (BCHLD): http://riskfactor.cancer.gov/tools/pharmaco/epi/british$_$columbia.html Accessed 23 October 2014
7. Cavoukian, A., Jonas, J.: Privacy by design in the age of big data. published online (2012), http://privacybydesign.ca/content/uploads/2012/06/pbd-big_data.pdf Accessed 23 December 2014
8. Centers for Disease Control and Prevention: Public-use data files and documentation. website, http://www.cdc.gov/nchs/data$_$access/ftp$_$data.htm Accessed 23 October 2014
9. Cox, L.: Linear sensitivity measures in statistical disclosure control. Journal of Statistical Planning and Inference 5, 153–164 (1981)
10. Duncan, G.T., Keller-McNulty, S.A., Stokes, S.L.: Disclosure risk vs data utility: The R-U confidentiality map. Technical Report LA-UR-01-6428, Los Alamos National Laboratory (2001)
11. Duncan, G., Elliot, M., Salazar-Gonzàlez, J.J.: Statistical Confidentiality. Springer, New York (2011)
12. Duncan, G., Pearson, R.: Enhancing access to microdata while protecting confidentiality: prospects for the future. Statistical Science 6, 219–239 (1991)
13. Ford, D.V., Jones, K.H., Verplancke, J.P., Lyons, R.A., John, G., Brown, G., Brooks, C.J., Thompson, S., Bodger, O., Couch, T., Leake.2009, K.: The SAIL Databank: building a national architecture for e-health research and evaluation. BioMed Central Health Services Research 9, 157 (2009)
14. Gećzy, P., Izumi, N., Hasida, K.: Service science, quo vadis? International Journal of Service Science, Management, Engineering and Technology 1(1) (2010)
15. Gill, L.: OX-LINK: The Oxford Medical Record Linkage System. Record Linkage Techniques. Tech. Rep. 19, University of Oxford, Oxford (1997)
16. Gomatam, S., Karr, A., Reiter, J., Sanil, A.: Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access systems. Statistical Science 20, 163–177 (2005)
17. Gouweleeuw, J., Kooiman, P., DeWolf, L.W.P.P.: Post randomisation for statistical disclosure control: Theory and implementation. Journal of Official Statistics 14, 463–478 (1998)
18. Holman, C.D.J., Bass, A.J., Rouse, I.L., Hobbs, M.S.: Population-based linkage of health records in Western Australia: development of a health services research linked database. Australian and New Zealand Journal of Public Health 23, 453–459 (1999)
19. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E., Spicer, K., de Wolf, P.P.: Statistical Disclosure Control. Wiley Series in Survey Methodology, John Wiley & Sons, United Kingdom (2012)
20. Kendrick, S., Clarke, J.A.: The Scottish Medical Record Linkage System. Health Bulletin (Edinburgh) 51, 72–79 (1979)
21. Kinney, S.K., Reiter, J.P., Reznek, A.P., Miranda, J., Jarmin, R.S., Abowd, J.M.: Towards unrestricted public use business microdata: The synthetic longitudinal business database. International Statistical Review 79(3), 362–384 (2011)
22. Little, R.: Statistical analysis of masked data. Journal of Official Statistics 9, 407–426 (1993)

23. Lucero, J., Zayatz, L., Singh, L., You, J., DePersio, M., Freiman, M.: The Current Stage of the Microdata Analysis System at the U.S. Census Bureau. Proc 58th Congress of the International Statistical Institute, ISI 2011 (2011)
24. Lusch, R., Vargo, S. (eds.): The Service-Dominant Logic of Marketing: Dialog, Debate, and Directions. ME Sharpe (2006)
25. Marley, J., Leaver, V.: A Method for Confidentialising User-Defined Tables: Statistical Properties and a Risk-Utility Analysis. Proc 58th Congress of the International Statistical Institute, ISI 2011, 21–26 August (2011)
26. Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D., Walford, N.: The case for samples of anonymized records from the 1991 census. Journal of the Royal Statistical Society Series A 154, 305–340 (1991)
27. Minnesota Population Center, University of Minnesota: Ipums international. website, `https://international.ipums.org/international/` Accessed 23 October 2014
28. Office for National Statistics: (website), `http://www.statistics.gov.uk` Accessed 23 October 2014
29. OKeefe, C.M., Gould, P., Churches, T.: Comparison of two remote access systems recently developed and implemented in australia. In: Domingo-Ferrer, J. (ed.) Privacy in Statistical Databases PSD2014. LNCS, vol. 8744, pp. 299–311. Springer (2014)
30. O'Keefe, C.M., Rubin, D.B.: Balancing the research use of health and medical data with confidentiality protection, preprint
31. O'Keefe, C.M., Westcott, M., Ickowicz, A., O'Sullivan, M., Churches, T.: Protecting confidentiality in statistical analysis outputs from a virtual data centre. Working Paper (29-30 October 2013), joint UNECE/Eurostat work session on statistical data confidentiality, Ottawa, Canada, 10pp, `http://www.unece.org/stats/documents/2013.10.confidentiality.html` Accessed 23 October 2014
32. Pitkänen, O., Virtanen, P., Kemppinen, J.: Legal research topics in user-centric services. IBM Systems Journal 47 (2008)
33. Population Health Research Network: `http://www.phrn.org.au/` Accessed 23 October 2014
34. Reiter, J.: Model diagnostics for remote-access regression systems. Statistics and Computing 13, 371–380 (2003)
35. Reiter, J.: Using CART to generate partially synthetic public use microdata. Journal of Official Statistics 21, 441–462 (2005)
36. Reiter, J., Kohnen, C.: Categorical data regression diagnostics for remote systems. Journal of Statistical Computation and Simulation 75, 889–903 (2005)
37. Robertston, D., Ethier, R.: Cell suppression: Experience and theory. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. Lecture Notes in Computer Science, vol. 2316, pp. 213–239. Springer-Verlag Berlin Heidelberg. (2002)
38. Roos, L.L., Wajda, A.: Record linkage strategies: Part 1: Estimating information and evaluating approaches. Tech. Rep. 28, University of Manitoba, Winnipeg (1990)
39. Rubin, D.: Discussion: Statistical disclosure limitation. Journal of Official Statistics 9, 462–468 (1993)
40. Sampson, S., Froehle, C.: Foundations and implications of a proposed unified services theory. Prod Oper Manag 15(2), 329–343 (2006)
41. Sax Institute: Secure Unified Research Environment (SURE). website, `http://www.sure.org.au` Accessed 23 October 2014

42. Skinner, C., Shlomo, N.: Assessing identification risk in survey microdata using log-linear models. Journal of the American Statistical Association 103, 989–1001 (2008)
43. Sparks, R., Carter, C., Donnelly, J., Duncan, J., O'Keefe, C.M., Ryan, L.: A framework for performing statistical analyses of unit record health data without violating either privacy or confidentiality of individuals. In: Proceedings of the 55th Session of the International Statistical Institute, Sydney. p. 4pp (2005)
44. Sparks, R., Carter, C., Donnelly, J., O'Keefe, C.M., Duncan, J., Keighley, T., McAullay, D.: Remote access methods for exploratory data analysis and statistical modelling: Privacy-Preserving Analytics^{TM}. Computer Methods and Programs in Biomedicine 91, 208–222 (2008)
45. Spohrer, J., Maglio, P., Bailey, J., Gruhl, D.: Steps toward a science of service systems. Computer 40, 71–77 (2007)
46. Thompson, G., Broadfoot, S., Elazar, D.: Methodology for automatic confidentialisation of statistical outputs from remote servers at the Australian Bureau of Statistics. Joint UNECE/Eurostat work session on statistical data confidentiality (Ottawa, Canada, 28-30 October 2013), 37pp
47. UK Data Archive: Secure data service (website), `http://ukdataservice.ac.uk/get-data/how-to-access/accesssecurelab.aspx` Accessed 23 October 2014
48. United States Census Bureau: (website), `http://www.census.gov` Accessed 23 October 2014
49. University of Chicago: NORC (website), `http://www.norc.org` Accessed 23 October 2014