Pacific Community
Communauté du Pacifique

Microdata Dissemination

1. Introduction and Main Objectives

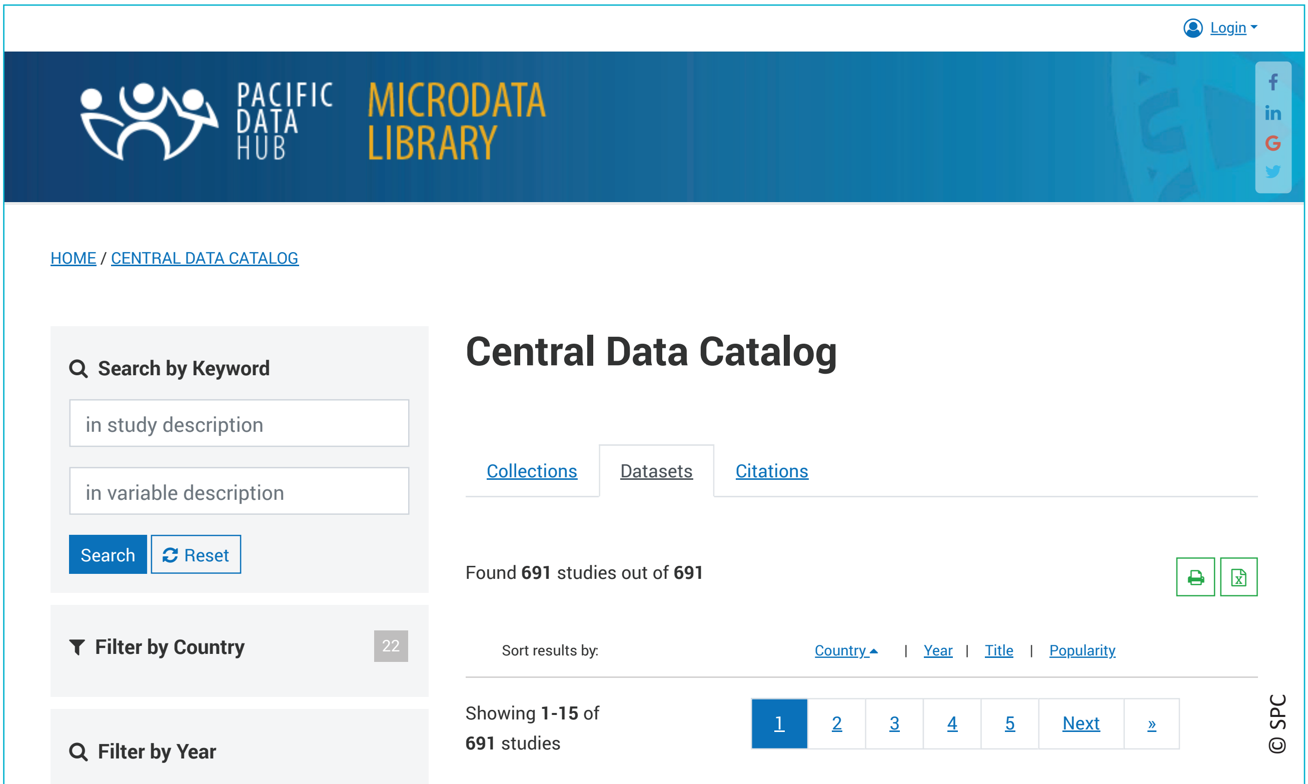
Microdata **dissemination** is a core activity in the statistics production process and a key goal of every survey programme. Data is most effective when it can be discovered and used by a broad range of users: making data more accessible adds value to existing data while reducing the need to collect new data.

In all countries, data producers are faced by expanding demand for microdata. Deciding the best way to disseminate microdata is not only a technical challenge – as data producers have to implement procedures for the documentation, cataloguing, dissemination and preservation of the data – but also a legal and ethical one. In fact, data producers must balance this demand with the need to keep respondent information confidential and reduce the **risk of disclosure** of individual information.

The **cost** of microdata dissemination, therefore, includes not only that of creating and documenting microdata files, but also the cost of creating access tools and safeguards, and of continuously supporting enquiries made by the research community. A certain level of **technical capacity** is also required to support dissemination of microdata files. Both costs and capacity can represent a limiting factor for an efficient data dissemination in small countries with limited financial resources, such as some of the Pacific Island countries and territories (PICTs).

The main goal of this publication is to describe all the concepts and the technical and legal aspects related with microdata dissemination, with a particular focus on the Pacific region, and to present possible solutions for improving and expanding microdata dissemination, such as the Pacific Data Hub Microdata Library (PDH-ML). The target audience is Government Statisticians of Pacific Island countries and territories and other statistics stakeholders from the region.

The main concepts and principles of microdata dissemination are described in sections 2 and 3; sections 4 to 6 provide information on the technical aspects of microdata dissemination – such as digitization of paper records, anonymization techniques, technical infrastructure and financial requirements – while section 7 outlines legal aspects. Section 8 presents the Pacific Regional Data Dissemination Strategy (PRDDS) and the Pacific Data Hub Microdata Library.



Central Data Catalog

Found **691** studies out of **691**

Showing **1-15** of **691** studies

© SPC

PACIFIC COMMUNITY
GUIDANCE NOTES SERIES

2. Main concepts of dissemination

Definitions: Microdata, Metadata and Documentation

Microdata: When statistical agencies or other data producers conduct surveys or censuses or collect administrative data, they gather information from each unit of observation. A unit can be a household, a person, a firm, an agricultural holding, or other entities. In this context, microdata are the electronic data files containing the information about each unit of observation – as opposed to macro data or aggregated data, which provide a summarized version of the information in form of means, ratios, frequencies or other summary statistics.

Typically, microdata is organized in **data files** in which each line (or record) contains information about one unit of observation. This information can be stored in **different formats**. Common formats include the non-proprietary ASCII format[1] and proprietary formats like those generated by specialized statistical software such as SAS, SPSS and Stata.

Public Use Files and Licensed Files: Preparing raw microdata files for dissemination involves processes that may adjust the content and/or number of records by suppressing information from direct and indirect identifiers to protect the anonymity of respondents. Suppressing information does not necessarily mean removing variables: in some cases, re-coding variables into less detailed categories to make them less informative is sufficient. Microdata files for dissemination, therefore, almost always differ from those strictly for use by staff of data producing agencies. **Public Use Files (PUFs)** are available to anyone agreeing to respect a core set of conditions about data use. In some cases, PUFs are disseminated with no conditions and made available on-line because the risk of identifying individual respondents is considered minimal.[2] **Licensed Files**, on the other hand, are restricted to users who have received authorization to access them after submitting a documented application and signing a Data License Agreement (DLA) governing the data's use by external bona fide users – trustworthy users with legitimate need to access the data. Direct identifiers such as respondents' names must be removed from a licensed dataset. The data files may, however, still contain variables that could indirectly identify respondents by matching them to other, external, data files.

Metadata: Data is of no value if it is not properly documented. ASCII files, for example, cannot be understood or used unless a data dictionary is provided as a separate file or document. To ensure microdata is used and appropriately preserved for institutional knowledge retention, it must be well-documented and include detailed **metadata**. Metadata is the descriptive information that accompanies the main data set (microdata); it is usually defined as 'data about data' and is intended to help researchers understand what the data is measuring and how it has been created.

Data Documentation: All relevant survey material that would allow the users to better understand the data and interpret the results should be attached to the microdata. These include questionnaires, interviewer manuals, descriptions of methodology, reports, publication citations, and so on. A **Basic Information Document (BINFO)**[3] is a critical part of the supporting documentation. It should contain information on all aspects of survey implementation, so that data users have all the information they need in one place, including: sample design and sampling weights; questionnaire design; dataset structure and notes on how to use the data; variables and values, coding and classification schemes; notes on how missing data were recorded and data cleaning undertaken; how consumption aggregates were created; and so on. Good **documentation** reduces the amount of user support that statistical staff must offer to external users of their microdata. Data not intended for dissemination must also be fully documented, to build institutional memory of data collection, aid in training new staff and improving data consistency over time. The international standard for the documentation of survey data is the **Data Documentation Initiative**

[1] ASCII data files only contain data, readable by most software applications. As they are not associated with software liable to obsolescence, ASCII files are optimal to guaranteeing long-term data preservation. To tabulate or analyse ASCII data, users must first import them into other software.

[2] Public Use Files can be called in different ways across countries; in New Zealand, for example, they are called "Confidentialised Unit Record Files" (CURFs), while in the USA they are referred to as Public Use Microdata (PUM).

[3] Sample BINFOs are available at http://go.worldbank.org/WFNY30UTJ0

(DDI),[4] which provides a structured checklist of survey documentation fully capturing the microdata and the entire survey lifecycle. Metadata prepared according to the DDI standard can easily be shared and transferred across systems and organizations5.

3. Principles of microdata dissemination

The UN Fundamental Principles of Official Statistics' principle on "Confidentiality" states: "Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly **confidential** and used exclusively for statistical purposes" (UN 2014).[6]

The "Five Safes" framework[7] that has become best practice for microdata access breaks down the decisions surrounding data access and use into five related dimensions:

1. **Safe projects. Is the data to be used for an appropriate purpose?**

   It is appropriate for microdata to be used for statistical purposes to support research as long as confidentiality is protected. Users wanting to access detailed microdata are expected to explain the statistical purpose of their project, showing it has a valid research aim and a public benefit.[8] A safe project, nonetheless, does not constitute per se an obligation to provide microdata: it is the NSO that has the final say about whether to provide microdata or not, as there might be other concerns that make it inappropriate to provide access.

2. **Safe people. Is the researcher appropriately authorised to access and use the data?**

   The procedures for researchers' access to microdata, as well as the uses and users of microdata, should be transparent and publicly available. Equitableness is a key principle of microdata dissemination. Microdata from publicly funded data collection, when it can be disseminated legally, should be made available to all potential users (policymakers; researchers employed by Governmental agencies, international agencies and academic institutions and other *bona fide* users) openly and on an objective basis. Access to research data should be easy, timely, user-friendly and preferably internet based. It is up to an NSO to decide whether, how and to whom microdata can be released, but their decision should be transparent and defined in a policy and in accompanying procedures or protocols.

3. **Safe data. Has appropriate and sufficient protection been applied to the data?**

   When releasing microdata, the privacy of respondents is paramount. All identifying information must be removed from any datasets that will be shared publicly, including names, addresses, telephone numbers, GPS coordinates, etc. Confidentiality protection is the key element of respondents' trust; if respondents believe or perceive that a NSO will not protect the confidentiality of their data, they are less likely to cooperate or provide accurate data. Access to non-controversial use of anonymised Public Use Files (PUFs), however, should be made as straightforward as possible.

4. **Safe settings. Does the access (IT and physical) environment prevent unauthorised use?**

   In data access contexts for publicly available data, safe settings are not required. On the other hand, sensitive data should only be accessed via secure research centres, having features such as: a locked room requiring personal authentication, IT monitoring equipment, auditing and other supervision.

5. **Safe outputs. Are the statistical results non-disclosive?**

   All statistical outputs – such as tables, graphs and maps – made available outside the data custodian's IT environment must be checked for disclosure before the release.

Sections 5 to 7 of this Note provide an overview of the technical and legal procedures to be followed to comply with the principles discussed above.

[4] https://ddialliance.org/

[5] For further guidance on properly documenting data, see the World Bank's Quick Reference Guide for Data Archivists (Dupriez et al., 2019).

[6] UN (2014). Available at: https://unstats.un.org/unsd/dnss/hb/E-fundamental%20principles_A4-WEB.pdf

[7] Available at: www.fivesafes.org, https://www.abs.gov.au/about/data-services/data-confidentiality-guide/five-safes-framework and https://ukdataservice.ac.uk/help/secure-lab/what-is-the-five-safes-framework

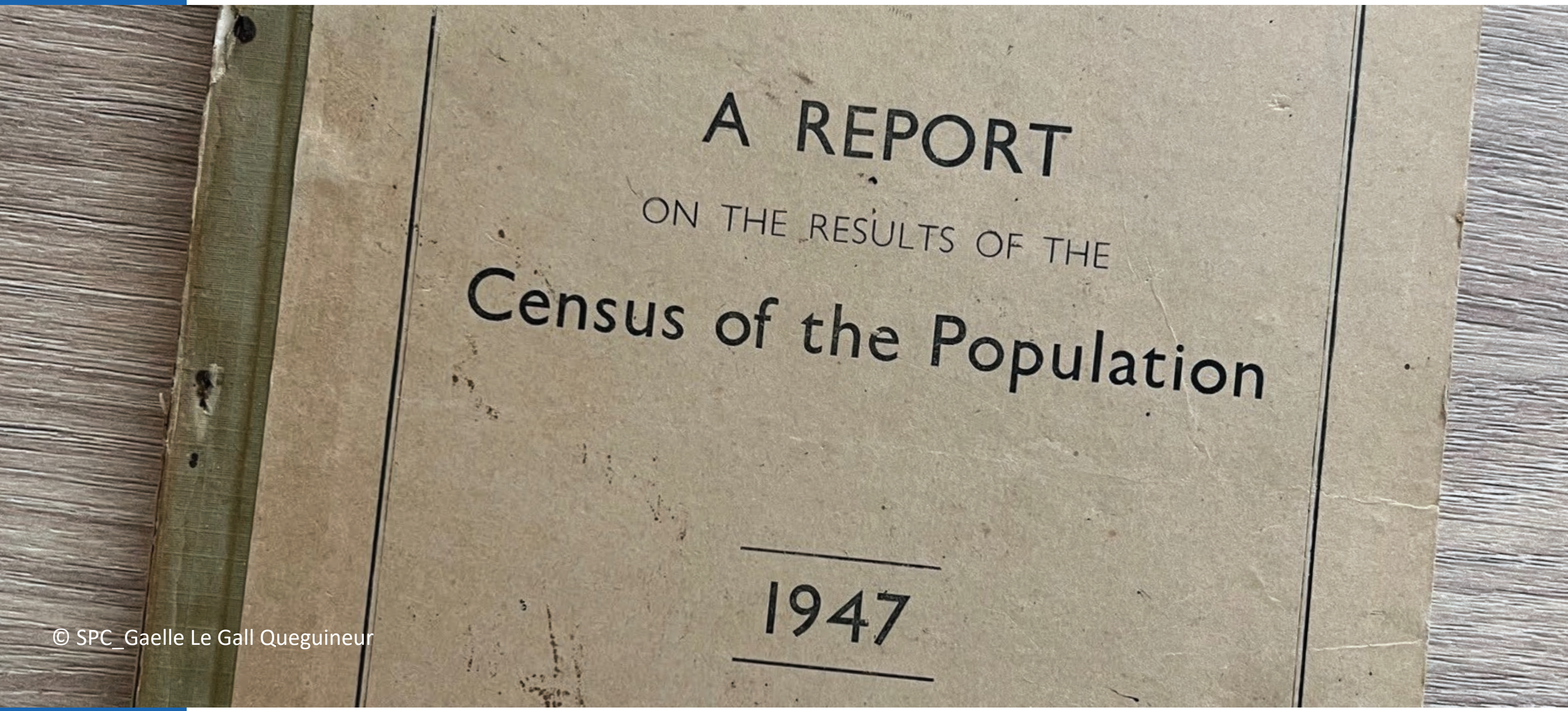
[8] Some countries have developed indigenous data sovereignty principles that govern access to microdata. New Zealand, for example, has developed the *Nga Tikanga Paihere framework, to ensure the use of microdata is respectful, ethical, and culturally appropriate. More information about this framework can be found at:* https://stats.govt.nz/integrated-data/apply-to-use-microdata-for-research/how-to-apply-nga-tikanga-paihere-to-microdata-research-projects

4. Digitisation of paper records

One of the obstacles to microdata dissemination, especially for old datasets, lies in the original paper-format of data. Document digitisation is the process of converting paper documents and records into structured, machine-readable digital files using scanning hardware and Optical Character Recognition (OCR) software. Digitising paper forms offers many benefits, such as easy data storage and access, increased data accuracy, reduced cost (once data is digitized, it can be easily shared) and increased efficiency.

Digitisation is a two-step process involving scanning the paper document with a scanner or a camera – to capture a digital image of printed or handwritten text – and then using OCR software to interpret the text and convert it to a text file format such as PDF or DOCX. Simply scanning a document, in fact, does not digitise it, but only creates an image file of unstructured data. Digitisation, instead, yields structured data and occurs at the point OCR extracts information and renders it in an indexed machine-readable format. Digitisation can also be performed manually, although this process can be lengthy and time consuming.

When considering the digitisation of sources where the content is not owned, attention needs to be paid to the copyright of the original material. The digitisation process should also be documented as this provides important information on the data quality, data source and purpose of the digitisation.



A REPORT

ON THE RESULTS OF THE

Census of the Population

1947

© SPC_Gaelle Le Gall Queguineur

5. Statistical Disclosure Control

5.1 Disclosure scenarios and disclosure risk

When releasing microdata, NSOs must evaluate the data to determine whether a public release would put the confidentiality of individuals or establishments at risk. This evaluation takes into consideration issues such as: 1) the level of detail for which data would be released (particularly as regards geographic specificity, and variables known to be held in common with outside data sources that serve as matching keys); 2) certain variables or combinations of variables that render respondents unique within the sample; and 3) other linkable data already available outside, such as those already released from the same or a related survey or information held by others from the same respondent.

A "disclosure" occurs when a person or organisation recognises or learns via released data something they did not know about another person or entity. There are two types of disclosure risks: identity disclosure and attribute disclosure.

Identity disclosure occurs when a respondent's identity is directly associated with a disseminated data record. This can occur easily when data records include **direct identifiers**, i.e., variables unambiguously identifying the respondent – name, address, telephone number, etc. It is essential that such identifying variables be removed from any microdata file before dissemination.

Attribute disclosure occurs when attribute values in the disseminated data are associated with a particular respondent. A combination of variables in a microdata record that can be applied to re-identify a respondent is referred to as a 'key' or **indirect identifiers**. Re-identification can occur when a respondent is: (a) rare in the population with respect to a certain key value; and b) this key can be used to match a microdata file to other data files that might contain direct or other identifiers.

a. The 'nosy neighbour' scenario occurs when a user has sufficient information about the attributes of one or more records that stems from personal knowledge. It is most likely when the sample target population is small – e.g., in household surveys conducted in small countries or sub-national areas with relatively few inhabitants. Detailed geographical information, in fact, is one of the main characteristics that can lead to identification of individuals, especially by users living in the same area who might know some of the respondents' characteristics.

b. The external archive scenario refers to cases where a user links disseminated microdata file records to those in another available dataset (or register) that contains direct identifiers, even though this is specifically forbidden in the data use agreement. The intruder does this by using identifying variables available in both datasets as merging keys (data matching). Prior to anonymising data, it is useful to define a disclosure scenario that describes which information potentially is available to a user, and how the latter could use it to identify an individual.

The essential part of dissemination of microdata files is avoiding both identity and attribute disclosures by applying **Statistical Disclosure Control (SDC)** or **Anonymisation** techniques to microdata files. Confidentiality is breached when a respondent is re-identified and an intruder – a user who is unauthorised or breaches the conditions set out in the data access and use agreement – can observe sensitive variables about the respondent.

There are two principal mathematical measurements of re-identification (or "disclosure") risk:

1. **Individual measurements:** assess the risk for each record. Usually these are expressed either as the probability of correctly re-identifying a respondent or by a measurement of uniqueness and rarity in the population sample.
2. **Global measurements:** assess the risk for the entire file. These are quantified as the expected number of correct re-identifications and can be derived by aggregating individual measures.

Besides disclosure issues, some NSOs might be concerned that providing microdata to researchers opens up the possibility that their results could contradict data producer estimates (due to errors by any of the two parties, use of different versions of the data – the full master file vs an anonymised/reduced public version – or different methodologies used). When the data producer is an official statistical agency, this may result in conflicting – official vs non-official – estimates, and lead to questioning of the data, with possible political implications. It is important for data producers to be able to defend their own estimates by fully documenting the collection, processing and analysis of the data in compliance with the replication standard[9].

5.2 Statistical Disclosure Control (SDC) techniques for microdata files

While accepting that disclosure risk cannot be eliminated entirely, there are methods that can reduce such risk. Processes to safeguard respondents' identity are referred to as "Statistical Disclosure Control" (SDC) or **Anonymisation** methods.

The first key step in SDC of a microdata file is to remove all **direct identifiers**, i.e., those variables that unambiguously identify the respondent. Thereafter, a microdata file can be further anonymised by applying **masking methods**, which generate modified version of the original raw microdata file. There are two types of SDC-masking methods: i) **non-perturbing-masking methods** reduce the amount of information

[9] For an overview of the concept of Reproducible Analytical Pipelines (RAP) see for example: https://analysisfunction.civilservice.gov.uk/support/reproducible-analytical-pipelines/

released by suppressing or aggregating data; ii) **perturbing-masking methods** edit and modify the data by purposedly introducing an element of error for confidentiality reasons.

5.2.1 non-perturbing masking methods

Data reduction or non-perturbing-masking techniques modify the microdata files by eliminating variables or records that can be associated uniquely with an individual. Alternatively, categories can be created in such a way as to increase the number of possible respondents in it (for example, an NSO may decide there must be a minimum number of responses in a category).

The most common non-perturbing-masking techniques are:

1. Global re-coding, which involves aggregation of the observed survey values into pre-defined classes in such a way that individual responses are not visible. This approach can be applied to continuous or discrete variables and to geographical codes. For example, age can be collapsed into age intervals and occupation, industry codes into broader categories, and geographical detail can be removed below the level at which the sample design is representative.

2. Top-coding and-bottom coding are applicable when, for numerical or ordinal variables, the highest and lowest values are very rare and could reveal the identity of respondents. Top coding involves creation of ‘catch-all’ categories such as ‘age greater than x’ or ‘income greater than y’. Bottom coding involves creating catch-all categories for small values.

3. Local suppression is a basic technique used when two variables taken together could lead to identifying a unique person. In other words, when combining these variables would result in a re-identification key for a particular record.

4. Removing variables from a microdata file for dissemination is necessary if information is regarded as too sensitive to be released, for example ethnicity or religion.

5. Removing records is sometimes necessary to protect the anonymity of respondents with a unique set of variables. When a record is removed entirely from a microdata file, it is necessary to compute and include adjusted weighting factors. Use of this approach should be minimised as removal of a significant number of records will distort the data.

5.2.2 Perturbing masking methods

These techniques involve modifying the data, so their matching becomes difficult and less certain. If re-identification is attempted, the values thus modified create uncertainty about whether the match is a true one. There are seven main perturbing techniques:

1. Additive noise is a technique involving the generation of random values that can be added to those reported by the respondent. Linear-programming techniques can help minimize the difference between the altered values and the true ones.

2. Data swapping methods transform a microdata file by exchanging values of confidential variables among individual records. Records are exchanged so that low-order frequency counts are maintained.

3. Rank-swapping is an approach whereby variables needing to be protected are sorted in ascending order, and groupings are constructed. Random pairs are selected from each group and their values swapped with values from other pairs within a pre-defined range. Creation of different group sizes leads to different data views.

4. Micro-aggregation involves replacing an observed value in the sample with the average of a small group of units (microaggregate), including the one under investigation. Units in the same group are represented in the released file by the same value. The groups contain a minimum pre-defined number k of units, with the k minimum accepted value being 3. The groups are constructed according to a criterion of maximum similarity between units.

5. Rounding techniques can be applied in a number of ways: to ensure that totals and certain summation properties are preserved; or, alternatively, randomly to ensure the cell counts in aggregate tables do not reveal counts of one or two observations.

6. Re-sampling involves taking a number of different independent samples of the values of the variables being masked. These are sorted using the same ranking criterion. The masked variables are created by taking, as first value, the average of the first values of the samples; and, as second value, the average of the second values, and so on…

7. Post-randomisation is a randomised version of data swapping. This technique induces uncertainty in the values of some variables by exchanging them according to a probabilistic mechanism. As with data swapping, data protection is achieved because users cannot ascertain with certainty if

a released value is true. Consequently, attempts to match the record to external identifiers can lead easily to a mismatch or attribute misclassification.[10]

5.3 Managing the SDC trade-off: disclosure risk vs information loss

Applying Statistical Disclosure Control techniques to a microdata file results in information loss. An NSO must strike a proper balance when trading information loss for reduced disclosure risk. In the same way that disclosure risk can be determined, the NSO may assess information loss associated with applying different SDC approaches. For continuous data, comparisons of mean squares, absolute means, and mean variation can provide a measure of information loss. However, as potential uses of microdata files are vast, it is impossible to undertake an exhaustive assessment of information loss. In practice, it is more useful to identify which subset of users will be most affected by SDC application measures that minimise disclosure. Typically, these will comprise researchers skilled in conducting advanced statistical analysis with microdata files. Since their research can result in important contributions of public benefit, it may be necessary to disseminate not only PUFs but also the less anonymised licensed files.

5.4 Documenting the SDC process

Dataset users should be aware if a disseminated dataset has been assessed for disclosure risk and whether methods of protection have been applied, including information on the technique used. They should be provided with an indication of the nature and extent of any modification due to the application of SDC methods, although the level of detail made available should not allow a user to apply reverse-engineering techniques to reconstruct the original microdata files.

6. Technical infrastructure and financial requirements

The most efficient and cost-effective way to ensure microdata dissemination is to enable users to search, discover, and download microdata – and its accompanying metadata – from a single online platform.

6.1 Microdata catalogue

The objective of a **microdata catalogue**[11] is to provide easy access to data and documentation in a format most convenient for users. From the data user perspective, a good survey catalogue provides tools for:

- Finding the data files most appropriate to the user's needs in a variety of formats, which typically include formats for SPSS, STATA, SAS and ASCII.
- Evaluating the data to ensure compatibility with the researcher's needs. This role is supported by the metadata and other documentation attached to the file (questionnaires, manuals, reports, etc.).
- Accessing the data. This involves an extraction and/or delivery system of some sort. Commonly, such files can be delivered via a website/portal and an FTP[12] server.

From the administrator's viewpoint, a good microdata cataloguing system provides: i) a secure environment for storing and sharing data and metadata; ii) the tools to manage the microdata access process; iii) solutions for sharing public use files and licensed use files; and iv) tools to collect information on users of the catalogue, the data they download and the purpose for which they are using the data.

6.2 Financing microdata catalogue

Building a microdata catalogue may incur in major managerial and financial issues. Scientific data infrastructure requires continued and dedicated budgetary planning and financial support. Moreover, a variety of skills are needed in meeting the needs of researchers, which might not necessarily be present in all NSOs. For some NSOs, establishing and maintaining such a data archive and dissemination service might be an unrealistic objective for budgetary or technical reasons.

An alternative option is to entrust an existing data archive, like the Pacific Data Hub (see Section 8), where anonymisation techniques are already in

[10] Further details on the SDC techniques can be found in Hundepool et al. (2010) and UNECE (2007).

[11] Software to build microdata catalogs are provided by the International Household Survey network (IHSN) at: https://www.ihsn.org/software

[12] File Transfer Protocol.

place, documentation is already provided according to DDI standards and assisting researchers' requests is facilitated. In fact, for researchers and development partners it is much easier to have a standardised set of licensing, conditions of access and way of applying for all the datasets available in the Pacific Region rather than having different standards across PICs. From the NSO perspective, moreover, SPC doing the administrative work of controlling access to researchers is a big advantage that reduces the amount of time and resources needed to address their requests.

7. Legal aspects

7.1 Legal base of microdata dissemination

The development of specific protocols is fundamental to comply with the essential principles for managing the confidentiality of microdata and with a country's legal framework. Issues related with authorship, producer credits, ownership, dissemination, usage restrictions, financial arrangements, ethical rules, licensing terms and liability should all be covered by such arrangements before microdata are released. The existence and visibility of these arrangements are necessary to increase public confidence that microdata will be used appropriately.

Enabling legislation is essential to enable:

(i) public confidence in microdata access and protection – that there are legal constraints that determine what can and cannot be done;
(ii) mutual understanding between NSOs and researchers on microdata arrangements;
(iii) consistency in how research proposals are treated; and
(iv) a basis for dealing with breaches.

The microdata dissemination policy for a country is shaped by its legislative framework. Some NSOs have no provision in their statistics acts for release of microdata files. In some cases, microdata dissemination may be explicitly forbidden; in others the act may not deal explicitly with the matter, so it is subject to interpretation. In all such cases, legislation needs to be revised before microdata files can be distributed. Legal advice should be sought to know exactly what limitations there are; and if there are restrictions, these should be addressed the next time the statistics act is updated. The legal base, however, does not need to be in primary legislation or law: regulations, ordinances and similar authorisations still have legal impact. It is important that the legislation (or authorisation) covers the following aspects:

(i) what can and cannot be done, and for what purposes;
(ii) the conditions of release; and
(iii) the consequences if these conditions are breached.

7.2 Data access arrangements

The NSO website is an effective place for providing information on how to access microdata. Following the **transparency** principle, detailed information on the procedures and conditions attached to the release of microdata files should be specified in **protocols** that will define:

- How users can request access to the data (online requests, request forms to be used, etc.);
- The permissions and restrictions attached to the various types of datasets;
- Who is responsible for decisions to grant access, and other practical information on the review
- process;
- What type of statistical disclosure control methods are in place;
- What information is required from the researchers, and what can be done with this information.

Conditions to be set are different for Public-Use Files (PUFs) and Licensed Files. Generally, data regarded as public is open to anyone with access to an NSO website. It is, however, normally good practice to include statements defining suitable uses for and precautions to be adopted in using the data, to sensitise the user. Prohibitions such as attempts to link the data to other sources can be part of the 'use statement' to which the user must agree, on-line, before the data can be downloaded.[13]

For licensed microdata files, terms and conditions must include the basic common principles plus some additional ones applying to the researcher's organisation. There are two options: i) data is

[13] An example of conditions for accessing and using PUFs is provided in Box 10 of Dupriez and Boyko (2010).

provided to a researcher or a team for a specific purpose; or ii) data is provided to an organisation under a blanket agreement for internal use, e.g. to an international body or research agency. In both cases, the researcher's organisation must be named, as must suitable representatives to sign the license.[14]

7.3 Managing breaches by users

Experience in countries with established microdata dissemination practices shows that breaches of data file confidentiality are limited (the researcher's reputation would be at risk, as would that of their organisation); nevertheless, whether intentional or accidental, they may occur. To maintain a sustainable policy of microdata dissemination, NSOs need to consider enforcement procedures, such as:

- If a legal offence has occurred, legal action should be considered.
- If researchers violate their undertaking, the NSO should consider suspending their access rights.
- If the undertaking is made by an organisation on a researcher's behalf, the organisation, rather than the NSO, may wish to consider the sanctions it should take towards one of its own.
- If necessary, the NSO should take steps to ensure further breaches do not occur;
- If the breach is minor, a warning should be considered as the only action necessary.

8. Dissemination in the Pacific Region

8.1 The Pacific Regional Data Dissemination Strategy (PRDDS)

The Pacific Regional Data Dissemination Strategy (PRDDS)[15] acknowledges the existence of two overriding considerations:

- the **ownership** of statistical data and associated metadata always resides with the national producers, in most cases the national statistics offices (NSOs); and
- the **confidentiality** of an individual respondent's information provided during a census, survey or other collection, must be protected, in line with requirements set out in each member country's Statistics Act or other relevant legislation.

The NSOs that produce the data are critical stakeholders in this regional strategy. Many NSO's

[14] See Box 12, Box 13, and Annex 1 in Dupriez and Boyko (2010) for examples of such agreements. Currently, SPC is developing DLAs to determine whether researchers' requests can be directly addressed without reverting to the NSOs.

[15] See SPC 2018.

© SPC

have National Strategies for the Development of Statistics (NSDS) or are in the process of establishing them. The NSDS should include an analysis of user needs and the actions needed for improved dissemination. Also, the NSOs need to have a clear set of policies, processes and standards to ensure that the dissemination activities are carried out effectively and consistently, and within national legislative mandates and authorities.[16] Should legal restrictions be in place, these should be addressed – seeking legal assistance if and when needed – within the context of a review of the national statistics act.

According to the regional strategy, all microdata access will respect the following principles:

- **Ownership:** PICTs maintain sovereignty over their datasets. SPC-SDD will store datasets on behalf of the owner. This will not involve a transfer of ownership of the data; the data will remain the property of the respective producers. SPC-SDD is a custodian to help protect long-term usability of datasets.
- **Public benefit:** Access to the data aims to maximise the use of the data to benefit Pacific people.
- **Protection:** All data will be protected, to minimise the risk of disclosure and meet legal requirements and ethical principles.
- **Documentation:** Metadata documenting the dataset and the methods of data collection will be open for all datasets, along with an assessment of the quality and completeness of the data and metadata, based on a standard assessment framework.
- **Confidentiality:** Researchers using the data will be expected, as a condition of use, to respect the confidentiality of the data and report their findings at an aggregate level.

Data License Agreements (DLAs) between NSOs and SPC, covering legal provisions, intellectual property rights, confidentiality and resolution of disputes, should set out the legal basis for sharing the data and making it available to external users.

8.2 The Pacific Data Hub Microdata Library (PDH-ML)

Given the limited resources available at national level to disseminate microdata in some PICTs, there is a well-established need for a regional repository and data library/archive that provides a focal point for users to access Pacific data – including controlled access to microdata for *bona fide* purposes under controlled conditions. This has proved to be a successful approach.

In the past, there have been many occasions when Pacific data has been lost because of natural disasters, computer viruses and system failures, fires and similar damaging events. Archiving the microdata in a central repository such as PDH-ML reduces the risk that such data losses happen again.

The Pacific Data Hub (PDH)[17] is a central repository of PICTs' surveys, censuses and administrative data and documentation. The platform serves as a gateway to the most comprehensive collection of data and information about the Pacific across key areas, including population statistics, fisheries science, climate change adaptation, disaster risk reduction and resilience, food security, public health and human rights. It is made up of four key components:

1. **Data Catalogue:** an open data repository which manages and publishes all data in the Pacific Data Hub. It is the central component which links to PDH.stat and the Microdata Library;
2. **PacificMap:** a geospatial data exploration tool providing map-based visualisation of spatial data;
3. **PDH.stat:** database explorer which contains the 132 Pacific Sustainable Development Goals (SDGs) Indicators as well as a range of economic, health, demographic and environmental;
4. **Microdata Library (PDH-ML):**[18] online census and survey documentation (reports, documents) and archiving application which also provides access to microdata for some collections. Through it, the Pacific Community (SPC) gives safe access to microdata to enable research and analysis that benefits Pacific Island people. As a pre-condition, PICT governments' data producers share their datasets once the data is fully processed and released.

Along with NSOs, the PDH-ML is therefore another critical stakeholder in the regional statistics strategy,

[16] For a possible outline of the dissemination strategy document (Current situation; Objectives; User needs; Actions; Expected results) see SPC (2018).

[17] See: https://pacificdata.org/about-us

[18] https://microdata.pacificdata.org

enabling researchers and development partners to have a standardised set of data licensing, conditions of access and way of applying, and by reducing NSOs' administrative work of controlling access to external researchers.

Statistical disclosure control (SDC) methods have been adopted to anonymize the datasets stored in the PDH-ML and therefore release microdata in a controlled way that protects the privacy and statistical confidentiality of individuals and other entities. SDC techniques include the removal of both direct identifiers (names, phone numbers, addresses, etc.) and indirect identifiers (detailed geographic information, exact dates of birth and marriage, etc.) from the data files. The adoption of such methods, along with the use of SDCMicro[19], make it possible to disseminate microdata to external researchers thus increasing its potential value for social research and policy analysis (Pontifex, Sharp 2020).

Access to micro data stored in PDH-ML can be provided in several ways. Public Use Files (PUFs) are available online for downloading and have been heavily screened for statistical disclosure of the respondent. There are explicit conditions of use that are agreed when access is granted but not actively monitored. Licensed Files, on the other hand, require more in-depth review before release and go through a greater process of deliberation. License application requests are not considered unless researchers satisfy eligibility criteria – such as the affiliation to a credible research or teaching institution or the proof of experience of analyses of large datasets. Researchers must accept and adhere to the SPC terms of use and the signed Data License Agreement (DLA) between SPC and the data owner. And researchers must agree to provide the final output of their work as a report, paper or otherwise, and to cite all data sources in the produced work.

Statistical organisations agreeing on the release of their microdata via the PDH-ML must: i) prepare adequate metadata to ensure that the data can be understood, and ii) prepare a data sharing protocol and a Data License Agreement (DLA) describing the terms of use.

Since its inception in 2019, the Pacific Data Hub – Microdata Library now contains over 150 fully documented and preserved Pacific Island microdata-sets according to international standards (DDI) and good practices. Lists of citations found in published works include more than 1000 citations collated referencing the use of these datasets in a variety of research, indicating the extent of knowledge generated by re-use of existing data sets (Pontifex, Sharp 2020). Making microdata available is even becoming a pre-requisite in some agreements for statistical assistance.

9. Concluding remarks

According to international standards and openness rankings, PICTs generally lag behind other comparable developing countries globally (Narsey 2022).[20] The Pacific Data Hub Microdata Library (PDH-ML) is a valuable platform for resource-scarce PICTs, because it not only helps preserve data for posterity, but also facilitates a greater level of access to the microdata for global, regional, national and independent researchers while maintaining high standards of data confidentiality. There is immense potential for generating evidence-based policy analyses by independent academic researchers analysing a wide variety of microdata sets.

The PDH-ML, however, can only store microdata sets if they have been given permission by the PICTs. Different PICTs have different dynamics that explain the relative under-utilization of their microdata sets and their lack of approvals to requests. In some, ambiguous legislation hinders decision-making; others see the data requests as not backed by reputable institutions or deem the proposed research of no benefit to their country and people. Risks of compromising confidentiality are also a concern for NSOs officers. Such issues are made worse by the scarcity in some NSOs of technical and analytical staff, and the need for assistance on data anonymization and analysis.

To address these issues, SPC can assist the PICTs to

[19] https://www.ihsn.org/software/disclosure-control-toolbox

[20] This paragraph is mostly drawn from the conclusions of Narsey, 2022.

be in control of the dissemination process through a tiered approach (Narsey 2022):

- **Tier 1:** The secure archiving of all microdata with the PDH-ML, with no necessary access to the public (unless agreed to).
- **Tier 2:** Microdata are made available by PICTs for a short-terms special project managed by SDD, for analysis and writing of reports on topics prioritized by the NSOs, such as, but not limited to, poverty, food security, climate change and gender inequalities. Following every publication of such a report, SDD and the NSO can facilitate a national workshop bringing together all the national and global stakeholders in the policy issues discussed. Eventually, an over-arching regional conference can be organized, to gather these Pacific-wide findings and publish them in a monograph.
- **Tier 3:** Microdata sets are made available by PICT NSOs upon application by external researchers, with the PDH-ML coordinating the process of approval by the PICTs, who maintain the sovereignty over that particular data set.
- **Tier 4:** Microdata sets that PICT NSOs make freely available for download, with strict conditions being met through binding terms of use.

In some cases, legislation needs to be updated and made more explicit in terms of its ability to make datasets available to the PDH-ML and external researchers. The establishment of "National Statistical Councils" or "National Advisory Boards" can guide NSOs in the development of national dissemination policies.[21]

Additional funds and capacity building activities, finally, might need to be provided to Pacific NSOs for the establishment of stable dissemination strategies over time. In this context, additional guidelines are going to be provided within the PACSTAT[22] project with reference to Statistical Disclosure Control (or "anonymization") techniques, which will benefit all the Pacific NSOs.

[21] See for example chapter 12 of the Cook Islands Statistics Act 2015-16.

[22] Statistical Innovation and Capacity Building in the Pacific Islands Project. Available at: https://sdd.spc.int/innovation-sdd/statistical-innovation-and-capacity-building-pacific-islands-project-pacstat

References

Dupriez and Boyko (2010). Dissemination of Microdata Files. Principles, procedures and Practices. IHSN Working paper No 5. Available at: http://ihsn.org/sites/default/files/resources/IHSN-WP005.pdf

Dupriez, Castro, Welch (2019). Quick reference guide for data archivists. Available at: https://guide-for-data-archivists.readthedocs.io/en/latest/

Hundepool et al. (2010). Handbook on Statistical Disclosure Control. Available at: https://ec.europa.eu/eurostat/cros/system/files/SDC_Handbook.pdf.

Narsey (2022). Maximizing the use of Microdata in PICTs. Paper presented at the 10th Meeting of the Pacific Statistics Method Board (PSMB). Available at: https://sdd.spc.int/events/2022/10/10th-statistics-methods-board-meeting-psmb

OECD (2007). OECD principles and guidelines for Access to research data from public funding. Available at: https://www.oecd.org/sti/inno/38500813.pdf

Oseni, G., Palacios-Lopez, A., Mugera, H.K. and Durazo, J. (2021). Capturing What Matters: Essential Guidelines for Designing Household Surveys. Washington DC: World Bank. Available at: https://www.worldbank.org/en/programs/lsms/publication/CapturingWhatMattersEssentialGuidelinesforDesigningHouseholdSurveys

Pontifex, Sharp (2020). Pacific Data Hub: Improved Data Dissemination and Use in Pacific Island Countries. Paper presented at the 2020 Asia-Pacific Statistics Week (15-19 June 2020, Bangkok, Thailand).

SPC (2018). Pacific Regional Data Dissemination Strategy. October 2018. Available at: https://spccfpstore1.blob.core.windows.net/digitallibrary-docs/files/bb/bbe3ffc85ffd67102306194aa27303d6.pdf?sv=2015-12-11&sr=b&sig=8zPIE4SghSuscXkuk%2FLsWHhQ7HtQIxnmQqloPKiA24o%3D&se=2023-06-19T13%3A07%3A32Z&sp=r&rscc=public%2C%20max-age%3D864000%2C%20max-stale%3D86400&rsct=application%2Fpdf&rscd=inline%3B%20filename%3D%22PSSC_Nov_2018_Doc7_Regional_Data_Dissemination_Strategy.pdf

UN (2014). Fundamental Principles of Official Statistics. Available at: https://unstats.un.org/unsd/dnss/hb/E-fundamental%20principles_A4-WEB.pdf

UN (2017). Principles and Recommendations for Population and Housing censuses. Revision 3.

UNECE (2007). Managing Statistical Confidentiality and Microdata Access: Principles and Guidelines of Good Practice. Available at: https://unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

UNSD (2005). Household Sample Surveys in Developing and Transition Countries. Studies in Methods, Series F No. 96. UN, New York.

World Bank (2020). Household Surveys at the World Bank: Protocol for Data Collection, Quality Assurance and Standard Setting (English). Washington, D.C.: World Bank Group. Available at: http://documents.worldbank.org/curated/en/848521606460880374/Household-Surveys-at-the-World-Bank-Protocol-for-Data-Collection-Quality-Assurance-and-Standard-Setting