THE PRACTICE OF SAMPLING IN THE DISPOSAL OF COMMONWEALTH RECORDS

Jenny Dean and Wendy Southern

The Australian Archives has recommended sampling be undertaken when there is tension between the need to preserve records for future research by agencies and the public, and the need to consider economic factors which often preclude retention of large quantities of records. Criteria for deciding whether sampling is appropriate, including research value, quantity and rate of accumulation of records, their homogeneity, and any duplication of information are presented. The most common sampling techniques are discussed.

Introduction

The Archives Act (1983) gives the Australian Archives the responsibility for selecting Commonwealth records for permanent retention. We select records that are likely to be of continuing value to the government or to the community, but we operate under constraints of storage space, the increasing rate of records production, and the costs to maintain and process them. The tension between the need to preserve archival resources while also giving attention to economic factors is at the heart of any discussion of sampling in an archival context.

While there have been arguments as to the validity of adopting sampling procedures for archival material[1], the general consensus is that in certain circumstances it is an appropriate and economical method of reducing the bulk of records required to be retained, while still meeting research needs. As a result the Australian Archives has developed sampling guidelines as part of its Disposal Manual. These guidelines are intended to aid appraisal staff in deciding whether or not sampling is appropriate for particular Commonwealth records and may be applicable in other circumstances. The discussion that follows is based on those guidelines.

Sampling

The purpose of sampling is to select a portion of a population which accurately reflects the characteristics of the whole population. Within an archival framework, sampling is used to select some part of a body of records from the same disposal class, which has been demonstrated to have continuing research value, but where the quantity of records precludes retention of the entire class.

EXAMPLE 1 - Selective retention within classes (extract from Patent, Trade and Designs Office authority, issued 1979)

Disposal class	Retention/Destruction
LIBRARY MATERIAL, INCLUDING TECHNICAL, SCIENTIFIC AND LEGAL LITERATURE	
which is of a unique nature	Retain permanently
which is of limited circulation	Retain permanently
which was originally received as an attachment to correspondence of a substantive nature	Retain permanently
which contains annotations by senior officers	Retain permanently
which has been used directly and extensively in administering the Patent function	Retain permanently
Other	Destroy when reference ceases

EXAMPLE 2 - Statistical sampling after selective retention of records (extract from Australian Development Assistance Bureau authority, issued 1980)

Disposal class	Retention/Destruction
SPONSORED STUDENT/TRAINEE PERSONAL CASE FILES, 1951-	
Those involving major welfare cases	Retain permanently
Those involving other special features	Retain permanently
Selected samples	Retain permanently
A sample of files retired annually is to be selected on the following basis:	
(a) 10% of files retired for each country	
(b) The sample each year is to cover as many courses as possible	
Remainder of case files	Destroy 5 years after action completed

In the widest sense, any decision to retain less than the whole population is a form of sampling. The disposal decisions the Australian Archives takes clearly fit this bill. We choose to keep only those records which have some permanent values, and the remainder are destroyed when their administrative uses or any legal requirements for their retention are fulfilled. However, this process is one of *selective retention* where the "samples" we keep are retained because they are special cases which we subjectively select as having some intrinsic value. On the other hand, *sampling* is a statistically based technique aimed at selecting a group which accurately reflects the general characteristics of the whole population. Items are selected because they are typical or representative, not because they are special in any way.

Criteria for deciding whether sampling is appropriate

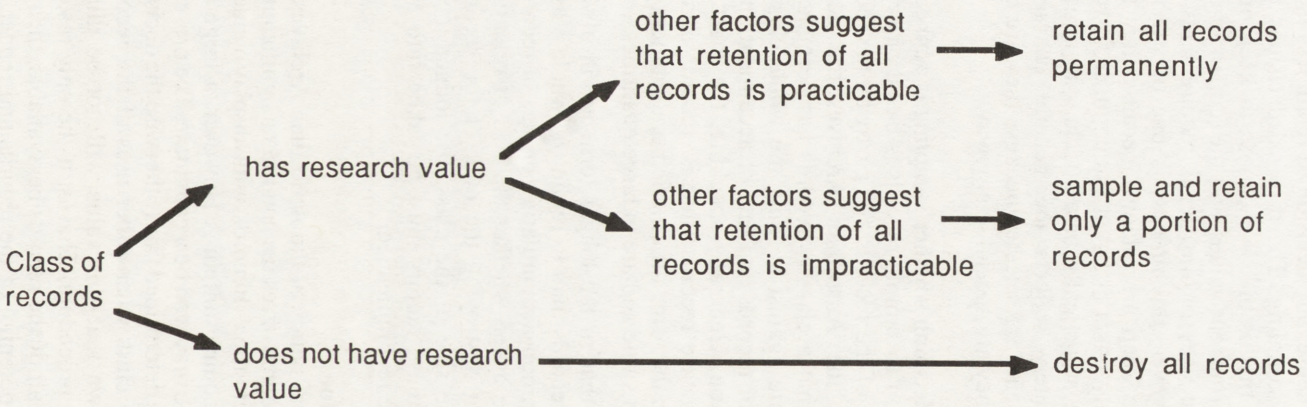
The decision to sample records is based on criteria, partly determined by the nature of the records and partly by the requirements for statistical sampling. In the Australian Archives the decision revolves around establishing that a class of records has permanent value and that other factors militate against keeping the whole class. These factors include the quantity of records and rate of accumulation, their homogeneity and any duplication. Generally speaking, the most suitable candidates for sampling are case records (that is, records relating to individual cases which contain the same information or document the same activity) which have research value and are in large quantities.

It is clear that in the longer term it will always be cheaper to store a sample of records than to retain the whole class indefinitely. However, when smaller, non-accumulating groups of records are being considered, it needs to be asked whether the cost of retention overrides the cost to researchers of destroying the bulk of the records. It is also clear that the larger the size of the class, the longer it will take to sample the records. All these factors must be taken into account before sampling provisions are recommended.

Research Value

The disposal policy of the Australian Archives states that records must have a high degree of research value, or significant display value to warrant permanent retention. Records with display value are rarely of sufficient quantity to require sampling. However, a large quantity of records with research value may well be a candidate. The presence of sufficient research value must be determined first (otherwise the decision should be destruction of the whole class when other uses of the records have ceased) and it must be shown that this value will not be diminished beyond use by the sampling procedure. That is, if the only research use for the records requires that all of the records be available, then sampling would destroy the research potential of the records and render them valueless. It would

TABLE 1 - Appraisal decisions relating to research value of records



be as logical in this case to destroy them all at the outset. Table 1 sets out the decision process.

Quantity and rate of accumulation

The Australian Archives does not have any prescriptive guidelines for balancing quantity and cost in the sampling decision. We recognise that accurate sampling requires a degree of technical skill not usually required in appraisal and sentencing, and thus the resources required to implement sampling are more intensive than usual. This has to be set against the costs of storage and the knowledge that sampling is not a wholly satisfactory alternative to retaining the whole class permanently. Therefore, the quantity of records in the class must be large before a decision to sample can be justified. As a general rule, we suggest that if a class of records with research value is greater than 100 metres nationally and is still accumulating, it is appropriate to consider sampling. On a regional basis, if a class of records exceeds 50 metres and is still accumulating, sampling should be considered.

Homogeneity

In statistical terms, homogeneity is of utmost importance. Increasing variation means the sample has to be larger to represent the full range of characteristics of the population, and as a consequence sampling is more difficult to implement. Therefore, if a class of records contains highly variable or dissimilar material, the sampled part is less likely to reflect the contents of the remaining portion of the class which has been destroyed. If there is high variation in the records with respect to those factors which are likely to be of research interest, the sample size would have to be so large in order to reduce the sample error to acceptable levels, that it may be more viable to retain the entire class. In a homogeneous class of records, on the other hand, a relatively small sample should reliably represent the class. Appraisers must therefore establish the homogeneity of any record class before proposing sampling procedures.

Records may vary in a number of ways. Variation may relate to information content (e.g. the amount of information on case files may vary according to the number of transactions a person had with an agency); to demographic factors (e.g. if 99% of the individuals documented in a record class were aged 50, it is extremely unlikely that any sample would greatly misrepresent the age variation across the class); or to the records' administrative context (e.g. records may have been created in a variety of regional locations). Records which are likely to be homogeneous include printed forms and some case records, whereas general correspondence files, unless they are of a very routine nature, woudl probably display too much variation.

Duplication

Because sampling is an imperfect alternative to retaining an entire class of records with research value, it is important to examine any alternative sources of information which duplicate the information in the records to any extent. The duplicates may not pose the same storage problem (e.g. if records have been copied onto microfiche). Even if these alternative formats do not exactly duplicate the information in the paper records, to retain them may be preferable to implementing sampling procedures in terms of both time and cost.

Similar records from alternative sources may not have the same type of research value as the records being appraised. For example, records may differ in the extent to which they contain aggregated information rather than raw, unanalysed data. Electronic records are usually more maniputable than other formats and therefore may have additional research uses to conventional format records. However, there are other problems associated with their retention. Such factors should be weighed up to assess the ultimate value of sampling against retaining alternative formats.

Selecting a sampling method

Once sampling has been determined as the best option for the retention of a particular disposal class, an appropriate sampling method must be chosen. The choice of method depends on the likely research uses the class of records has, administrative considerations, such as how and where the records are held, the difficulty of implementation and the statistical accuracy required. The methods of sampling which are most commonly used in the archival context are:

- exampling
- random sampling
- systematic sampling
- combined sampling.

These are discussed in detail below and summarised on Table 2.

Exampling

The method of exampling is not based on statistical principles, but is included here because it is widely used in archival circles. It involves the selection of one or a few specimens from within the whole class to illustrate administrative practice at a particular time. The major reason for keeping examples is to retain evidence of what was done in an agency.

It must be stressed that exampling has limited value for research purposes. Possible research uses are restricted to documenting how certain functions of the agency were carried out, but the example can only be cited as an indicator and cannot be used in statistical or comparative studies.

TABLE 2 - Different Sampling Techniques

	Research use - suitability	Administrative suitability	Difficulty in implementing	Statistical accuracy
Exampling	Very limited, only used to document functions of agency and to indicate that class existed and was once in use	No particular administrative considerations	Involves extensive registering of series for very limited research uses	Not statistically representative
Random sampling	Particularly suitable for quantitative research e.g. demography	Suitable for records which are held in one place, or are controlled centrally	Usually difficult to implement	Very statistically accurate, if carried out correctly
Systematic sampling				
Numerical	May be suitable for quantitative or other types of research	No particular administrative considerations	Relatively easy to implement	Approximates a random sample if records have no inherent numerical ordering
Alphabetical	May be suitable for quantitative or other types of research	Suitable for personal case records arranged in alphabetical order	Relatively easy to implement	Statistical accuracy relatively good if selection of letters is non-biased
Chronological	May be suitable for quantitative or other types of research	Suitable for records where activity documented is repetitive or continuous	Easy to implement if records are arranged chronologically	Statistical accuracy relatively good if records not subject to chronological fluctuation
Regional	May be suitable for quantitative or other types of research	Suitable for records arranged by location	Relatively easy to implement	Statistical accuracy relatively good if records not subject to regional variation
Combined methods of sampling	May be suitable for quantitative or other types of research	No particular administrative considerations	Relatively difficult to implement	Statistical accuracy is improved through using a combination of methods in this way

EXAMPLE 3 - An authority which includes exampling provisions

Disposal class	Retention/Destruction
STUDENT EDUCATION ALLOWANCE FILES	
Examples	Retain permanently
5 examples are to be selected from the files	
Remainder of files	Destroy 5 years after action completed

Random sampling

Random sampling means that every element making up the whole class has an equal mathematical chance of being selected, so that the chance of bias and error is dramatically reduced and the chance of the sample being representative of the whole class is maximally increased. The method is effective only if the population has low variability. It does not mean a haphazard selection of records, rather the accurate taking of a random sample involves several steps. The appropriate unit of selection has to be identified and then each unit is given a unique number in consecutive order. Units must then be selected by means of a random numbers generator or table. It is most useful for research of the kind which conducts quantitative analysis on the information contained in the records.

These procedures can obviously be time consuming, expensive and difficult to administer, thus it is not a commonly used technique within archives. The only exception is likely to be if the class of records is controlled by computer index, where it may be an easy task to identify and select the appropriate units on-screen and then retrieve them using normal registry procedures.

Systematic sampling

This form of sampling involves selecting records according to a particular pattern, on a numerical, alphabetical chronological or regional basis. For example, every tenth record or all records where the title begins with the letter 'k' are selected. If care is taken in the selection process, the mathematical difference between some types of systematic sample and a random sample can be negligible. In addition, because it is relatively easy to administer, systematic sampling is a favoured technique among archivists.

Before proposing this method of sampling, however, it must be clear that there is no inherent order or bias in the class of records being sampled. For example, if a group of case files were ordered in such a way that every second file had a female subject, and every other file a male subject, a decision to sample every 10th file would be inappropriate. Similar biases

could emerge from case files ordered by name, where using the letter "m" may select a higher proportion of individuals of Anglo-Saxon origin than is representative of the whole series.

Chronological sampling can be undertaken where records are arranged in chronological order, or where records document activities undertaken on a repetitive or continual basis. For example, if surveys are conducted annually it might be appropriate to retain every second or third survey. Finally regional sampling can be undertaken where records are arranged by region or location (e.g. by country, state or town). It is generally easy to administer, but it can be applied only where there is uniformity across regions. It is also a form of sampling which is valid for research only for the areas which are sampled, and not for the remainder which are destroyed.

Combined methods of sampling

Combining methods of sampling may result in a more representative sample, however, it is likely to be more difficult to implement. It does mean that the level of variation according to two parameters (e.g. time and region) can be controlled. In some cases it may be appropriate to combine a sampling method with selective retention processes, where it can be shown that the class has varying types of research value, although in this case the apparent lack of homogeneity probably precludes sampling.

Sample size

In all types of sampling, except exampling, the size of sample is an important consideration. Specifying the appropriate sample size will ensure the statistical validity of the sample and enhance its research use. To retain too few records within a class will mean that the sampling error will be too high to be of use in most types of research (i.e. the sample will not accurately reflect the characteristics of the population); to retain too many records begs the question of keeping a sample at all. The sample size should be determined before finalising the method used to sample the records being appraised.

The accuracy with which a sample represents the class from which it is drawn is determined by the size of the sample. This in turn is determined by the variability of the parent class and the amount of error acceptable in generalising from the sample to the class.

As a general rule the size of the sample increases with the degree of variability in the parent class. This is independent of the size of the parent class. Where records are concerned, we need consider only variation in aspects which are likely to be of research interest and the smallest sample will derive from a class where these aspects are homogeneous.

It is possible to employ statistical tests of significance to find out what

sample size is necessary to counter the variation in a class of records, and thus to determine whether the sample has statistical validity. Any sampling method which claims to be statistically accurate should proceed only after this has been done. These tests are not difficult to carry out and in the case of records would rely on identifying and quantifying the parameters within the records which vary, e.g. the number of transactions, length of time covered by the record.

Complexity of the sampling method

As a general principle, implementing a sample should be no more difficult than sentencing other classes. When proposing to sample, therefore, it is important to have a clear idea of how the records in the proposed class should be sampled and to give unambiguous sampling instructions. Most importantly, whoever is actually sentencing the records must realise that sampling is not an arbitrary process, but one based on statistical principles needing careful implementation.

Application of these guidelines in the Australian Archives

A number of Disposal Authorities issued by the Australian Archives have included sampling provisions for some classes. The application of the sampling procedures has meant that we are faced with a less substantial storage burden, while allowing for future research. It is obvious that sampling involves an extra step in the appraisal and sentencing process, and requires a degree of expertise to implement. It therefore has to be fully justified before we consider it an appropriate action to undertake.

It follows from this that sampling is never a "best" option in appraisal, and should be considered for implementation only in the most exceptional circumstances. It should not be used as a convenient way to avoid difficult appraisal decisions relating to the research value of the records, or as a means of retaining some of the records "in case" they have research value. In the Australian Archives, sampling is an alternative only to complete retention and not to destruction.

FOOTNOTE

1. See for example, "The use of sampling techniques in the retention of records: A RAMP study with guidelines" (1981) prepared by Felix Hull.