BACKGROUND FOR AN ADMINISTRATIVE RECORD CENSUS

Wendy Alvey and Fritz Scheuren, Internal Revenue Service

> And it came to pass in those days that there went out a decree from Caesar Augustus that all the world should be taxed... And all went to be counted, every one unto his own city...
> (Luke, Chapter 2)

Throughout history, including during the Roman Empire, the concept of a census has often been closely tied to the collection of revenues for tax purposes. In fact, the English word "census" comes from the Latin "censere," meaning "to enroll, tax, or assess" [1]. In these times of financial austerity, the idea of using tax records in the decennial population census is resurfacing as a possible viable alternative to what the General Accounting Office has estimated might otherwise be a $4 billion effort for 1990 -- up from $1 billion for 1980 [2].

In order to keep those costs from continuing to escalate at such a rapid rate, a whole new orientation towards the traditional decennial census may be in order. First of all, the necessity for direct enumeration should be carefully reconsidered. Are there other means of counting the population? For instance, much of the personal data actually collected every ten years (or items quite similar) are already available from existing microdata systems established for purposes of Federal program administration. Perhaps these files could serve as an alternative source of basic census information.

Second, traditionally accepted orientations with regard to census data may need to be reassessed. For example, are current conceptual definitions the most useful regardless of the approach employed? The "household" has been the longstanding basic unit of the conventional census. Part of its appeal, however, is related to the way the measurements of population are taken. Under different circumstances, some other definition, such as an income tax filing unit, might prove useful.

Finally, in order to reduce costs, the importance of each data item now collected should be reweighed. Are all of the items on the current schedule necessary on a 100 percent basis? or would sample responses be adequate? If so, how large a sample? All of these issues, and many more [3], are being considered by the U.S. Bureau of the Census as part of the decennial planning effort.

Although not members of the Census Bureau, we felt it might be worthwhile to examine one possible alternative that our experience suggested to us: the administrative record census. In particular, we would like to provide some background information for exploring an approach which begins with existing administrative record systems, instead of direct enumeration, to conduct a basic population census. The research issue this raises, of course, is whether or not such an approach would lead to a successful "count" of the number of individuals in the country by prescribed geographic area, as defined by law, while obtaining limited demographic information for the whole population.

More specifically, the proposal is to link Internal Revenue Service (IRS) data (Forms 1040 and 1040A) to wage and retirement earnings (Forms W-2 and W-2P), unemployment compensation records (Forms 1099UC), and Health and Human Services' benefit (Old-Age, Survivors and Disability Insurance, Supplemental Security Income, and Medicare) administrative files to obtain a "bare bones" population census. It is conceivable that, by so doing, cost reductions (in terms of money, respondent burden, and timeliness) could be achieved, while still satisfying the Constitutional mandate, obtaining basic enumeration by local area, and collecting some "characteristics" on small area units. It is conjectured that a 98% coverage rate could be expected. Evaluations and improvements in the coverage and content could be attempted with supplementary canvasses. Undercount adjustments or imputations, while controversial, might also be relied on to complete the enumeration.

The paper examines in a general way the Federal administrative record systems that might be employed in conducting a population census. The focus of the description given is on the extent to which these systems could be used together to obtain population census counts. Sections I through V describe the various microdata files, touching on coverage and content issues. Section VI discusses the basic methodology proposed, raising some of the major linkage issues to be considered. Finally, in Section VII, an agenda for researching the proposal is suggested. Supplemental tables and facsimiles of forms are available upon request.

I. THE INDIVIDUAL INCOME TAX MASTER FILE

The law requires that all U.S. citizens and resident aliens, regardless of age, must file an individual income tax return if their income was above the prescribed level or, regardless of their income, if they owed such taxes as Federal Insurance Contribution Act (FICA) tax on tips not reported to an employer, penalty tax on premature distributions from an Individual Retirement Account, "minimum tax," or tax from recomputing a prior year investment credit [4]. These rules also apply to nonresident aliens and resident aliens married to citizens or residents of the United States at the end of the prior year who filed a joint return. In addition, there are two main reasons why Forms 1040 and 1040A are also filed by persons who do not have to file: (1) if they had taxes withheld and they want to collect a tax refund, or (2) if they want to qualify for a refund based on an earned income credit.

Processing

The individual income tax returns are mailed by taxfilers to one of the ten IRS service centers throughout the country, where revenue processing begins. This involves sorting the returns; checking for missing schedules, forms, signatures, etc; editing items for keying; data entry; and math-verification [5]. Once the detailed information from each return is on tape, weekly transaction files are sent to the main IRS computer center in Martinsburg, West Virginia, where they are subjected to further testing.

Much of this additional testing involves doublechecking the entity data (composed of name, address, and social security number or SSN). One of the hundreds of tests run concerns zip code analysis. First, the five digits of the zip code and the city and State parts of the taxpayer's address are compared to a computerized zip code directory. If there is no agreement at the five-digit level, then testing is carried out on the first three digits. If no agreement is found there either, then the first three digits of the computerized zip code for that city and State are picked up instead.

Another test conducted at this stage in the processing involves the validation of name and social security number, as provided on the tax return. To do that, the transaction tape name and SSN are run against the "Data Master One" (DM-1) file of validated account numbers provided to IRS by the Social Security Administration (SSA). (The IRS version of the DM-1 is updated quarterly by IRS to reflect changes, additions and deletions--e.g., name changes due to marriage, new SSN's issued, decedents, etc.--received weekly from SSA.) For primary taxpayers [6], returns with no SSN and SSN's which fail to validate are sent for microfiche research. This involves searching IRS files up to three years old for a match on name and, if partial number is present, number. If a match is made and a number is present, it is picked up. If no number is found there, a temporary SSN (beginning with 9 or T) is assigned and the record is put on an invalid file. The service center is then asked to attempt to obtain an account number from the taxpayer.

Once testing is completed, the transaction tapes are used to update the IRS Individual income tax Master File or IMF. For each taxpayer, the IMF contains an entity section and a number of tax modules. The entity section has the taxpayer identification information, which is updated each year with the validated identifying information just described. Each tax module contains an extract from the transaction tape data for a given tax year. The updating is accomplished by matching the SSN and the first four letters of the surname and then posting the transaction tape information to the current IMF file. All valid account numbers not previously posted to the IMF -- i.e., first-time filers -- are posted, as well. All invalid and temporary social security numbers are kept on a separate invalid file, where the data are retained until the SSN has been corrected and validated. Revalidation with the DM-1 file is carried out quarterly, to purge the now-postable cases from the IMF invalid file.

Coverage

For Tax Year (TY) 1979, the period which essentially coincides with the 1980 census, 92.7 million Forms 1040 and 1040A were filed [7]. This represents 224.7 million exemptions. It is these exemptions, rather than the number of returns filed, which give a count of the number of individuals actually covered by the tax filing system.

There are, however, some adjustments that have to be made to compensate for over- and undercounting. First, extra exemptions for age (65 or older) and blindness have to be subtracted out. For TY 1979 there were 11.3 million old age exemptions and 173,000 blindness exemptions. That brings the total number of exemptions other than age and blindness to 213.2 million.

Figure 1.--Exemptions for All Individual Income Tax Returns, TY 1979

Total exemptions	224,691,732
Exemptions for age 65 or over	11,322,713
Exemptions for blindness	173,096
Exemptions other than age and blindness	213,195,923

Adjustments are also needed to eliminate duplicate counting of dependents with unearned income (i.e., individuals -- usually minor children -- who filed a return for income from interest and dividends, but who were claimed as dependents on another's return). As shown in Figure 2, about 2.1 million such cases exist. That brings the total number of adjusted exemptions to 211.1 million.

Figure 2. -- Adjustments for Exemptions of Dependents with Unearned Income, TY 1979

Total exemptions, other than age and blindness	213,195,923
LESS	
Exemptions for dependents with unearned income	2,053,760
Exemptions claimed on Forms 1040	513,440
Estimated exemptions claimed on Forms 1040A	1,540,320
EQUALS	
Adjusted total number of exemptions	211,142,163

NOTE: Information on the number of exemptions for dependents with unearned income is not available from Forms 1040A for TY 1979. It was, however, for TY 1976. Therefore, we assumed that, to make the estimate, the same ratio of 1040/1040A exemptions observed for TY 1976 would hold for TY 1979, as well.

Similarly, adjustment is needed to compensate for overcounting of certain deceased taxpayers, dependents of divorced taxfilers and, especially, dependents with earned income. Undercounting of some other individuals, including children alive less than a year, must also be factored in.

Figure 3. -- IRS and Census Coverage Comparison Before Final Adjustments

Estimated total IRS taxfiling population, TY 1979 ----------	211,142,000
1980 Census total population --	226,505,000
IRS as percent of Census population ----------	93.22

NOTE: The IRS figure shown here still requires additional adjustment, as mentioned in the previous paragraph. Also, the 1980 Census total has not been corrected for undercount.

All in all, rough figures indicate that IRS coverage for Tax Year 1979 may be as much as 93 percent of that for the 1980 census. However, because of the need to further adjust the IRS figure down, for overcounting, and probably "correct" the census number up, to compensate for census undercounting, it is conjectured that, in actuality, IRS coverage using the IMF data would be closer to 90 percent or so of the "true" total population [8].

Content

In addition to counting the population, the census collects characteristic information for each household. For the 100 percent enumeration, the basic census form for 1980 consisted of ten demographic questions for each individual and nine housing items. An additional 22 housing questions and 24 personal items were also included on the extended questionnaire, which a sample of the population were required to answer.

Of those questions, the main ones obtainable from the IRS Forms 1040 and 1040A are the names, home address [9], list of persons in the tax filing unit, and extensive income information -- plus some data on family relationships, marital status and occupation.

It almost goes without saying that major definitional differences exist between the census data and income tax returns. The most notable of these is the basic conceptual entity (mentioned earlier): the household vs. the tax filing unit. There is, of course, considerable overlap between the two, particularly for "married, spouse present" couples, as discussed in a paper by Richard Irwin and Roger Herriot of the Bureau of the Census [10]. However, the two terms are by no means synonymous--many single tax filers live in households with other tax filers and many married persons filing separately live in the same household.

II. UNEMPLOYMENT COMPENSATION INFORMATION FILES

Unemployment compensation paid under Government programs was subject to taxation for the first time in 1979. Since then, a Form 1099UC, Statement of Recipients of Unemployment Compensation Payments, has been filed annually by each State for each individual who received $10 or more in unemployment compensation during the calendar year. These forms are filed with the ten IRS service centers. Most of the data received are provided on magnetic tape. It is estimated that only about 30,000 of the 10,505,000 1099UC's received in 1980 (for TY 1979) came in as paper documents. (Most of these paper records represent corrections to the magnetic tape files already received.)

Processing

All data on magnetic tape are forwarded to the IRS National Computer Center (NCC) in Martinsburg, to be merged with the other information returns that form the IRP or Information Returns Processing file. First, however, the returns must pass through a validation process. To do that, like the individual returns, they are run against an IRS copy of Social Security's DM-1 file of validated account numbers. Returns with no SSN and SSN's which fail to validate are dumped onto an invalid file and sent for TIN (Taxpayer Identification Number) perfection. This entails computer analysis, based on address components, to try to find a match; microfilm research, in which IRS files up to three years old are searched in an attempt to find a name and number that match; and, if all else fails, correspondence with the taxfiler to request a valid SSN. (Data which comes in on paper copy are sampled and only those documents selected are keyed and sent for validation and further processing.)

After validation and TIN perfection, the 1099UC's are sent to underreporter analysis, where they are matched to the Individual Master File to see if a tax return was filed for each individual. If so, the return is checked to see if the unemployment compensation was reported by the individual taxfiler on Line 20 of his Form 1040. If a mismatch occurs (i.e., the line item is blank and a 1099UC is present, or visa versa) or a discrepancy exists in the amount reported, the return undergoes further underreporter processing. This entails validating the identifiers, to be sure that the returns are for the same person; checking the 1040, to ensure that the income was not reported under another line item; and, if needed, corresponding with the taxfiler, to obtain either a valid explanation or the underreported taxes. Similar follow-up is carried out for nonfilers with 1099UCs.

Coverage

Virtually all of the State Unemployment Compensation recipients are covered on IRS' Information Returns Processing file. It is speculated, however, that the addition of these cases to the IMF population can be expected to

augment the administrative record data base by a significant number (though only a small percent) of cases. The reason for this is that most insured unemployed workers are already filing tax returns, as illustrated in Figure 4.

Figure 4. -- Coverage of Insured Unemployed in the Taxfiling Population, TY 1979

Total taxfilers with Form 1099UC --	9,799,038
Number who filed a tax return ---	9,112,064
Number who did not file ---------	686,974

Content

Forms 1099UC contain full name, mailing address, social security number, and total amount of unemployment compensation paid to each individual by each State. Therefore, the addition of these records to the administrative record file does not add substantially to the content items included in the traditional census. Its use in augmenting the tax filing population, however, thus improving overall coverage, makes the relatively simple task of including these records worthwhile. Furthermore, for those individuals who had also filed a tax return, comparison of the entity items is useful as a verification measure.

III. WAGE AND PRIVATE PENSION FILES

The law requires that a Form W-2, Wage and Tax Statement, must be filed by each employer for each employee to whom any of the following items apply:

(a) income tax or social security tax was withheld;
(b) income tax would have been withheld had the employee not claimed more than one withholding allowance;
(c) wages of more than $600 were paid; or
(d) services were received by a trade or business in exchange for payments of any kind, including noncash.

Forms W-2P, Statement for Recipients of Periodic Annuities, Pensions, Retired Pay, or IRA Payments, are required to be filed by employees' trusts or funds; Federal, State or local governments; life insurance companies; and other payers of such payments. A Form W-3, Transmittal of Income and Tax Statements, is a summary form which accompanies Forms W-2 and W-2P.

Processing

Wage and Tax Statements are required to be filed by February 28th of each calendar year. They are processed for IRS by the Social Security Administration. Employers are encouraged to submit these forms on magnetic tape; however, smaller employers often transmit them in paper form.

All paper copies are sent to SSA's data operation centers, where they are put on microfilm, a copy of which is forwarded to the Internal Revenue Service for general revenue purposes. The W-2's and W-3's are then keyed, either through optical scanners or, if unreadable, manually, and put on tape. This tape copy of the employer wage reports is merged with the W-2 and W-3 data received on magnetic tape. A file of self-employment earnings data from Forms 1040 Schedule SE, Computation of Social Security Self-Employment Tax for individual income tax filers, provided by IRS, is also merged at that time.

The merged files then go through a balancing and validating process. Taxable earnings fields on the W-2's and W-3's are summed to balance within a tolerance. Federal Insurance Contribution Act and non-FICA items on the Forms W-2 are also validated for IRS at that time. (As the self-employment data come from IRS originally, they are not passed through the validation process by SSA.) The validated file is then sent to IRS for use in revenue processing.

Coverage

The W-2 and W-2P population is, in large measure, included in the group which appears on the IMF, namely those who filed tax returns for the previous calendar year. There is, however, a small increase in coverage among those who received wages but were not required to file and did not apply for a refund.

Figure 5. -- Coverage Overlap of W-2 and Taxfiling Population, TY 1979

Number of W-2's filed, TY 1979 [11] ----------	175,328,000
Number of returns filed, TY 1979 ----------	92,694,302
Estimated number of returns with W-2's ----------	82,056,000
Estimated number of returns with one W-2 ----------	40,780,000
Estimated number of returns with two or more W-2's ---	41,276,000

NOTE: TY 1974 was the last year for which data on W-2's were published in the Statistics of Income series. Estimates for TY 1979 are based on the assumption that the ratio of returns with TY 1974 wages and salaries to TY 1979 wages and salaries would hold for returns with W-2's, as well. Complete W-2 data for TY 1979 are currently being matched to the IMF and counts should become available later this year.

Content

Forms W-2 contain name, address, SSN and wage information for each employee, all items available on their tax returns. They also contain employer information--name, address, and employer identification number. These items are most useful in improving the occupation data provided by the primary tax filer (and spouse) on their Form 1040 or 1040A. This could prove especially beneficial when more than one job is involved. Information on locality is also available, which might be helpful for geographic

coding (including migration and journey-to-work analyses). Forms W-2P contain payer information, plus recipient's name and SSN.

IV. SUMMARY EARNINGS RECORD FILES

The Social Security Administration has a summary earnings record (SER) on file for each individual to whom a social security number has been issued since 1937. In order to apply for an SSN, a Form SS-5, Application for a Social Security Number, must be completed. These applications are received at the local Social Security District Office and transmitted to the main SSA office in Baltimore, Maryland. There, they are keyed and an SER is created, which is added to the file of all account number holders. This file is then used to post earnings information for all reported social security covered earnings and serves as a basis to establish entitlement to the various Federal Insurance Contribution Act programs.

Processing

Since 1978 social security covered earnings have been reported by employers, through the Forms W-2 and W-3, described above. After the identifying information on the balanced W-2's has been verified, as detailed in the previous section, earnings data are posted to the SER's. This is done by matching the wage reports to the earnings records, using social security number and the first six characters of the surname. Any wage activity reported during that year for social security covered earnings is recorded on the SER for each individual. However, no non-FICA items are posted. If unpostable, the data are stored on a "suspense" file and correspondence with the employer is initiated to obtain valid identifying information.

Posting to the SER file is done on a flow basis, several times a year. While the filing of wage reports is required by the end of each February, intermittent reporting occurs throughout the year, due to terminations, changes, and corrections. On the average, current lag-time in posting to the earnings files is running about 12 to 18 months.

Coverage

Since the administrative record census is based on the social security number as the primary matching key and since all SSN holders are present on the SER, the earnings record file is a virtual duplicate of the sum of the proposed linked administrative record systems. With the possible exception of some children, who may not be represented on any of the matched files, the addition of these records should not serve to augment the covered population. The demographic items available from the SS-5, however, do add to the content portion of the administrative record census. Therefore, the inclusion of the SER is of considerable importance.

Over 263 million social security numbers have been issued between 1937 and 1979, all of which are found on the SER. To date, no social security numbers have been intentionally reissued. Therefore, the SER file represents all persons to whom an account number was ever issued -- living and dead.

Since social security numbers are now required as identification numbers for the civil service and the armed forces, as well as for all covered employment, virtually all of the U.S. adult population has an SSN and, hence, an SER. In fact, in addition to child beneficiaries, tax advantages and requirements by some school districts have resulted in increased use of SSNs for persons under age 16. Approximately 36 percent of individuals under 5 years of age are currently on the SER file, as are virtually 100% of those 18 years or older.

There are, however, some coverage issues which must be addressed in using these earnings records. The main one is a problem of overcounting in two areas: decedents and multiple account number holders. Good estimates are needed to adjust for their over-representation on the SER file. Adjustments are also needed to compensate for undercounting caused by persons who use someone else's SSN instead of obtaining their own [12].

Content

The linkage of the SER (and, hence, the SS-5 system) to the administrative record census introduces several content items that are not available on the other administrative record files. In addition to name and social security number, the SER (SS-5) file contains sex; race/ethnic description; age at time of application and date of birth; city and State or country of birth; and maiden name (useful for matching purposes). The file also contains longitudinal earnings history data for all FICA-covered employment.

V. MASTER BENEFICIARY RECORD FILE

In addition to the summary earnings record file, SSA maintains several other administrative record files. The most notable of these is the Master Beneficiary Record file (MBR). Prior to 1977 the MBR contained a record on file for each individual who had been awarded a monthly old age, survivors, or disability insurance (OASDI) benefit. In 1977 the system was revised to include all persons for whom an application for such social security benefits had been filed. Furthermore, since 1974, "interface" records [13] have been added to the MBR for individuals whose sole benefits come from Medicare, Railroad Retirement, Supplemental Security Income or the Black Lung program. Also present on the MBR are individuals whose only survivor's benefit was the award of a one-time lump-sum death payment to help defray funeral costs.

Processing

In order to create an MBR record an individual must contact the local Social Security district office to file a claim. The district office completes an SSA Form 450, providing basic identifying information, such as

name, SSN, sex, and date of birth. The SER which is extracted is then used to establish whether the filer is eligible for benefits on the claim. This decision is based principally on the number of years of social security covered earnings accrued. Once the SER has been drawn and insurability established, the MBR is created. If the claimant is eligible, beneficiary data are entered onto the record. If the claim is denied or disallowed, the reason for disallowance is noted. Posting of new benefit records and of changes to the existing benefit file is done on a daily basis.

Coverage

The MBR, like the SER, is a permanent longitudinal file. As of December 1979, it contained data on 77.9 million OASDI beneficiaries, including terminations, denials, and disallowances. (Only about 35.1 million of these were in current pay status, however.) The MBR also contained interface records for 607,000 railroad retirees, 21,000 Black Lung beneficiaries, 1.9 million Supplemental Security Income recipients, 966,000 active Medicare cases, and 2.9 million "lump sum only" claimants.

Because social security benefits are not subject to the income tax, many OASDI beneficiaries do not have to file tax returns. Hence, including the MBR in the administrative record census increases the coverage substantially, especially for older people, as shown in Figure 6. In fact, for persons age 65 or over, the file is virtually complete, since Medicare establishes a record on file for each individual (about 30 million in December 1979) who has an SS-5 age of 64 years and 9 months, in order to notify them of their eligibility for health insurance benefits. On an overall basis, therefore, it is speculated that total coverage may grow by 5 percent or more, since less than half of the population age 65 or older are taxfilers and many of the social security beneficiaries under 65 may be nonfilers.

Figure 6. --Coverage Overlap of Master Beneficiary Record and Taxfiling Population

Beneficiaries in current payment status, December 1979	35,124,856
Less than 65 years old	11,892,022
65 years or older	23,232,834
Exemptions for age 65 or over on 1040's and 1040A's, TY 1979	11,322,713

NOTE: Not shown in these figures are coverage improvements which would be obtained from the inclusion of other groups on the MBR, notably those Supplemental Security Income recipients who are not also receiving social security benefits.

Content

For the most part, the Social Security benefit files will provide additional income items, disability information, improved age data, more timely notification of death, and an up-to-date mailing address. They also establish a tie-in for auxiliary beneficiaries who are collecting payments under the SSN of the primary beneficiary and have none of their own (e.g., a widow who never worked and is receiving benefits based on her deceased husband's past earnings).

VI. METHODOLOGICAL ISSUES

The basic methodological approach proposed involves exact data linkage of the income tax returns, earnings, and benefit files just described, followed by evaluation procedures and supplementary activities, to augment the coverage of the matched administrative record system. This section focuses on the first phase of that effort: the initial data linkages.

Basic Assumptions

Several basic assumptions underlie this proposal:

(A) First and foremost is the presumption that the administrative records can be put into a one-to-one correspondence with all living persons in the United States at a given point in time and at a specific geographic location. Needless to say, that hypothesis is false, if for no other reason than the fact that it is known that the files being used do not cover the entire population. However, other experience with administrative record linkages suggests that it is reasonable to assume that this holds true for the bulk of the population [14].

(B) Due to the nature of the administrative record systems involved, compliance is good -- a reasonable supposition based on what we know about the taxfiling population [15];

(C) There is only one person for each active SSN to be matched -- there is reason to believe that, with the exception of a small number of cases involving fraud, use of more than one social security number by an individual in a single year is very infrequent [16];

(D) Only modest changes in the administrative record systems are needed to carry out this proposal and they can be achieved at relatively low cost -- such as minor changes in the tax forms, possible reassignment of the "Census Day," and certain legislative amendments to clarify the Bureau of the Census' access to the various files for these specific research purposes; and

(E) That the resulting data would be adequate to meet the Constitutional mandate; i.e., that the administrative record approach, as supplemented, would be considered "enumeration."

Basic Methodology

Since over 90 percent of the U.S. population is covered on IRS individual income tax returns, the Internal Revenue Service's Individual income

tax Master File would be the logical core for an administrative record census. Other files would then be matched in to increase the covered population. This would be done by applying the kinds of exact matching techniques employed in joint Census, Social Security, and IRS studies conducted over the past 30 years [17]. Data from one file would be linked to that for the same person on another file, using social security number as the primary matching key. If a match occurs (i.e., the same individual is verified as being present on both files), the record will be appended to the 1040 data for that individual. If no match exists, a new record will be created on the administrative record system, subject to all subsequent matches. Confirmatory information on the files will be used to guard against mismatching (linking records for apparently different people with the same SSN).

Linking Issues

In order to link these files, however, several major matching issues need to be resolved. These include data access, matching rules, coverage, content, completeness, timing, and, especially, geographic coding. Each of these points is touched on below.

Access. -- First of all, confidentiality considerations need to be addressed. All of the agencies involved have very strict rules governing the privacy and confidentiality of their identifiable record files. While provisions are already written into the regulations under the Internal Revenue Code to provide the Census Bureau with certain specified tax return information, legislation might be needed to give them broader access to the tax return data in order to conduct an administrative record census. The Census Bureau use of such files would also, of course, have to adhere to the disclosure safeguards set forth in the Internal Revenue Code [18]. Similar arrangements would be needed to assure the Census Bureau access to Social Security Administration data.

Matching Rules. -- Another important linkage issue is that of how the actual match will be effected. Since social security number is present on all files to be matched, that is the logical starting point for linking records for the same individuals from different files. In order to obtain an accurate count for dependents, however, SSN's would be needed for them, as well. This would require a change in the tax forms, to request numbers, if available, for each of the dependent exemptions provided under item 6d of the TY 1979 Forms 1040 (or 5d of Forms 1040A). These and other form changes need to be explored, pretested, and resolved.

Next, matching procedures would need to be established to specify agreement rules for confirmatory variables, in order to ensure that the correct individuals have been matched. Name and address are the most universal items available to verify the matches. Race, sex, and marital status could also be used, if present on both files. A scheme to establish the quality of the matches would, then, be needed, so as to provide some basis for estimating the likelihood of mismatching [19]. Some of this, of course, is already being done routinely as part of Social Security and IRS administrative processing procedures. Their rules, however, may not be stringent enough to meet the needs of the Census Bureau in conducting an administrative record census.

Coverage Adjustment. -- Not only do linkage procedures and matching rules need to be designed to prevent overcounting due to duplicate representation among the different administrative record systems being matched; procedures are also needed to compensate for over- and undercoverage within the linked systems themselves (i.e., individuals having more than one record in a particular administrative file). We have already noted, for instance, that simply counting exemptions would lead to an overcount of the taxfiling population. This is due to the fact that a person can file his own return, while being claimed as a dependent on another's return. Therefore, adjustment procedures would be needed to unduplicate those cases. This would involve matching the file on social security number and then looking at all duplicate SSN's to ascertain (using name and address) whether they refer to the same person.

Similar efforts will be needed for other administrative files, as well. For example, matching of W-2's to their respective tax returns is done routinely for administrative purposes. Hence, mismatching should not be a serious problem. However, wage reports filed by different employers for the same employee may present a linkage concern, especially where confirmatory variables are inconsistent (e.g., name has changed, address is different, etc.). The same holds true for individuals with more than one Form 1099UC, such as those who have suffered more than one period of unemployment or collected unemployment compensation from more than one State.

A slightly different coverage problem arises when using the Social Security data files. Since, as noted, both the Summary Earnings Record and Master Beneficiary Record files are permanent systems, they contain information on decedents, as well as living individuals. For records on file prior to 1977, dropping the decedents is a fairly straightfoward operation. Up until then, the date of death was posted to the SER. After that date, however, a death indicator is not provided on the earnings record. There is, instead, a death indicator present on the MBR for all decedents for whom a claim has been made. Hence, most of the decedents can be removed from the "countable" population by dropping records for all those who died prior to the Census Day [20]. Further adjustments can be made for those for whom a claim was not filed (mostly children and female nonearners) by employing date of death information provided on the Form 1040. (IRS is notified of the death of each taxpayer on a tax return.)

Content. -- The three most important content

issues which need to be explored are differences in concept, definition and actual item availability. Notable among these conceptual differences is "address," mentioned earlier; on administrative files this is most likely to be a mailing address as opposed to residential address. In some cases compensation -- possibly through the use of tax form changes along the lines of the revenue sharing question which appeared on the 1980 tax returns -- could be used to overcome these differences. In other events, the user would have to be made aware of the impact such changes will have on the data.

Similarly, differences in definition exist between current census data and those collected for administrative purposes. Take ethnicity, for instance; information on hispanic origin might be based on hispanic surname, rather than on a respondent's reply and census-taker observation. Nuances of this sort would need to be clarified [21].

Finally, some of the characteristic data made available by an administrative record census (such as age and income items) are probably more complete and reliable than has traditionally been true of the decennial census. However, there is bound to be concern for the impact of missing items -- those not collected for administrative purposes (e.g., education attainment information). Supplementary activities could be used to obtain these additional characteristic items through the use of sample surveys. (The Census Bureau could even explore the possibilities of increasing the sample size or item content locally on a reimbursable basis.)

Completeness. -- Two completeness issues need to be examined. First what is the overall coverage under the proposed scheme and how might one improve it? Supplementary surveys would be needed to study the coverage achieved, perferably beginning immediately with current survey programs (especially the Current Population Surveys) already underway at the Census Bureau. To improve on the current proposal, further data linkages could be considered with such files as the National Death Index, the National Welfare Index [22], or such State- or locally-based systems as welfare rolls, drivers license files, and school system enrollments. The second question to be addressed, naturally, is: what is the completeness of the characteristic data within the matched files? For example, a dependent without an SSN will have less demographic information than someone whose SS-5 record is available for matching. Similarly, a Medicare recipient who is not required to file a tax return may not have residential address information (assuming a change in the tax form to request that data). In such cases, supplementary mailings could be employed to improve the completeness of the item content.

Timing. -- Another major concern in conducting an administrative record census is the problem of timing. In recent decades, "Census Day" has been April 1st; "Income Tax Day" is April 15th. However, income tax returns can be filed as early as January and, with extensions, even as late as a year or two later. This problem, plus the difference in reference periods -- tax returns are filed for the previous calendar year [23]-- raises the possibility of changing the Census Day -- perhaps to January 1 -- so that, at least, the same reference period is being covered. Then, in order to compensate for lag-time, pre-Census Day activities could begin by linking the administrative record files for the previous year. As returns come in, they could be matched in to update the pre-Census system and supplementary activities would only be needed for those who had moved, died, or been born into the population since that time.

Geographic Coding. -- Perhaps one of the areas of greatest concern in using administrative record data to conduct a population census is the need to assign geographic codes down to the tract or block level, for purposes of political redistricting. A major part of this problem could be "solved," (again, as noted earlier) through a tax form change similar to that introduced to obtain revenue sharing information in 1980. Residential address and locality of residence could be obtained. Nine-digit zip code (which is expected to become available sometime during the current decade) could, then, be used to "geo-code" down to blocks. While conceptually satisfactory, this approach has many implementation difficulties to overcome. For example, problems would exist with the information provided, particularly in rural parts of the country. More traditional census techniques may be needed to obtain adequate address information in those areas. Further research is needed to get a "feel" for how extensive these geo-coding problems are, particularly in the rural South and West.

VII. FURTHER RESEARCH

A lot of questions remain to be answered if an administrative record census is to become a reality. Certainly, pretesting and evaluation of any approach under consideration for 1990 needs to begin now. In fact, it may already be too late to implement an effort as vastly unconventional as an administrative record census, at least on a full scale, in time for the next decennial enumeration.

As described in the last section, a few of the areas needing exploration include the technical feasibility of such a data linkage, the form and procedural changes necessary to carry out an administrative record census, and the impact of differences in content and coverage on the resulting data and their uses. Access, timing and, especially, geographic coding issues need to be resolved, as well. Finally, the various supplementary activities required to augment the population count and the accompanying characteristic information have to be devised, pretested, and evaluated. Perhaps the best starting point for addressing many of these issues may be to incorporate the needed research into the redesign of the Census Bureau's ongoing survey efforts [24]. Detailed comparisons to the last decennial census would

provide another basis for assessing how well this proposal accomplishes the goals it sets out to achieve [25].

Needless to say, an administrative record approach to the decennial population census is, unquestionably, quite a radical departure from tradition. Perhaps this paper has set the groundwork, however, for concluding that the idea is not as farfetched as it may sound [26]. The paper has also attempted to sketch the many difficult implementation problems that will need to be addressed.

In conclusion, we feel that, given continuing tightened budget constraints and rapidly rising costs, the administrative record population census may be an idea whose time has come. At any rate, it is certainly encouraging to note that the Bureau of the Census has seen fit to consider this, along with other alternatives, in their planning efforts for 1990 [3].

ACKNOWLEDGMENTS

The authors would like to take this opportunity to extend special thanks to Richard Irwin, Bureau of the Census; Warren Buckler and Erma Barron and their staffs, in Social Security's Division of OASDI Statistics; and Roger Hicks, Health Care Finance Administration. Within the Internal Revenue Service, we would like to thank the Information and Miscellaneous Returns Branch, Returns Processing and Accounting Division; Dale Gustavson, Information and Returns Processing Branch, Management Systems Division; and the Individual Statistics Branch, Statistics of Income Division, for their assistance in researching this paper. Much appreciation is also due to Denise R. Herbert and Nancy J. Robinson for their extensive efforts in typing this report.

NOTES AND REFERENCES

[1] Webster's New World Dictionary, College Edition, The World Publishing Company, 1962.

[2] General Accounting Office, "Report to the Congress by the Comptroller General of the United States: A $4 Billion Census in 1990? Timely Decisions on Alternatives to 1980 Procedures Can Save Millions," Washington, DC, 1982.

[3] Many such issues were raised by Peter Bounpane, U.S. Bureau of the Census, in his paper," Planning the Planning for the 1990 Census," presented at the Census Advisory Committee Meeting of the American Statistical Association, April 16, 1982.

[4] Internal Revenue Service, 1979 Package X: Informational Copies of Federal Tax Forms, Washington, DC, 1981.

[5] Weiner, Leonard, "What Happens to Your Return Now," U.S. News and World Report, April 20, 1981, pp. 38-41.

[6] The primary taxpayer is the first person listed on the return.

[7] See Internal Revenue Service, Statistics of Income--1979 Individual Income Tax Returns, Washington, DC, 1982, p. vii.

[8] Some further discussions of data limitations are provided in the Internal Revenue Service's Statistics of Income Supplemental Report--Area Data for Individual Income Tax Returns, 1974, pp. 440-441.

[9] It should be noted that, while the income tax return specifies "present home address, city, town or post office, State and zip code," the information provided is often mailing address rather than residential address.

[10] See "An Initial Look at Preparing Local Estimates of Households and Household Size from Income Tax Returns," in the 1982 American Statistical Association Proceedings, Section on Survey Research Methods, by Richard Iwrin and Roger Herriot, U.S. Bureau of the Census.

[11] For purposes of program administration, there were a total of 188.3 million Forms W-2 and W-2P (13 million) received in time for Information Returns Processing correlation. This includes 8.8 million duplicates. Another 4.3 million records were received after January 1981, the IRP cut-off date.

[12] Intentional multiple use of SSNs is particularly a problem in the case of illegal aliens, who need an account number in order to work but cannot obtain one because of their improper residence status. The most important cases where more than one person is unintentionally using the same SSN arise because social security numbers have been employed on occasion by advertisers in promotional schemes. Perhaps the best known such instance is the number 078-05-1120. It first appeared on a sample social security number card contained in wallets sold nationwide in 1938. Many people who purchased the wallets assumed the number to be their own. It was subsequently reported thousands of times on employer's quarterly reports; 1943 was the high year, with almost 6,000 wage earners listed as owning the number. Even today the number is still being reported by a few people. (For more information on the role of the social security number in matching administrative and survey records, see the session by that title in the 1974 American Statistical Association Proceedings, Social Statistics Section, pp. 126-156.)

[13] An interface record is a short record for non-OASDI beneficiaries entitled to benefits under one of the other programs administered by Social Security.

[14] For a review of the literature up to 1975, see the bibliography provided in

Studies From Interagency Data Linkages, "Report No. 4: Exact Match Research Using the March 1973 Current Population Survey -- Initial Stages," Social Security Administration, July 1975. For more recent work on record linkage, see also "Generalized Iterative Record Linkage," by Martha E. Smith and John Silins, in 1981 American Statistical Association Proceedings, Section on Survey Research Methods and "Development of a National Record Linkage Program in Canada," by Martha E. Smith, in the 1982 American Statistical Association Proceedings, Section on Survey Research Methods.

[15] While the General Accounting Office report on "Who is Filing Income Tax Returns? IRS Needs Better Ways to Find Them and Collect Their Taxes," (GGC-79-79, July 11, 1979) points out some problems with nontaxfilers, the widespread use of social security numbers and the requirement for employers and States to report wages (on W-2's) and unemployment compensation (on 1099's), respectively, make this assumption a reasonable one. For some information on estimates of illegal aliens -- one of the groups for which it is unclear how adequately they would be counted in an administrative record census -- see also "Counting the Uncountable Illegals: Some Initial Statistical Speculations Employing Capture-Recapture Techniques," by Clarise Lancaster and Fritz Scheuren, in 1977 American Statistical Association Proceedings, Social Statistics Section.

[16] While some cases exist where a person may use more than one SSN in the same year, administrative procedures exist which lead to eventual cross-referencing in most cases.

[17] The Social Security Administration's series on Studies from Interagency Data Linkages documents work on the 1973 CPS-IRS-SSA Exact Match Study. For a summary of that project, see also "The 1973 CPS-IRS-SSA Exact Match Study: Past, Present, and Future," by Beth Kilss and Fritz Scheuren, with Faye Aziz and Linda DelBene, in Policy Analysis with Social Security Research Files, Social Security Administration, 1978, pp. 163-194.

[18] Complete Internal Revenue Code of 1954, September 1, 1981 Edition, Prentice-Hall, Inc., Englewood Cliffs, NJ.

[19] To get a flavor of what is required here, see "Fiddling Around with Mismatches and Nonmatches," by Fritz Scheuren and H. Lock Oh, in 1975 American Statistical Association Proceedings, Social Statistics Section.

[20] For more information on the quality of death reporting for Social Security administrative record files, see "Mortality Coverage in Social Security's Earnings and Benefit Record Systems," by Linda DelBene and Faye Aziz, in 1980 American Statistical Association Proceedings, Section on Survey Research Methods and "Further Investigation into Mortality Coverage in Social Security Administrative Data," by Linda DelBene and Faye Aziz, in the 1982 American Statistical Association Proceedings, Section on Survey Research Methods.

[21] Some recent work on hispanic origin is reported on by Jeffrey S. Passel and David L. Word, U.S. Bureau of the Census, in "Constructing the List of Spanish Surnames for the 1980 Census: An Application of Base Theorem," presented at the 1980 annual meetings of the Population Association of America in Denver, Colorado.

[22] The National Welfare Index is a system proposed by Presidents Nixon, Ford, and Carter, as a means of centralizing data on State welfare recipients. At the present time, welfare information is maintained on separate systems by each of the 50 States and the District of Columbia.

[23] Exemptions on tax returns can be claimed for anyone alive between January 1 and December 31 of the previous calendar year, with dependents required to meet tests for income, support, married dependents, citizenship, and relationship, as detailed in the accompanying instructions for Forms 1040 and 1040A. This means that a person who lived only a few minutes during that year can be claimed as a dependent and would be countable in an administrative record census.

[24] Administrative record matching can profitably be integrated routinely into the estimation used in the Current Population Survey (CPS). Such an endeavor would not only reduce the mean square error of CPS employment (and probably unemployment) statistics, but it would provide a natural link with the administrative records used by the Census Bureau to make local area estimates for revenue sharing and other purposes during the intercensal period. (For some additional details, see Scheuren, et al., Studies from Interagency Data Linkages, "Report No. 10: Methods of Estimation for the 1973 Exact Match Study," Social Security Administation, January 1981.)

[25] Siegel, J.S. and Jones, C.D., "The Census Bureau Experience and Plans," Conference on Census Undercount, 1980, pp. 15-24.

[26] The Economic Censuses, it might be noted, have relied on a mixture of administrative and survey data for over two decades.