V. DESCRIPTION OF SPECIFIC TASKS

The following tasks are necessary to develop the March 2000 CPS-Florida FSP linkage:

- Agency Negotiations **(Task 1)**;
- Tabulations on Matching Elements **(Task 2)**;
- Complete Census Proposal Process **(Task 3)**;
- Negotiate Access to State(s) Administrative Files **(Task 4)**;
- Establish Memorandum of Understanding (MOU) to Access and Match the CPS and Florida FSP Data **(Task 5)**;
- Develop Computer Space at Research Data Center (RDC) **(Task 6)**;
- Extract and Transfer State FSP Administrative Data **(Task 7)**;
- Extract and Gain Familiarity with the State Administrative Data Files at Data Access Center **(Task 8)**;
- Deterministically Link State Administrative Files to CPS Files **(Task 9)**;
- Assess the Reliability of the Match and Develop Final Files **(Task 10)**;
- Data Access **(Task 11)**;

The agency negotiation task (Task 1) represents a planning stage for all of the agencies that will be involved in this initiative. Upon completing these negotiations, the next step is to generate basic tabulations on data elements in the CPS and state administrative files (Task 2). These tabulations will allow the USDA and Census to assess the risks of data linkage (e.g.. will we be able to successfully link records? how many records are missing?). The next four tasks (Task 3-6) outline the methodology to obtain access to restricted CPS and state administrative files and to store these files at an RDC. The data processing tasks (Task 7 and 8) highlight the important data manipulations that are necessary to become familiar with the data prior to the actual linkage. The data linkage tasks (Task 9 and 10) provide a description of the methodology to link the files and to assess the reliability of the match. Finally, the data access task (Task 11) builds on the existing agreements outlined in Tasks 3-6 and set up a system that allows restricted access to researchers.

A. Task 1: Agency Negotiations

The most important step in developing a link between the CPS and FSP data will be a set of negotiations between the USDA, Census, and state agency officials that lay the foundation for the data linking process. Our initial meeting at the Census in October 2001 (mentioned above) should provide a foundation for bringing together representatives from the Census and USDA. All participants were very interested in the possibility of creating a data linkage and agreed that further negotiations were necessary if there was enough interest for eventual implementation.

The negotiations under this task will need to address three basic questions:

- Which state data should be linked to the CPS or other Census survey (Question 1)?
- Which methodology should be used to link the data (Question 2)?
- How will the data be accessible (Question 3)?

Because several options exist, it is important for the USDA and Census to identify specific responses to these questions in the upcoming months.

We have made specific assumptions to address these questions. First, we assume that the March 2000 CPS will be linked to Florida's state FSP administrative data (Question 1). We assume that two research files will be created using a deterministic linking process (Question 2). The first will be used to address issues assess the reliability of self-reported questions in the CPS and the second will be used to address research questions related to FSP program dynamics.[10] Finally, we assume that the data will be accessible from a Census RDC (Question 3).[11]

The USDA can take three discrete steps in preparations for this task. First, using the findings from the Abt (2001) and Hotz, et al. (1999), the USDA can assess which state data files are of most interest for the data linkage, including Florida. We suggest developing a brief memo that summarizes several promising data sources. Second, to gain further background information, it will be critical to conduct teleconference calls with administrative officials in "promising states" to discuss the feasibility of using state data at a Census approved site. Third, the USDA should identify the specific research questions they would like to address with the linked data. Our summary of research questions in Section III should provide some guidance in this area.

We suggest two rounds of meetings. The USDA and Census will conduct the first meeting to identify the most promising data sources and research questions for the match. The USDA will conduct a second meeting with the Census and DCF to identify specific implementation barriers, especially those associated with the legal and technical issues in linking data, and discuss how the data could be accessible on an on-going basis. Based on the results from the meetings, the USDA (and/or another agency) would then decide whether to fund this particular initiative.

Agency	Cost
Census	$4,448
USDA	$4,448
DCF	$2,224
Total Task	$11,119

[10] In Task 10 below, we discuss the differences between these two files.

[11] Unfortunately, based on our discussions with Census, there do not seem to be any alternatives to data access to storing the data at an RDC. However, it is important to note, the memorandum of understanding, described in Task 5 below, will govern how the data can be accessed and used on an on-going basis. Consequently, it is critical that USDA and Census create a strong foundation to ensure that these data can be accessible to address research questions of interest to both agencies.

B. Task 2: Tabulations on Matching Elements

Two potential major issues could significantly reduce the value of the linkage. First, the actual data matching elements (i.e., SSN, name, and date of birth) in the CPS or state administrative records could have several duplicative or missing values. If either file contains missing or incomplete information, the quality of the match will be significantly reduced. Second, the outcome information from the administrative records may include several missing or incomplete values. While the Census can readily assess the survey elements in the CPS files (e.g., race, gender, and income), it would be important to ensure that the Florida program history elements from the state file are also complete.

To minimize the costs of a "poor linkage", we suggest calculating basic tabulations on the matching elements and outcome information in the CPS and administrative data. On the CPS side, tabulations will be necessary for the matching elements to identify the percent of missing and/or duplicate values. On the state side, tabulations will be necessary for the matching elements and the FSP histories.

Researchers and administrators could use this information to weigh the costs and benefits of proceeding with the match. For example, assume that the CPS includes information on 90% of the matching elements, whereas the state data includes information on 80% of the matching elements. Based on this information, we would approximate that 72% of the cases (0.9 x 0.8) could be matched. The USDA could assess whether this match rate "is reasonable" before committing additional funding to the project. It is possible, for example, that the match rates would be significantly higher using an alternative state database.

We assume that the tabulations for both the CPS and state files are readily available from preexisting projects. Consequently, the costs of this task are relatively low for the Census and DCF.

Agency	Cost
Census	$1,931
USDA	$0
DCF	$1,931
Total Task	$3,862

C. Task 3: Complete Census Proposal Process

To begin the linkage process, it is necessary to obtain permission to use restricted files of the March 2000 CPS data files, which include essential information on the matching elements (SSN, name, and date of birth). To address the inherent privacy issues, the proposal will need to meet the Census's "Criteria for the Review and Approval of Census Projects that Use Federal Tax Information." These guidelines, (which are summarized at - http://www.ces.census.gov/download.php?document=50) require any research project to meet prior approval from the Census. Specifically, the project must meet a host of security controls including physical and computer security safeguards, approved methods

of data transferal, site approval, oversight of personnel using the data, and approved disclosure protections applied to products.

The USDA will develop the proposal to access these files. The Census will incur smaller costs related to reviewing the proposal. Presumably, the USDA and Census will outline a strategy to meet these guidelines during their initial negotiations in Task 1.

Agency	Cost
Census	$1,779
USDA	$4,448
DCF	$0
Total Task	$6,227

D. Task 4: Negotiate Access to State(s) Administrative Files

The next step is to obtain access to the raw state administrative FSP files from the DCF. The DCF has access to several types of administrative data extracts. In many cases, the FSP extracts can also be linked to other program records. While there are several potential files of interest, the initial linkage will rely on the DCF's longitudinal individual history file, which includes a full program history of any individual who ever participated in a Florida state program since 1993.

The DCF has a formal proposal process that requires researchers to submit a letter asking permission to use the state administrative file. Based on our conversations with the DCF, authorization will require a written request to the Director of Florida's Work And Gain Economic Self-Sufficiency (WAGES) program outlining the specifics of the request and the goals of the research. In general, these goals must illustrate how the research will benefit the state's programs. For example, it is possible that a better understanding of the dynamics of FSP participants and non-participants could inform state outreach efforts.

The USDA will be responsible for writing the letter requesting permission for data access. The DCF will review the letter and process the request. Based on our conversations with the DCF, we anticipate that this proposal process should be relatively straightforward.

Agency	Cost
Census	$890
USDA	$4,448
DCF	$2,669
Total Task	$8,006

E. Task 5: Establish Memorandum of Understanding (MOU) to Access and Match the CPS and Florida FSP Data

The USDA will summarize the agreements in Tasks 3 and 4 in a MOU. The MOU is a critical component of the data linking exercise because it outlines the provisions that govern data use and access. In short, it represents a summary of the confidentiality agreements established by the Census, DCF, and USDA to use restricted files of the CPS and state administrative records. After the records have been linked, the MOU will also summarize the guidelines that researchers must follow in using the data.

The USDA, with guidance from the Census and DCF, will establish the MOU. The USDA can use a template from previous a MOU between the Census and state agencies as a guideline (see ***Appendix B*** for a sample MOU).[12]

Agency	Cost
Census	$8,895
USDA	$26,686
DCF	$8,895
Total Task	$44,476

F. Task 6: Develop Computer Space at Research Data Center (RDC)

After obtaining rights to the restricted CPS and FSP files, the next step is to establish an infrastructure to match and store the data at a Research Data Center (RDC). The Census data programs are confidential, and may be used for statistical purposes only at an RDC by Census employees or by individuals who have obtained special sworn status from the Census. The Census has established RDCs at six sites, though we assume that the primary site to house these data will be at the site in Washington DC.[13] RDCs include full security systems and access protocols for data access.

The only cost associated with this task is in developing computer space at the RDC for the data linkage and storage. We anticipate that the match will require a mainframe system to handle the large state administrative files. The final matched file could be

[12] Ronald Prevost of Census Bureau provided the template that appears in Appendix B.

[13] Currently, there are six operating RDCs: Washington, DC (Census Center for Economic Studies), Boston (Boston RDC), Pittsburgh (Carnegie Mellon University), Los Angeles (University of California at Los Angeles), Berkeley (California Census Research Data Center), and Durham (Duke University Triangle RDC). For more information on these RDC, see http://www.ces.census.gov/ces.php/rdc.

stored on a secured drive of a personal computer. The Census will establish and monitor this personal computer, including the purchase of any necessary hardware and software to manipulate and store the data.

Agency	Cost
Census	$32,603
USDA	$0
DCF	$0
Total Task	$32,603

G. Task 7: Extract and Transfer State FSP Administrative Data

The DCF will provide the state FSP history files dating back to January 1993.[14] By linking these files (using a recipient ID common across both files), a full history file can be creating that includes identifiers for the match (from the demographic file) and program history information from 1993.

The DCF will assemble the full history file and extract it to a data cartridge to a Census RDC. We expect the file to be quite large based on the large number of FSP participants that exist on the full history file. During the data extraction process, Census programmers may want to conduct site visits to discuss the tape layout with the DCF. The data could be transferred using IBM 3480 data cartridges.

Agency	Cost
Census	$0
USDA	$0
DCF	$1,562
Total Task	$1,562

H. Task 8: Extract and Gain Familiarity with the State Administrative Data Files at Data Access Center

Because of the intricacies associated in processing state administrative data, it is recommended that Census agents responsible for creating the CPS-Florida FSP linkage become familiar with the state administrative data extracts. Specifically, it will be critical to have a full understanding of the data's historical development, state practices for overwriting, purging, and archiving data, and program rules for the documentation. In addition, it will be important to understand the obstacles that other researchers faced in

[14] Specifically, the FSP history can be linked with the matching elements from the state files by linking Florida's Demographic administrative file to the Individual Eligibility full history file. The Demographic file includes a record for each person who has received public assistance and shows name, SSN, recipient ID, date of birth, race, and gender. The individual eligibility file includes a record for each month for which a person was eligible for public assistance. It includes family number, recipient ID, month, and year.

using previous state extracts, including potential issues related to data integrity (e.g., missing data, duplicative observations).

This check will involve running simple cross tabulations that check for outliers on variables from the state files. The tabulations will be cross-checked with the tabulations from the DCF tape layout.

The Census will be in charge of extracting the data and converting it into a useable format for the data linkage. Specifically, they will extract the state files and transform them into a useable format at the RDC. Florida's state FSP records are in DBS - Fox Pro, which is the database management file used in Florida. These files can be easily converted into a useable ASCII or SAS file using a database conversion program, such as STAT Transfer. The format of the state files will match the format of the restricted access CPS files, which are presumably in SAS format. The DCF will provide technical assistance on specific data elements.

Agency	Cost
Census	$13,016
USDA	$0
DCF	$2,603
Total Task	$15,619

I. Task 9: Deterministically Link State Administrative Files to CPS Files

The files will be linked using the available matching elements in both the CPS and FSP. The link will start by merging records that have a common SSN in both the CPS and FSP records. FSP state records that were in the CPS will be added to the linked files. FSP records that did not contain a CPS SSN will be excluded from the match. The next stage of the link will use the name and date of birth information to link other files that may have incorrect SSNs in either the CPS or FSP state file. SAS algorithms can manipulate name information into various forms to complete the match.[15]

The Census will be in charge of all matching activities associated with this task and will rely on the DCF to provide technical assistance on any state data questions.

Agency	Cost
Census	$17,464
USDA	$0
DCF	$4,448
Total Task	$21,911

[15] Matches can be made on names that may be misspelled in either the CPS or FSP records by using algorithms that link phonetically similar records.

J. Task 10: Assess the Reliability of the Match and Develop Final Files

Any matching procedure must assess potential problems that arise due to "false links" and "false unlinks." False links and false unlinks arise because of errors in the matching variables. For example, if a transpositional error exists in an SSN on one record, the record will remain unlinked. Alternatively, two people with similar records (e.g., they have the same last name) may be falsely linked.

Most of the methods to adjust for false links and false unlinks depend on the research question being addressed. For example, if researchers are interested in understanding the differences between FSP patterns observed in the CPS and the FSP records, then the matching procedure will likely focus on potential false unlinks. Alternatively, if researchers are primarily interested in understanding the behavior of a representative sample of FSP participants, then either imputation (using FSP information on the FSP and CPS records) or reweighting procedures can be employed to ensure that the population is representative. In addition, some methods may just rely on subjective judgements of the researcher. For example, if the date of birth in the CPS does not match the date of birth in FSP records, it may be preferable to disregard the linkage. Alternatively, researchers may choose to use a more expansive matching algorithm that uses several combinations of a SSN (to adjust for potential transpositional error) or a person's name to increase the number of matches. It is important to note, however, that the subjective choice of the researcher will likely be influenced by the content of the research question.

This task builds on the "first stage" of data linking summarized in Task 9 by generating two research files that serve potentially different purposes. The first linked file will be used by the USDA to examine representative populations of FSP participants in Florida. To assess the representativeness of the file, it will be important to compare the CPS-FSP estimates to available state administrative estimates on the overall size and composition of the FSP caseload. It is likely that reweighting and/or imputation procedures will be necessary to produce a representative sample of FSP participants in the CPS. In addition, separate weights may be necessary to produce representative samples in other years. We assume that a base weight will be created for 2000 and separate weights could be generated in the future.

The second file will be used by the Census to examine differences in FSP reporting in the CPS survey and actual participation patterns in the FSP. Hence, this match will not employ any type of imputation or reweighting procedure based on CPS FSP survey responses. Rather, researchers will use these files to examine differences in reporting across the CPS and administrative data.

The Census will coordinate activities in developing the two research files. The DCF will provide estimates of FSP participant characteristics that will be used by the Census in the benchmarking process.

Agency	Cost
Census	$24,730
USDA	$1,302
DCF	$2,603
Total Task	$28,635

1. Task Alternative: Probabilistic Matching

An alternative to the approach outlined in Tasks 9 and 10 is a more complex probabilistic matching procedure designed to improve the reliability of the data linkage. Probabilistic record linkage assumes that no exact match between fields common to the source databases will link a person with complete confidence. Instead, probabilistic record linkage calculates the likelihood that two records belong to the same person, by matching together as many pieces of identifying information as possible.

This approach seeks to limit the probability of false links and unlinks in the data matching procedure. The precision of the match improves with the addition of common data elements that uniquely identify each individual. For example, this procedure could use all of the same matching elements described above to generate probabilities of matches.

A major advantage of probabilistic matching is that it allows researchers to use consistent criteria in calculating the probability of the match. Before starting the data linkage, researchers assign "weights" for the probabilistic linkage that place more emphasis on certain direct linkages, such as SSNs, than other variables. Once the data are linked, researchers can evaluate the reliability of the match by examining the probabilities that the records are "exact matches." For example, assume that a record includes a transpositional error in the SSN that precludes an exact match. A probabilistic procedure will calculate the probability that the record should be linked based on the SSN, as well as other identifying information. In a deterministic linkage, the researcher must make a judgement of the reliability of the individual record linkage, and, hence, must reexamine several types of "mismatches." In a probabilistic linkage, however, a researcher can make a decision on the weights to assign for every match and chose to only include matches that have, say, 80 percent likelihood of being a match.[16]

The drawback of the methodology is that it is far more costly. For the purposes of illustrating costs, we assume that the Census will develop the algorithms in SAS to conduct the data match.[17] Unlike the deterministic linking process, the algorithms for probabilistic matching tend to be very complicated and labor intensive.

[16] It is important to note, however, that in both the deterministic and probabilistic linkage, a researcher must make some assumptions on the reliability of individual matching elements. Consequently, both data linking processes contain some degree of subjectivity.

[17] Census could also purchase commercial matching software, such as Automatch, for the linkage.

Agency	Cost
Census	$70,508
USDA	$0
DCF	$0
Total Task	$70,508

K. Task 11: Data Access for Linked CPS-FSP file

The resulting linked files will be accessible at a Census RDC. The rules governing data access and use will be specified in the MOU. Researchers will be able to use these files by submitting a proposal to the Census and USDA that satisfies the guidelines state in the MOU.

The Census will provide on-going support for monitoring the data and ensuring its confidentiality. These responsibilities include monitoring the data and assessing proposals to use the data. The USDA will also provide input on the proposal process. Our cost estimates for the Census and USDA both assume a one-year time frame, though it is likely that these costs will be applicable in future years.

Agency	Cost
Census	$18,375
USDA	$2,224
DCF	$0
Total Task	$20,599

1. Task Alternative: Data Access for Linked CPS-FSP file and FSP Administrative-only Extract

In addition to storing the linked data file, the USDA may consider storing the administrative extract used to create the linked file. The administrative file could be stored on the same computer as the linked file, though the Census would need to develop a larger platform to store the administrative extracts.

Researchers could use the administrative extract to examine issues that may require much larger sample sizes. For example, an analysis of a particular subgroup, such as elderly FSP participants, using the CPS-FSP file could be limited because there is not a large sample of elderly FSP participants in the CPS. Presumably, researchers could generate a large enough sample of elderly FSP participants using administrative-only information from the state file.

The costs of storing the extra file should be limited to the purchase of additional storage space on a PC.

Agency	Cost
Census	$3,000
USDA	$0
DCF	$0
Total Task	$3,000