UNITED STATES OFFICE OF PERSONNEL MANAGEMENT

An Assessment of the Accuracy of Data Elements in the U.S. Office of Personnel Management's EHRI-Statistical Data Mart

August 2013

a New Day for Federal Service

An Assessment of the Accuracy of Data Elements in the U.S. Office of Personnel Management's EHRI-Statistical Data Mart

August 2013

This document provides an assessment of the accuracy of data elements in OPM's EHRI-Statistical Data Mart, the successor system to the Central Personnel Data File.

U.S. Office of Personnel Management
Planning and Policy Analysis
Data Analysis Group

TABLE OF CONTENTS

Introduction to Enterprise Human Resources Integration - Statistical Data Mart (EHRI-SDM)

The EHRI-SDM is an Office of Personnel Management (OPM) database used to analyze the Federal workforce and support decision making on human resources issues across the Federal government. It is based on data submitted to the Enterprise Human Resources Integration Data Warehouse (EHRI-DW). Agencies make biweekly dynamics and monthly status submissions to the EHRI-DW. Dynamics submissions reflect personnel actions while status submissions are snapshots of the workforce. As submissions are received by the EHRI program office and placed in the staging area of the warehouse they are submitted to numbers quality checks and edits. Invalid file names, blank files, duplicate files, corrupted files, and data failing edits are noted. This process is overseen by the Records Management division within the Chief Information Officer's office, who is responsible for maintaining data standards for the data that gets submitted to EHRI along with maintaining data edits. OPM's *Guide to Data Standards(http://www.opm.gov/policy-data-oversight/data-analysis-documentation/data-policy-guidance/#url=Data-Standards)* and *The Guide to Edits (http://www.opm.gov/policy-data-oversight/data-analysis-documentation/data-policy-guidance/reporting-guidance/ge60.pdf)* are helpful documents in terms of understanding what data resides in the EHRI data warehouse and how data quality is assured.

Submission data in the EHRI-DW is not necessarily well suited to analysis. Primarily, corrections and cancellations to dynamics data have not yet been applied and "generated" data elements have not yet been generated. Data from the EHRI-DW is subjected to a process wherein the cancellations and corrections are applied and various data elements are generated, placing the EHRI-DW data into a format better suited for statistical analysis and then moved into the EHRI-SDM as the reporting repository. The EHRI-SDM also provides the data for OPM's online tool FedScope (http://www.opm.gov/policy-data-oversight/data-analysis-documentation/fedscope/).

The EHRI-SDM database replaces the former Central Personnel Data File (CPDF) as the source of Federal workforce data. The last study of CPDF accuracy was conducted by the General Accountability Office (GAO) in 1998.

Purpose of Study

The aim of this study is to assess the accuracy of data elements in the Statistical Data Mart. Since the EHRI-SDM is the source of Federal workforce data, it is important that the data elements within it are as reliable as possible. These data elements are used by OPM, The Office of Management and Budget, the General Accountability Office, and other Federal agencies to make decisions regarding the management of the workforce. In order to accomplish this task,

data on dynamics and status records from the EHRI-SDM were compared with data on Standard Form 50s (SF-50s) included in the Official Personnel Folders (OPFs) of a sample of employees.

Survey Populations

The initial survey population for the dynamics portion of the review was comprised of all SDM dynamics records from FY 2010, excluding those documenting award actions. SF 50s for award actions are not required to be placed in the OPF. Only data elements appearing on the SF 50 could be checked for accuracy. The initial survey population for the status portion of the review was comprised of all records from the September 2010 SDM .

Since the eOPF system has not been fully implemented in all agencies, the population for the study was revised to consist of two components: (1) records of non-DOD employees in agencies with fully implemented electronic Official Personnel File systems (eOPFs) and (2) records of all employees in Defense agency components that submit data to EHRI-SDM. Major agencies without fully-implemented eOPF systems were the Departments of State, Commerce, Justice, and Treasury, as well as the General Services Administration. Thus records for these agencies and several smaller agencies were not part of the study population.

The population was further revised to include only those employees present in the September 2010 file who were also present in the September 2009 file. Another refinement to the population was to remove any employee who was not active as of March 31, 2011, the most recent status file at the time of the sample selection.

In order to obtain a sample of transactions covering a wide assortment of personnel actions, the population was stratified so that it would include a significant number of transactions for promoted and transferred employees. By examining the grades and agencies in the 2009 and 2010 status files, employees with grade increases or with different agencies were assumed to have been promoted or had transferred agencies.

Non-DOD employees thus identified in the status files were sorted into the following strata:

(i) Stratum 1 - Employees transferring to another agency during the year who were not promoted

(ii) Stratum 2 - Employees promoted during the year who did not transfer agencies

(iii) Stratum 3 - Employees who transferred agencies and were promoted during the year

(iv) Stratum 4 - Employees who neither transferred nor were promoted during the year.

Due to sample size considerations, DOD employees were stratified into the following two strata:

(v) Stratum 5 - Employees who transferred or were promoted during the year.

(vi) Stratum 6 - Employees who neither transferred nor were promoted during the year.

Survey Design and Sample

EHRI-SDM Dynamics

A random sample of employees in each of the six strata was selected. For each employee all EHRI-SDM dynamics records were selected with effective dates during FY 2010. This sample design is referred to as a stratified random single-stage cluster sample. The cluster refers to multiple dynamics records being selected for each individual.

EHRI-SDM Status

Unlike the dynamics file, there is no directly corresponding action in the eOPF corresponding to a status record. For determining the accuracy of the EHRI-SDM status file, only the most recent personnel action in the eOPF was compared to data from the Status file. Status records as of September 30, 2010 were compared with the most recent personnel action in the eOPF on or prior to September 30, 2010. Data element values on this personnel action should be reflected in the EHRI-SDM status file for September 2010. The status record sample design is a stratified random sample since we only examine one record per individual.

Data Elements Reviewed

Dynamics data elements reviewed were pay plan, basic pay amount, locality pay amount, adjusted basic pay, prior pay plan, prior basic pay, prior locality pay adjustment, prior adjusted basic pay, occupation series, grade, veterans preference, date of birth, legal authority, tenure, pay rate determinant, retirement plan, work schedule, position occupied, duty station, agency subelement, personnel office identifier, prior grade, step, prior step, prior pay basis, and prior pay basis. Data elements checked in the status file were pay plan, basic pay amount, locality pay amount, adjusted basic pay, occupation series, grade, veterans preference, date of birth, legal authority, tenure, pay rate determinant, retirement plan, work schedule, position occupied, duty station, agency subelement, personnel office identifier, and step.

Data Weighting

Since the sample design yields stratum sample sizes disproportionate with the population counts, data results were weighted to account for this discrepancy. This procedure results in unbiased estimates of data element accuracy.

Data Collection

Non-Defense Agencies

Data from the EHRI-SDM were compared with transactions in the eOPF. A listing of records for each sampled employed was provided in EXCEL format which displayed all data elements in the dynamics and status files that are included on the SF 50. Records were sorted by agency, social security number (SSN), and effective date. A column for each variable was given that indicated whether the EHRI-SDM data element matched data on the form SF 50. The default character in the match field was set to "0" which indicated a match. If the data did not match data on the SF 50 then this field was set to "1" indicating a non-match.

In order to gain access to the electronic personnel folders, the Office of the Chief Information Officer set up user accounts and passwords on the eOPF system for selected employees in the Data Analysis Group. For each sampled individual, electronic copies of the SF 50 for the selected personnel actions were viewed and data were compared with SDM data elements to determine if the values of the data element were the same. Data mismatches were coded as "1" as described above.

Defense Agencies

Listings of sampled employee records were provided and data comparisons were conducted by DoD employees using the same procedure as described above. Completed data comparisons were sent back to OPM and merged with non-DoD data.

In some cases SF 50s corresponding to actions taken during FY 2010 were missing from the electronic personnel folder. This situation happened in many cases when an employee transferred agencies and the eOPF was not updated in a timely fashion. In some instances paper copies of the SF 50 had not been scanned and added to the eOPF. Of the 1,961 records in the final sample, 303 records had no corresponding SF 50 in the eOPF. Thus the final dynamics sample was 1,658 records.

Population and Sample Counts

Population and sample counts are displayed in the following table.

STRATUM	Dynamics Population	Final Dynamics Sample	Status Population	Final Status Sample
(i) Non-DOD with Transfer	8,463	322	2,321	116
(ii) Non-DOD with Promotion	520,355	473	156,338	139
(iii) Non-DOD with Transfer and Promotion	4,829	110	1,285	45
(iv) Non-DOD Other	1,406,038	309	750,981	176
(v) DOD with Transfer or Promotion	197,709	261	65,320	89
(vi) DOD Other	1,147,438	183	572,322	91
TOTAL	3,284,832	1,658	1,548,567	656

Accuracy Results

The table in Appendix I on the next page presents estimated data element accuracy rates for the Statistical Data Mart and for the last accuracy study of Central Personnel Data File (CPDF) that was conducted by GAO in 1998. Columns 2 and 4 contain weighted percent accuracy estimates for SDM dynamics and status data, respectively. Columns 3 and 5 display 99 percent lower confidence bounds for accuracy. These bounds are included to account for sampling variability and provide a conservative estimate of the data element accuracy. Columns 6 and 7 display the GAO results.

GAO measures of CPDF status file accuracy were determined from a questionnaire delivered to each sampled employee. The final sample size for this the GAO status accuracy was 407. Estimates of dynamics accuracy were determined by comparing CPDF data with the OPF of sampled employees. The sample size for this GAO comparison was 113. Lower bounds are not provided for the GAO estimates but would be considerably less reliable than those for estimates from the OPM survey.

Current Survey

Based on the lower bounds, 19 of the 26 dynamics data elements have an estimated accuracy rate of 98 percent or higher, with most closer to 100 percent. These estimates are shaded in green. The seven remaining data elements have an estimated accuracy rate of at least 95.7 percent. These are shaded in yellow. For status data, 10 of 17 data elements have an estimated accuracy rate of at least 98 percent. The seven remaining status data elements have an accuracy rate of at least 95.5 percent.

Comparison with GAO Survey

Estimated CPDF accuracy rates for the prior GAO study are very similar to rates computed for the SDM in this study. Although we did not perform a formal statistical test of differences in data element accuracy, the survey results are so similar that, with the exception of SCD, differences would not be significant.

The reason that formal tests for significant differences were not performed is due to the relatively small sample sizes for the GAO survey. The sample size for the GAO dynamics comparison was 113, while the status sample was 407. For this survey, 1658 dynamics actions and 656 status records were compared with the eOPF. For the majority of data elements whose accuracy rates were assessed in each survey, accuracy estimates were within 2 percent of each other.

Evaluation and Recommendations

Data elements with estimated accuracy rates of 98 percent or greater should be considered to be acceptable. Accuracy rates for data elements between 95 and 97.9 percent are acceptable, but not ideal. Step is one of these data elements. The other six are the pay-related variables; basic pay, locality pay, adjusted basic pay, prior basic pay, prior locality adjustment, and prior adjusted basic pay. The accuracy rates of these six pay variables are correlated since they are dependent on each other. For most employees the pay values are driven by basic pay -- locality adjustment is a percentage of basic pay and adjusted basic pay is the sum of basic and locality adjustment -- so if basic pay is wrong then any percentage based on it will be wrong and sum that includes it will be wrong.

Further research is warranted into these data elements. For example, accuracy rates may vary by agency. The sample size in this survey is insufficient to make agency specific inferences with a high degree of reliability. However, examination of the data may provide possible indications of agencies or types personnel actions with higher error rates for these variables. It must be noted that EHRI-SDM is not a payroll system. OPM plans on assessing the veracity of payroll feeds EHRI-DW receives in the near future.

Appendix I– Accuracy Rates By Data Element (With Sample Sizes)

Data Element	Percent Accuracy Estimate – Dynamics (1658)	99% Lower Bound For Accuracy – Dynamics (1658)	Percent Accuracy – Status (656)	99% Lower Bound For Accuracy – Status (656)	GAO Accuracy Status (407)	GAO Accuracy Dynamics (113)
Pay Plan	99.8	99.4	99.9	99.8	99.3	97.3
Basic Pay	98.0	96.5	98.6	97.3	N/A	N/A
Locality Pay	97.6	95.7	97.5	95.5	N/A	N/A
Adjusted Basic Pay	98.1	96.6	97.7	95.8	98.8	93.8
Prior Pay Plan	99.8	99.5	N/A	N/A	N/A	N/A
Prior Basic Pay	97.9	96.6	N/A	N/A	N/A	N/A
Prior Locality Adjustment	97.9	96.7	N/A	N/A	N/A	N/A
Prior Adjusted basic Pay	98.1	96.9	N/A	N/A	N/A	N/A
Occupation Series	99.7	99.2	99.9	99.6	100.0	99.1
Grade	99.4	98.4	98.2	96.9	99.3	97.3
Step	98.8	97.8	98.2	96.9	N/A	N/A
Veterans Preference	100.0	100.0	100.0	100.0	99.3	96.5
Date of Birth	100.0	100.0	100.0	100.0	100.0	99.1
Legal Authority 1	100.0	99.9	N/A	N/A	N/A	N/A

Data Element	Percent Accuracy Estimate – Dynamics (1658)	99% Lower Bound For Accuracy – Dynamics (1658)	Percent Accuracy – Status (656)	99% Lower Bound For Accuracy – Status (656)	GAO Accuracy Status (407)	GAO Accuracy Dynamics (113)
Legal Authority 2	100.0	100.0	N/A	N/A	N/A	N/A
Tenure	99.9	99.8	98.2	96.5	N/A	100.0
Pay Rate Determinant	100.0	100.0	99.2	97.9	N/A	98.2
Retirement Plan	100.0	100.0	100.0	100.0	100.0	100.0
Work Schedule	99.9	99.5	100.0	100.0	100.0	100.0
Position Occupied	100.0	99.9	99.9	99.6	N/A	100.0
Duty Station	100.0	100.0	99.5	98.6	99.7	100.0
Agency/Subelement	100.0	100.0	100.0	100.0	100.0	100.0
Personnel Office Identifier	100.0	100.0	99.7	99.1	N/A	99.1
Prior Grade	99.3	98.8	N/A	N/A	N/A	N/A
Prior Step	99.2	98.6	N/A	N/A	N/A	N/A
Prior Pay Basis	99.7	99.5	N/A	N/A	N/A	N/A

For Further Information

If you have any questions about this report or any other aspect of the Statistical Data Mart, please contact the Data Analysis Group at FedStats@opm.gov.

UNITED STATES
OFFICE OF PERSONNEL MANAGEMENT
Planning and Policy Analysis
Data Analysis Group
1900 E Street, NW
Washington, DC 20415

PPA/DAG-08/2013-01