APPENDIX B:
WHISARD DATA MATCHING PROCEDURES USING THE ESTABLISHMENT MATCHING APPLICATION

This appendix provides details on how the WHISARD data was matched over calendar years using the OSHA Establishment Matching Application (EMA). Specifically, this appendix discusses:

- The general overview of the EMA,
- The use of the EMA in matching the seven years of WHISARD data together

B.1 General Overview of the OSHA EMA

The OSHA EMA is a Microsoft Access-based software application that can be used to match a database containing a set of establishment records with another database set that also contains establishment records. The EMA will attempt to find matches between the two databases (i.e., the same establishment appearing in each database). The EMA is designed for situations where a direct link between the two databases either does not exist or, if it does exist, is considered imperfect. This is the situation that fits the seven years of WHISARD data used in this analysis. Specifically, there is no link between each year of the WHISARD data (nor within each year of data) that can be used to analyze a single establishment or employer's investigation data over time.

The EMA compares key identifier fields for each establishment in each database to determine which establishments are in each database. The fields compared by the EMA are:

- Establishment name,
- DUNS,
- Street number,
- Street name,
- City,
- State,
- Zip code,
- SIC code, and
- Number of employees.

In performing the matching, the EMA looks for exact matches as well as similarities between records in each database for each field. For example, when comparing two establishment records, the EMA first looks for whether the two establishment names are identical. If the names are not identical, the EMA compares the first six letters of the name. If the first six are not identical, the EMA then compare the first three letters.

In performing the comparisons, the EMA assigns value (positive or negative) to reflect the degree to which each field between the two records match. For example, in the name comparison above, if the complete name matches, a value of nine is assigned; if only the first six letters match, a value of six is assigned; if only the first three letters match, then a value of four is assigned; and if there is no match, a value of –2 is assigned. These values are called t-scores, with higher values for the score reflecting a more exact match. The largest possible t-score is 38. Thus, each record in one database is compared to each record in the other database, and each of these paired comparisons is assigned a t-score to reflect the exactness of the match.

The EMA also assigns each paired comparison to one of two groups: high category matches or low category matches. High category matches are assumed to be exact matches and do not require further review. The high category matches reflect paired comparisons that match exactly on five key criteria:

- Establishment name
- Street number,
- Street name,
- City, and
- Five-digit zip code.

The low category matches are all paired comparisons that are not designated as high category ones. For the low category matches, it is necessary to review t-score values to determine which paired comparisons to retain as "true" matches. The process of determining which low category matches to retain can be complex and may involve visual review of the matches. The decisions made for this analysis are discussed below.

B.2 Using the EMA to Match the Seven Years of WHISARD Data Together

The analyses for this project required the use of a dataset consisting of establishment-level records for the most current investigation data available in WHISARD for WHD cases from FY 2005-2008, and related investigation data from the most recent previous investigation (if there was a previous investigation within the past 3 years) In order to create an analysis file with current investigation data and related investigation data from the most recent previous investigation, ERG separated the WHISARD data sets WHD provided into seven individual files for each fiscal year 2002 through 2008.

After individual FY files were created, for each of the FY 2005 through FY 2008 files, ERG used EMA to generate a set of matches between the FY of the file and the 3 previous FY files. For example, for the FY 2006 file, EMA generated a set of matches between the FY 2006 file and itself, and between the FY 2005, FY 2004, and FY 2003 files. The WHISARD fields the EMA compared were:

- Establishment trade name
- Establishment legal name
- Street number,
- Street name,
- City, and
- Five-digit zip code.

As described above, the matches generated by EMA were assigned to either the high category match or low category match group. Those in the high category match group were assumed to be matches between the FY files. For those in the low category, t-scores were reviewed. Every match in the low category generated when the 2005 file was matched to itself and when the 2006 file was matched to itself was reviewed. An analysis of these matches was conducted to determine the percentage of possible matches that were determined to be actual matches after visual review for each t-score. Any t-score where over 97.5% of the possible matches did indeed prove to be actual matches in both the 2005 and 2006 files was determined to be a reliable indicator of a match. Table B-1 presents the percent of possible matches that were actual matches by t-score for both files. Based on these results, it was decided that any matched pair with a t-score of 21 or more would be designated a match, while those with a 20 or less would need to be visually reviewed and be designated a match or not as appropriate on a case-by-case basis.

Table B-1. Percent of Possible Matches Determined to Be Actual Matches After Review for the FY 2005 and FY 2006 Files.

t-score	Percent of possible matches that are actual matches		
	2005 file	2006 file	Both files
31	100%	100%	100%
29	100%	100%	100%
28	100%	100%	100%
27	100%	100%	100%
26	98.68%	99.09%	98.90%
25	100%	100%	100%
24	98.81%	99.46%	99.08%
23	100%	100%	100%
22	98.11%	97.92%	98.02%
21	100%	100%	100%
20	14.22%	51.47%	22.67%
19	60.32%	72.34%	65.45%
18	14.34%	14.79%	14.52%
17	54.55%	18.00%	32.53%
16	15.88%	10.28%	13.10%
All	49.01%	54.54%	51.44%

Once each of the four FY files used in the analysis (FY 2005 – FY 2008) were matched to themselves and the previous three years using the procedure described above, the most recent prior match (if one exists) was saved. The final analysis file contained a total of 62,532 records of investigations, 8,255 of which had related investigation data from the most recent previous investigation in the past 3 years (different numbers of records were available for each analysis depending on availability of values in other data fields included in each analysis). Table B-2 presents a tabulation of records by fiscal year and whether or not previous investigation data were available.

Table B-2. Tabulation of Records in Analysis File by Fiscal Year and Whether or Not Previous Investigation Data Were Available.

Fiscal Year	Number of Records with No Previous Investigation Data	Number of Records with Previous Investigation Data	Total
2005	11,011	1,446	12,457
2006	12,148	1,814	13,962
2007	16,228	2,590	18,818
2008	14,890	2,405	17,295
Total	54,277	8,255	62,532