UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
WASHINGTON, D.C. 20460

DEC 18 2002

OFFICE OF
ENFORCEMENT AND
COMPLIANCE ASSURANCE

MEMORANDUM

SUBJECT: Results of the Random Audit of FY 01 Inspection Data

FROM: Frederick F. Stiehl, Director
Enforcement Planning, Targeting & Data Division

TO: Regional Enforcement Division Directors, Regions 1, 2, 4, 6, 8
Regional Media Division Directors, Regions 3, 5, 7, 9, 10
Regional Enforcement Coordinators
Lead Compliance and Enforcement Data Stewards

This memorandum reports the results of the Random Audit of FY 01 Inspection Data. begun on April 2, 2002 and completed in July 2002. (See my attached April 2, 2002 memorandum to the Enforcement Coordinators and Data Stewards.) I want to express our appreciation for your support on this project: 100 percent participation across state and Regional programs made this a successful data quality initiative that gives us an objective baseline on key data used to measure the national compliance monitoring program. I will also email this memorandum and final report to all participants who logged onto the audit Web site and extend my gratitude for their effort in completing the audit for their program. In designing the audit, care was exercised to ensure that the audit minimized the efforts required of respondents while preserving participant interest in the results. We feel this helped solicit such widespread participation. The introduction to the report describes how this was achieved. Even so, we recognize that this took scare resources to complete. Our ability to characterize for the first time in a statistically valid manner the quality of our inspection data makes this a worthwhile investment.

As a result of the audit the following three statements concerning the quality of inspection data in the AFS, PCS, and RCRAInfo data systems at a national level can be made:

1. The inspection information maintained by EPA and the states in AFS is 87% accurate[1].
2. The inspection information maintained by EPA and the states in PCS is 87% accurate[1].

---

[1]Based on the survey methodology used, we are 95% confident that this error rate is within 4% of the true value for AFS and PCS, and within 2% of the true value for RCRAInfo.

3. The inspection information maintained by EPA and the states in RCRAInfo is 97% accurate[1].

In order to keep sample sizes at a minimum the audit was designed to provide estimates of accuracy at the national level only. These results answer the question: "What percentage of facilities have completely accurate inspection information?" This national level information provides us with an objective measure of how well we are doing in maintaining important information on our monitoring program; and provides OECA with a baseline from which improvements in data quality and the impacts of data quality initiatives can be measured. (Although states and Regions may request their specific data, the audit was not designed to provide assessments of data accuracy at the state or Regional level.)

Table 2 in the report breaks down the results into percentages of facilities with: 1) a missing inspection, 2) an erroneous inspection, and 3) an error in an inspection record field. The following conclusions can be made from the data:

- missing inspections account for the majority of errors in all three data systems;
- for the inspections in the data systems, inspection record information: date, lead agency, and type, are more than 95 percent accurate in all three data systems;
- across all three data systems, less than 2 percent of facilities contained erroneous inspections; RCRAInfo was 100 percent accurate in this respect.

Table 3 in the report makes statements regarding the accuracy of inspection coverage rates (i.e., percentage of facilities having been inspected at least once in a given year).

There are two likely reasons why the inspection data in RCRAInfo can be expected to be more complete and accurate: 1) RCRA determines violations through inspections. In contrast, violations and Significant Noncompliance (SNC)/High Priority Violator (HPV) status in the Air and NPDES programs can be determined by other reporting methods (especially self reporting); and 2) RCRAInfo is a modernized system. Since self reporting can trigger non-compliance and SNC/HPV status in AFS and PCS, there may be less of an incentive to get inspections in these two systems. Section 2.1 of the report provides a description of the modernized aspects of RCRAInfo which promote data completeness and accuracy. In particular, data business rules are enforced in the RCRAInfo system itself so that an inspection/evaluation record is required to be entered before a violation can be entered (AFS and PCS does not enforce this restriction.)

We expect to use the results of the audit in at least two ways: 1) the results of the audit will accompany important reports and analyses that rely on inspection data so that the reader can better understand the underlying accuracy of the analyses; and 2) the results help to target data quality efforts such as data entry and data system modernization.

I would like to encourage states and Regions to make sure inspections at all federally reportable sources are entered/uploaded into AFS and PCS. Without complete data entry of inspections, states and Regions are receiving less credit for their compliance monitoring

activities in national reports than were actually conducted. Complete and accurate data will be even more important as enforcement and compliance data is released to the public. Section 2.1 of the report also describes PCS and AFS modernization efforts and other data quality initiatives from headquarters which will promote complete and accurate reporting of data to PCS and AFS.

In FY 2003 we are proposing to conduct a similar audit of state and federal enforcement action data. This audit should be straightforward and build on the success of the inspection audit. Please see section 3.4 of the report which provides a list of lessons learned from this inspection audit which we will apply to improve the process for the enforcement action audit.

CC:

Director, Office of Compliance, OECA
Director, Office of Regulatory Enforcement, OECA
Environmental Council of the States (ECOS)
Regional CAA, CWA, and RCRA Program Data Stewards
AFS Database Managers
PCS Database Managers
RCRAInfo Database Managers
Random Audit Participants
FY 2002 and FY 2003 Data Quality Strategy Workgroups

RESULTS OF THE RANDOM AUDIT
OF
FY01 INSPECTION DATA

Prepared for:
U.S. Environmental Protection Agency
Office of Enforcement and Compliance Assurance
Office of Compliance
Information Utilization and Targeting Branch

Prepared by:
Abt Associates Inc.

EPA Contract 69-W-039

Table of Contents

1.0 Introduction

Conclusions drawn in analyses and reports by the Office of Enforcement and Compliance Assurance (OECA) as well as those of external groups using the Environmental Protection Agency's (EPA) data have, at times, been challenged based on claims of data quality concerns. Inspector General reports have stressed the need for improved data quality in EPA efforts to monitor and measure enforcement and compliance. The Agency's recent increased commitment to creating systems to ensure the highest quality of information possible has led to a number of new data quality initiatives and investigations. In FY2002, OECA convened a Data Quality Strategy Workgroup (hereafter known as the Workgroup) to coordinate and implement quality initiatives within OECA. The Workgroup is composed of staff from Headquarters, Regional Offices, all programs with relevant data systems, and includes state representatives. In FY2002, the Workgroup was involved in a number of data quality activities ranging from developing analytical tools for detecting data entry problems to nominating studies of data quality in various data systems.

In this first year, the Workgroup proposed to audit quality mission-critical information -- specifically, the inspection information entered into and maintained in three EPA data systems:

- AIRS Facility subsystem (AFS);
- Permit Compliance System (PCS); and
- RCRAInfo.

The audit addressed both federal and state inspection information for a sample of eight facilities per state per media.

The methodology used to design the random audit was developed based on decisions made by the Workgroup as well as input from media program staff and Regional compliance managers. A summary of the audit-related issues discussed by the Workgroup can be found in Appendix A. The audit methodology was finalized as of April 2002 and administered over the next three and one-half months to accommodate the schedules of state and Regional staff. Care was exercised to ensure that the audit minimized the efforts required of respondents. This was accomplished by:

- simplifying the approach so that instructions would be simple and readily understood by state and Regional EPA staff;
- minimizing respondent burden by using the smallest sample that would yield statistically relevant results;
- using an on-line survey form pre-populated with inspection data from the respective data systems to facilitate distribution and administration of the audit; and
- providing user support throughout the period the audit was active.

This first audit, completed in July 2002, is considered a successful data quality initiative because of the level of participation across delegated state programs and Regional programs. There was 100 percent participation across state and Regional programs. It was reported that the level of participation was due in part to the fact that the time required by participants was minimal and the participants' own interest in the results of the audit.

This inspection audit provides OECA with statistical support for statements concerning the quality of inspection data in the PCS, AFS and RCRAInfo data systems at a national level. The audit results also provide OECA with a baseline from which improvements in data quality and the impacts of data quality initiatives can be measured.

2.0 Results

Each data system was analyzed independently at the national level. One year (federal fiscal year 2001) of inspection records of 408 randomly selected facilities for each program, or eight facilities per state per media (including Puerto Rico) were audited. This approach was designed to provide estimates of accuracy at the national level only and does not support an objective assessment of data accuracy at the state or Regional level. The audit verified both the accuracy and the completeness of each facility's inspection records. For each facility, participants were asked to:

- verify that each of the inspections in the federal data system did occur;
- identify any missing inspections; and
- correct the "date," "lead agency" and/or "inspection type" fields for the presented inspections of sampled facilities.

Detailed information about the way the on-line audit was designed and administered, how facilities were sampled, and the inspection types included is in the Methodology section of this report (Section 3).

Two different estimates are derived that will assist EPA, states, and other data users to understand accuracy of inspection information in the federal data systems. The first approach answers the question, "What percentage of facilities have completely accurate inspection information?" The second view answers the question, "How accurate are inspection coverage rates (i.e., percentage of facilities having been inspected in a given year)?" Along with the point estimate error rates for these two analyses, a 95 percent confidence interval is presented, the meaning of which is that we are 95 percent confident that the true error rate of the entire population (if a census of every facility were conducted) is within this range denoted by " ± [the confidence interval value]." Such contextual information is critical to a wide range of downstream analyses and decision-making, including but not limited to: media program administration, data quality assurance programs; state and federal resource planning, compliance and enforcement targeting, and other environmental analyses.

2.1 Accuracy of Inspection Data

Table 1 shows the percentage of facilities in each data system that have an error in their associated inspection data.

Table 1: Facilities With An Error In Inspection Data

Data System	Percentage With An Error*
AFS	**13.40%** ±3.80
PCS	**13.24%** ±3.78
RCRAInfo	**2.81%** ±1.65

* The variance (±) is a 95% confidence interval.

The audit results provide policy makers, planners, and data systems users critical information with which to characterize the quality of data in the three data systems examined. These results also support statements about the accuracy of specific data systems, such as:

- The inspection information maintained by EPA and the states in AFS is 87% accurate. Based on the survey methodology used, we are 95% confident that this error rate is within 4% of the true value.
- The inspection information maintained by EPA and the states in PCS is 87% accurate. Based on

the survey methodology used, we are 95% confident that this error rate is within 4% of the true value.

- The inspection information maintained by EPA and the states in RCRAInfo is 97% accurate. Based on the survey methodology used, we are 95% confident that this error rate is within 2% of the true value.

To better understand the types of errors that are occurring in the database, Table 2 breaks down the error rates from Table 1 into three different components:

Facilities with a missing inspection: An estimate of percentage of facilities in each database with one or more missing inspections.

Facilities with an erroneous inspection: An estimate of percentage of facilities in each data system with an inspection record that did not in fact occur.

Facilities with an error in an inspection record field: An estimate of the percentage of facilities in each data system with an error in any of the following inspection record fields: date, lead agency, and inspection type. (A date must be off by more then seven days to be considered an error.)

Table 2: Detailed Error Rates

Data System	Facilities With A Missing Inspection*	Facilities With an Erroneous Inspection*	Facilities With An Error in an Inspection Record Field*
AFS	**10.03%** ±3.31	**1.51%** ±1.47	**4.44%** ±2.34
PCS	**9.36%** ±3.30	**1.39%** ±1.27	**4.02%** ±2.06
RCRAInfo	**2.48%** ±1.57	**0.00%** -	**0.89%** ±1.16

* The variance (±) is a 95% confidence interval. The confidence intervals are not calculated for error rates of 0% or 100%, as is the case for RCRAInfo records with erroneous inspections.

From Table 2, it is apparent that missing inspections account for the majority of errors in all three data systems. The inspections record information: date, lead agency, and type, are more than 95 percent accurate in all three data systems. Across all three data systems, less than 2 percent of facilities contained erroneous inspections; RCRAInfo was 100 percent accurate in this respect. These error rates are not additive to the values in Table 1 because there are instances where these three types of errors can and do occur in the same facilities.

Data system staff for AFS, PCS, and RCRAInfo were asked to review draft results of the inspection audit. There are two likely reasons why the inspection data in RCRAInfo can be expected to be more complete and accurate: 1) RCRA determines violations through inspections. However, violations and Significant Noncompliance (SNC)/High Priority Violator (HPV) status in the Air and NPDES programs can be determined by other reporting methods (especially self reporting); and 2) RCRAInfo is a modernized system. Since self reporting can trigger non-compliance and SNC/HPV status in AFS and PCS there may be less of an incentive to get inspections in these two systems. The data system staff provided the following additional details about their respective systems as relevant to the quality of information within the federal data systems. They also presented some of the near-term changes which are expected to affect data quality.

The following aspects of RCRAInfo are believed to promote both data completeness and accuracy:

- The new graphical user interface (GUI), drop-down lists, and other "point-and-click" GUI tools simplify viewing, entering, and retrieving RCRA data for occasional and veteran users alike.
- Code values are always presented with clear explanations.
- Data business rules are enforced by the software through data entry edits, therefore partial records

are not allowed. For example, an inspection/evaluation record is required to be entered before a violation can be entered (AFS and PCS does not enforce this restriction.)

There are several initiatives that are expected to promote complete and accurate reporting of data to both AFS and PCS:

- PCS is currently undergoing modernization as Phase II of ICIS.[1] Changes are needed both to address outdated technology and new program requirements. The current estimate for modernized PCS is late 2004.
- An AFS Needs Analysis report is being finalized. This report will prioritize the needs for the next generation of AFS and also to identify action items that need addressing before design of this modernized system.
- Under its Enforcement and Compliance Data Quality Strategy, the Office of Compliance of OECA is conducting several data quality projects per year. These projects are developed in consultation with Regions and states. In addition to the random audit projects, these projects include:
  - < Periodic, comparative analyses of particular data fields across organizational units with delegated authority to identify potential data quality problems (particularly incomplete data entry);
  - < Analyses of key enforcement and compliance activity data fields to determine if there are discrepancies in Regional or state usage. Discrepancies found are documented, and guidance developed to assist program implementers and database users in how to use the codes for nationally consistent reporting.
  - < Currently available Internet error correction tools are being applied to particular records in need of correction. For example, OC is using information from the Headquarters, Regional, and state data steward networks, as well as the (Online Tracking Information System) OTIS site and its error correction process to pinpoint data records within OC's national databases in need of correction. This error correction process is expected to be available to the public shortly through the Public Access Internet site.

[1]The Integrated Compliance Information System (ICIS) supports the information needs of the National Enforcement and Compliance program. ICIS will integrate data that is currently located in more than a dozen separate data systems. The Web-based system will eventually enable individuals from states, communities, facilities, and EPA to access integrated enforcement and compliance data from any desktop connected to the Internet. EPA's ability to target the most critical environmental problems will improve as the system integrates data from all media.

2.2 Accuracy of Inspection Coverage Estimates

The information collected in the audit can also be used to estimate the accuracy of national inspection coverage estimates. That is, the percentage of facilities that were inspected at least once in FY01. This information, important to program and compliance management, is often used to ensure equitable and effective distribution of resources and inspection outcomes (ranging from detection of violations to deterrence of noncompliance behaviors to public health protection). Table 3 contains the baseline percentage of facilities inspected according to each data system along with the percentage of facilities that are falsely considered inspected when no federally-reportable inspections occurred, and those falsely considered "not inspected."

The following table indicates that if a facility has been inspected at least once in FY01 this coverage data is accurate 99 percent of the time for all three data systems. Whereas statements asserting that a facility has not been inspected during the time period can be made with 92 to 99 percent accuracy depending on the particular data system. Again, these results are driven by the missing inspections.

Table 3: Errors in Inspected Designation*

Data System	# Facilities in the System	% Facilities Inspected (FY01)	% of Facilities Falsely Considered Inspected	% of Facilities Falsely Not Considered Inspected
AFS	**42,326**	**43.46%**	**0.08%** ±0.12	**8.07%** ±3.17
PCS	**6,615**	**71.61%**	**0.06%** ±0.11	**5.41%** ±2.62
RCRAInfo	**22,239**	**23.48%**	**0.00%** -	**1.12%** ±0.73

* The date range for a facility to be inspected is 10/01/00 to 9/30/01. See Table 4 for details on which types of facilities were included in the audit.

3.0 Methodology

The methodology is presented in three parts: selecting the sample, verifying the inspection information, and estimating the percentage of facilities with an error. As stated earlier, issues addressed by The Workgroup in the formulation of this methodology may be of interest to readers and are located in Appendix A. Also, detailed statistical methodologies that supplement this discussion are presented in Appendix B.

3.1 Selecting the Sample

One of the first steps in the audit was to determine a suitable sample size for an effective national survey. As the media programs are often delegated to states and burden per state and per Region was a concern, per state sample sizes were presented for discussion during the planning phase of this project as documented in Appendix B. For this audit, margins of error, expressed as a 95 percent confidence interval, were considered. Based on previous data quality investigations, it was assumed that the true error rate (how often the national data systems contain information that does not match the field records) in the inspection records was in the neighborhood of 5 percent.[2] Several sample sizes were considered by the Workgroup taking into account the tradeoffs between accuracy of the results and respondent burden. After considering these factors, a sample size of eight facilities per state per media was used. Facilities from all fifty states and Puerto Rico were sampled, totaling 408 facilities per program nationally.[3]

Table 4: Sample and Universe Sizes*

Data System	# of Facilities in the Universe	Sample Size	Universe Description
AFS	42,326	408	The RECAP universe: Class A, Synthetic Minor, and NESHAP minor sources with an operating status of: Operating, Temporarily Closed, or Seasonal
PCS	6,615	408	Active Major NPDES permits
RCRAInfo	22,239	408	The RECAP universe: TSDs and LQGs. The LQGs selected reported to BRS that they were LQGs or they were flagged in RCRAInfo as LQGs and had activity (evaluation, violation or enforcement) in the last 5 years.

* AFS and PCS data was pulled from IDEA on 1/31/02, RCRAInfo data was pulled 2/20/01

The sample size does lead to over-sampling of states with smaller numbers of facilities, but it also leads to equitable distribution of burdens. Moreover, the over-sampling of small states decreases the overall confidence interval. Statistical weighting of the audit results, adjusted for over representation of these states in the sample. In simplistic terms, the result for each sampled facility is weighted by the number of facilities in the universe that sample facility represents, which is proportional to the fraction of U.S. facilities in the state. A detailed explanation of the statistical weighting method is located in Appendix B.

[2] Sector Facility Indexing Project - 1998-2001; Region 10's EC-Online - 2000.

[3] Facilities Washington D.C. were not sampled because there were not 8 facilities in each program in D.C.

The inspection records to be reviewed were limited to federally-reportable inspection types. A list of the types of inspections that were audited is in Appendix C. A comparison between the distribution of inspection types in the universe and in the sample is presented in Appendix E. The comparison shows that the sample is similar to the national distribution among inspection types.

Finally, the sample within each state was selected to also account for variation of the number of inspections per facility. The distribution of the number of inspections per facility in the sample was the same as the universe for each state. A national comparison of the number of inspections per facility in the universe and in the sample is presented in Appendix E. It shows that the sample and universe are virtually identical in terms of the distribution of inspections per facility for each program.

3.2 Verifying the Inspection Information

The verification of the inspection information took approximately four months to complete. Every one of the selected facilities was audited by appropriate staff. In most cases, the facilities were audited by both state and Regional staff. The state staff reviewed all facilities and verified all inspections where a non-federal entity was or should have been the lead agency. Regional staff reviewed all facilities and verified all inspections where a federal entity was or should have been the lead agency. In some cases, where the Regions maintain state data, the state lead inspections were verified by the Regional staff.

A memorandum from Frederick Stiehl announcing the start of the inspection audit was sent out on April 2, 2002 to the Regional Enforcement Division Directors and Coordinators, the Regional Data Stewards, and the Database Managers. The memorandum and attachments provided instructions to the Regional leads for delegating responsibility for the audit and for the executing the audit. The Regional leads forwarded the instructions on to Regional and state program staff who completed the audit. Over four hundred individual users logged on to the Web site to participate in the audit.

An interactive Web site was created to allow respondents to easily participate in the audit. Respondents used the site by logging on and viewing the selected facilities for which they were responsible for auditing. After printing or reviewing the list of sampled facilities, the participants located the corresponding hard copy inspection records. They then logged back on to the Web site and either verified that the records in the federal data system reflected the hard copy records or reported differences. Examples of pages from the Web site are available in Appendix D.

Once the audit was underway, telephone and email support was provided to the users' questions. Most questions related to the technical use of the Web site and the logistics of the audit. Phone and email prompts were sent out by Headquarters staff intermittently to remind participants of the schedule, solicit feedback and offer assistance, thereby increasing completion rates.

3.3 Data Analysis

After every facility was audited, the data collected by the Web site was cleaned. The cleaning process involved scanning the reported errors to ensure that they are in fact errors. This cleaning included:

- Checking to make sure that the dates of added inspections were within FY01.
- Checking to make sure that the type of added or changed inspections were federally reportable.
- Checking to make sure that reported differences in fields were in fact differences that should be reflected in the federal data systems. For example, if the lead agency field was "State" and the respondent changed it to "RIDEM" this would not be considered an error because "RIDEM" would be resolved to "State" in the federal data systems.

Once the data was cleaned, it was analyzed to determine the error rates. Detailed statistical methods are located in Appendix B. An overview of the method used to determine the error rate is outlined below.

1. Each facility was flagged to designate it as having an error either due to an erroneous inspection, a missing inspection, a false positive inspected status, a false negative inspected status, or corrections to the content of an inspection (date, lead agency, or inspection type fields).
2. The "weight" of each facility was determined by dividing the number of facilities in the sample by the universe size in each state. This "weight" quantifies the number of facilities that each sampled facility represents in the overall population.
3. For each error type, the total weight of the facilities with an error was totaled. This number was then divided by the total number of facilities in the country to determine the error rate.

3.4 Potential Improvements to Future Audits

In the course of the Inspection Audit, the Workgroup, audit participants, and data system staff made a number of suggestions regarding the design and implementation of the Inspection Audit. Summarized below, these recommendations will be useful in modifying the methodology, design and implementation of any future audit.

- Streamline and shorten the instructions to make it easier for people to find their way to the site and log on.
- Create context-specific help links on each page. This way users can get help on specific parts of the site instead of going back to the instructions.
- Change the web site so users can more easily verify that facilities have no activity (e.g., no inspections).
- Make it easier for auditors and managers to determine the completion status of the audit for any program and geographic area(s) of interest.
- Begin telephone prompts earlier to a) provide user support answers, b) answer individual questions about the audit and c) update users and manager on the status of particular state/Regional programs.

The Workgroup will continue to gather recommendations and feedback as part of its review and follow-up to this audit.

Appendix A: Workgroup Issues

The methodology was developed based on decisions made at the Identification of Data Problems Workgroup meetings and internal discussions. Care was exercised to ensure that any methodology minimized the effort required of respondents.

Workgroup Issues

The following is a list of issues the Workgroup discussed and the decisions made:

Issue # 1: What universe should be analyzed for each of the three media?

Decision: Only federally-reportable universes should be studied. In particular, for RCRA we are considering only including TSDs and LQGs in the universe.

Question: Should SQGs also be included in the RCRA universe? Region 1 argues that SQGs are both federally-reportable and are, in many cases, among the most environmentally significant inspections that Regions and States do. While including SQGs will not necessarily increase the sample size, it will mean statements made about the accuracy of inspection information in RCRAInfo will not be representative of the TSDs + the LQG universe. This is because of the large number of SQGs (200,000, very few of which have been inspected) compared to the 42,000 TSDs and LQGs.

Decision: Exclude SQGs for this iteration unless further comment is received. EPA considers SQGs to be a distinct and separate sub-population receiving different levels of inspection attention. On a small effort such as this, co-mingling audit results of TSDF and LQG populations with that of the SQG population would not allow any statements of accuracy to be made about inspection data for the LQG/TSDF population.

Issue # 2: In choosing the random samples, should the combined universe of all three programs be analyzed or should each universe be analyzed separately? Studying the combined universe of 71,791 facilities would reduce the number of samples needed. Studying each of the universes separately would increase the number of samples needed, but would allow separate statements to be made about the accuracy of inspection data in each of three media databases.

Decision: Each media database should be studied separately. Even though this increases the number of samples needed, it is important for us to be able to make separate statements about the accuracy of inspection data in each of three media databases.

Issue #3: What inspection records do we want the Regions and States to compare to the data in our National databases (i.e., should they compare against their own paper records or against their own internal database, or both).

Decision: We went back and forth on this issue. It is crucial that the Regions and States compare our data to the information on their own paper records. However, it is also important that when differences are found with State data, that the records in question be checked against the state/local data systems. That will enable us to determine the route cause of the problem (either data entry or the translator program). Please see Issue #6 for further discussion.

Issue #4: What system could be developed to make the comparison process as easy as possible for the Regions and States?

Decision: A Inspection Data Quality Audit Web site will minimize the time Regions and States spend responding to the verification request (see attachment).

Issue #5: There is a need to minimize the number permit/facility ID's given to the Region and States to review so as not to overburden, but still choose enough samples to have a robust methodology and to take into account some non-response.

Decision: Workgroup members thought the maximum number of permit/facility ID's a state would be willing to review for each of the three media is eight (8). A sample size of eight (8) facility records per state per media will allow us to be confident in the results for each data system within a reasonably small (95 percent) confidence interval. This sample size does lead to over-sampling of the smaller states. This over-sampling of some states should provide a cushion by helping to address the possibility that some states will not participate. Statistical weighting of the audit results, however, will adjust for over-sampling in these states. A description of the statistical methodology will be prepared in a separate document.

Issue #6: If any discrepancies are found between the hard copy records and the federal data system, it will be important to ascertain where the discrepancy occurred. The question is whether to ask the same personnel participating in the audit of field records to check their state/local data systems (for all identified discrepancies) or to undertake this follow-up verification separately.

Decision: The retrieval of identified discrepancies from the state/local data systems will be undertaken as part of a separate request once all field records have been reviewed.

Appendix B: Statistical Methodology

Introduction

The approach and methodological details reflect the decisions and input of OC's Data Quality Strategy Workgroup, the Environmental Council of the States (ECOS), and program leads knowledgeable about CAA, CWA, and RCRA programs and data that are the subject of this audit.

The random audit evaluates separately inspection data in AFS, PCS, and RCRAInfo. The goal of the audit was to be able to state how accurate inspection data are in each of the three media databases. Therefore, a random sample of inspection records from each data system was required. Facility-level records were sampled according to the distribution of the number of inspections during a one year period (the program leads have concurred on using the one-year period 10/01/2000 - 9/30/2001). A sample was drawn that contains the same distribution of inspections per facility as occurs in the national database. This strategy is discussed in more detail below.

Determining the Sample Size

One of the first steps in designing the audit was to develop estimates of the sample sizes that would yield estimates with various margins of error. The margin of error is expressed as a half-width of a 90 percent or a 95 percent confidence interval. Based on previous data quality investigations, it was assumed that the true error rate (how often the national data systems contain information that does not match the field records) in the inspection records is approximately 5 percent.[4] In reviewing the sample sizes, the Workgroup considered the tradeoffs between accuracy of the results and respondent burden for each state, Region, and local agency with delegated authority who would be participating in the audit.

Error Rate and Variance Calculations

In the random audit, population proportions or percentages of some outcomes of interest are estimated. The variable of interest takes either a value of zero or one. For example, the response to a question is "yes - the data system accurately reflects the field record" or "no - the federal data system does not match the field record." Essentially, the population consists of zero and ones.

Let "P" denote the proportion of "ones" in the population. For example, if 10 percent of the 100 inspections in the population are accurate, the population can be considered to consist of 10 ones and 90 zeros.

The variance of the characteristic (which takes a value one or zero) in such a population is given by:

$$P(1-P)$$

If a simple random sample of size "n" is selected from such a population and a sample proportion is computed, the variance of the sample proportion "p" is given by:

$$P(1-P)/n$$

The standard deviation of the sample proportion is given by:

[4] Sector Facility Indexing Project - 1998, 2001; Region 10's EC-Online - 2000.

Square root of P(1-P)/n or Sq[P(1-P)/n]

A 95 percent confidence interval for the population proportion is given by:

p plus or minus 1.96 x Sq[P(1-P)/n]

The half-width of the confidence interval is given by:

1.96 x Sq[P(1-P)/n]

As shown in Tables 1 and 2, the confidence intervals for AFS and PCS were generally larger then for RCRAInfo. This is because the estimated accuracy of RCRAInfo was closer to 100% then the other data systems. The closer a point estimate is to an extreme (100% or 0% in this audit) and not the midpoint (50%) the smaller the confidence interval. This is because if most every response is lining up behind one value, the random selection of the sample is less likely to affect our point estimate at the extremes.

Accounting for Variation in the Number of Inspections per Permit

In the data quality audit, permits containing varying numbers of inspections were reviewed and verified. To ensure that the range of inspections per permit were adequately represented in the sample, permits were sampled proportional to the number of inspections per permit. To do this, the following steps were taken:

1. The permits were sorted by the number of inspections for each state separately;
2. The total number of permits was counted;
3. The number of permits was divided by eight (the number of permits to be verified in each state);
4. A fractional sampling interval was applied to determine which permits were selected in each state to form the random sample.

An example is given below.

1. Assume that in the state of Alaska there are a total of 50 permits.
2. Dividing the 50 permits by eight yields a sampling interval of 6.25.
3. Select a random number between 0 and 1, say .3. A different random number was selected for each state.
4. Multiplying the sampling interval (6.25) by .3 yields the fractional sampling interval of 1.875.
5. Adding the sampling interval (6.25) to the fractional sampling interval (1.875) yields a new fractional sampling interval of 8.125.
6. Continue adding the new fractional sampling interval a total of eight times.

The results will be as follows:

1.875	14.375	26.875	39.375
8.125	20.625	33.125	45.625

Round up each interval to determine which observation to include in the sample. In this example, we would select the:

2nd	15th	27th	40th
9th	21st	34th	46th observations.

Sampling Weights

For producing population-based estimates, each sampled inspection for which complete data is available was given a sampling weight. This weight was the number of observations in the population represented by the sample observation. The sampling weight adjusts for unequal probabilities of selection and provides unbiased estimates. The final weight combines a basic weight which is the inverse of the probability of selection of the record and an adjustment for non-response.

REFERENCES:

Cochran, W. G., Sampling Techniques, John Wiley & Sons, Third Edition, pg 85.

Appendix C: Inspection Types Audited

Clean Water Act (CWA)

- Compliance Evaluation (Non-Sampling)
- Compliance Sampling
- Performance Audit
- Compliance Biomonitoring
- Toxics Sampling Inspection
- Diagnostic
- Reconnaissance
- Legal Support
- Concentrated Animal Feeding Operation (CAFO)
- Stormwater
- Sanitary Sewer Overflow (SSO)
- Sludge
- Combined Sewer Overflow (CSO)

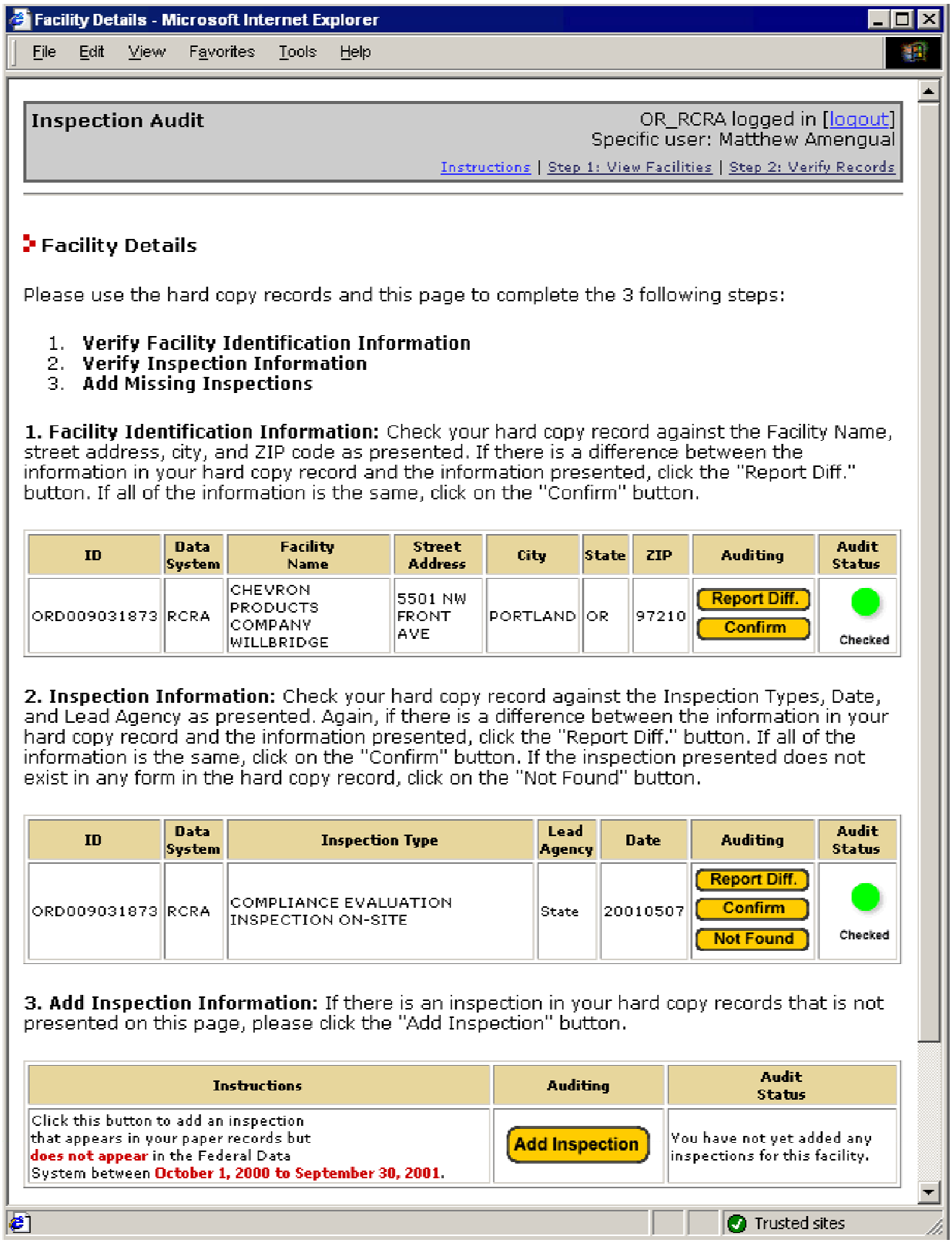
Clean Air Act (CAA)

- 1A - EPA inspection, level 2 or greater.
- 2A - EPA source test conducted.
- 3A - Owner/operator-conducted source test.
- 5C - State inspection, level 2 or greater.
- 6C - State source test conducted.
- FS - State conducted FCE/On-site
- FF - State conducted FCE/Off-site
- FE - EPA conducted FCE/On-site
- FZ - EPA conducted FCE/Off-site
- TO - EPA req (O/O cond) stack test/observed & reviewed

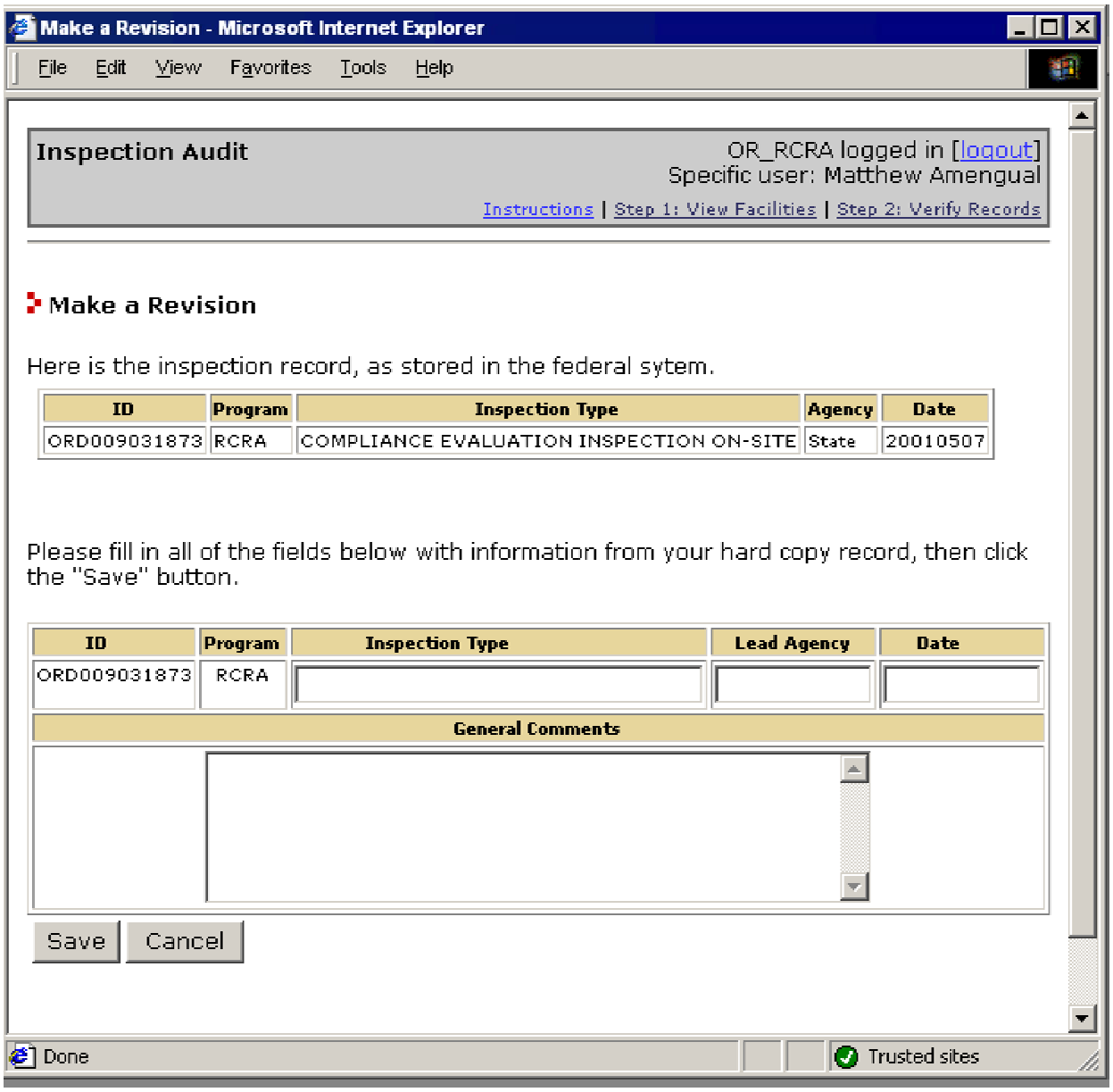
Resource Conservation and Recovery Act (RCRA)

- On-Site Inspection of Corrective Action Activities
- Case Development Inspection
- Compliance Evaluation Inspection
- Compliance (Groundwater) Monitoring Evaluation
- Sampling Inspection
- Financial Record Review
- Non-Financial Record Review
- Compliance (Groundwater) Monitoring Evaluation Without Sampling
- Compliance Schedule Evaluation
- Operation and Maintenance Inspection
- RCRA Compliance Evaluation Inspection Performed with Screening Checklist
- Comprehensive and Coordinated Inspection
- Detailed Multimedia Inspection
- Other Evaluation

Appendix D: Inspection Audit Web Application - Example Screens



Appendix D: Inspection Audit Web Application - Example Screens



Appendix E: Representativeness of Sample

Number of Inspections Per Facility						
Number of Inspections	**AFS**		**PCS**		**RCRAInfo**	
	Sample	Universe	Sample	Universe	Sample	Universe
0	53.19%	56.35%	29.17%	28.39%	72.30%	76.52%
1	39.95%	37.49%	46.32%	45.77%	18.87%	17.05%
2	4.66%	4.51%	13.48%	13.33%	5.88%	4.35%
3	0.98%	0.95%	4.66%	4.13%	1.96%	1.07%
4	0.25%	0.34%	0.98%	2.42%	0.98%	0.50%
5	0.25%	0.13%	0.98%	1.42%	0.00%	0.19%
6-9	0.25%	0.14%	2.21%	2.69%	0.00%	0.20%
10-19	0.49%	0.06%	1.72%	1.71%	0.00%	0.08%
20-39	0.00%	0.00%	0.25%	0.09%	0.00%	0.02%
40 and up	0.00%	0.00%	0.25%	0.05%	0.00%	0.02%

AFS Inspection Types		
Inspection Type	**Sample**	**Universe**
EPA CONDUCTED FCE / ON-SITE	0.00%	0.18%
EPA INSPECTION - LEVEL 2 OR GREATER	8.17%	1.99%
EPA REQ (O/O COND) STACK TEST/OBSV &	0.00%	0.03%
EPA SOURCE TEST CONDUCTED	0.00%	0.01%
OWNER/OPERATOR-CONDUCTED SOURCE TEST	10.51%	5.01%
STATE CONDUCTED FCE / ON-SITE	7.39%	9.53%
STATE CONDUCTED STACK TEST	2.72%	2.79%
STATE INSPECTION - LEVEL 2 OR GREATER	71.21%	80.46%

PCS Inspection Types		
Inspection Type	**Sample**	**Universe**
CSO Inspection	0.16%	0.24%
Compliance Sampling	21.01%	17.68%
Compliance bio-monitoring	0.98%	2.39%
Compliance evaluation (non-sampling)	32.74%	36.53%
Diagnostic	0.00%	0.07%
Enforcement case support	0.00%	0.09%
Performance Audit	2.12%	2.70%
Reconnaissance	41.69%	38.57%
Sanitary Sewer Overflow SSO	0.33%	0.27%
Sludge	0.65%	0.46%
Stormwater	0.16%	0.26%
Toxics Inspection	0.16%	0.73%

RCRAInfo Inspection Types		
Inspection Type	**Sample**	**Universe**
Case Development Inspection	3.03%	1.50%
Compliance (GW) Monitoring Inspection	0.00%	0.30%
Compliance Evaluation On-Site	67.27%	64.14%
Compliance Schedule Evaluation	4.24%	6.41%
Comprehensive & Coordinated Inspct With CEI	0.61%	0.74%
Detailed Multimedia Inspection With CEI	0.61%	0.71%
Financial Record Review	3.64%	4.91%
Non-Financial Record Review	10.30%	6.56%
On-Site Inspection of Corrective Action Activities	0.00%	1.24%
Operation and Maintenance Inspection	0.00%	0.19%
Other Evaluation	9.09%	12.31%
RCRA CEI Performed W/ Screening Checklist	0.61%	0.25%
Sampling Inspection	0.61%	0.73%

AFS: Lead Agency		
Inspecting Agency	**Sample**	**Universe**
Fed	8.17%	2.21%
Non-Fed	91.83%	97.79%

PCS: Lead Agency		
Inspecting Agency	**Sample**	**Universe**
Contractor	0.33%	0.94%
EPA (Regional)	4.89%	5.03%
Joint EPA & State (EPA Lead)	0.65%	0.71%
Joint EPA & State (State Lead)	0.33%	0.29%
NEIC	0.00%	0.01%
State	93.81%	93.01%

RCRAInfo: Lead Agency		
Inspecting Agency	**Sample**	**Universe**
EPA	12.12%	7.09%
EPA Contractor	1.21%	0.53%
Oversight Inspection	1.21%	0.83%
State	85.45%	91.52%
State Contractor	0.00%	0.04%

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
WASHINGTON, D.C. 20460

APR - 2 2002

OFFICE OF
ENFORCEMENT AND
COMPLIANCE ASSURANCE

MEMORANDUM

SUBJECT: Random Audit of Inspection Data

FROM: Frederick F. Stiehl, Director Frederick F. Stiehl
Enforcement Planning, Targeting & Data Division

TO: Regional Enforcement Coordinators
Lead Compliance and Enforcement Data Stewards

The Enforcement and Compliance Data Quality Strategy serves as the guiding document for ensuring the Agency's enforcement and compliance data is of the quality to support program decisions, accurately reflect activities and accomplishments, and ensure public confidence. The FY 2002 Data Quality Strategy Implementation Plan outlines several data quality projects for implementation in FY 2002. These projects are an essential component of the Data Quality Strategy and have been designed in such a way as to provide a credible assurance of data quality with a limited strain on existing resources and those of our state partners. This memorandum announces the start of one of the projects for FY 2002: a data quality audit with a focus on inspection data. This audit was developed by a workgroup consisting of Headquarter and Regional personnel, in conjunction with contractor support. The Strategy calls for a continuation of these audits, one conducted per year, for the various enforcement and compliance data fields.

The purpose of these audits is to assess the quality of information maintained in EPA's national enforcement and compliance data systems. The audit also compliments the data review that will be conducted in preparation for release of enforcement and compliance data via the Internet. The audit assess the quality of the inspection data for the Clean Air Act (CAA), Clean Water Act (CWA), and Resource Conservation and Recovery Act (RCRA), as well as identify policy issues that may be affecting the quality of the data in inspection fields by comparing federally-reportable inspection data in EPA data systems (AFS, PCS and RCRAInfo) with the actual inspection records in state and EPA Regional files.

For each state in each Region, 8 randomly selected facilities have been selected in each of the three programs. To facilitate this record review, a website has been created where state and EPA Regional staff will be able to 1) retrieve inspection information for the selected program

identifiers (IDs) and 2) enter their findings as to whether their hard copy records match the national databases.

We are asking the Regional Enforcement Coordinators in conjunction with the lead compliance and enforcement data stewards to distribute via email the instructions attached to this memorandum to the appropriate state and Regional staff who have access to the hard copy inspection records for the CAA, CWA, and RCRA. The first attachment to this memorandum provides background information for the audit, including: issues considered in the development of the audit, the audit methodology, the audit instructions, the federally reportable inspection types, and the statistical methodology. The second attachment to this memorandum provides concise instructions and login information. You may distribute both attachments. The recipients can decide whether they want to read the background information, or simply read the instructions and log onto the website.

The recipients of your emails will be responsible for execution of the audit -- that is, pulling the hard copy records for the selected facilities; checking them against the records in the federal data system; and entering their findings for each record on the audit website.

The 3 step process is as follows: 1) when program staff log into the audit site, they will be provided with the list of program IDs that are subject to the audit in their program. 2) Staff will be asked to pull their program's hard copy files for the audited records which will include information about inspections. 3) The final step is for staff, with hard copy files in hand, to log back onto the website and confirm or revise the inspection data being displayed from the national data system. More detailed instructions and login information are provided in attachment 2 and at the audit website. The audit site will be activated for approximately one month, ending on May 15, 2002, at which time the audit review should be completed.

Thank you for your help. We look forward to sharing the results of this important work with you upon completion. If you have any questions, please call me at 202-564-2290 or David Sprague at 202-564-4103.

Attachments

cc: Regional Enforcement Division Directors, Regions 1, 2, 4, 6, 8
Regional Media Division Directors, Regions 3, 5, 7, 9, 10
Regional CAA, CWA, and RCRA Program Data Stewards
AFS Database Managers
PCS Database Managers
RCRAInfo Database Managers
Data Quality Strategy Workgroup