IFES

Macedonia Project

General Voters List – Database Integrity Analysis

Final Report

Presented by Dale Leake

Database Consultant

March 2001

INTRODUCTION

This paper has been developed as a result of the Database Integrity Test performed against the Macedonia General Voter's List (defined throughout the remainder of this document as the GVL) during the period of 23 February – 12 March 2001.

The approach and methodology used in the GVL database integrity analysis was defined in the **General Voter List – Database Integrity Analysis** document published during the week of 19 February 2001 with final amendments dated 23 February 2001. A copy of this methodology document can be found in Appendix A.

The notification and agreement to the terms of the database integrity analysis by the Ministry of Justice dated 23 February 2001 can be found in Appendix B.

The processing described and executed in the document has been done in complete cooperation of the Ministry of Justice and State Bureau of Statistics. In addition, this analysis has been done completely independent of these entities to assure a transparent analysis process.

Data used in this testing has been limited to the State Bureau of Statistics GVL database only. This data was loaded onto an IBM NetVista computer owned by the Bureau of Statistics and remained in a secured environment at the Bureau throughout the testing period.

No data contained on the GVL that is considered private data protected under any data protection regulations has been extracted from this IBM computer. Information extracted for further analysis by IFES contains statistical collections of information only.

All scripts and data files moved between the IFES laptop computer and the IBM computer were verified and documented in logs and signed by both the Bureau of Statistics representative and me. A copy of each signed log has been provided to the Bureau of Statistics for their records.

A more comprehensive description of each test, including the SQL scripts and Visual Basic code used to execute the integrity tests, can be found in the **General Voters List – Final Report Part II Technical Analysis** document provided with this report.

Detailed analysis graphs and charts of various database integrity test results can be found in the **General Voters List – Final Report Supporting Documentation** document provided with this report.

EXECUTIVE SUMMARY

The Ministry of Justice has initiated a review of the current legislative framework for elections. One area of review is the Macedonia GVL. The Database Integrity Tests performed are a result of an agreement of the Ministry of Justice to provide assistance in conducting such tests.

The government of Macedonia, political parties, and all citizens of Macedonia have a stake in ensuring that the GVL is as accurate as possible. A fundamental basis for a completely free and fair election begins with a valid GVL.

A database integrity test, like an audit, is designed to find problems or anomalies within the database. This analysis cannot give a definitive endorsement of the accuracy of the database. A more accurate verification process would include random field sampling or house-to-house canvassing which would be a more time consuming and expensive process.

With this database integrity analysis limitation acknowledged, this analysis testing is a quick and inexpensive way to identify possible flaws in the GVL. This testing can give a definitive statement whether any significant flaws are discovered depending upon the number and depth of the questions asked within the testing. The deeper the questions probe into the database, the higher degree of confidence of accuracy or inaccuracy can be determined.

As a result of database analysis testing and in-depth observations of random data records contained on the GVL, I have found enough probable anomalies to raise questions on the integrity of the GVL database. Based upon the GVL database alone, it is impossible to prove the problems without further analysis.

As my analysis criteria changed in order to sample various unique data element combinations, my results continued to identify the same result data sets along with additional duplication possibilities[1]

Based upon what I have tested and observed in my review of the GVL, I believe that there is a possibility for thousands to tens of thousands of duplications included on the GVL. This duplication appears to be due to invalid source data input and/or intentional (or unintentional) manipulation of the EMBG national identification number.[2]

[1] A result data set is the collection of GVL records that uniquely matched the query selection criterion. As testing progressed and the selection criterion broadened, the same duplicate records continued to appear with additional duplicate possibilities identified in each subsequent query result data set.
(Amended 26 March 2001)

[2] The EMBG number is the Unique National ID number. The breakdown of the EMBG is defined in the Law on Unique Personal Number of a Citizen. Excerpts of this law are included in Appendix C.

My analysis at this time cannot determine if 'suspected duplicates' are actually fraudulent or erroneous.

An example of what I have seen includes a duplicate GVL entry containing the exact same name and address, inhabited place and municipality codes with EMBG numbers that are similar, yet different and unique. Without further analysis, I cannot determine this example as an error in the data entry process or an example of intentional fraud. I believe, based upon my observations, there is a good probability that there are many examples of both.

Recommendations

In order to reach a reliable conclusion with respect to the input data source, I believe the following additional analysis should occur as an independent process of any and all of the Macedonian Government entities involved. This analysis must be done as objectively and as transparently as possible in order to achieve a reliable conclusion.

- An audit, or technical review, of the Ministry of Interior's Information Systems processing systems should be done to help identify holes in the process that would allow invalid data to be entered or passed along until it reaches the GVL database. This review may also include Field level processing of data.

- A Database Integrity Analysis, similar to what was just performed on the GVL, would help identify similar anomalies and provide further justification for the comparative analysis.

- A Comparative Analysis of the data contained in the GVL with input data source information at the Ministry of Interior, Ministry of Defense, General and Supreme Courts and other data source providers.

Estimated time for project completion is approximately 3 months.

2-4 weeks	Technical Audit Review Process
2-4 weeks	Database Integrity Analysis Process
4-6 weeks	Comparative Analysis Process

The estimates quoted are for the recommended analysis to be completed by an independent party. These estimates do not imply that the data cleanup has been completed in this time frame. The Information Systems processing modifications and the cleanup processing of invalid data identified as a result of this analysis will in fact take longer to implement among the various governmental bodies involved.

I believe the end result of this work will be an accurate GVL with integrity that will provide the basis and confidence for free and fair elections for all of the people of Macedonia.

METHODOLOGY EXECUTION

The process began with the installation of a new, out the box, IBM NetVista personal computer provided by the Bureaus of Statistics on 23 February 2001. This computer was preloaded with the Windows 2000 operating system. I personally installed the Office 2000 Professional software set and the SQL*Server 7.0 database software on the new machine.

To secure and limit access to the database content on the computer, I set up hardware and administrator level passwords on the machine. I placed these passwords in a signed and sealed envelope to protect, as well as guarantee access to the computer in the event I was not available. This envelope was given to Dan Blessington, IFES Macedonia Project Director, and placed at the IFES offices.

To verify that the GVL database content did not change, a row count query was executed daily to verify the total number of rows in the GVL.[3] In addition, a checksum process routine was developed to run against GVL database.[4] At the completion of the checksum process, a single line entry was automatically written into a text file on a floppy disk carried by me each day offsite. This process ran both at the start and end of each day of the testing cycle period.[5] The row count query and the checksum result was exactly the same each time these processes were run verifying that the GVL database was not altered in any way during the testing period.

Signature checklists were developed to document all activity that occurred during the testing cycle period. All items moved from the IFES laptop computer, all queries including the checksum validation routine, and all result data sets removed from the Bureau of Statistics IBM computer were documented and signed by the Bureau of Statistics personnel and myself. The original log documents have been bound and remain at the IFES Macedonia offices. The Bureau of Statistics verified all query executions and movement of scripts and data between computers.

On Monday, 12 March 2001, I visited the Bureau of Statistics offices to remove the MS Office 2000 Professional, MS SQL Server 7.0 database software and related database integrity testing files from the IBM NetVista computer. The hardware level password cannot be removed unless the Windows 2000 software is completely reinstalled. It was decided not to do this reinstallation and the password was given to Ilija Gjorgjevich.

[3] The GVL contains 1,634,859 total records.

[4] A checksum process evaluates the data as a mathematical equation resulting in a numeric hash value. Our process calculated a hash value of 114650724. The checksum process always calculated this same hash value throughout the integrity analysis testing. If any data changed during the testing, the checksum process would identify this change by calculating a different hash value.

[5] One exception to this was on the second day when it was decided to run additional queries rather than stopping for the last hour of the day on the checksum validation process.

DATABASE INTEGRITY TESTING ISSUES AND ANOMOLIES

The Database Integrity testing is broken down into 3 distinct categories of testing.[6] These categories include

- **Validation Analysis** – These tests are designed to validate the data values of the GVL. These tests checked for valid EMBG breakdown components, valid gender, municipality and inhabited place code values.

- **Demographic Analysis** – These tests are designed to identify possible anomalies in regards to Size of Household, Age and Gender breakdowns by Municipality as compared to National trends.

- **Duplication Analysis** – These tests are designed to identify potential duplicate occurrences of unique GVL data element combinations.

Important information has been identified and collected by each of these specific areas of analysis. Highlights of these tests are described below. A more comprehensive description of each test, including the SQL scripts and Visual Basic code used to execute the integrity tests, can be found in the **General Voters List – Final Report Part II Technical Analysis** document.

VALIDATION ANALYSIS

The tests for the validation of the GVL data components prove that the GVL database from a data structure perspective is fairly well intact. The validation issues listed below are not surprising to me.

The minor validation issues that were identified in the testing include:

- Identified 8 Invalid EMBG numbers.
- Identified 73 records with improper EMBG gender code usage.
- Identified 944 records containing invalid Inhabited place code and Street code combinations.
- Identified data elements that contained leading spaces. These included surname, name, street name, house number, and additional house number.

I suggest that this information be used to evaluate the processing systems at the Bureau of Statistics. If it is determined that validation checks for these issue are not included in the routines used to validate the input data used to update the GVL, I

[6] As defined in the General Voter List – Database Integrity Analysis approach and methodology document contained in Appendix A.

recommend that changes be implemented to the Bureau of Statistics' processing systems to prevent future occurrences of similar errors.

I also suggest that this information should be presented to the Ministry of Interior, and other appropriate government entities providing input data, for review and correction within their processing and database systems.

With these suggestions in mind, I see no need for further analysis in this area of testing.

DEMOGRAPHIC ANALYSIS

The intent of these tests is to get a view of the various demographic models by municipality as compared to the national demographic view of Macedonia as a whole, where it is possible and where it makes sense.

The demographic tests were broken down into 3 specific areas of analysis:

- **Analysis by Size of Household** – This test is used to provide an analysis of the number of Voters living in a uniquely defined place of residence.
- **Analysis by Gender** – This analysis shows the gender breakdown of persons listed on the GVL both nationally as well as by individual municipalities.
- **Analysis by Gender and Age** – This analysis shows the spread of voters by gender and age both nationally and by municipality.

Demographic Analysis by Size of Household

My tests in this area were based upon a definition of a unique residence that used a combination of data elements for inhabited place code, street code, house number, additional house number, entrance and apartment. The Bureau of Statistics provided this definition to me.

When I ran these tests, I found some rather large anomalies. Of the 123 municipalities, I found 7 that contained unique households containing 10s of voters, 60 municipalities containing households with 100s of voters and 56 municipalities showing thousands of voters within a unique household.

Some of these anomalies can be explained. The Bureau of Statistics identified that the place of residence identified in small villages throughout the country takes on the same value within the GVL. Small villages do not have street names, house numbers, etc. to uniquely identify a place of residence. I believe that the anomaly of the very large hundreds and thousands identified within the data sets is explained with this lack of street and house indicators in the local villages

I believe that there is still an unexplained anomaly on returned data sets indicating household sizes in the tens and hundreds. I understand a common practice in Macedonia is to have multigenerational families living in multilevel houses or housing units. The addresses to these units potentially would be common showing larger average numbers of voters within these types of units. Further analysis may help explain some of these household living arrangements, but not all.

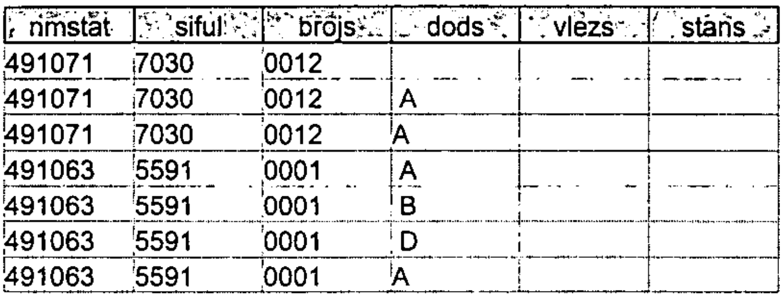
I believe further analysis should be done against the GVL database to identify data sets where these smaller sizes of household anomalies exist. My analysis in municipalities within the Skopje region where known populations reside in dwellings that should contain valid street and house number information produced results with households of questionable sizes in the tens and hundreds.

Additional input data source information from the Ministry of Interior is necessary to complete a more in-depth analysis in this area of analysis. I would suggest a comparison of current address information contained within the civil registry with the GVL address information for each EMBG national id to identify any inconsistencies or errors.

Distinct Unique Residence Identifier

Another analysis that was done included the extraction of the combination making up the unique place of residence. In this analysis I have found a potential issue where the additional house number, when used, contained a letter (i.e. A, B, C, etc.). In several instances, I found that the code generally left justified in the data column contained a leading space before the letter identifier. This leading space resulted in 3,542 additional unique places of residence to be identified within the GVL database. Further analysis of this data from sources at the Ministry of Interior is recommended to resolve any duplication caused by this anomaly.

Unique Residence Identifier Sample



nmstat	siful	brojs	dods	vlezs	stans
491071	7030	0012			
491071	7030	0012	A		
491071	7030	0012	A		
491063	5591	0001	A		
491063	5591	0001	B		
491063	5591	0001	D		
491063	5591	0001	A		

Demographic Analysis by Gender

The scope and intent of these tests performed against the GVL is to identify the distribution of the gender both nationally and at a municipality level.

The data results from the national view indicate a break down the middle at 50% for both the Male and Female category.

The view from the municipality perspective varies within the municipalities as much as 58% Male to 42% Female. My analysis of the GVL at this level does not provide enough information whether I should be concerned. Is a particular region heavy in manufacturing, mining or farming that would increase the likelihood that the male population would be greater? Does the makeup of a municipality by ethnicity or religion affect the gender distribution in that area? I do not know the demographics of particular regions within the country to determine whether disparities within gender distributions are valid or possibly flawed with invalid GVL information.

I suggest further analysis from the municipality level should be done to take into consideration the region of the country, the work environment and other population demographic factors to validate the anomaly seen at some of the higher discrepancy levels. Other data sources such as the census information may help validate these findings or show inconsistencies within the GVL.

Demographic Analysis by Gender and Age

The scope and intent of these tests performed against the GVL is to identify the possibility of data irregularities within age and gender groups identified on the GVL database.

Analysis has been done on both National as well as Municipality levels. The charts created using National data show trends that that would be expected in a normal population range. Both gender categories hover around the 50% mark of the population base. From a National perspective, these numbers do not immediately indicate that there are any irregularities or anomalies within the GVL database. The picture takes on a different viewpoint when you begin to look at similar analysis at each individual municipality.

The importance of looking at this demographic both nationally and by municipality is to identify any trends that indicate population groups that appear 'out of the ordinary'. This analysis identifies local trend anomalies that differ greatly from the national trends. It identifies gender/age groups that spike within the local demographic trends. These anomalies identify data within the GVL that could possibly be in error or intentionally altered or placed into the GVL.

Roughly half of the municipalities take on a similar trend as the national demographic analysis. With the other municipalities, analysis shows irregular distributions between the Male and Female populations within the municipality. Based upon this information, additional analysis should be done to further validate the voter population demographics within selected municipalities.

The data provided within the GVL is the only source of data used to perform this analysis. To provide further analysis, more in-depth breakdown analysis of specific municipality data sets will need to be done against the GVL. In addition, validation from input data sources maintained at the Ministry of Interior used to provide information to the GVL will need to be analyzed to provide verification of these potential irregularities.

DUPLICATION ANALYSIS

This analysis is intended to identify potential duplicate unique occurrences of data element combinations.[7]

These tests are designed to count the number of GVL data records that match the unique combination of selected data elements. If the database was perfect, the most specifically detailed combinations should result with a count that equals the total number of records contained in the GVL database. As the combinations of unique data elements start to change and broaden, the resulting counts returned identify variables of unique records within the database. The difference in these resulting counts as compared to the GVL total population indicate the number of records that are possibly duplicated on the database.

The significance of these tests is to provide an indication that duplicates exist within the GVL as well as provide the potential number of duplicates for a particular unique combination of data elements. More in-depth analysis looking at the detail of these possible duplicates on the GVL helps provide the real meaning and value to the analysis.

Some of the duplicates that were identified during these tests include:

- Identifying 519 unique Name, Birthday combinations
- Identifying 9,400 unique Name, Inhabited Place, Municipality, Gender and Age combinations.
- Identifying 22,000 unique Name, Street code, and Street name combinations.
- Identifying 148,000 unique Name, Municipality, and Inhabited place combinations.

Anomalies found in further analysis of the mentioned tests:

- EMBG numbers with birth month and day transposed or changed by days or weeks with all other data exactly the same.
- Duplicate names with similar address information, e.g. house number off by one or two digits, or missing on one entry completely.
- Appearance of improper input source data maintenance causing duplication, e.g. duplicate names with similar, but different address information and EMBG numbers.
- Identified various anomalies with data positioning, e.g. leading spaces on some data elements.

[7] The unique combination of data elements used to identify possible duplicates are defined in the General Voter List – Database Integrity Analysis approach and methodology document contained in Appendix A

What do the results of these tests indicate? For instance, what does identifying 519 unique Name and Birthday combinations mean? How significant are 9,400 unique Name, Inhabited Place, Municipality, Gender and Age combinations?

The tests run against the GVL to identify unique Name and Birthday combinations indicate that on the GVL there are 519 entries or, a total of 1038 records, that contain a shared name and EMBG birthday. Looking at the details of the duplicated data set, I noticed several familiar patterns within these data records.

One pattern that I noticed on some data records was that the data elements were exactly the same with the exception of the EMBG number. Within the EMBG number, I noticed that each component of the EMBG, with the exception of the gender ordinal birth number, was also the same. The gender ordinal birth number on suspected duplicates was sequentially in order in some cases or just one or two digits apart in other cases.[8] Another pattern with this analysis showed the similar patterns with the EMBG gender ordinal birth number along with an inhabited place code (also referred to as the settlement code) and/or address difference.

There were enough occurrences to indicate a pattern that two EMBG numbers were assigned at the time of the birth registration. Most likely the second number was assigned by an error in the processing. At some time since the original EMBG assignment, a person had an address change that was reflected on the valid EMBG used by that person. The second record referenced by the incorrect EMBG has probably never had any changes applied and has since resided dormant on the input source data files. Because both of these records are contained on the civil registry as valid EMBG entries, both records have been loaded into the GVL.

This analysis is based upon observations within the GVL database only. Further analysis to validate my theory should be done using data contained within the civil registry, birth and death registry and/or census input data sources. I believe further analysis of birth records, parent's names, and current address information should help identify and correct all of these 519 duplicates discovered on the GVL.

Another duplication test query showed that 9.400 duplicates occurred on the GVL when identifying unique entries based upon name, inhabited place code, municipality, gender and age. Extraction of the duplicates resulted with 18,448 total records identified for this combination of data components.

[8] For example, suspected duplicates were found with all data columns equal and EMBG numbers like 1308961450018 and 1308961450026. This example illustrates a Birthdate of 13 August 1961, register office code of 45, and sequential ordinal birth numbers of 001 and 002. Other suspected duplicates that were found contained similar EMBG numbers with ordinal birth numbers of 501 and 503, for instance.

NOTE: The last digit is used only for electronic validation and will most likely calculate to a different value. It was not be used as a part of this duplication analysis. The breakdown of the EMBG is defined in the Law on Unique Personal Number of a Citizen. Excerpts of this law are included in Appendix C. (Amended 26 March 2001)

Looking into the detail of these duplicates identified some of the most interesting analysis I had come across so far in the testing. In this analysis I observed anomalies within the EMBG number itself. The data identified by the test tells me that everything about the person's name, gender, age, inhabited place or settlement, and municipality is the same. I found that in many cases, the address information was exactly the same as well. The only difference in these cases was the EMBG national id number.

Looking closer at the EMBG number, I noticed the birth month and day were transposed in some cases. In other cases, I noticed birth month and days very close to each other, e.g. within days or weeks. I noticed birth years within 2-5 years of each other. There didn't appear to be a particular group of birth years affected with these anomalies; it spread through all of the decades since the 1920's. This anomaly gives the appearance that EMBG national id numbers were manufactured to provide for more than one EMBG for the same person.

As a result of these and all of the other duplication database tests along with my observations of random data records contained on the GVL identified by this analysis, I believe I have found enough probable anomalies to raise questions on the integrity of the GVL database.

Further analysis of all of the input data sources is recommended to provide a reliable conclusion to these findings. Unfortunately, until additional comparative analysis with the civil registry, birth and death registry, the census and all other input data sources is done, the exact number of duplicates is hard to determine. But in my opinion, based upon what I observed with data that appeared to be in error due to data entry or processing problems and data that appeared to be intentionally (or unintentionally) manufactured, I believe it is highly probable that the total number of duplicates found on the GVL could total into the tens of thousands of records. This can only be reliably concluded and validated with further analysis of all GVL input data sources.

PROPOSED FOLLOW-UP ANALYSIS

My database analysis testing and observations found enough probable anomalies to raise questions on the integrity of the database. But as mentioned earlier, based upon the GVL database alone, it is impossible to prove any problems as erroneous or fraudulent without further analysis. I believe there is a high probability that there are both.

I believe the following additional proposed analysis should occur in order to reach a reliable conclusion. This analysis should remain independent of the Bureau of Statistics or the Ministry of Interior. Only with an independent view of the data will the process continue to be as objective and as transparent as possible to achieve this reliable conclusion.

- An audit, or technical review, of the Ministry of Interior's Information Systems processing systems should be done to help identify holes in the process that would allow invalid data to be entered or passed along until it reaches the GVL database. This review may also include systems involving field level processing of data.

- A Database Integrity Analysis, similar to what was just performed on the GVL, would help identify similar anomalies, I believe, and provide further justification for the comparative analysis.

- A Comparative Analysis of the data contained in the GVL with input data source information maintained at the Ministry of Interior, Ministry of Defense, General and Supreme Courts including:
  Civil Registry
  Birth Registry
  Death Registry
  Any other data sources feeding the GVL

In addition to this proposed input data source analysis, it may be determined that the data used as input cannot give a definitive endorsement of the accuracy of the GVL database. It may be necessary to establish a definitive correlation between the names contained on the GVL and the eligible voters. This could only be accomplished by physically locating the voters, either by random sample field tests or by a comprehensive house-to-house canvass. This process would be slow and expensive, but provide a definitive endorsement of the accuracy of the GVL database.

Estimated time for analysis completion (not including any field tests) is approximately 3 months.

2-4 weeks	Technical Audit Review Process
2-4 weeks	Database Integrity Analysis Process
4-6 weeks	Comparative Analysis Process

It is important to note that these estimates do not imply the data cleanup has been completed in this time frame. The Information Systems processing modifications and the cleanup processing of invalid data identified as a result of this analysis will in fact take longer to implement between the various governmental bodies involved.

In addition, these estimates do not include any time required if field level tests are determined necessary to definitively endorse the accuracy of the GVL database.

Appendix A

General Voter List - Database Integrity Analysis

Stakeholders

The Ministry of Justice has initiated a review of the current legislative framework for elections. One area of review is the voters' list. Ministry of Justice has been engaged to provide assistance in conducting a series of database integrity tests.

The government of Macedonia, political parties, and all citizens of Macedonia have a stake in ensuring that the general voter list is as accurate as possible.

An attempt will be made to invite questions from a broadly representative sample of these groups. Ministry of Justice will solicit input from the Working Group, while IFES will seek input from political parties and from the international community.

The methodology for this testing, once developed, could serve as the basis for conducting regular audits of the voters' list in the future, either by the body in charge of maintaining the voters' list, or by the body responsible for certification of the list.

Methodology

A methodology should be carefully defined for three different phases of the integrity testing:

- Identifying the questions to be answered by the process
- Set up of the testing environment
- Categories of questions

Each of these will be covered in the following sections.

Identifying the Questions

It should be acknowledged that a database integrity test, like an audit, is designed to find problems. As such the database integrity test cannot give a definitive endorsement of the accuracy of the database. The accuracy of a voters' list can only be definitively verified by establishing correlation between the names on the list and eligible voters, and this can only be accomplished by physically locating voters, either by random sample field tests or by a comprehensive house-to-house canvass. These are expensive and slow processes.

While acknowledging the limitations of a database integrity test, such testing is a relatively quick and inexpensive way to identify possible flaws. It can only give a definitive statement concerning whether any significant flaws are discovered. If no flaws are discovered the degree of confidence in the database is determined by the number and types of questions asked. If only a few questions are asked that do not probe deeply into the data, the testing does not result in a very high degree of confidence in the accuracy of the database. If a large number of questions are asked, analyzing the database from a wide variety of perspectives, and no problems are found, the result is a higher degree of confidence in the database. It is, therefore, in the interest of all to ask as many deeply probing questions as possible.

To this end, we invite input from all sources. This document is being distributed to representatives of Ministry of Justice, Ministry of Interior, and Bureau of Statistics, to representatives of political parties, and to interested international organizations who have been involved in providing assistance or observing elections. We welcome all input as to additional questions that should be addressed during the testing.

The Test Environment

Principles

1. No list data will be removed from Bureau of Statistics office.

2. Data will be protected from alteration by any party during the progress of the tests.

3. Only those tests authorized by the Working Group will be performed on the data. The Working Group will publish a list of all questions considered, including explanations of the reasons for disallowing any question.

Procedures

1. The State Bureau of Statistics will provide a brand new in-box IBM NetVista computer. The computer will reside in an office that is locked outside of normal working hours. The computer will be password protected at the BIOS startup level and Administrator logon level to prevent access by anyone in the absence of the IFES consultant. Passwords will be known only to the IFES consultant, and will be stored in a sealed envelope at the local IFES office.[9]

2. The IFES consultant will not be allowed access to the computer without a representative from the Steering Group, or other monitor approved by the Steering Group, present.

3. All setup and operations of the notebook computer will be done by at least two persons, and a log will be kept of every action performed. Software to be installed is Microsoft SQL Server 7.0 and Microsoft Office Professional 2000.

[9] Amended 23 February 2001

4. Data will be transferred from Bureau of Statistics to the notebook computer through a formal handover process, and a receipt will be signed by all parties present.

5. Before conducting any tests, a row count will be done on all tables in the database, and the result will be recorded on the receipt. Also, a checksum will be calculated for each table, and the result will be recorded to allow detection of any alteration that may occur. These row counts and checksums will be used as a baseline to ensure that the data is not altered at any time during the course of the testing.

6. At the beginning of each day of testing a row count will be done and a checksum will be calculated on all tables in the database. These will be compared with the starting baselines; any deviation will be noted and no further testing will be performed until the discrepancy is corrected.

7. At the end of testing a row count will be done and a checksum will be calculated on all tables in the database, and these will be compared to the baselines. Any deviation will invalidate all tests.

8. The IFES consultant will carry no data identifying any individual by name, EMBG, or any other personal details from the Bureau of Statistics office. The consultant may take statistical information offsite on floppy disk for purpose of creating reports and/or charts. The monitor will inspect all files on any floppy before the consultant takes it off premises.

9. All data will be removed from the IFES notebook computer at the conclusion of testing. The hard disk will be formatted, and overwritten with a large file containing random text to ensure that the data is unrecoverable. Staff at Bureau of Statistics will monitor this step, and a representative of the Working Group will sign a receipt acknowledging that no data has been altered during testing or removed from the premise.

10. All data will be removed from the IBM NetVista computer at the conclusion of testing. Microsoft Office 2000 Professional and SQL Server 7.0 software sets will be removed from the IBM NetVista computer at the conclusion of testing. The hard disk will be formatted, and overwritten with a large file containing random text to ensure that the data is unrecoverable. Staff at Bureau of Statistics will monitor this step, and a representative of the Working Group will sign a receipt acknowledging that no data has been altered during testing or removed from the premise.[10]

Categories of Questions

In defining the types of testing that can be conducted, we are restricted to those tests that can be accomplished using only the data. It may be useful at some point to do additional

[10] Amended 23 February 2001

testing using outside sources of information such as the upcoming census, or field testing of a random sample of the database, but this is beyond the scope of the current tests which will only analyze the data itself. The following fields of data are available for conducting this analysis:

EMBG	Unique Registry Number of the Citizen
PREZIME	Surname
IME	Name
POL	Gender
NMSTAT	Code of the inhabited place
SIFUL	Code of the street
BROJS	House number
DODS	Additional house number
VLEZS	Entrance
STANS	Apartment
IMEULS	Name of the street
IMU	Polling Station Number
SERISKIBR	Serial number of the Voter ID card
KBRM	Control Number
SIFPROM	Code of the last change
RBR	Ordinal number of the issued ID card

1. Validation

Before any other testing, we will determine whether there are any inaccuracies in the data, indicated by impossible EMBG numbers, or codes for non-existent Municipalities, inhabited places, polling stations, etc.

- Number of numerically invalid EMBG's (invalid Date of Birth, invalid Issuing Authority, invalid Serial Number, or invalid checksum)
- Check for validity of all reference codes (NMSTAT, SIFUL, IMU)
- Count of voters by Polling Station compared to number of voters on list in last election

2. Demographic

These queries will analyze distribution of voters across age, gender, and household size, in order to determine whether there are any significant variations in trends.

- Percentage of voters by Gender (National and by municipality)
- Count of voters by Age, Gender, Municipality
- Voters by Gender, Birth Month, as a percentage of the total number of voters. (National and by municipality)
- Size of household (e.g. There are X households with 1 voter, Y households with 2 voters . . . Z households with 20 voters, etc.)

3. Analysis of possible duplicates

These queries will attempt to identify unusually high incidences of duplication in any of the following combinations (both at National and Municipal level):

- Surname, Name, Inhabited Place
- Surname, Name, Inhabited Place, Age
- Surname, Name, Inhabited Place, Gender
- Surname, Name, Inhabited Place, Age, Gender
- Surname, Name, Inhabited Place, Street Code
- Surname, Name, Inhabited Place, Street Name
- Surname, Name, Inhabited Place, Street Code, Street Name
- Surname, Name, Municipality, Age
- Surname, Name, Municipality, Gender
- Surname, Name, Municipality, Inhabited Place
- Surname, Name, Municipality, Inhabited Place, Age
- Surname, Name, Municipality, Inhabited Place, Gender
- Surname, Name, Municipality, Inhabited Place, Age, Gender
- Surname, Name, Age
- Surname, Name, Age, Gender

Appendix B

Republic of Macedonia

MINISTRY OF JUSTICE
Number 10-1390/3
February 23rd, 2001

TO :
THE INTERNATIONAL FOUNDATION FOR
ELECTION SYSTEMS (IFES)

SKOPJE

SUBJECT: Notification

The Ministry of Justice, in the framework of the Working program for 2001, also anticipated changes and amending of the laws regarding the elections. For that purpose, the Ministry of Justice formed a Working Group, in which there are included experts in this area and representatives from appropriate organs and organizations.

The Ministry of Justice also established cooperation with the International Foundation for Election Systems (IFES) for the draft-reforms in the electoral legislation. Besides the other activities that will be commonly exercising between the Ministry of Justice and IFES, testing of the data in the General Voters' List is also anticipated to be performed.

In the period from 23rd February, IFES has engaged an expert who will work on testing the data for the questions that will be determined by the Ministry of Justice, the Working Group, the Ministry of Interior, the State Bureau of Statistics, as well as other open questions that may occur during the General Voters' List testing procedure.

The testing of the General Voters' List database will be performed in the State Bureau of Statistics. The following subjects will be included in this operation:

From the Ministry of Justice:
1. Zagorka Tnokovska, and
2. Ilija Petrovski

The State Bureau of Statistics:
1. Ilija Gjorgjevich,
2. Liljana Vlaich,
3. Slobodan Karajovanovich, and
4. Aleksa Petrevski

From the International Foundation for Election Systems:
1. Dale G. Leake,
2. Translator

1. The testing of the General Voters' List data will be performed in presence of the above mentioned persons, and the legal provisions for protection of the personal data will be taken into consideration.
2. The IFES expert will be allowed access to the General Voters' List database for the purpose of its testing.
3. All open questions will be agreed upon by the representatives of the Ministry of Justice, the State Bureau of Statistics and IFES representatives.
4. Once the testing is done, a common report will be prepared for the questions that had been subject of process and it will be submitted to the Ministry of Justice, the Working Group and the State Bureau of Statistics.
5. Other persons, also, can be included in the procedure of testing the General Voters' List, if needed, after previous agreement from the Ministry of Justice.

/RP State Secretary,

Appendix C

Excerpts from the Law On Unique Personal Number of a Citizen

(This Law is published in Official Gazette of the Republic of Macedonia, number 36/92)

Article 1

Unique personal number of a citizen (hereinafter: personal number) represents an individual and unrepeatable mark of identification data on the citizen

Article 2

The personal number is composed of thirteen figures categorized in six groups:

I group: date of birth (two numbers),

II group: month of birth (two numbers)

III group: the year of birth (three numbers)

IV group: the number of the register (two numbers),

V group: combination of the gender and the ordinal number for persons born on same date (three numbers). Men from 000 to 499, women from 500 to 999, and

VI group: control number (one figure)

Article 3

The segment of the EMBG number that shows the date of birth (I group), the month of birth (II group), the year of birth (III group), and the gender (V group) is being determined on the basis of the data in the data birth register.

The control number (VI group) is being determined by an electronic processor.

If the data from paragraph 1 of this article is changed through a procedure envisaged by the law, a new unique number will be determined on the basis on the Decision on the basis of which the correction has been carried out of the data birth register.

Article 4

There are nine registration regions in the Republic of Macedonia, with the following register numbers: Bitola - for the municipalities of Bitola, Resen and Demir Hisar-41; Kumanovo – for the municipalities of Kumanovo, Kratovo and Kriva Palanka – 42; Ohrid – for the municipalities of Prilep, Krushevo and Brod – 44; Skopje – 45; Strumica – for the municipalities of Strumica, Valandovo and Radovish – 46; Tetovo – for the municipalities of Tetovo and Gostivar – 47; Titov Veles – for the municipalities of Titov Veles, Gevgelija, Kavadarci and Negotino – 48 and Shtip for the municipalities of Shtip, Berovo, Vinica, Kochani, Probishtip and St. Nikole – 49.

Article 5

The Ministry of Interior determines the unique number of the citizen.

The determination of the unique number and the registration of the same are automatic.

The Ministry of Interior provides preservation, usage and protection of the data from unauthorized access in compliance with the law.

Article 6

Unique number is determined according to the place of registering the newborn child in the data birth register that is led for the territory of the Republic of Macedonia.

Unique number for a newborn child that is born abroad is determined according to the parents' place of residence in the Republic of Macedonia.

Article 7

The Unique number for the foreigners residing in the Republic of Macedonia in compliance with the law and for whom records are kept and public identification documents are being issued on the basis of the law, the Ministry of Interior issues them a unique number for foreigners.