MAKING THE CASE

Investigating Large Scale Human Rights Violations Using Information Systems and Data Analysis

MAKING THE CASE

Investigating Large Scale Human Rights Violations Using Information Systems and Data Analysis

Edited by Patrick Ball
Herbert F. Spirer
Louise Spirer

HURIDOCS Cataloguing in Publication Data
TITLE: Making the Case: Investigating Large Scale Human Rights Violations Using Information Systems and Data Analysis
PERSONAL AUTHORS: Patrick Ball, Herbert F. Spirer, and Louise Spirer
CORPORATE AUTHOR: American Association for the Advancement of Science (AAAS) Science and Human Rights Program
PLACE OF PUBLICATION: Washington, DC, USA
PUBLISHER: American Association for the Advancement of Science
ADDRESS: 1200 New York Avenue, NW, Washington, DC 20005, USA
TELECOMMUNICATIONS: Tel: 1-202-326-6790; Fax: 1-202-289-4950; e-mail: shrp@aaas.org; World Wide Web: http://shr.aaas.org
DATE OF PUBLICATION: Spring 2000
PAGES: viii, 300
ISBN: 0-87168-652-X
LANGUAGE: ENG
STATISTICAL INFORMATION: Y
INDEX: Information systems / Human rights / Human rights violations
GEOGRAPHICAL TERMS: Universal
GEOGRAPHICAL CODES: 0000
FREE TEXT: Ten experts who worked on data processing, database representation, and generating analytical reports documenting large scale human rights violations tell how they created and used information systems.

ISBN: 0-87168-652-X

This report is a product of the American Association for the Advancement of Science (AAAS) Science and Human Rights Program which operates under the oversight of the AAAS Committee on Scientific Freedom and Responsibility (CSFR). The CSFR, in accordance with its mandate and Association policy, supports publication of this report as a scientific contribution to human rights. The interpretations and conclusions are those of the authors and do not purport to represent the views of the AAAS Board, the AAAS Council, the CSFR, or the members of the Association.

Printed in the United States of America

Table of Contents

Preface

In May 1999, the American Association for the Advancement of Science convened a weeklong meeting of ten information system experts in Washington, DC. These experts had all worked on creating and using information systems to document large-scale human rights violations in El Salvador, Guatemala, Haiti, and South Africa from 1992-1999. The combined experience included three truth commissions, a United Nations mission, and three non-governmental organizations.

During this meeting, they shared their experiences by presenting papers that were then jointly analyzed in detail, discussing the nature of the lessons learned, and developing recommendations for future work.

There were two purposes for investing the time and effort to achieve this free and open exchange. The first was to provide all attendees with a clear understanding of the issues and raise the group level of expertise. The second was to make available to those who will follow them their considerable experience and findings about information systems methodology for documenting large-scale human rights violations.

In presenting these papers, we hope to provide a history of the development of the technological and managerial processes used in our organizations. Our anecdotes and lessons learned may guide others who will want to build on these methods. Accordingly, we have edited them for uniformity and readability to make the proceedings a manual of how to determine *who did what to whom* (see Ball 1996). The reader can learn how to collect testimonies from a wide range of deponents, standardize concepts and vocabularies to create common categories across thousands of testimonies, design the computer data entry screens, structure the data into relational databases, and then how to adapt a database to meet the changing criteria imposed by changing circumstances. There are discussions about how to create statistical tables and charts and innovative methods to make supportable inferences about the magnitude of violence and its characteristics in time and space. The development of thesauri of vocabulary for use in reducing narrative information to coded form is discussed in several contexts. The appendixes provide sample pages from the working documents used on several projects.

Every paper includes or references a section on “Lessons Learned,” discussing problems, solutions, and recommendations for others. The Lessons Learned sections and the cited resources provide a guide to running large-scale databases with a high level of effectiveness and efficiency.

The experts who came together for that week in May 1999 are Patrick Ball, Themba Kubheka, Oliver Mazariegos, Rocío Mezquita, Gerald O’Sullivan, Eva Scheibrethner, Humberto Sequiera, Herbert Spirer, and Sonia Zambrano.

The editors would like to acknowledge the assistance of the following people: Priscilla Hayner, Neil Kritz, Brinton Lykes, Fritz Scheuren, and Audrey Chapman for sharing their time and insights; and Elisa Muñoz, Gretchen Richter, Eric Wallace, Matthew Zimmerman, and Margaret Weigers for helping with organizational matters. The editors are indebted to David Banks and Julie Carlson for their painstaking reviews of the final manuscript.

The AAAS Science and Human Rights Program would like to express its gratitude to the donors that have made this work possible: the Institute for Civil Society and by the John D. and Catherine T. MacArthur Foundation.

Finally, we would like to acknowledge the United Nations missions, truth commissions and non-governmental organizations with whom we have worked: in El Salvador, the non-governmental Human Rights Commission (CDHES); in Guatemala, the Commission for Historical Clarification (CEH), the International Center for Human Rights Research (CIIDH), the UN Verification Mission for Guatemala (MINUGUA), and the Catholic Church's Interdiocesan Project for the Recuperation of Historical Memory (REMHI); in Haiti, the National Commission for Truth and Justice (CNVJ); and in South Africa, the Truth and Reconciliation Commission (TRC). On behalf of the experts, we would like to say that we have felt honored to have had the opportunity to contribute to these projects, and we wish our future colleagues in human rights information management all the best.

Patrick Ball, Herbert F. Spirer, and Louise Spirer, editors
June 2000, Washington DC and Stamford CT.

Author Biographies

Patrick Ball, Ph.D., is Deputy Director of the American Association for the Advancement of Science (AAAS) Science and Human Rights Program. Since 1991, he has designed information management systems and conducted quantitative analysis for large-scale human rights data projects for truth commissions, non-governmental organizations, tribunals and United Nations missions in El Salvador, Ethiopia, Guatemala, Haiti, South Africa, and Kosovo. His 1997 Ph.D. dissertation "Liberal Hypocrisy and Totalitarian Sincerity" examined the roots of the non-governmental human rights movements in Ethiopia, Pakistan and El Salvador. AAAS has published three previous books by Dr. Ball: *Policy or Panic? The Flight of Ethnic Albanians from Kosovo, March-May 1999* (2000), *Who Did What to Whom? Planning and Implementing a Large Scale Human Rights Data Project* (1996), and *State Violence in Guatemala, 1960-1996: a Quantitative Reflection* (1999, with Paul Kobrak and Herbert F. Spirer).

Herbert F. Spirer, Ph.D., is Adjunct Professor of International Affairs at Columbia University, Professor Emeritus of Operations and Information Management of the University of Connecticut, and consultant to the AAAS Science and Human Rights Program. He has been a consultant to many NGOs on data analysis for human rights, and is a past Chair of the American Statistical Association's Committee on Scientific Freedom and Human Rights, and a former Vice President of the Institute for the Study of Genocide. He is co-author of the AAAS publication *Data Analysis for Monitoring Human Rights*. He was made a Fellow of the American Statistical Association in recognition of his achievements in applying statistics to human rights.

Louise Spirer is an independent scholar, editor, and author in the field of human rights. Co-author of articles on human rights, she is the editor of newsletter of the American Statistical Association's Committee on Scientific Freedom and Human Rights, a member of the Board of Directors and Treasurer of the Institute for the Study of Genocide, and a co-author of the AAAS publication *Data Analysis for Monitoring Human Rights*.

Sonia L. Zambrano Gómez is a Colombian anthropologist and lawyer. She has worked in this country as human rights researcher, and she has written publications about this subject. She also worked for the Historical Clarification Commission of Guatemala, as Director of the Database.

Themba Kubheka is Deputy Director in the Information Technology of the Department of Land Affairs in the South African Government. His main function is to empower regional management to participate in the broader Information Technology plan. Themba has worked for Macro International Inc – a US based multinational funded by USAID - as their Management Information System Specialist. From April 1996 to February 1998, Themba worked for the South African Truth and Reconciliation Commission (TRC) as its Information Coordinator. Later he also served in the position of the Documentation Officer. In his 15 years in IT, Themba has conceptualized, designed and written numerous computer applications. In his most recent experience with the TRC, he assisted in the development of the database and the processing of the Human Rights Violations statements.

Lic. Oliver Mazariegos was the programmer and systems administrator for Guatemalan Archbishop's Human Rights Office "Proyecto Interdiocesano Recuperación de la memoria Histórica" (REMHI).

Rocio Mezquita, B.A., has worked on human rights projects as a data processing professional REMHI as well as in the Guatemalan Truth Commission. She was previously an intern with Amnesty International/USA and worked as an election observer in the former Yugoslavia with the Organization for Security and Cooperation in Europe (OSCE). She is presently working as a researcher in Guatemala, in a human rights project at the Center for Legal Action in Human Rights (CALDH).

Gerald O'Sullivan was the National Information Systems Manager for the South African Truth and Reconciliation Commission. He has been in the IT industry since 1981, working primarily on financial and management information systems in South Africa and abroad as an exiled war resister. He is currently the Director of Information Systems in the Department of Land Affairs, implementing GIS technology to facilitate the redistribution of land.

Humberto Sequeira, is the senior software programmer and database designer at Solo Software Development in Panamá where specializing in Point of Sale software for Hospitality environments. The most challenging and rewarding job he has been part of is the Truth Commission (CEH) in Guatemala. He is very interested in how new technologies in software, communications and database development will fill the technical gap between data and Human Rights researchers.

Eva Scheibreithner, as a student of international economics in Austria went to Guatemala in 1996 and started working as a volunteer human rights worker with national NGOs and communities of returnees. In 1997 she joined the CIIDH project in Guatemala working as a data processor and analyst. The Guatemalan Truth Commission (CEH) in 1998 was her second human rights data project, she worked on statistics there.

Ken Ward, B.S., is a computer database consultant. He has designed Human Rights Violations database systems for the United Nations Mission in Guatemala and various non-governmental organizations in Cambodia. He has designed several systems related to the Central American Peace Process in El Salvador and Guatemala and has also worked as a Human Rights investigator in Guatemala.

Introduction

Patrick Ball and Herbert F. Spirer

> A truth commission can promote reconciliation, outline needed reforms, allow victims a cathartic airing of their pains, and represent an important, official announcement of a long-silenced past.
>
> Priscilla B. Hayner
> *Commissioning the Truth.*
> *Third World Quaterly, Vol 17, pp. 19-29, 1996*

Overview

Telling the truth in such a way that it cannot be denied is the first need of a truth commission established in the aftermath of gross human violations. The magnitude of violations is often so great that individual researchers cannot apprehend the complex nature and multiple patterns of such crimes, building an official history from a *collective* memory is essential to truth telling. This is our concern in these proceedings: building such a collective memory, and the analysis of the past through examination of that memory.

To the reader:

This introduction summarizes our concept of the relationship of the information management system issues to the truth telling process. In the course of this summary, we frequently reference sections and chapters in these proceedings. To facilitate your use of this introduction as a guide, we have given the relevant references in boxes such as this, associated with the related text.

While the primary goal of truth telling is to provide massive and objective support for historical facts and patterns that cannot be denied, it also serves an "internal" role for those who analyze the past to make the official record. Without an accurate and precise collective memory that can be readily accessed, they will not be able to check their assumptions about the process of violations, or provide credible analyses.

The official record is derived from the collective memory, and the collective memory is based on *information* and *data*. The systematic arrangement of the information and data is the basis of the *information management system.*

These proceedings are about all aspects of **how** to build, manage, and generate analyses from such a system. They provide an accessible handbook to guide truth tellers who want to build on the lessons learned in these several information systems.

Fundamental to our concept of truth telling in human rights is determination of *who* did *what* to *whom* and *how.* You will find this concept discussed in detail In Chapters 1, 3, 4, 6, and 9.

In this introduction, we discuss the conceptual issues pertaining to the use of information management systems in the truth telling process. The discussion is grounded in the theory and application presented in the papers in these proceedings.

Purposes

When an organization concerned with truth telling in cases of gross crimes against humanity – an official truth commission or a non-governmental organization – sets out to write official histories, it often undertakes massive research projects. These projects may use hundreds of people working in thousands of communities to acquire information. The organization may be charged with gaining an overall understanding by generalization based on the entire body of evidence in addition to reporting on individual cases.

Chapters 3 and 4 discuss the South African Truth and Reconciliation Commission (TRC), the largest human rights data project ever conducted.

These tasks require bringing all the collected information together and analyzing it. By so doing, what all the many individuals in the organization have discovered becomes the organization's understanding of the truth.

Through these projects, the organizations that document large-scale human rights violations collect much more information than any one investigator can remember or fully encompass. Further, they may perform general analyses or correlate information from geographically dispersed sources. Information about a given case could be given to any member of the teams of investigators, who may number in the hundreds. In a given case, partial information could be given by people in the southwest of the country (where the case happened), while other information about the case is given to investigators in the northeast (where survivors fled after the incident). An investigator working on this case may not know that other investigators in a different part of the country have found complementary information.

The information management system provides a collective memory and the ability to relate information from different sources. By so doing, it allows anyone in the organization to access information collected by any investigator, without restriction. An information management system used for these purposes is a *process* by which information is collected, standardized, represented in a database, and then analyzed by a variety of methods. The database – the computers and software in which the data reside and by which it is processed – is not "the system," it is a major component of that system. The human rights narratives collected by the organization are complex, as are the legal and social science processes used to classify components of human rights stories. The complexity of the information management system and in particular, the database, reflects the complexity of the narratives and the legal and scientific concepts necessary to serve the cause of truth telling.

To effectively make information widely available with precision and consistency, the information management system must standardize the classification and categorization of information. For example, if a witness reports to the commission that a person was tortured, the appropriate information system personnel decide whether the acts described by the deponent fit the organization's definition of torture. When witnesses and victims describe where events occurred, they often describe the location in casual terms. To convert this narrative information to data that will represent the truth in the database, the data processors must, for example, decide where on a map the events happened, and classify the events by suitable location designations. Painstaking and precise classification is necessary to assure that the data are of high quality, but not sufficient to do so. The entire system must also be of high quality for the system outputs to be credible and valid.

Standardization, classification, and categorization are discussed in all the chapters. Particularly detailed examples of both the technical and managerial issues involved appear in Chapters 3, 4, 6, 9, and 12.

Credibility and Validity

Once an organization has collected data and presented its analyses based on those data, critics may argue that the data do not support the organization's conclusions or analyses. Our experience shows that criticisms fall largely into three categories.

First, critics may argue that the methods are flawed. The structuring of human rights data is a complex process and there are many possible sources of data errors that ultimately lead to statistical results that distort the truth. Aside from the usual errors that plague statistical work (reliability of data processors and investigators, bias in the interview process, numerical and typographical errors, etc.), the most egregious errors result from oversimplification. This latter category of errors is often difficult to fully comprehend and may become apparent only in the process of analysis. One example of such an error occurs if a victim suffers multiple violations in one event but only the "worst" violation is reported. To find the balance between a simplification that makes the data easy to analyze without distortion, and oversimplification that seriously distorts the results is an ongoing challenge. Shortcuts are dangerous, and the structuring of the data in the database calls for care and open debate, not haste. Oversimplifications invariably distort the results.

The consequences of such practices are discussed in Chapters 5 and 6.

Second, critics may argue that the chosen interview subjects are not representative of the population of all victims. Even if a group has taken testimonies from many thousands of subjects, there are probably many others who were victims or witnesses of human rights violations but were not interviewed. The data might therefore be biased, reflecting only the knowledge of those who were subjects. In this context, bias means that in some way the patterns shown by the data are a systematically distorted reflection of the historical reality. We discuss bias in more detail later in this introduction.

Third, critics may argue that the data are inadequate substantiation for the organization's arguments. For example, an organization might find in their data there were 100 killings reported for the year 1978, yet only 10 killings were reported for the in the prior period from 1960 to 1977. On this basis, the organization might want to argue that 1978 was a watershed year of dramatically increased violence. A critic might respond that showing only 10 killings in the prior seventeen-year period reveals that the organization failed to adequately investigate that period. If the critic is able to show even a few killings from the 1960-1977 period that were excluded from the original analysis, the entire argument might be doubted.

If interview subjects have been chosen by appropriate probability sampling methods, all three criticisms may be rigorously evaluated (and hopefully rejected). The use of probability sampling allows the analyst to scientifically determine that the results are valid within a measurable margin of error (the *confidence interval).* In practice, few human rights projects can use probability sampling. Such sampling can be technically complex and is time-consuming, costly to administer, and difficult to carry out in the chaotic conditions that follow gross human violations.

Some human rights projects assume that conducting an interview with a witness may help that witness come to terms (psychologically) with what happened. Thus, those projects invest resources in taking more interviews, rather than of obtaining fewer interviews by scientifically rigorous methods. Also, in the event of large numbers of deaths – many of which were not witnessed by any survivor – the sampled population is not the same as the target population.

Some human rights projects claim that their data are valid because they collected "very large" numbers of interviews. On the surface, "very large" is scientifically meaningless, for who is to decide what is "very large"? Should this term be referred to an absolute number, such as "several thousand" interviews, or "more than 5,000." The numbers of testimonies collected for three of the projects described in these proceedings are 7,000 for the CEH and the Haitian National Commission for Truth and Justice, and 21,000 for the TRC. Or should it be based on a relative amount, some percentage of the estimated total number of witnesses, survivors, or victims? And once again, who sets a satisfactory threshold for a "sufficiently high" percentage? And furthermore, how does the project estimate the total number of witnesses, survivors, or victims?

It is possible to answer the question as to how large is large enough. The critical assumption is that the project has collected enough interviews to merit the statistical findings, if it is unlikely that an equal or larger number of interviews would tell systematically different testimonies. It is certain that there are some interviews that tell different stories, but if enough interviews have been collected, it may be implausible that there are enough potential (but omitted) witnesses whose stories are so different that the findings would change substantially if the omitted witnesses were included. After collecting thousands of testimonies, and if other kinds of data are available about the patterns of gross human rights violations, we can test for bias using certain analytical methods. We describe some of these methods in the analytical objectives and bias sections below.

It is basic to the process that in practice a human rights organization cannot document every violation that may have occurred, if for no other reason than the fact that many victims may have been killed without witnesses and without any remains. Thus, the truth-telling human rights organization must define its broad analytic objectives explicitly and with attention to the needs and resources. Despite resource limits on the depth and scope of the work of the organization, the organization's sponsoring bodies may mandate that it gets a "complete" picture. To the non-scientific personnel on the body that makes this mandate, this might mean that the organization is to document every violation. Even recognizing the above limitation on collecting complete data, this is enormously expensive. With limitations of time, of availability of skilled personnel, and of jurisdiction, it is undoubtedly impossible. In their negotiations concerning their objectives and in their final report, the organization must clearly explain these limitations. The organization may only be able to ascertain patterns and trends, and cannot enumerate every possible violation. Given a

general mandate, the organization must be prepared to explicitly state its analytical objectives. Typical objectives are listed in the next section.

Analytical objectives of large-scale human rights data collection

Once an organization has collected large-scale data, processed it, and represented it in a database, it can choose among many analytical options. Four broad categories of analytical uses of large-scale data are listed below.

Filing and searching

The database is an efficient filing system that allows the use of complex criteria to access the equivalent of hundreds of thousands of hard copy pages of interview records. Thus, the organization can quickly search for particular people and events and combinations of people, events, times, locations, and so forth.

Description

Building on filing and searching, the organization can seek answers to questions such as these: How many acts of severe ill treatment occurred in May 1983? Did the number of people detained increase or decrease from 1986 to 1987? Were the monthly numbers of people tortured during states of emergency greater or less than months in which there was no state of emergency? In Nebaj, Guatemala, were a higher proportion of indigenous people or non-indigenous people killed? Questions like these can be answered by querying the database and obtaining flat data sets from which an analyst can create highly informative charts. These charts describe patterns and trends in the historical reality being studied, and give a full *picture* of the findings.

The process of querying the database to answer such questions is discussed in detail in Chapters 1, 2, 7, 10 and 11.

The criteria for effective charts are given in the Chapter 7, Graphs: The Visual Display of Information.

Inter-sample validation

If an organization has access to multiple databases about human rights violations, each database can be used to check the others. For example, at the National Commission for Truth and Justice in Haiti (CNVJ), data on killings were collected by more than 7,000 interviews. The CNVJ also collected the records kept by the hospital morgue in Port-au-Prince on violent deaths. Analysts for the American Association for the Advancement of Science (AAAS) compared the number of violent deaths in each month reported by the morgue to the number of killings reported by the interviews. Although only a few killings were reported in both sources, the monthly numbers of deaths in the two sources were highly correlated. This is strong support for the hypothesis that the two data sources measured the same social phenomenon of repression, validating both measures. Analysis of this kind can also be used to measure bias, or to reject the hypothesis that the data were biased.

The discussion of inter-sample validation for the CNVJ appears in Chapter 2.

Projection

It is impractical to interview every potential witness and victim to obtain a count of the total number of violations; but it may be possible to estimate the total number of violations by use of multiple independent data sources. With multiple data sources, each violation that occurred during the historical period being studied may be reported in one or more sources, or may not have been reported to any project. We can derive an estimate of the total number of violations -- those reported plus those not reported -- from the proportion of cases that occur in more than one of the data sources (the *overlap rate*). The higher the overlap rates the smaller the number of cases that we can estimate to have been missed by all of the projects. Such an estimation is important in situations of gross human rights violations because a scientifically informed estimate of the total number of violations can be given without either interviewing everyone in the country or taking a probability sample.

The use of multiple data sources to derive an estimate of total violations is described in Chapter 11.

These are a few of the basic techniques. There are many others and many variations of each. Essential to any use of these techniques is the availability of researchers capable of formulating

meaningful questions in terms that can be answered by analysis of the data, and analysts who can implement the relevant statistical methods.

Collecting Information

The first step in information management is data collection, the process of getting information to manage. For most truth telling organizations, the primary source of information is interviews with victims and witnesses of gross human rights abuses. Other sources are documentary records of non-governmental organizations and reports in the various forms of public media.

The successive steps involved in an information management system are Collecting Information, Data Processing (Classification and Coding), Database Representation, and Generating Analytical Reports. All chapter titles reflect this structure.

Assuming that the dominant source is interviews, the first priority is to design an interview *process* (forms, approaches to the subjects, training programs for data collectors, and so forth). A primary goal of this design is to assure that the person giving testimony (the *deponent*) will feel that his suffering has been acknowledged and made a part of the public record. As mentioned earlier, many people in truth telling organizations believe that giving the deponent an opportunity to be heard is a cathartic process. Although recent research has questioned these premises, it is still clear that a conversational interview mode, in which power is shared between the interviewer and the statement giver, is much less likely to re-traumatize people relative to an interrogation using closed-ended questions and an aggressive or police-style interrogatory style. In addition, and the quality of data obtained by interrogation methods is not as good as that obtained by conversational methods. While researchers have questioned these premises as general principles, in any given case they may apply.

You will find this issue discussed in Chapters 3 and 6, with reference to the collection of information in South Africa and Guatemala.

However the interview is structured, the information must be gathered so that the data processors can determine *who* did *what* to *whom* from the interview notes. The interview process must be designed to manage even the most complex stories. The narrative is often complex because each narrative can contain from one to many victims, violations, and perpetrators, and they may be related to each other through complicated relationships. Because individuals remember in different ways, important questions should be asked several times in different ways, via direct questions and in open narratives.

For a flow chart of a data model that reflects these relationships, see Figure 4 in the section The Data Model of Chapter 4.

The basic elements of a human rights narrative are:

Many victims

A deponent may speak about gross or associated violations that happened to one victim, or that happened to many victims. Her story, for example, may discuss only her own detention and subsequent torture. However, in addition to her own story, she may speak about her son's killing and her husband's disappearance. The witness may or may not herself be a victim.

Many violations

Each of the victims described in the statement may have suffered one or more gross violations. For example, the witness's son may have been detained and tortured on several separate occasions before he was killed. These violations may have been connected to other violations that occurred at the same time and place (e.g., several different people who were detained and tortured together), or they may have been isolated incidents.

Many perpetrators

Each of the violations described in the narrative may have been committed by one or more identifiable perpetrators, or by one or more unidentifiable perpetrators. The witness may or may not have seen the violation occur. For example, she may have been notified that her

The UN Verification Mission in Guatemala (MINUGUA) used method 1) in reports prior to 1996, but then reformulated their system (see Chapter 5). TRC statements after August 1996 were based on method 2). The data processors used qualitative information to recover uncoded additional violations (see Chapter 4) The TRC statistics probably underestimate violations that occur more than once to the

son's body had been found. In such a case, she might be unable to identify any perpetrators. If the witness was herself a victim, she may be able to describe the organization to which the perpetrators of her violations belonged. She may also have personally recognized one or more of the perpetrators or the identity of the perpetrator's organization. Furthermore, each of the identified perpetrators in the narrative may have been responsible for one or more violations. For example, the witness may identify the individual responsible for both her torture and her son's killing.

In the interview process and all subsequent steps of data processing and representation, the information system must maintain the identity of *who* did *what* to *whom*, without simplifying the witness's story in ways that distort it or systematically conceal certain kinds of information. The decision either in the design of the system or in the implementation of the interviewing process to accept a reduced version of a complex story is a frequent cause of this kind of distortion. For example, 1) a system might choose to represent only one of the violations that happened to a particular victim, or 2) to represent only one of each kind of violation. Both of these choices distort the data, and quantitative analyses based on these simplifications are not reliable. Fortunately, if there is sufficient narrative information in the form of qualitative descriptions of what happened, data processors usually can recover good information from distorted interview forms, but at considerable effort.

Data Processing

Data processors receive the essentially raw data from the interview narratives and prepare it to be entered into the database. In so doing, they extract the names of victims, perpetrators, and organizations, and apply standard definitions of types of violations and geographic locations. For example, consider the following narrative:

Detailed descriptions of how data processing worked at the CEH and at the TRC, respectively, can be found in Chapters 3, 8, and 12.

> Two days ago, heavily armed men in green uniforms came to my house and demanded to see my son. I asked if they had a warrant and I didn't want to call my son but they ignored my questions and threatened to fire their weapons into the house if I didn't open the door. My son heard them and came near the door. They broke through the door, grabbed my son and were hitting him. Then they took him outside and put him on a truck and drove away. I am pretty sure I recognized some of the guys from the local police station, but when I went there, they claimed not to know anything about it. But a neighbor of mine heard from his cousin who is a police officer that they had my son and they took him to the military detachment over by the highway.

Data processors may take the information above and put it in a structured form as in the tables of Figures 1a and 1b, below. Of course, the exact nature of the tables used depends on the design of the particular information management system.

Figure 1a. People Table.

ID code	Name	Sex	Birth Date	Ethnicity	Profession
P001	Jaime Raimundo	M	26 April 1972	Ixil	Student
P002	Catarina Raimundo	F	5 May 1950	Ixil	Housewife

Figure 1b. Violations Table.

ID code	Date	Place	Violation type	Alleged perpetrator	Source testimony ID code
P002	11 Sep 1999	Victim's house, Nebaj, El Quiché	Threat	Local police	P002
P001	11 Sep 1999	Victim's house, Nebaj, El Quiché	Abuse of authority	Local police	P002
P001	11 Sep 1999	Victim's house, Nebaj, El Quiché	Illegal detention	Local police	P002
P001	13 Sep 1999	Police station, Nebaj, El Quiché	Disappearance	Local police, military detachment	P002

Figure 1 reveals several characteristics of the structuring of the data. First, as discussed earlier, each victim can suffer one or many violations. Catarina (P002) suffered one violation (threat), while Jaime (P001) suffered three violations (abuse of authority, illegal detention, and disappearance). One perpetrator may commit some violations (such as the threat against Catarina), while more than one perpetrator may commit other violations (such as Jaime's disappearance).

See Chapter 3 for a detailed discussion of a creative approach to the process of defining categories and the resulting tables of definitions.

Second, the data processors are the people in the organization who take each story and decide whether the evidence is sufficient to classify the acts described in the story as violations according to the agreed definitions of the organization. Was the beating the perpetrators gave to Jaime Raimundo sufficient to be considered an abuse of authority? The data processors apply the organization's rules and classifications to make this decision. By applying these rules and standardizing the disparate information, the data processors create an organizational memory that can be accessed by any member or part of the organization. The classification rules determine what the commission will be able to analyze. Thus, "What constitutes a violation?" is a question the commission should address at the earliest possible moment

Chapters 2, 3, 6, 8, and 12 give extensive listings of human rights violation categories and associated definitions.

Many of the concepts about human rights violations are hard to define, such as severe ill treatment or massacre. These two concepts were central to the work of the South African and Guatemalan commissions. In the Haitian National Commission of Truth and Justice, extortion emerged as one of the primary human rights abuses committed under the *de facto* regime. After all of the data had been processed once, the data processors had to revisit every case to re-code for extortion.

If after all the data processing has been done, a category turns out to be important, the data must be re-coded. Although this is time-consuming, re-coding is much faster the second time. However, neither organization had a clear definition of these concepts until several months after data processing work had started. The data processors' work is to apply definitions. Hence, when definitions are unclear, the data processors are the first to initiate demands that the organization establish clear working concepts. Unfortunately, such determinations involve many actors and are often influenced by political factors. When the organization cannot obtain consensus on the definitions of key concepts, the data processors must develop provisional working definitions in such a way that they can later re-code the data when the debates are finally settled.

See Chapters 6, 9 and 12 for discussions of the development of the concept of massacre in the CEH information management system.

The data processors' work prepares the information to be represented in a computer-based database, usually in a relational structure.

Database Representation

There is a common tendency to conceive of the total process in terms of the computer hardware and software components. However, specifying the hardware and writing the software are the easiest parts of the work. A qualified database programmer can implement and test a human rights database in about one month. In our experience, human rights projects are so different from each other that it is ineffective and inefficient to develop a standard software program that must be customized for each project. In the six projects we personally have worked on in the last eight years, none of them could have shared their database software with the others. This is the case even though they all shared certain design characteristics. Today, all that is needed is that the software supports relational structures; the computer language in which it is written does not matter. Good human rights databases have been written in Paradox (in 1991-1993), Oracle, Access, and FoxPro.[1]

The need for customization of the database representation and its implementation is discussed in Chapters 4, 5, 6, 9, and 12. It is a primary concern in system design.

However, it is important for organizations to recognize that they will need a full-time staff programmer to write and maintain the software and to use queries to extract data in formats appropriate for the analysts. Organizations too small to hire a programmer should contract with a private-sector firm to write and maintain the software they need, or they may be unable to carry out their essential functions in a timely manner.

When making decisions about software, decision-makers often think in terms of *compatibility*. In human rights data projects, compatibility depends on the classification structures used by the data processors much more than on the computer software used to store the data. If two systems share the concepts and definitions about what human rights violations are, then a programmer can transform the data from one software package to another no matter what software was used originally to implement the systems. In fact, analysts may transform the data into three or more different formats to use different packages that offer different tools. If the systems have differences in their concepts and definitions, then even if the databases are both written in the same program, the data are incompatible.

Thus, from the perspective of an organization's leadership, the critical questions about the database are: What does the database contain? What is the meaning of the information contained there? We discuss these issues in the next section.

What is the Database?

A human rights database has two principal functions. First, the database preserves, standardizes, and represents information that the organization gathers. This is true even if the same information is represented many times, which human rights organizations often refer to as the *problem of duplicated cases*. Second, the database represents a unique set of incidents (involving people, places, violations, and organizations) that **in the group's judgement** happened in the situation of interest. The database must fulfill both objectives, but it can be difficult to design the system so that both functions are achieved concurrently.

The organization collects data drawn from hundreds, or possibly thousands of testimonies, press clippings, secondary materials, documents, and physical evidence, which are collectively called evidence when discussing a particular case or victim. The relationships among the entities stored in the database may be many-to-many, many-to-one, one-to-many, or one-to-one. For example, a violation may be documented by one or more pieces of evidence (one-to-many), or a victim may have a unique official identification number (one-to-one).

More specifically, the killing of Juan Perez in County Y in May 1983 may be documented in three testimonies (e.g., from Mr. Perez's son, his priest, and his widow). There may be forensic evidence of the killing from an exhumation, and the killing may have been reported in the contempora-

[1] Ball, Patrick, Ricardo Cifuentes, Judith Dueck, Romilly Gregory, Daniel Salcedo, and Carlos Saldarriaga. 1994. *A Definition of Database Design Standards for Human Rights Agencies.* Washington, DC: American Association for the Advancement of Science and Human Rights Information and Documentation Systems International, a discussion of human rights database design, is available at http://shr.aaas.org/dbstandards/cover.html.

neous press from which we have two clippings. When all the evidence has been collected, the organization must decide how to save the information about the killing. If the evidence comes to the organization in independent streams, the researchers may not recognize until later that all of these pieces of evidence relate to the same incident. Confounding the issue is that the facts are often slightly different among different sources. But if we save all the different pieces of evidence documenting Mr. Perez's killing, we will have six distinct representations of this one incident. Simple statistics done on this information would count Mr. Perez six times, which is obviously an error. Groups that choose to keep all the accounts simultaneously are deciding that the database is primarily serving the first principal database function, as a faithful representation of the sources, and not the second function, establishing the "true" event.

In the above example, an organization might try to eliminate the duplication by choosing one of the sources and deleting the others. By keeping only one reference to Perez's killing, the organization can make sure that their statistics are correct and clear – Mr. Perez will only be counted once. Cleaning the data in this way is deciding that the database is to be a true representation of the historical events, and thus deciding not to represent all the data that has been collected. This is a use of the database in its second principal function, representing what is believed to have really happened. In effect, the database that has been created looks like what is shown in Table 3, below.

Table 3. Sample database of multiple reports of the killing of Juan Perez

Name	Date	Place	Violation	Source
Juan A. Perez	May 83	County Y	Killing	Son's testimony
~~Juan Perez~~	~~May 83~~	~~County Y~~	~~Killing~~	~~Priest's testimony~~
~~Juancito Perez~~	~~May 83~~	~~County Y~~	~~Killing~~	~~Widow's testimony~~
~~Juan Perez~~	~~May 83~~	~~County Y~~	~~Killing~~	~~Forensic evidence~~
~~Juan Perez~~	~~May 83~~	~~Unknown~~	~~Killing~~	~~Newspaper 1 (story)~~
~~Juan Perez~~	~~June 83~~	~~County Y~~	~~Killing~~	~~Newspaper 2 (story)~~

Note that although six records were created in this database, five of them have been deleted (displayed by the crossed-out lines). These records are effectively lost, and are not available for any organizational use.

This strategy has several drawbacks. First, the audit trail from analysis to Mr. Perez and back to the source information will be broken. If a statistical finding that included this killing were challenged (for example, by attorneys for the alleged perpetrators), the database must be able to link the statistic in question with all the source information that provided evidence for the statistic. Suppose that the human rights organization has reported that there were six killings in County Y in May 1983. One of the six reported killings is Mr. Perez, and so the database must now show how the group knows that Mr. Perez was killed by connecting the statistic with all the source material. Mr. Perez's killing was quite widely documented, and the argument that this killing really happened is relatively strong. However, if five of the six sources were deleted, we are now faced with a massive paper search for the original sources, and having to do a paper search indicates that the computerized system has failed.

A second problem is that by deleting five of the six representations of the killing, we lose the ability to look at exactly what was coded from each source. If we want to check the data processing by reviewing the exact data that was coded and entered from Mr. Perez's son's testimony, we may not be able to see the data because it was deleted in the data cleaning. Losing the connections between sources and information they plan to report can seriously affect the effectiveness of the organization.

For example, at the CEH, there is no stable count of how many interviews actually were conducted. Field investigators took information from various interviews and composed "cases" which were passed to the database team – the interviews were therefore merely raw material used by the

field investigators to make cases. But from the point of view of the database, the interviews have now been hidden behind the cases, and so it was impossible to count the interviews or to measure which violations appeared in many interviews compared to violations that appeared in only one interview. This limitation eliminated several additional layers of analysis that might have strengthened the projection of the total number of killings.

The third and most serious problem with deleting multiple points of information about the same violations is that we also destroy the information that certain violations are more frequently reported than others are. Perhaps Mr. Perez's neighbor, Mr. Raimundo, was killed with Mr. Perez, yet appeared only in one of the press clippings. But Mr. Raimundo was not mentioned in any other source. What was different about Mr. Raimundo that led to his being nearly missed by this process? Perhaps Mr. Raimundo was of a different ethnic group than Mr. Perez, and people of Mr. Raimundo's group have less access to the media. If we can identify what kinds of victims are less frequently reported, then we may be able to assume that we have not documented many more victims of this kind. If, when people of Mr. Raimundo's group appear in our database with a clear pattern of less systematic reporting than people of Mr. Perez's group, we may suspect that there other people in Mr. Raimundo's group who are being missed by our investigation. The numbers of such people might be quite large. We might therefore direct investigative resources to Mr. Raimundo's group, or we might use a statistical correction to increase the number of killings projected to have occurred to people of Mr. Raimundo's group relative to Mr. Perez's group.

The right way to handle multiple reports is to create two databases: the first includes all the information faithfully from the sources, and the second encodes the organization's judgements about what is true. Computer hard disks are inexpensive, and most of this work can be done by appropriate software. Keeping the database in two different forms involves no more work than doing it once and then deleting all the multiply reported violations. But instead of deleting the violations that are judged to be the same, the user creates one record in the second database for this violation; and this step can be automated to be a single mouse click for each new record. This new record is linked to all the constituent original records in the first database that in the "delete the extras" method would have been deleted. The resulting form of the records in the source and judgment datasets is shown below in Tables 4a and 4b.

Table 4a. Sample source database of multiple reports of the killing of Juan Perez

Name	Date	Place	Violation	Source	Link to judgement ID
Juan A. Perez	May 83	County Y	Killing	Son's testimony	SV01
Juan Perez	May 83	County Y	Killing	Priest's testimony	SV01
Juancito Perez	May 83	County Y	Killing	Widow's testimony	SV01
Juan Perez	May 83	County Y	Killing	Forensic evidence	SV01
Juan Perez	May 83	Unknown	Killing	Newspaper 1 (story)	SV01
Juan Perez	June 83	County Y	Killing	Newspaper 2 (story)	SV01
Jaime Raimundo	May 83	County Y	Killing	Brother's testimony	SV02

Table 4b. Judgement database linking to source database of multiple reports of the killing of Juan Perez

Name	Date	Place	Violation	Judgement ID
Juan Perez	May 83	County Y	Killing	SV01
Jaime Raimundo	May 83	County Y	Killing	SV02

Note that it takes no more work to link the records (by creating the records in the judgement database and linking them to the source data via the Judgement ID field) than it did to delete them. For statistical analysis, we use the second database to check coding and audit trails. We use the first database to measure *reporting density* (the relative frequency with which certain categories of data are reported). Both structures serve important purposes.

Bias

In the statistical sense we are using here, *bias* does not imply that data have been chosen to support an ideology, or that the data reflect implicit prejudice against ethnic or political groups. In the statistical sense, bias refers to an effect, which deprives a statistical result of accuracy by *systematically* distorting it. This is different from a random error, which may distort on any one occasion but balances out on the average. The random errors effect precision, but not accuracy. There could be many sources of bias, including systematic technical errors or strategic misdirection that led the organization to miss some parts of the reality they purported to study.

Oversimplification is the most common cause of bias introduced by technical errors. For example, the South African Truth and Reconciliation Commission (TRC) decided to represent only one of each kind of violation that happened to each victim. The system, for example, recorded only that each victim suffered one act of torture, one act of severe ill treatment, etc. For killing, this is not a problem, since a person can only be killed once. But victims who are persecuted by their political opponents may be detained and tortured on multiple occasions, or suffer repeated acts of severe ill treatment. In the TRC's representation, the count of the number of violations that could have happened to each victim on multiple occasions (severe ill-treatment, torture) was biased downward relative to the count of killings. That is, the statistics on killings were a better representation of the real patterns and trends in killings than the statistics on non-fatal violations. This bias is hard to detect after the fact, but it is relatively common.

These applications of overlap are discussed in Chapter 11.

Often, when a critic charges that a human rights study is biased, s/he means that the study is too intently focused on violations committed by one perpetrating group. This is taken to imply that the analysis has ignored or undercounted violations committed by some other perpetrating group.[2] For example, in Guatemala some critics claimed that the various large-scale human rights data projects had overstated the proportion of violations for which the state was responsible relative to the proportion for which the guerrillas were responsible. Because there were three independent projects surveying the same human rights situation, it was possible to test the hypothesis that the data were biased in this way. The data in each of the three projects were divided into the cases attributed to the state and those attributed to the guerrilla. The overlap rates among the three projects were measured for the state cases and the guerrilla cases. If overall the projects had focused more on the state cases than on the guerrilla cases, then there should have been a higher overlap rate among cases attributed to the state because the investigations would have covered a higher proportion of the universe of cases. However, there was no significant difference in the overlap rates of state cases and guerrilla cases, which implies that the coverage rate was roughly the same over both perpetrators. In this example, it was possible to say that taken together, the proportions of responsibility attributed to the state and to the guerrillas were not biased relative to the proportion of violations in the universe of all violations.

There is generally no way to argue that data are completely unbiased in every way. The best defense against the charge of bias is to take scientific samples of people who will be interviewed. If this is not feasible, and if the organization has access to different kinds of data from different sources, comparisons can be made between analyses from different sources. If the sources agree, then either they share the same biases or they are all roughly unbiased. If the sources disagree, additional research would be required to explain how one or more of the sources might be biased.

[2] A related form of this bias results when a critic challenges the objectivity of an organization's work arguing that "violations were committed on both sides" when in truth nearly all violations were committed by one side. Such claims are based on the attribution of moral equivalence, and are often made by diplomats, the press, commissions of inquiry, and other quasi-official processes professing objectivity.

Conclusion

The sum total of our experts' experiences are that if an organization effectively uses a well-designed and properly supported information management system, the organization will find that the credibility of their report's conclusions is high enough that critics will prefer not to challenge the scientific conclusions. This was the case for the final report of the CEH.

Clearly, the information management system is the critical element in achieving the ultimate goal of a truth telling organization: To produce accounts of crimes against humanity that cannot be denied.

Chapter 1

The Salvadoran Human Rights Commission: Data Processing, Data Representation, and Generating Analytical Reports

Patrick Ball

Introduction

In this paper, I describe the work I did as while working for the Salvadoran Human Rights Commission (*Comisión de Derechos Humanos de El Salvador,* CDHES).[1] Between 1979 and 1991, the CDHES took more than 9,000 interviews that were recorded in written form as testimonies. We planned to begin work in May 1992, in conjunction with other organizations which, like the CDHES, were part of the Coalition of Non-governmental Human Rights Groups *(Coordinadora de Organismos de Derechos Humanos)*. The organizations included among others, Legal Aid *(Socorro Judico)*, the Human Rights Institute of the University of Central America, and the Human Rights Office of the Lutheran Church. For a variety of reasons, among which were political issues and the perceived lack of adequate data, all except the CDHES withdrew from the group in June 1992.

This was one of the earliest large-scale human rights information systems projects. By and large, the other projects discussed in this handbook, which came later, had fewer of the problems experienced in this project. However, this project is important to gaining an understanding of the issues involved in planning and implementing large-scale data projects for human rights violations.

Even today, there are many organizations which do not have database and analytical expertise and which may be working through similar problems. They may find the discussion in this paper helpful in their current work.

The goal of this project was to target individual perpetrator responsibility. Only a modest fraction – about 125 – of the total of 9,000 testimonies were entered into the full data base and used to provide reports targeting individual perpetrator accountability. Note that the fully processed testimonies were thoroughly documented and were the most important cases identifying individual accountability. These cases were presented in their entirety to the truth commission by the CDHES.

Because it proved impossible to follow the planned process and enter very many cases into the full system, we developed a parallel process into which we entered the entire set of CDHES testimonies. This second process formed the basis for the analysis that gave this project its impact.

Data Processing, Part 1

I was not involved with the data collection. As mentioned above, the CDHES had collected over 9,000 testimonies and it was our task to carry the project forward from data processing (coding) through data representation, and to end with the generation of analytical reports. I initially read several hundred testimonies in preparation for this phase of the project. At that time, it appeared that it would be possible to code the cases quickly enough and with a reasonable level of data processing effort.

Database and Reporting, Part 1

I wrote a database program to meet the project needs. By July 1992, the completed FoxPro database[2] and user interface was operational. There were two major problems: (1) the output capabilities of FoxPro were quite limited, and (2) data entry was slow.

I solved the first problem of providing output by brute force, manually creating the output routines and managing all the fields with variable quantities of information. We printed to a Hewlett-

[1] This was before I worked at AAAS. In fact, I first met then-AAAS Senior Program Associate Dan Salcedo when he visited the CDHES in September 1992.

[2] The database was written in the fourth normal form, which enabled a number of powerful search methods. See (Ball et al. 1994). Note that several earlier databases were implemented in El Salvador

Packard laser printer which required us to write inline escape sequences to define font selection, bold, italics, etc.

The second problem, slowness of entry, was not so easily resolved and was linked to the data processing. When data processors were set to the task of coding the testimonies it was apparent that we had greatly underestimated the time needed for data processors to extract the relevant data elements from the testimonies (victim identifications, perpetrator identifications, violation types, locations, etc.) and subsequently to enter the data.

A typical output from the automated report process is shown in Appendix 1.[3] This case, number 85 from the set of 110 cases presented to the Truth Commission for El Salvador in October 1992, is identified by the date of the complaint and the date of the event. The complexity of preparing this report is concealed by the apparent simplicity of its presentation. Although it appears to be a document that a user could type while reviewing the data manually it is, in fact, structured output generated by a database. Since each case has a different number of victims, violations, etc., a complex process is needed to generate this report. Among the tasks that a database can do for an organization, this kind of reporting can be very helpful to synthesize repetitive, detailed information in easy-to-digest reports. The final presentation to the Truth Commission included about 600 pages of text generated in this manner.

In case 85, shown in Appendix 1, the three victims are named in the "VICTIMS" section. Note that there might be any number of victims, from one to several hundred. In the next block ("AGE," "SEX," and "OCCUPATION", personal data about each victim is reported.

The "TYPES OF VIOLATION" section lists all violations that were reported as being committed against each victim. Each victim could have suffered one or several violations, and different victims might suffer different combinations of violations. The violation type is listed, followed by the identification of the perpetrator(s) alleged to have committed it. Torture was listed separately by type of torture and notes about each torture act were reported.

The database provided links to the officials alleged to have had command responsibility for the units that committed the violations. These individuals are listed in the "PERPETRATOR" section. Note that the number of perpetrators can vary according to the number of units alleged to have participated in the event.

Lawyers who worked on the case drafted a narrative describing each event. Their legal work is presented in the "LEGAL ACTIONS TAKEN" and "AVAILABLE DOCUMENTATION" sections. Those witnesses willing to be identified appear in the final section, "WITNESSES." The objective of the CDHES for this presentation was to show the Truth Commission that the Salvadoran judiciary had taken essentially no action despite nearly 15 years of continuous legal activities on the part of the human rights NGO community.

Coincident with this work, we had entered the entire command structure of the Salvadoran military and security forces into a database structure like that defined in Ball et al. (1994).

Data Processing and Database, Part 2

In late August 1992, the CDHES leadership informed me that they felt that the data processing and associated database entry was moving too slowly. This was a reasonable criticism. The data were being entered into my complex FoxPro database too slowly to get the work quickly enough to have a significant impact in assigning perpetrator responsibility. The resolution of this problem by the CDHES was to code cases and enter them into Word Perfect 5.1 tables. This gave them summary sheets that they could use for manual review but left no possibility for relating data elements or performing analyses using the capabilities of the computer. Appendix 2 shows a typical page resulting from this process. It is clear from this page that they were entering data into a Word Perfect table.

In early September 1992, CDHES staff members were reading the Word Perfect documents one line at a time, looking up the commander of the perpetrating unit in the military career structure database according to the unit alleged to have committed the violation and the date of the violation.

[3] This presentation format has been used subsequently by other NGOs. In July 1997, the International Center for Human Rights Research (CIIDH) presented to the Commission for Historical Clarification (CEH) in Guatemala about 140 of their 17,000 cases, along with lists of the people the CIIDH had registered as killed or disappeared. The volume containing this information was more than 700 pages long.

The military career structure database showed which officers held which jobs in this unit at the time the violation was committed. Then they typed the commander's name into an eighth column of the table showing his command responsibility for this violation.

CDHES had tried to save time by avoiding entry into my FoxPro database and putting the data into the Word Perfect table. Now, they were paying the price for that decision and investing a large amount of time because they could not use a database program to perform this next phase of the process. The magnitude of the problem was roughly this: They were able to enter about 15 cases per hour. For the 7,000 cases we had identified as within the mandate of the various commissions who wanted the results, this amounted to about 470 hours; the estimated total effort amounted to almost ten person weeks with six 10 hour days.

At this rate, we would not complete the analysis in time to present the results to the *Comisión Ad-Hoc*. I realized that the Word Perfect tables could be parsed and wrote a program that read an ASCII-versions of the Word Perfect document. The program then broke the data down into fields and tables. This was not a simple process because there could be any number of values in each cell of the table, and the victim values had to be matched to changing date, violation, and perpetrator values by counting lines within each cell. This parsing program created as its output a database whose structure included three related tables (case, victim, and perpetrator). The victim table included a field for each of the 15 violation types we coded, and the value in each field indicated whether or not the victim suffered that violation. Table 1 shows the 15 violation types and their codes.

Table 1. Violation types and codes.

Arbitrary execution	EA
Forcible disappearance	DF
Torture	Tt
Massacre	Mc
Illegal detention	DI
Sexual violation	VS
Threatening	Az
Persecution	Ps
Allanamiento	Am
Destruction or theft of property	Db
Displacement of population	Dp
Disappeared	Dd
Stabbing or wounding	Hd
Robbery	Ro
Other violations	Ot

With this structure each victim can suffer each violation *type* only once in the context of each "incident," or time by place combination; with repeated incidents within a case, other violations against the same victim could be repeated. Note that this does not mean that each victim suffered only one violation in each case. Rather, for example, the victim could only be recorded as having suffered detention and torture in a given incident, rather than detention, torture, torture, and torture if there were three torture types employed.

This limitation is not realistic and may distort the data. [4] However, it is much less severe than other distortions due to simplification, such as one victim-one violation, as discussed in Ball (1996). Quick checks of the testimonies showed that it was rare for witnesses to report more than a single instance of the same violation against the same victim (e.g., multiple illegal detention incidents). Appendix 3 shows the summary statistics drawn down from this database.

Now we faced the need for standardization of the non-standard spellings and other references to perpetrators. To resolve this problem, I made a list of all the non-standard perpetrator names from the original data and matched all names (by a combination of computer and manual methods) to a set of standard codes. I created tables that translated between all the possible non-standard spellings of the perpetrator names (e.g., "National Guard," "NG," "Guard," "Nat.Gua." and so forth) to a desired standard code (e.g., in this case, "NG"). With standard perpetrator codes applied to each violation, I could use the dates (which were also non-standard and had to be extensively edited) and the codes to match to the perpetrators' career histories.

[4] Many other systems suffer from this oversimplification, notably that of the South African Truth and Reconciliation Commission (TRC), although the TRC data processors used narrative data recorded by the interviewers to recover from the error. See Chapter 4 for details

The results of the parsing were 7,150 cases, including 9,346 corporate perpetrators involved in 11,940 incidents. More than 17,000 victims who suffered 29,000 violations were documented by these data.

Appendix 4 shows the results of the matching, titled "Responsible Military Individual."[5] It is unfortunate that that the full set of testimonies was not fully captured in the format shown in Appendix 1 as a result of resource limitations. However, the political impact of the Indices of Individual Accountability and the more limited system was great. The overall lesson is that if the analytic and political objectives are clear, the systems designer should build a system that is just adequate for those objectives. More complexity can cause many problems while not adding much value from the additional capability.

Lessons Learned

Entity, Function	Lesson	Recommendation	Issues
Conversion of non-standard input to standard codes	Editing is never done; users are always working on data. If you change the original data, and users come up with a new version, all of the changes must be redone from start.	Use a two-way table to translate changes from the original data to a cleaned output. Do not make changes to original data. Learn how a) to parse raw text files into structured data, and b) to standardize uncontrolled entry into controlled structures	Table must be set up and used at the initiation of work on data, although it will be modified constantly throughout the project. Establishing the rule that all changes to source data come from users and automated processing must be robust enough to deal with uncontrolled entries.
Achievement of goals.	Scientifically optimal outcome may not be feasible with time and resources available.	Information system personnel must plan and re-plan as necessary to fit results to resources. In designing and implementing the database use the smallest possible components to accomplish the organization's core goals.	Effective communication to mission leadership by information system personnel of resource-based limits on achievements. Good working relationships between parties. Ability and willingness of personnel to plan and design to meet constraints.
Functionality	Complex methods and procedures difficult to execute.	Simpler is better; less is often more.	Self-discipline, planning.
Replacement of manual methods by automated methods.	Manual procedures can fail late in project.	Be prepared for late-term rush projects to automate manual procedures.	Flexibility in response. High level of skill required of system and program designers.

[5] The strategic aspects of this project are described in more detail in (Ball, 1996).

Appendix 1

Case Description Document

Date of complaint: 13/02/89 **Date of Event:** 12/02/89 **Document no.:** 85

VICTIMS: 1. COLINDRES PANAMEÑO, Manuel Antonio
2. GARCIA, Hernán
3. PINEDA ALVAREZ, Eduardo

AGE:	1. 18 years	**SEX:**	1. Male	**OCCUPATION:**	1. Laborer
	2. 38 years		2. Male		2. Unknown
	3. 40 years		3. Male		3. Farmer

TYPES OF VIOLATIONS:

1. Illegal detention (DMIFA)
1. Illegal detention (DefCiv)
(Units not authorized to detain persons.)
1. Illegal detention (PMun)
Detained in the jail of *Santiago Nonualco y Zacatecoluca* without cause.
1. Torture (DMIFA)
1. Torture (DefCiv)
2. Illegal detention (DMIFA)
2. Illegal detention (DefCiv)
(Units not authorized to detain persons.)
2. Illegal detention (PMun)
Detained in the jails of Santiago Nonualco and Zacatecoluca without cause.
2. Torture (DMIFA)
2. Torture (DefCiv)
3. Illegal detention (DMIFA)
3. Illegal detention (DefCiv)
(Units not authorized to detain persons.)
3. Illegal detention (PMun)
Detained in the jail of Santiago Nonualco y Zacatecoluca without cause.
3. Torture (DMIFA)
3. Torture (DefCiv)

TORTURES:

1. Hung by the testicles (DefCiv)
In the commander's office of Santiago Nonualco, by members of the Civil Defense and of the DMIFA.
1. Pretend to kill by asphyxiation (DefCiv)
1. Stripped of clothes (DefCiv)
In the commander's office of Santiago Nonualco, by soldiers of the Civil Defense and of the DMIFA.
... [descriptions elided to save space]
2. The "airplane" (DMIFA)
Hung by the hands (behind the back), and beaten with a stick.
2. Tied up, with shackles on hands and/or feet. (DefCiv)
2. Tied up, with shackles on hands and/or feet. (DMIFA)
...
3. Beaten over whole body. (DMIFA)
...

CURRENT LEGAL STATUS:

1. Freed with no charges.
2. Freed with no charges.
3. Freed with no charges.

PERPETRATOR

INDIVIDUAL

Col. CANJURA ALVAYERO, Benjamín Eladio (Commander, DMIFA)
Col. GOMEZ, José Humberto (Director, National Guard)

Col. HERNANDEZ CASTRO, José Dionisio (Assistant Director, National Guard)

...

ORGANIZATIONAL

DMIFA
National Guard
Municipal Police
Civil Defense

LOCATION OF THE EVENT:

Near the Chincuco bridge, Santiago Nonualco, Department de La Paz

NARRATIVE OF THE EVENT:

The three victims were captured by members of the Civil Defense working in conjunction with the DMIFA, who accused Eduardo of being a collaborator of the FMLN and Hernán and Manuel of being guerrillas. They took them to the local headquarters at Santiago Nonualco, where they were each brutally tortured for several hours. At 5 PM of the 13th of February, they were transferred to DMIFA jail at Zacatecoluca, where the torture was continued.

...

The commander of the DMIFA, in an official statement to the CDHES, confirmed that the capture of Manuel was carried out by personnel under his command. This case was presented to the CIDH[6] along with dossier 15/89C CIDH, which deal with the murders of the parents and a brother of Manuel Antonio Colindres Panameño, only eleven days after he was freed.

LEGAL ACTIONS TAKEN:

Writ of *habeas corpus* 15/02/89

Before whom: The Supreme Court
For whom: PANAMEÑO DE COLINDRES, María Luisa
In reference to: 1

...

AVAILABLE DOCUMENTATION:

Photographs

Of victims 1 and 3.

Letters

To DMIFA 15/02/89
From DMIFA 20/02/89

Other

Videotape with the testimonies of victims 1 and 3. Demonstration of the stigmata of torture (25 minutes).

Copy of *habeas corpus* 15/02/89
Statement of COLINDRES PANAMEÑO, Manuel Antonio 20/02/89

WITNESSES:

COLINDRES PANAMEÑO, Manuel Antonio
COLINDRES VASQUEZ, Andrés
GARCIA, Hernán
PANAMEÑO DE COLINDRES, María Luisa
PINEDA ALVAREZ, Eduardo

SOURCE:

Human Rights Commission of El Salvador (CDHES), NGO

[6] CIDH is the acronym for the Interamerican Commission for Human Rights *(Comision Interamericana para Derechos Humanos)* of the Organization of American States.

Appendix 2

Typical Manual Case Entries

Human Rights Commission of El Salvador (CDHES)

1988 Cases

Col. 1[1]	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6	Col. 7
Document no.	**Victim Name**	**Event Date**	**Violation Type**	**Place**	**Perpetrating Unit**	**Comments**
01/88	PEÑA PALACIOS, Raúl Guillermo	25-02-88	6,9	San Marcos, San Salvador	Death Squad	
03/88	CHAVEZ DIMAS, Salvador PANIAGUA, Francisco LOPEZ CARTEGENA, Antonio CHAVEZ HERNANDEZ, Luis AGUILAR ANZORA, José Alex CHESONI, José Luis MARTINEZ, Lucio	22-12-88 13-01-88	6 10	Colón, La Libertad	Death Squad	Sr. Chávez Dimas assured that he was in the jail of the Cavalry Regiment because at the time of the sounds of the shots there were military present from that military organization.

And so forth ….

[1] Column numbers did not appear in original documents. They are given here for reference.

Appendix 3

Human Rights Commission of El Salvador (CDHES)
Summary of Presented Documents, by Type of Violation and Year of Event[7]

Year	EA	DF	Tt	Mc	DI	VS	Az	Ps	Am	Db	Dp	Dd	Hd	Ro	Ot	Vt	Pb	In	Cs
1973	3	1	5	1	5	0	3	3	2	0	0	0	0	0	2	5	0	1	1
1974	1	0	0	0	1	0	1	1	0	0	0	0	0	0	1	1	0	1	1
1975	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	2	0	2	1
1977	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	2	0	2	2
1978	0	0	2	0	4	0	0	0	0	0	0	0	0	0	1	4	0	2	2
1979	20	13	13	1	21	0	2	1	1	1	0	1	1	1	13	29	0	14	13
1980	496	262	238	12	494	34	44	33	72	22	7	36	295	24	95	1237	15	388	370
1981	1610	327	328	18	692	23	87	21	173	77	5	50	10	58	81	2221	10	481	464
1982	419	471	297	9	1000	31	54	13	260	56	10	105	16	30	177	1488	18	722	681
1983	234	172	113	6	467	7	16	6	46	26	23	82	10	25	26	626	1	353	346
1984	96	154	188	15	566	9	31	10	76	14	4	115	13	9	80	835	2	557	541
1985	60	90	159	1	863	5	63	15	98	28	7	44	32	36	86	1012	13	668	650
1986	97	45	188	2	514	3	87	86	131	71	56	38	64	32	20	724	15	367	349
1987	73	55	204	3	410	12	165	63	90	41	12	15	43	20	96	558	10	293	260
1988	91	68	351	3	834	9	273	137	123	53	8	42	63	66	114	1203	44	611	500
1989	115	119	1003	3	1753	19	539	147	330	134	40	39	79	132	233	2209	45	1012	924
1990	86	90	378	2	770	15	320	122	178	55	15	40	66	55	103	1180	35	678	622
1991	46	24	340	0	959	8	571	135	148	159	25	36	98	105	257	1446	87	693	597
1992	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1
Unknown	11	4	16	0	27	0	19	13	5	2	5	0	4	5	9	55	4	40	34
Total	3460	1896	3825	76	9383	175	2275	806	1733	739	217	644	794	598	1394	14838	299	6886	6359

[7] For meaning of violation type codes, see Table 1. The other codes are as follows: Vt, total number of victims; Pb, collective victims; In, total number of events; Cs, total number of cases. No data are given for 1976 since none was available.

Appendix 4

Human Rights Commission of El Salvador (CDHES)
Individuals with Alleged Command Responsibility, Typical entries

ACEVEDO, Mario Enrique

1982 1a. Infantry brigade, Executive

Document numbers:
487/82

EA	DF	Tt	Mc	Dl	VS	Az	Ps	Am	Db	Dp	Dd	Hd	Ro	Ot	Cs	Vt	Pb
2	0	2	0	2	0	0	0	0	0	0	0	0	0	0	1	2	0

==========

PINEDA VILLALTA, Humberto

1981 National Police, Personnel Chief
Document numbers:
100/81, 103/81, 105/81, 1079/81, ..., 76/81, 90/81, 97/81

EA	DF	Tt	Mc	Dl	VS	Az	Ps	Am	Db	Dp	Dd	Hd	Ro	Ot	Cs	Vt	Pb
230	24	35	1	73	0	2	0	35	0	0	0	0	0	28	39	321	1

1978-1981 Navy, Commander
Document numbers:
325/80, 417/80, 264.1/1984, 271.a/1983, 67/85, 82/85

EA	DF	Tt	Mc	Dl	VS	Az	Ps	Am	Db	Dp	Dd	Hd	Ro	Ot	Cs	Vt	Pb
0	2	1	0	8	0	1	0	2	0	0	0	0	0	1	6	8	0

1980-1984 Navy, Unknown
Document numbers:
117/82, 325/80, 417/80, 535/82

EA	DF	Tt	Mc	Dl	VS	Az	Ps	Am	Db	Dp	Dd	Hd	Ro	Ot	Cs	Vt	Pb
0	3	0	0	6	0	1	0	2	0	0	0	0	0	0	4	6	0
total:	230	29	36	1	87	0	4	0	39	0	0	0	0	0	29	49	335	1

References

Ball, Patrick. 1996. *Who Did What to Whom? Planning and implementing a Large Scale Human Rights Data Project.* Washington, DC: American Association for the Advancement of Science.

Ball, Patrick, Ricardo Cifuentes, Judith Dueck, Romilly Gregory, Daniel Salcedo, and Carlos Saldarriaga. 1994. *A Definition of Database Design Standards for Human Rights Agencies.* Washington, DC: American Association for the Advancement of Science and Human Rights Information and Documentation Systems International.

Chapter 2

The Haitian National Commission for Truth and Justice: Collecting Information, Data Processing, Database Representation, and Generating Analytical Reports

Patrick Ball and Herbert F. Spirer

Introduction

In early May 1995, Francoise Bouchard, President of the Haitian National Commission for Truth and Justice (*La commission nationale pour la vérité et justice,* CNVJ), invited the American Association for the Advancement of Science (AAAS) to advise the commission on how to develop a large-scale interviewing project to take the testimonies of a several thousand witnesses of human rights violations. The AAAS team[1] met with the CNVJ commissioners and planned a project to include 40 interviewer teams, ten data processors, and five data entry specialists. The interviewing was to be done in July and August 1995, and the report produced by mid-December 1995. The commission's final report was given to President Aristide in February 1996, but because of policy disagreements in the Haitian government, it was not published until September 1996, and then in a printing of only 75 copies. A second edition was published in February 1997.[2]

The CNVJ team took 5,453 interviews. In all, they identified 8,667 victims who suffered 18,629 violations. The CNVJ interviewing was quite good by scientific standards. A data processing group composed of eleven of the interviewers applied standard definitions to the raw interview data and produced detailed regional analyses, incorporating qualitative material from the interviews, as well as historical, economic and demographic analysis. Unfortunately, in the last stages of the process, the commissioners discarded almost all the work the field investigators did and substituted a chronology of the *de facto* regime. The commissioners never informed the AAAS of their reasons for not using the regional data; although the statistical analyses were presented, the tables omitted most of the content and the translations into French were inadequate. Thus, observers should not judge the quality of the field research by the AAAS team for the CNVJ on the basis of the published official report.

A statement of the prevailing atmosphere appeared in *Le Monde Diplomatique*[3]:

> To the great disappointment of all, this report [the final CNVJ report submitted to President Aristide on 4 February, 1996], was for some strange reason was hidden in the files of the minister of justice, M. P- Pierre-Max Antoine, for many months. After many protests, only small parts of this report were published. The public and the many victims still wait for its publication in Creole. The majority of the final recommendations were never enacted. Former perpetrators occupied positions in the new national police or as prison guards: one of them was even in the security guard of the national palace even though his name appeared in Appen-

[1] Drs. Patrick Ball and Daniel Salcedo comprised the AAAS team.

[2] A version of the report that does not include the appendices describing the work of the AAAS team is available at www.haiti.org/truth/table.htm

[3] Roussiere, D. and Danroc, G., "Soif de justice en Haïti," *Le Monde Diplomatique,* May 1998, pp. 22-3. The original text follows:

> Ce rapport, à la grande déception de tous, est étrangement resté, durant de très longs mois, caché dans les tiroirs du ministre de la justice, M. Pierre-Max Antoine. Après de nombreuses protestations, celui-ci ne l'a publié qu'au compte-gouttes. La population et les nombreuses victimes attendent toujours sa diffusion en créole. La majorité des recommandations finales n'ont pas été mises en oeuvre. D'anciens bourreaux ont occupé des fonctions dans la nouvelle police nationale ou encore comme gardiens de prison : l'un d'eux était même dans le corps de sécurité du palais national alors que, pourtant, son nom figurait dans l'annexe 4 du rapport final CNVJ (page 1-b code P 0402). Pourtant une Commission vérité, sans compétence pénale, ne peut être efficace qu'en informant largement la société civile et en transformant véritablement le système judiciaire ainsi que le fonctionnement des administrations. Rien n'y a fait, la paralysie, l'inertie et le laxisme demeurent.

> dix 4 of the final CNVJ report (Page 1-b, code P 0402). However, a truth commission without the ability to punish can only be effective if it can fully inform society at large, and transform the judicial system and the functioning of administrative bodies. Nothing has been done; paralysis, inertia, and inactivity reign.

The failure to publicize the report was the responsibility of the administration of President René Préval, not of the commissioners.

Interviews, Data Processing, and the Database

Interviews

The sequence of data collection in the CNVJ interviews was as follows. A person, the *dénunciateur*[4] comes to the interview team to give information about an abuse. The violent events being reported may have happened to the *dénunciateur*, and no one else. Or, the *dénunciateur* may be reporting abuses that happened to other people. Thus, the *dénunciateur* may or may not be a victim. Furthermore, there may be other victims. Thus a single interview may yield information about one, two, three, or any number of victims.

Each victim may have suffered one or many violations. The violations may have happened at one or many points in time. That is, a victim may have been detained and tortured on one date in one place, but raped and murdered on a subsequent date in a different place.

Furthermore, one or many identifiable perpetrators may have committed each violation. That is, "Antoine" and "Pierre" were responsible for the hypothetical detention and torture in the previous paragraph, but "Pierre" and "Michel" committed the rape and murder. Complex relationships among these various entities existed, and were captured in the interviews.

The teams conducted more than 7,000 interviews. Interview teams consisted of one Haitian and one international team member. This pairing was largely to satisfy an explicit mandate of the CNVJ that interview teams comprise both Haitians and internationals. Few internationals – even Francophones – speak Haitian Créole so that it was an absolute necessity that each interview team includes a Haitian.

Data Processing

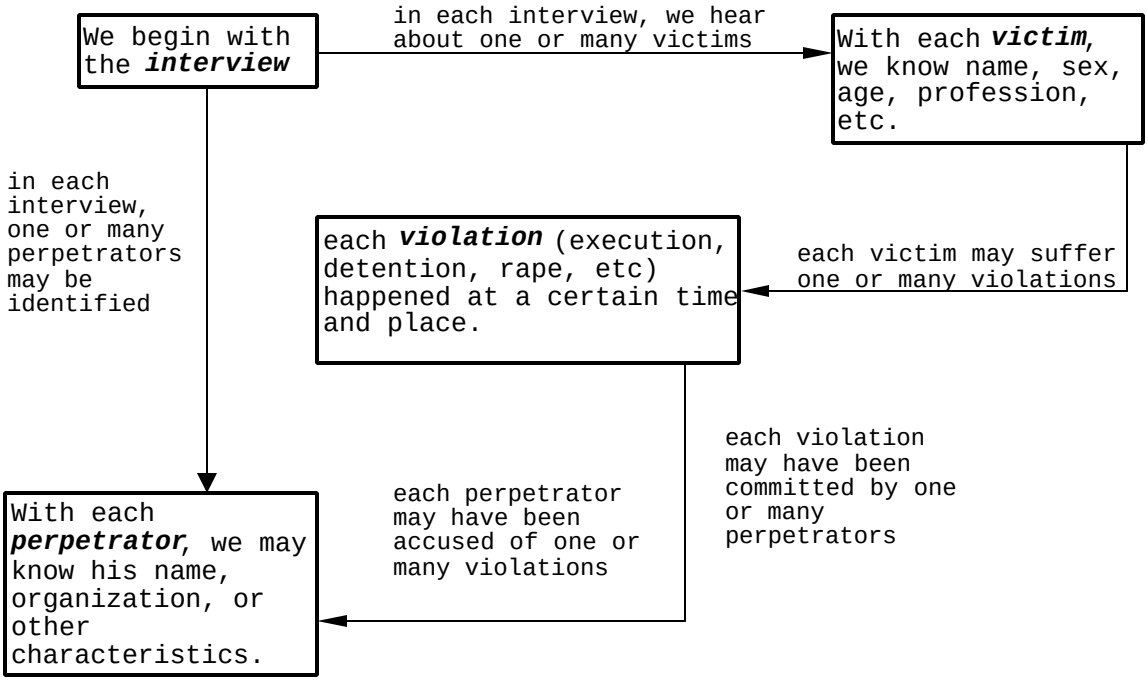
A serious handicap to the work of the CNVJ was the late start of data processing, which did not begin until the teams had completed interviewing at the end of August 1995. This was in large part due to a significant leadership vacuum that took some months to fill. This lapse occurred because of serious political differences and the lack of staff experience with research of this kind. As a consequence, the directors of the interviewing teams resigned in late August 1995.

Analysts were chosen from among the interview teams, and they applied the codes following the methods that most projects use. Again, like most projects, the definitions of key concepts changed and new ideas were added to the analysis after most of the interviews had been coded. The analysts re-coded the entire set of interviews at least three times, although during the second and third reads on each interview they were re-coding only for specific themes.

All data entry was done by means of FoxPro "browse windows," with field and record level validation. For example, all the codes (for types of violation, geography, victim or perpetrator references) were checked as the users typed them to at least assure that the codes were valid. The six workstations were all freestanding. There was no sharing or serving of files across the network (although the machines did share a printer).

[4] The standard terminology for a person complaining about a human rights violation in the Central American and Caribbean regions is *denunciador* (Spanish) and *dénunciateur* (French and Créole). The word is more closely related to the English "complain or "report" than to "denounce."

All entered data were aggregated into a single database on a central machine. We achieved this result without duplication or loss of entered data by assigning a unique block of key values to



each workstation. Before aggregation, the relational integrity of the data was checked by tracing each foreign key to the primary table from which it originated.[5] This was necessary because the database software used (FoxPro for Macintosh) did not do internal integrity checking. The data were merged into the common database on which the analysis was run. Our major programming task was to carry out the preceding functions essential to creating a common database with assured data integrity.

As part of the aggregation process we standardized several codes that were not originally controlled in the data entry, such as political affiliation of the victims. We achieved this result by creating a unique list of all the phrases and abbreviations found in the free text field. An analyst reviewed these selected phrases and abbreviations and assigned a code to each of the text fields. We then merged the code and text combination back into the original data, thereby assigning a code to each of the previously uncontrolled fields.

But even with the recoding, many analytic categories remained poorly reported. For example, many deponents were reluctant to report their political affiliation, and so the only analysis that could be done on this field used only a small fraction of the data and was therefore unstable.

Database

We designed the CNVJ's database to the standards established in (Ball, et. al., 1994). In accordance with those standards, we followed two fundamental rules:

[5] When a primary key from a table is incorporated into another table to form a relationship between the tables, it is called a “foreign” key.

1. The database must not introduce additional ambiguity into the data. That is, to the extent that the original sources permit, the database must be absolutely precise regarding who committed which violations against whom.

2. To represent a wide range of abuses, interventions, people, organizations, and to unambiguously represent the complex relations among these entities the database must be as extensive and amenable to change as is consistent with available resources.

As we have mentioned, any of the entities in a human rights violation may have complex relationships with none, some, or all the other entities; it is important that the data model enables an appropriate representation in the database. The diagram below shows in schematic form the data representation model that we used for the CNVJ.

The database designed to represent the processed data in accordance with this model permitted the CNVJ database to represent the complex stories the *dénunciateur* reported to the CNVJ interview teams.

Analyses

The teams conducted 5,453 interviews, in which we heard about 18,629 violations that were committed against 8,667 victims. During the course of the project, we carried out many statistical analyses using the relational structure described above as a basis for analyzing counts of violations by type, time, and geographic location.

A discussion of some of the statistical analyses follows. Analysis at the CNVJ met with a set of problems similar to those dealt with at the TRC and CEH.[6] Continuous efforts to establish and maintain data quality at all stages kept the database itself in a state of change until hours before the results had to be reproduced for distribution. The challenge for the system designers – and to some extent, those that implemented their design – was to define the entire analytic process in ways that supported dynamic updating.

As we have discussed at greater length elsewhere, our experience has convinced us that every human rights project is unique and has different attributes from the others.[7] Some analyses are general and likely to be common to most human rights projects (such as victim, perpetrator, witness attributes, and violations by time and place), but many of the analytical issues are particular to a given project.

Table 1. CNVJ Human Rights Violations Categories and Types

Category (Right)	Violation types
Life	execution disappearance
Liberty and Integrity	torture detention rape
Property	theft attacks on goods attacks on property

In the following sections, we first describe the nature of many of the analyses performed that are of the general category, and then discuss several of the analyses that were particular to this project.

Victims and Violations

We analyzed victims by age category (infant, child, adult, elderly), profession, sex, and affiliation.[8] As we mentioned earlier, the affiliation data are not reliable because they were not recorded consistently. Each of these analyses is repeated in a monthly time series for each of the CNVJ's categories (Life, Integrity and Liberty, and Property). Table 1 shows these categories and the types of violations within the categories.

A subset of analyses reported on the violation of rape. These analyses included a histogram of victims' ages, affiliation, the number of total female victims by department and month and corre-

[6] At this point, we recommend that the concerned reader read or reread relevant sections of Chapters 3, 4, and
[7] See the introductory chapter in this volume.
[8] In French, denoted *appartenance*.

sponding proportion raped, and the proportion of rape victims assaulted in a named place such as a barracks or a military post.[9] (Note that the affiliation data are weak for the reasons already stated.)

To understand the nature of the problems with affiliation and to extract such limited information as we could from the incomplete and inconsistent data, we determined the proportion of victims with some affiliation. For this purpose, we coded a victim as having affiliation if there was any text in the relevant descriptive field. In most of the analyses for which there is sufficient data for reliable estimates, the percent of victims with an affiliation was from 50-70%. Other analyses include victims driven into internal refugee status, [10] finding their proportions by department, age and sex of the victim.

Perpetrators

We analyzed perpetrators with monthly and annual time series by CNVJ human rights violation category (Life, Integrity and Liberty, and Property) and by the affiliation of the perpetrator. Table 2 shows the five categories of perpetrator that we tracked (FadH, FRAPH, *Police rural, Attaché,* and Other). Violations without perpetrators identified by category were not included even if there were perpetrators identified for other violations against the same victim at the same time and place. We also did these analyses by department.

Table 2. CNVJ Human Rights Violation Perpetrator Categories and Types

Acronym or name	Meaning	Comments
FadH	Armed Forces of Haiti	The Haitian army
FRAPH	Front for the Advancement and Progress of Haiti	Paramilitary enforcers for the *de facto* regime. The acronym is a pun on "blow" or "beating"
Police rural	Militia	
Attaché	in urban areas, a semi-legal deputy to the police in rural areas; to the militia	
Other	Other	

There are special considerations in analyzing data about victims and perpetrators in combination. One or many violations could have happened to each victim. Thus, sums of the numbers of violations are usually significantly different from sums of numbers of victims. This disparity is logical, since in a given interview, a violation may have been committed on the same victim more than once at several dates. Similarly, none, one, two, or many identified perpetrators may have committed each violation. Consequently, no count of perpetrators from one or more given organizations can be summed with counts for other perpetrating organizations unless the perpetrators are combined in categories, as we describe below.

Our final analysis in this section looked at combinations of perpetrators. Since any violation may have been committed by one or more perpetrators who were not identified, or one, two, or many identified perpetrators, we had to combine categories of perpetrators to analyze how actual violations were committed. For example, it is clear that the Haitian army alone committed the bulk of violations in which a perpetrator is identified. Also, it is much easier to identify "two soldiers" than to identify random civilians. Substantial numbers of violations were committed by the *Attachés* working in conjunction with the FRAPH, by the militia working with the Haitian army, and by the militia. However, the single largest category is "no identified perpetrator."

[9] In French, *caserne* and *avant-poste*, respectively;

[10] For convenience, we refer to these internal refugees by the French term, *marronage*, which literally means "runaway."

National and Departmental Trends

To determine national trends, we analyzed the number of CNVJ human rights violation categories by month and by department. We also analyzed the individual and joint occurrence of detention and extortion by month and by department, and the relative proportion of all the possible combinations of torture, detention, and extortion. We found that extortion happened much more frequently in the presence of detention or the combination of detention and torture than it does alone or in conjunction with torture.

However, our principal finding was that detention, detention and torture in combination and torture are the most frequent combinations of abuse. The number of violations of these three types was much greater than the number of killings.

We also analyzed the counts of the CNVJ human rights violation categories and summarized the counts of different violation types by commune for each department. In this analysis and in the subsequent analyses in which departmental breakdowns were used, Port-au-Prince was separated from the Ouest and Raboteau is separated from l'Artibonite.

Validity of the Survey with Respect to Time

Findings concerning the distribution of violations in time were the next section of the report. For example, we noted that for the national and departmental breakdowns, October and November 1991, and October 1993, were exceptionally violent periods. They showed significant, large and increased numbers of all categories of human rights violations. To be sure of the validity of these findings, we wanted to assure that the non-probabilistic sampling and the nature of the interviewing process had not led to any non-representative selection of interviewees. One way to confirm or disconfirm this hypothesis was to compare the CNVJ team's time distribution of violations in time with other human rights violation data for the period September 1991 to October 1994.

Fortunately we were given access to a dataset based on cadaver data collected at the University Hospital by Mercedes Dorretti of the Argentine Anthropological Forensic Team (EAAF; see Dorretti and Cano, n.d.).[11] From this EAAF dataset, we determined the number of cadavers found each month, and plotted these counts against the monthly counts of summary executions from our survey data.

The two data series measured different variables and covered different geographical regions. Thus it is extremely unlikely that they will be alike in their values or time series unless they have a common underlying source. If the process of violations generated both series, we expect them to increase and decrease at the same time. When the level of human rights violations increases, there should be more cadavers in the morgue and more *dénunciateurs* reporting killing of members of their family or friends. The goal was to measure how much correlation there was between the monthly movements (rises and falls) of the EAAF and CNVJ killings data.

We measured this effect by calculating the increases or decreases of killings from each month to the subsequent month, which are statistically called *first differences.* The first data point was calculated as the difference in the number of killings in September 1991, from October 1991 in the CNVJ data. The next point is the difference between October and November in the same data, the third between November and December, and so on. We repeated these calculations for the months up to October 1994.

To determine if these two series represent different views of the same phenomenon (and thus confirm that the CNVJ data are representative), we used the first differences of both series to determine the extent to which the two monthly series move together despite the large difference in their absolute values.

We determined correlated movement in two ways:

1. Graphical, by visual examination of the scatterplot of the CNVJ and the EAAF first differences.
2. Analytically, by computing the correlation coefficient to numerically express the extent of the correlation.

[11] Dr. Mercedes Doretti and her international team of forensic experts were a second component of AAAS assistance to the CNVJ.

From the scatterplot, a viewer can readily ascertain that there is a strong tendency for a rise in the EAAF series to be coincident with a rise in the CNVJ series, and similarly, for coincident behavior of the declines. In addition, there is an obvious tendency for the magnitude of the differences to correlate: the greater the rise in one series, the greater the rise in the other.

For the analytical measurement, we calculated the Pearson Product Moment Correlation Coefficient (*r)* these series to be r=.865 (r^2=.748), which is quite high. If these data were from random samples from two series in which there was no correlation among first differences, the probability of such a high correlation occurring by chance is vanishingly small ($z = 10.4$, $p < .000000$).[12]

This was strong evidence of a similarity in the two series, confirming the representative validity of the CNVJ team data's representation of the trends of violations.

Validity of the CNVJ Survey with Respect to Location

The sample of *dénunciateurs* was not distributed in exact proportion to the number of people living in each location. For example, approximately 33% of Haitian people live in the Ouest department, where the team performed only 18% of its interviews. We felt it necessary to deal with the possibility that our findings regarding inter-departmental differences in absolute numbers of violations are due in part to our area sampling bias.[13]

Our technique for estimating the amount of over- and under-sampling by department is to compute the ratio of the proportion of all Haitians living in a given department to the proportion of interviews collected in that department. This ratio is denoted *sampling weight*. The departmental sampling weights ranged from .273 to 2.009 (mean value 1.11, standard deviation .63).

Using these sampling weights, we recalculated the numbers of different types of violation by department at the national level, including Raboteau and Port-au-Prince. The principal result of applying these sampling weights is that the Ouest department becomes even more extreme in its number of violations, having more violations than other departments. Even in the raw data, the Ouest (usually with Port-au-Prince separated) had been one of the departments with the most numerous violations. This effect is consistent across different violation types, but is most pronounced for execution and disappearance. In violations of the rights to life and property, Nord and l'Artibonite seem also to have more violations than the other departments in the weighted data, but violations are more evenly distributed across departments in violations of the rights to integrity and liberty.

We concluded that controlling for area sampling bias by this method did not affect our findings.

Correlation Analysis of the Violation Time Series

Summary

The analysis in this section deals with the tendency of the numbers of violations of different types to increase or decrease at the same times and the tendency of numbers of a particular violation type to rise and fall together in time across different departments. As we discussed above, we argue that finding high correlation among a number of types of violations and different departments is strongly suggestive that the perpetrators of different types of violations, and the perpetrators of the same type of violation in different departments, are responding to similar influences that lead them to act at the same times. In both analyses, we found the hypothesized high correlations in these data and conclude that the human rights violations were consistent in time across different kinds of violation, and that the same kinds of violation were consistent in different departments.

Similarity of trends among different kinds of violations

How similar are the trends in time of the different types of violation? By similarity, we mean the tendency of the trends to rise and fall together. If different types of violence tend to occur at nearly

[12] In this case, n=38. For n>30, when the population correlation coefficient is zero, the standard error is sqrt($(1-r^2)/(n-2)$), and the sampling distribution is normal. In this case, $(1-r^2)$ = .252, $(n-2)$=36, and the standard error is .084; z=.865/.084=10.4.

[13] Our analyses of trends in time are independent of differences in the absolute magnitudes of the series being considered. Accordingly, issues of regional coverage in the sampling do not affect the trend analyses.

the same times, and to rise and fall together, then we have a strong suggestion that people who perpetrate violence are responding to the same influences. If, however, some kinds of violence follow very different patterns in time, then we would have to conclude that influences to commit violence are different for different kinds of violence or that the data that we have do not support the contention that they shared a common influence.

As we did previously (in Validity of th**e** Survey with Respect to Time), we compared the EAAF University Hospital cadaver data and the CNVJ summary execution data. We obtained the correlation of the first differences of the monthly counts of violations for each of ten types of violation. Table 3 shows the violation codes and Table 4 shows the correlation coefficients for these violations.

Table 3: Violation codes.

VBN	Attacks on goods	**VMP**	Threats and persecution
VDS	Disappearance	**VDT**	Arbitrary detention
VES	Arbitrary execution	**VMS -**	Massacre
VEX	Extortion	**VSX**	Rape and sexual abuse
VLB	Attacks on liberty	**VTE**	Attempted murder
VTT	Torture		

Table 4. Correlation matrix of monthly first differences between different types of human rights violations at the national level in Haiti, September 1991 to October 1994.

Type of violation	VDS	VDT	VES	VEX	VLB	VMP	VSX	VTE	VTT
VBN	.66	.88	.88	.81	.88	.93	.74	.93	.94
VDS		.80	.87	.61	.74	.72	.74	.78	.80
VDT			.84	.92	.93	.93	.84	.90	.96
VES				.64	.81	.84	.78	.90	.90
VEX					.85	.87	.69	.81	.88
VLB						.97	.88	.91	.95
VMP							.88	.94	.98
VSX								.83	.86
VTE									.96

We performed hypothesis tests on each of the coefficients to determine which of the coefficients were significantly different from zero at the $\alpha = .01$ level. This is a more stringent requirement than the usual $\alpha = .05$ level. Of the 45 correlation coefficients, 45 were statistically significant at the $\alpha = .01$ level.

Because we made multiple hypothesis tests, we could not assume that the significance level is truly $\alpha = .01$. To determine the possibility that this many correlation coefficients could have been found significant by chance, we determined the probability that out of the 45 "trials," all 45 of these hypothesis tests were truly significant.

Since the probability of a hypothesis test turning out to be significant by chance is .01 (by our choice of a = .01), we modeled this process with the binomial distribution for 45 trials and 45 "successes" each with a probability of success of p=.01. The result is p<.000000; it is essentially impossible to have this many significant correlation coefficients by chance. The violation counts are measuring an underlying, common phenomenon. Additional support for this contention comes from the magnitude of the correlation coefficients. The coefficients range from .61 to .98. The mean is .85, and the median, .87. These coefficients are not only statistically significant (i.e., not zero), they are *practically* significant; these are strong correlations. Additional support from this statement is found by looking at the r^2 values in Table 5.

Table 5: Matrix of square of the correlation coefficients (r^2) for the monthly first differences between different types of human rights violations at the national level in Haiti, September 1991 to October 1994.

Type of violation	VDS	VDT	VES	VEX	VLB	VMP	VSX	VTE	VTT
VBN	.44	.77	.77	.66	.77	.86	.55	.86	.88
VDS		.64	.76	.37	.55	.52	.55	.61	.64
VDT			.71	.85	.86	.86	.71	.81	.92
VES				.41	.66	.71	.61	.81	.81
VEX					.72	.76	.48	.66	.77
VLB						.94	.77	.83	.90
VMP							.77	.88	.96
VSX								.69	.74
VTE									.92

Similarity of trends between departments

For any given type of violation, how similar are the trends in time for different departments? That is, does the number of arbitrary detentions in the Nord-Ouest increase and decrease in the same periods as the number of arbitrary detentions in l'Artibonite?

We sought an answer to this question in the same way that we did for the similarity of trends between different types of violations. Table 6 shows the number of significant coefficients by violation type. The second column of Table 6 presents the probability of obtaining that many significant coefficients by chance. Note that in all but the case of arbitrary executions, it is essentially impossible that we would find this many significant coefficients by chance.

Table 6. Number of significant correlation coefficients for the monthly first differences between different types of human rights violations by department in Haiti, September 1991 to October 1994.

Type of violation	Number of significant correlation coefficients	Probability of this many or more out of 45 occurring by chance
Torture	29	.0000000000
Arbitrary Detention	29	.0000000000
Attacks on goods	21	.0000000000
Extortion	14	.0000000000
Arbitrary execution	8	.0000000006
Rape and sexual abuse	7	.0000000016

For all of these kinds of violations, it would be extremely unlikely to find so many significant correlation coefficients by chance (e.g., for rape, seven significant correlation coefficients in 45 pairs would occur by chance on average 16 times in ten billion trials). We find that the departments are much more consistent for the violations for which there are many more instances of this kind (torture, arbitrary detention, attacks on goods). More than half of all possible pairs of departments has significant, nonzero correlation coefficients between their monthly first differences of instances of torture and arbitrary detention.

Although the findings are weaker, it is also true that instances of rape and sexual abuse, extortion, and arbitrary execution are consistent across departments. Across all the kinds of violations examined here, the number of significant correlation coefficients found is sufficient to find that for these kinds of violations, the violence was committed consistently in time across different departments.

A concern of the project was the extent to which data could support the hypothesis that there was national influence on local perpetrators.[14] We took the approach of seeing if the time series for the several departmental series for relevant variables were in fact, more correlated in their trends than would be likely by chance. Similarity of movements of a variable in all departments would be taken as evidence of a national influence. For example, if in most departments rape rose during the same several months and fell during others, this would be an indication of some kind of national control.[15]

[14] It was beyond the scope of the AAAS scientific component of the CNVJ activities to attempt to determine that nature of that influence.

[15] While national motivation or stimulation of actions is a likely cause, it is also possible for apparently disorganized local perpetrators to communicate about their activities or be the recipients of information coming from other departments.

Lessons Learned

Entity, Function	Lesson	Recommendation	Issues
Commission level decisions affecting the effectiveness and efficiency of information system work	Conflicts over goals related to political issues can affect the work of the database area by increasing mid-course changes, and delaying imperative operational decisions.	Project personnel must develop ways to effectively work with commissioners to communicate the consequences of commission-level decisions.	Project personnel below the leadership level may have no voice at the commission level. Some "political" choices, despite their effects on the information system, may be consistent with the commission's mission. Lack of understanding of information systems and research at the commission level. The commission has the final word.
Decision-making at the project administration level	Leadership counts. Lack of leadership means delays affecting effectiveness and efficiency of project	Project personnel must inform commission management of the dangers of delays.	Developing the communication skills needed to speak to commission management in terms and ways that enable them to understand the consequences of leadership vacuums.
Release of scientific findings	Suppression or non-release of scientific findings.	Negotiate for controlled release of scientific findings, based on meeting security conditions, limiting output, access to data, etc.	Negotiation of these terms and conditions should be done at the start of the project.
Responsibility of technical functions	Lack of clear responsibility leads to wasted effort, either because of doing unnecessary tasks or doing work over.	Project leadership must obtain clear and relevant definitions of responsibility.	Access to management. Need for high level of interpersonal skills for project (technical) leadership. Need for a "champion" among the commissioners.
Lack of understanding of technical work	Lack of understanding of technical work at management levels leads to bad decisions.	Project personnel and technical leadership must invest the time and effort to present the technical case.	Access to management.
Variables in data collection	Variables that seem relevant at start may not draw responses, or such responses as are obtained are not useful	Pilot test proposed questionnaires for field interviewing, and put through a pilot data processing and analysis cycle.	Unwillingness of some personnel to invest the effort up front to save effort in the future, when there is pressure for immediate results.

Missing data for certain variables	Despite pilot testing, field interviews may develop high rates of non-response on certain variables	Detect the problem quickly and determine corrective action	Early detection means concurrent processing of interviews

References

Ball, Patrick. 1996. *Who Did What to Whom? Planning and implementing a Large Scale Human Rights Data Project*. Washington, DC: American Association for the Advancement of Science.

Ball, Patrick, Ricardo Cifuentes, Judith Dueck, Romilly Gregory, Daniel Salcedo, and Carlos Saldarriaga. 1994. *A Definition of Database Design Standards for Human Rights Agencies*. Washington, DC: American Association for the Advancement of Science and Human Rights Information and Documentation Systems International.

Dorretti, Mercedes, and Ignacio Cano, n.d. "Violations of the Right to Life in Haiti: 1985 to 1995." Paper available from authors.

Chapter 3

The South African Truth and Reconciliation Commission: *Data Processing*

Themba Kubheka

Introduction and Objectives

The objectives of the Truth and Reconciliation Commission of South Africa, as set out in the Promotion of National Unity and Reconciliation Act, are to:

- Give as complete a picture as possible of the gross violations of human rights that took place as a result of past conflicts.
- Restore to victims their human and civil dignity by letting them tell their stories and recommending how they can be assisted.
- Consider granting amnesty to perpetrators who carried out the abuses for political reasons, and who give full details of their actions to the Commission.

The Act defines a *gross violation of human rights* as

(a) the killing, abduction, torture or severe ill-treatment of any person; or

(b) any attempt, conspiracy, incitement, instigation, command or procurement to commit an act referred to in paragraph (a), which emanated from conflicts of the past and which was committed during the period 1 March 1960 to 10 May 1994, within or outside the Republic [of South Africa], and the commission of which was advised, planned, directed, commanded or ordered, by any person acting with a political motive.

To gather information about the gross violations of human rights suffered by South Africans, a Gross Human Rights Violation (GHRV) statement was developed and referred to as *the protocol*. The GHRV statement gave victims an opportunity to relate the violations they suffered, and in so doing, provided the information for data processing. As the commission went about its work, the GHRV statement went through several conceptual stages as ordered below:

> **Tell your story.** It started as a narrative statement, but developed into a questionnaire to make it easy for victims to understand.
> **Give the deponents the emotional space to tell the story in their own way**. This meant presenting the events and highlighting the issues as perceived by the statement-giver. However, some regional officers believed that the TRC had to serve the deponent's emotional needs.
> **For many people the act of giving a statement was a mini-hearing**. The GHRV statement fell into two main groups of deponents' statements: those made by victims themselves and those made on behalf of victims.

Information also came to the TRC by letter. Initially, letters were screened and the letters which were accepted were those that provided narratives that were within the mandate of the commission. Later in the course of the work, a *Designated Statement Program* helped in-house statement takers reach out to thousands of South Africans who suffered gross human rights abuses. This program was administered by non-governmental organizations.

In the next section, I give a chronological history of the Data Processing Unit of the TRC, and a summary of its functions and work practices as they progressively developed during the project.

History and Operations of the Data Processing Unit

The Data Processing Model

Prior to the establishment of the TRC, a database development group was formed to establish the goals and deliverable objectives of the database, along with a work plan. The overall design

was derived from the "who did what to whom?" model, based on the experience with databases for other truth commissions.

The database consultant[1] applied a “Who did what to whom?” model. In this model, which is based on his experience with other truth commissions, the first principle is to record “who did what to whom.” For example, in “John hit Jane on the head,” the victim and perpetrator are linked through the recorded act of violence. “When and where” place the act in the context and are recorded through the narrative that links acts into a coherent whole. Recorded acts are specific and are the building blocks of the system. For example, a number of acts, e.g., blow to the head, electric shock to the genitals, make up an incident of torture. An event comprises several incidents, such as “The arrest, detention and torture of Mr. Masinga.”

The rationale for detailed recording of each act is the complexity of incidents of human rights violations. During a single event such as the Boipatong massacre, there can be many linked perpetrators, victims and acts separated from each other at various times and places. To make sense of this massive amount of information, it is important to break down the event into its component parts in the greatest possible detail.

This conceptual model of a human rights violation protocol is the basis for the training of statement takers and their subsequent work with deponents.

The Role of Data Processing

Following the taking of a statement, a *data processor* was employed to decide which protocol information would be processed for entry into the database. To consistently and reliably make these decisions we needed *controlled vocabularies* for violations as shown in Appendices 1 (Initial Action Types and Codes) and 2 (The Final Set of HRV Categories). These controlled vocabularies are coded for entry into the database.

Every data processor had a copy of the controlled vocabularies and copies were given to researchers, investigators and other concerned parties. The head data processors had final approval over any additions or deletions. To assure that the data processors cannot change these codes, the HRV categories and their types of violations are hard-coded. Appendix 3 (Example of Coded Killings) shows the appearance of these codes when they are coded for particular killing violations.

In addition to violations, there are other items of data for which controlled vocabulary is required, such as organization, locations, etc. These items are described later.

Chronology of Events

We then initiated the process of collecting supporting information. This process involved contacting a wide range of organizations and institutions to assemble comprehensive lists of necessary and potential data sources. In this process, additional data processors and data entering personnel were employed by the TRC and were involved in the task of information collection. Because there was a large amount of information that was not available in machine-readable form, it was decided that the data processing unit would compile a range of resources. These could be stored in hard copy as part of a resource pack or included in the computerized information on the database. At this time, we started collating some of the information collected, such as cross-checking various lists of trade unions.

We organized and conducted a one-day workshop to train new data processors and data entering personnel. Training materials included an overview of the legislation governing the work of the TRC and the “Who did what to whom?” model. In this workshop, we brainstormed acts of violence, adding additional acts of violence to the list already developed and discussing the hierarchical organization of the acts, in particular how this structure would relate to the TRC’s legislative categories of human rights violations. The consensus was that some of the enlarged categories were too broad to be analytically useful and that the hierarchies should be based on acts of violence; for example, the category of asphyxiation would include tear-gassing. How to link acts of violation to the legislative categories was left open. Once the draft list of acts of violence had been

[1] Patrick Ball of the American Association for the Advancement of Science acted as methodological advisor to the TRC and led the database development group. The "who did what to whom?" model is outlined in (Ball et. al., 1994) and (Ball, 1996).

completed, a list of synonyms for the acts was compiled, as well as a non-hierarchical, alphabetical list, as shown in Appendix 4 (Acts of Violence).

As the discussion of acts of violence continued, researcher Lydia Levin and systems analyst Gerald O'Sullivan worked with the database software vendor, Oracle, to develop a database model. Data processors brainstormed, proposed and reviewed possible questions to which the database would supply answers. For almost three months the data processors and data entering personnel met daily to review every possible act of violence one person could inflict on another. Data processors provided ongoing and *ad-hoc* assistance to the research department, particularly in compiling a legislative chronology. The results are tabulated in Appendix 4 (Acts of Violence). Every act was written on the board and debated in detail.

Under the four categories, Killing, Abduction, Torture and Severe Ill-treatment, described in the Act, we ended up with about 200 types of violations (Appendix 1, Initial Action Types and Codes) which were later reduced to 90 (Appendix 2, Final Set of HRV Categories). The TRC added the two categories, Attempted Killing and Associated Violations.

Using the initial codes of Appendix 1 (Initial Action Types and Codes) we could only detect the act of violation from the outcomes, such as: did the victim die, or become injured or miscarry. Hence, we would code one of the following outcomes: death, injury, damages to property, pregnancy, disappearance, abduction and forced removal as the violation.

The Associated Violations category, which is not a gross violation of human rights, is important for understanding the context of the act. There are also two more categories for unclassified cases: Other and Unknown violations. Each of these categories has several sub-headings which explain how the violations took place (a person can be killed by different methods, so we need to identify how they were killed). By breaking the categories into sub-headings, we can then do meaningful counting for the final report.

In September of 1996, the three head data processors from Johannesburg, Durban and East London met their counterparts and the head researchers in Cape Town. The CEO of the TRC and the TRC's methodological advisor decided that we should reduce the 200 acts of violation to about 50. After three days we could not reach consensus and returned to our regional office to consult with our respective data processors. Following that consultation, consensus was mandated, and achieved. At our second workshop, we were told to produce the final product and we did.

The data processors who were by this time using the initial codes of Appendix 1 (Initial Action Types and Codes), put forward a number of observations, critical comments and objections, expressing their concerns about the categories of Appendix 2 (Final Set of HRV Categories). They felt a detailed violation was more meaningful in describing an act. For example, beating a victim with a gun was different from beating the victim on the soles of feet or whipping with a Sjambok (initial action codes in Appendix 1). In the new codes of Appendix 2 (Final Set of HRV Categories) these different types of assaults all were subsumed under BEATING.

Of course, with the new codes, the data processors' task was easier and we could process more statements than before and the new codes of Appendix 2 (Final Set of HRV Categories) were retained. To synchronize the acts already captured with the new ones, the system analyst Gerald O'Sullivan created an Excel spreadsheet for each regional office and instructed head data processors to change the old acts to the new ones, line by line. The Johannesburg office had over 10,000 lines of codes of violations to be modified, but the job was done and the corrected spreadsheets loaded into the database.

What Was Data Processing at the TRC?

Critical questions in an information management system include, How will the information be processed? How will raw data from a large number and wide variety of sources be transformed into a body of information to be used to generate an analysis based on a database of a problem, issue, or situation?

Integrating the Data

How will these raw data be integrated? Each source has common and unique problems. The primary information collected from witnesses is vulnerable to bias and inaccuracy. Documentary sources of evidence such as commission reports contain their own assumptions. Different methods

are needed to handle each information source. Dealing with these input issues to produce useful, integrated results is *data processing.*

But how outputs are defined -- what will be done with the data once it is processed -- also determines the nature of data processing. The TRC uses information from the data base for many purposes: to conduct research, to facilitate investigation, to record the testimony of victims of human rights violations, to record the evidence of amnesty applicants and to formulate a reparation and rehabilitation policy. Not all of the output of the system will be tangible. The understanding of dynamics, conflicts, and so forth, that investigators, and possibly statement takers develop, is information that will not necessarily be easily put into a protocol. To satisfy the objectives of the TRC, data processing must make it possible for the database to serve all these needs.

Not only will data come from different sources but also it will be gathered at different regional and national levels. It is crucial that all these information-gathering processes are not carried out in parallel to each other but are part of an integrated whole. At this level of integration, data processing is the key element. It is the point at which the incoming data and information is managed and organized, and where the analytical process begins. Data processors make a great number of decisions about how to define the information. Such decisions might include the answers to questions such as these: “Is this truly a gross human rights violation?” “Was it part of the Boipatong massacre?” “Is Colonel Swanepoel the same man already implicated in numerous other torture cases?”

Data processing is where the investigation begins. Data processors deal on a daily basis with the full spectrum of incoming information. They should be the first to pick up on discrepancies between the stories of amnesty applicants and their victims, and first to identify the trends of violations in particular areas or perpetrated by particular people, units, sector of society, and so forth. A structured means of feeding this information into the research and investigation processes on an ongoing basis is crucial or these insights will be lost.

Also, data processing is the first point of contact between the national process of amnesty applications and the regional processes of human rights violations reporting. It is the skills of data processors that assure that amnesty applications can be cross-checked with reports of human rights violations. National investigations and research processes are meaningless unless they can draw on the full range of information available from different regions.

None of these processes necessarily happen sequentially. As the TRC does its work, verification of new information, or contradiction of information, adding to or complicating information already collected, is an ongoing process. Hearing evidence for amnesty applicants or victims of human rights violations must be recorded and linked to the original statements of witnesses or applications for amnesty. Discrepancies and additions must be identified and fed into the research, investigation and reparation processes.

In addition to its research and investigative functions, the TRC is also attempting to deliver reparations to victims. It is crucial to this process to accurately record the individual consequences of violations of human rights and the needs resulting from those violations. This information must be systematically gathered and processed to generate a national policy on reparation and rehabilitation and to ensure full attention to the needs of every victim of a reported human rights violation to the TRC.

Capturing the basics

The database was designed to accept statements and pass the information through stages on the way to making findings. Later at the TRC, I had two jobs, that of being the Information Coordinator and the Documentation Officer. I received all statements and I reviewed them to check whether they fell within the TRC mandate. Based on the mandate, my criteria for accepting a statement were:

1. Time period - 1 March 1960 to 10 May 1994.
2. Politically motivated.
3. Gross human rights violations.

My guidelines were as follows:

? Do not analyze a bundle of statements as a group and then capture them as a group. In so doing, you may confuse the statements.

- Finish one statement entirely before moving onto another. You must code exactly what is on the statement, even if you believe the statement is inaccurate, is full of contradictions, etc.
- You must be very careful not to allow any biases to creep in, and further, not to allow any of your own commentary or observations to enter the coding. What is captured is exactly what the statement says, but in a coded form.
- Remember that you are trying to extract as many acts and victims per statement as possible. Even if you have scanty details about a particular event, code and capture what is there. You may be able to gather more information through investigations, research, etc.

A step-by-step description of how data was captured is given in Appendix 7 (Data Processing User Guide).

Problems

To show the nature of the work of data processing, the following is a list of some of the problems we encountered and resolved. (Note: The lessons learned in this aspect of work on the TRC information management system have been integrated with those of the database representation, and are found in Chapter 4.)

Data flow

- How will new information collected through hearings, statements, informally, and through the work of the investigators, be entered and again made available for further research and investigation?
- How will the database or processing be used to handle the problem of various types of information coming from the different sources?
- How will the information gathered be fed into the investigation and research processes?
- How to maintain a high rate of document processing in view of potentially lengthy verifications and statements that are difficult to code?

Definition and naming

- How to name categories? For example, we have DEGRADATION. But the purpose of torture is degradation. Should we then use SHAME or EMBARRASSMENT?
- How do we deal with inconsistency, such as between “death” and the omission of “attempted killing”
- Should we state whether a statement met the TRC criteria using the words ACCEPTED or REJECTED?
- In what category do you put assassination, circumstances of death and capital punishment?
- What is the purpose of all the synonyms?

Quality control

- How to minimize coding errors?
- How to establish and enforce consistent coding practices?
- How to check for typing errors?
- How to code ambiguous information?
- What is the impact of serious errors on the quality of data?
- What are the implications of changing information and updating the database?

Overall issues

- What is role of the database in reparation and rehabilitation?
- How can the database be used as an integrative tool? For example, how do we link national and regional processes, amnesty, HRV and reparations?
- What will be the benefits of feeding information to investigation and research processes?
- What is the desired role of the database in supporting the objectives of the TRC?
- How will the database be used as a corroborative tool (what for and how)?

- What is the role of the database in tracking various processes happening in the TRC?
- How will the database assist in making HRV findings?
- If we look for trends, how will we identify which trends to look for and how will these trends be tracked?

System design

- How to incorporate the free text field
- How to group acts into events and mega-events, and finding connections between persons places, acts, and vehicles.
- Should the data processor attempt a preliminary corroboration of the file material and if so, how?

Getting the Relevant Data

After the major design problems were resolved, assignments were given to both data processors and data enterers to assemble relevant literature and statistical documents from multiple sources to capture into the database.

Examples of the kinds of literature and documents that we captured on the database included community policing, policing in general, military and private army uniforms, private army structure, police dockets, police station daily registers, NGO databases.

We obtained such information from NGOs, police authorities, library reference materials, military museums, university archives, purchased books (for example, from the Institute of Race Relations), public libraries, the National Archivist, and so forth. Our personal experiences with the apartheid system also provided valuable information.

Specific types of information we obtained included the names of all the political parties, the names of all liberation movements, new far right parties and armies, non-governmental organizations, trade unions, civic, women's, students' societies and organizations, military and paramilitary forces and their ranks, all the languages spoken in South Africa, lists of other countries, and regional divisions (breaking down each region to its smallest entity).

Some of our breakdowns of information may be of value to future data processing projects. These include:

Relationships of victim to interviewee. Brother, sister, father, mother, husband, wife, step-mother, -father, -brother, -sister, -grandfather, -mother, aunt, uncle, other relative, friend of the victim, neighbor of the victim, colleague of the victim, student of the victim, teacher of the victim, person living in the same house, an employer of the victim, member of the same organization as the victim, lawyer of the victim, priest of the victim, other relationship to the victim.

Exact location. Victim's home, home of a colleague or work associate, home of a friend, victim's place of employment, business owned by the victim, land owned or rented by the victim, shopping center, military headquarters, police station or post, court, union office, open veld, vacant lot, park, graveyard, road, taxi rank, train station, parking lot, embassy, border crossing, refugee camp, guerrilla camp, inside a toilet, in a police cell, in the bedroom, in the street, etc.

Military and paramilitary forces and their ranks. Examples are South African Defense Force (SADF), South African Police (SAP), General, Lieutenant General, Major General, Brigadier, Colonel, Commandant (now Lieutenant Colonel), Major, Captain, Lieutenant, Second Lieutenant, Candidate officer, CCB - covert organization.

Coding Using the Data Processors' Coder's Sheet

When coding, the first step is to decide on the HRV category into which category the violation falls. Appendix 2 (Final Set of HRV Categories) shows the categories with their definitions (the words in boldface should be used in the description of the act when they are applicable).

We developed a coder's sheet to enable data processors to code acts in a chronological order since most statement narratives are not chronological. Some protocols had only one act belonging to one victim, others had several acts dating back as far as 1960 and others had two or more victims with one or several acts. Some abuses may have happened at one time or at several different times and at different places. If several victims were mentioned in the same protocol, we tried to group

together those acts belonging to that particular victim in a chronological order. Appendix 5 shows a completed coder's sheet and Appendix 6 is a complete TRC statement.

TRC System User Guide

A TRC user guide with the various input screens was provided by Oracle, the database designers. Later I simplified this guide to use in training new data processors. To show the nature of this guide, in Appendix 7 (Data Processing User Guide), I show the main screens used by data processors in their daily capture of statements into the database.

Problems with data processing

Despite our many efforts to train, standardize, and simplify, we still had many problems to solve. We did solve them with varying degrees of effort, and we list some of them here as a guide to others working on similar projects.

- Incomplete statement by statement-takers, e.g., no acts of violation, no narrative, no address, no victims, etc. Difficult to recall deponent to correct errors especially those who live in rural areas.
- Duplicate statements, i.e., statements taken twice and the narratives differ, same narrative but statement has been registered a second time.
- Incorrect registration and statement has already been processed.
- Illegible statements.
- Forged statements by statement takers.
- Untrue stories by victims.
- Unclear TRC mandates in statements.
- Not within the mandate and already registered.
- Statements taken or delivered after the official closing date.
- Victim applying for both HRV and Amnesty in the same protocol.
- Lost and missing statement i.e., in the office, post office. Registered but statement is missing. Difficult to re-take statement.
- Unsworn statement. Faxed, letters, sent by post.
- Undelivered statement, e.g., the Designated Statement Takers (DST) program - DSTs would not deliver statements until paid direct.
- Reference number designed to enable extraction of statement from various areas within the old boundaries. Difficult to extract statements from former homelands.
- Person ID number - create one number for perpetrator, e.g., SAP-UNKNOWN and use same number in other statement rather than create a new one for each statement.
- Bad data processing - incorrect/missing acts of violations, biased/wrong summaries, no witnesses/perpetrators, etc.
- Tracking of statements.
- Bad photocopying of copy statements.
- How many statements per day per processor. Long, winding statements take a long time to process.
- Number of data processors per region.
- Vetting - quality control.
- Daily queries from victims. Who must attend to them?
- Attending to request by other sections, e.g., investigators, researchers, etc.

Appendix 1

Initial Action Types and Codes

This is the initial set of action types and their codes, as described in the section, Chronology of Events. Data Processing used these types and codes in the initial coding work, but in September 1996, teamwork among the head data processors and researchers resulted in the final set of categories and types shown in Appendix 2.

Top Level

Abduction	ABD	Harmful substances	SBS
Abuse using animals	AUA	Improper burial	BRL
Assault	ASS	Incarceration	INC
Assault using vehicles	AVE	Life threatening situations	LFT
Bombing	BOM	Physical stress	PHY
Burns	BRN	Psychological torture	PSY
Capital Punishment	CPP	Sexual abuse	SEX
Deliberate spreading of disease	ILL	Shooting	SHT
Deprivation	DEP	Staged accident / suicide	STG
Drowning	DRW	Stress to the senses	STR
Electric shock	ELS	Suffocation	SUF
Financial impropriety	FIM	Theft and vandalism	THF
Framing	FRM	Threats	THR
Harassment	HRS	Violation after death	VLN

Subsidiary

Abduction

Forcible abduction	ABD_FRDC	Other	ABD_OTHR
Unknown	ABD_UNKN		

Abuse using animals

Animal abuse	AUA_ANML	Unknown	AUA_UNKN

Assault

Hitting / Kicking / Slapping / Punching	ASS_HKSP	A Kaffir Klap	ASS_KKLP
Stabbing and/or Hacking with a panga, knife, sharp object	ASS_SHRP	Banging head against wall	ASS_BHDW

Beating with blunt object e.g. baton, hosepipe, knobkerrie, cane	ASS_BLNT	Scalping	ASS_SCLP
Sjambokking and/or Whipping	ASS_SJMB	Removal of nails	ASS_RMNL
Beating with a gun e.g. rifle, butt, pistol-whipping	ASS_BGUN	Arms and/or wrist twisted	ASS_TWST
Stoning and/or Objects thrown at victim	ASS_STON	Breaking bones	ASS_BRBN
Spitting	ASS_SPIT	Pulling out of hair	ASS_PLHA
Beating on soles of feet	ASS_SOLE	Amputation	ASS_AMPU
Beating pregnant woman on stomach	ASS_PREG	Unknown	ASS_UNKN
Clapping on ears with both hands	ASS_EARB	Other	ASS_OTHR
Pulling out of teeth	ASS_PLTE		

Assault using vehicles

Dragged / pulled behind / attached to a moving vehicle	AVE_DRAG	Thrown out of moving car / taxi / bus / etc.	AVE_MCAR
Put in boot	AVE_BOOT	Rough ride	AVE_RIDE
Driven over	AVE_DRIV	Unknown	AVE_UNKN
Thrown out of moving train	AVE_MTRN	Other	AVE_OTHR

Bombing

Bomb	BOM_BOMB	Letter / parcel bombs	BOM_LPBM
Land mine	BOM_LMNE	Car bomb / bomb placed to go off in a car	BOM_CARB
Petrol bomb	BOM_PBOM	Other booby trap / disguised bombs / bomb is hidden in equipment so that when you use it, the bomb goes off	BOM_BOOB
Hand grenade	BOM_HGRN	Unknown	BOM_UNKN
Grenade / mortar bomb / shell	BOM_GRMS	Other	BOM_OTHR

Burns

Chemicals	BRN_CHEM	Necklacing	BRN_NKLC
Cigarettes	BRN_CIGR	Set alight with petrol	BRN_PTRL
Scalding (with water)	BRN_SCLD	Unknown	BRN_UNKN

Live fire	BRN_FIRE	Other	BRN_OTHR

Capital punishment

Judicial hanging	CPP_JHAN	Other	CPP_OTHR
Unknown	CPP_UNKN		

Deliberate spreading of disease

Disease spread	ILL_ISPR	Other	ILL_OTHR
Unknown	ILL_UNKN		

Deprivation

Deprivation of medical attention / treatment	DEP_MEDC	Deprivation of privacy	DEP_PRIV
Deprivation of food and/or water	DEP_FOOD	Deprivation of sanitary facilities	DEP_SNTR
Deprivation of sleep	DEP_SLEP	Deprivation of darkness	DEP_DARK
Deprivation of light	DEP_LITE		
Unknown	DEP_UNKN		
Other	DEP_OTHR		

Drowning

Total submersion in water	DRW_TSBM	Unknown	DRW_UNKN
Head submersion in water	DRW_HSBM	Other	DRW_OTHR

Electric shock

Electric shock to the genitals	ELS_GNTL	Unknown	ELS_UNKN
Electric shock to the body	ELS_BODY	Other	ELS_OTHR

Financial impropriety

Bribery	FIM_BRIB	Blackmail	FIM_BLML
Extortion	FIM_XTRT	Unknown	FIM_UNKN
Pay off	FIM_PYOF	Other	FIM_OTHR
Ransom	FIM_RNSM		

Framing

Person framed	FRM_PRSN	Other	FRM_OTHR
Unknown	FRM_UNKN		

Harassment

Surveillance	HRS_SRVY	Telephone harassment	HRS_TELE
Dismissal from employment	HRS_DSMS	Unknown	HRS_UNKN
Frequent unwanted visits	HRS_VSTS	Other	HRS_OTHR

Harmful substances

Poison	SBS_POSN	Unknown	SBS_UNKN
Medication / Drugs	SBS_DRUG	Other	SBS_OTHR
Common household materials	SBS_HHLD		

Improper burial

Buried in shallow grave	BRL_SHLW	Anonymous burial	BRL_ANON
Buried alive	BRL_LIVE	Unknown	BRL_UNKN
Mass grave	BRL_MASS	Other	BRL_OTHR

Incarceration

Detention (if victim reports act as Arrest AND detention, only enter as DETENTION)	INC_DETN	Banning	INC_BANN
Arrest	INC_ARST	Unknown	INC_UNKN
House arrest	INC_HRST	Other	INC_OTHR
Banishment	INC_BNSH		

Life threatening situations

Left for dead	LFT_LEFT	Unknown	LFT_UNKN
Released into hostile environment	LFT_HENV	Other	LFT_OTHR
Released into an unknown place	LFT_UNPL		

Physical stress

Forced stationary postures	PHY_FRSP	Suspension - hanging victim by arms, legs, etc.	PHY_SUSP
Forced exercise and/or labor	PHY_FREX	Unknown	PHY_UNKN
Stretching of limbs and/or trunk	PHY_STRL	Other	PHY_OTHR
Helicopter - hanging the victim from a stick between knees and arms	PHY_HELI		

Psychological torture

Simulated execution	PSY_EXCU	Victim is forced to watch and/or listen to torture of others	PSY_WTCH
Detention of significant other people	PSY_DTEN	Victim is forced to participate in the torture of others	PSY_PART
False and alarming information	PSY_FLSE	Victim shown other torture victims	PSY_SHOW
Russian Roulette - gun against the head with one bullet left	PSY_RUSS	Unknown	PSY_UNKN
Solitary confinement	PSY_SOLI	Other	PSY_OTHR
Pistol placed in mouth	PSY_PIST		

Sexual abuse

Forced sexual acts	SEX_FRSX	Pumping water into the uterus	SEX_PUMP
Introduction of objects into the rectum / vagina / urethra	SEX_OBJT	Nakedness	SEX_NKED
Rape by someone of the opposite sex	SEX_RAPE	Abuse with bodily fluids	SEX_DBFL
Rape by someone of the same sex	SEX_RPSS	Abuse using animals	SEX_ANIM
Gang rape	SEX_RPGA	Suspension of weights from genitals	SEX_SPWG
Assault and/or touching	SEX_ASLT	Unknown	SEX_UNKN
Body searching	SEX_BODY	Other	SEX_OTHR
Genital mutilation	SEX_GTMU		

Shooting

Rubber bullets	SHT_RUBB	Unknown	SHT_UNKN
Live ammunition	SHT_LIVE	Other	SHT_OTHR
Birdshot and/or buckshot	SHT_BIRD		

Staged accident / suicide

Staged accident	STG_ACCI	Unknown	STG_UNKN
Staged suicide	STG_SUIC	Other	STG_OTHR

Stress to the senses

Loud noises and/or music	STR_LOUD	Confined to a small space	STR_SMAL
Powerful lights	STR_LITE	Bad smells	STR_SMEL
Blindfolding	STR_BLND	Complete immobilization	STR_IMMO
Exposure indoors to extreme heat or cold	STR_HEAT	Handcuffed	STR_HAND
Exposure to elements	STR_ELEM	Unknown	STR_UNKN
Bound or tied up	STR_TIED	Other	STR_OTHR

Suffocation

Hanging	SUF_HANG	Wet towel or bag over the head	SUF_WETT
Gagging – forcing object into mouth e.g. tube, bottle, sock	SUF_GAGG	Unknown	SUF_UNKN
Strangling / throttling - strangling with hands, rope or other means	SUF_STRG	Other	SUF_OTHR

Theft and Vandalism

Arson	THF_ARSN	Unknown	THF_UNKN
Theft	THF_THFT	Other	THF_OTHR
Vandalism	THF_VAND		

Threats

Death threats	THR_DETH	Derogatory language and/or insults	THR_INSL
Threats of violence and/or torture	THR_VIOL	Unknown	THR_UNKN
Threats of detention	THR_DETN	Other	THR_OTHR
Pointing / brandishing guns	THR_GUNS		

Appendix 2

Final Set of HRV Categories

As discussed in the section, Chronology of Events, teamwork by the head data processors and researchers resulted in this final set of HRV Categories, types and codes. Data Processing used the types and codes of Appendix 1 in the initial coding, but all those codes were replaced by the appropriate codes defined in this appendix.

HRV Category	Code	Definition
Killing	KILLING	A killing is when a person dies, in one of the three ways: **Assassination** – killing of a *targeted person* by a person or group who *developed a secret plan or plot* to achieve this. Person is targeted because of his political positions. **Execution** - capital punishment (death sentence) imposed and carried out by a legal or authorized body such as a court of law or a tribunal. Victim is aware of death sentence. Perpetrators are the state, homeland governments or security structures of political movements. **Killing** - all other deaths including a killing by a crowd of people.
Attempted Killing	ATT KILLING	This category is the same as that for killing. In attempted killing the victim does not die but there was a clear intent to kill him/her.
Torture	TORTURE	Torture happens in *captivity* or in *custody* of any kind, formal or informal (for example: prisons, police cells, detention camps, containers, private houses or anywhere while tied up or bound to something). Torture is usually to get information or to force the person to do something (for example to admit to a crime or sign a statement). It includes mental or psychological torture (for example: sleep deprivation or telling the person that their family is dead).
Severe Ill-treatment	SEVERE	Severe Ill-treatment covers all forms of inflicted suffering that did NOT happen in custody (for example: injury by a car bomb or beaten up at a rally).
Abduction	ABDUCTION	Abduction is when a person is forcibly and illegally taken away (for example: kidnapping). It does NOT mean detention or arrest. It is not a gross violation of human rights to be arrested (see Associated violations). If the person is never found again, it is disappearance.
Associated Violation	ASSOCIATED	These are not gross violations of human rights but are important for understanding the context of the violation (for example: detention, harassment, framing and violating a corpse after death)
Other violations	OTHER	Violations, which are described but which, do not fit into any of the above categories.
Unknown violations	UNKNOWN	Unspecified violations.

The Violation types

The next step is to decide how the violation took place (for example: was it a beating, an electric shock and so on). The tables below show the HRV categories and their types of violations within each.

1. KILLING	Code	Definition
Beaten to death	BEATING	Person is beaten to death by being **hit, kicked, punched**. State, which part, of the body was assaulted if known. Example: **feet, face, head, genitals, and breast.**
Burnt to death	BURNING	Victim is killed in a fire or burnt to death using **petrol, chemical, fire, scalding, and arson** but does NOT include Necklacing or Petrol bomb. The last two are separate codes.
Killed by poison, drugs or chemical	CHEMICALS	Killed by **poison, drugs** or **household substance** such as bleach or drain cleaner.
Killed by drowning	DROWNING	The person is drowned in a **river, swimming pool** or even in a **bucket** of water.
Killed by electrocution	ELECTRIC	Killed by an electric shock.
Hanged or executed	EXECUTE	**Hanging** or **shooting** as decided by a formal body (court or tribunal) such as the state, homeland state or a political party. It is the consequence of a death sentence.
Killed in an explosion	EXPLOSION	Killed by a manufactured explosive or bomb but NOT a petrol bomb (see below). Explosives include **dynamite, land-mine, limpet mine, car bomb, hand grenade, plastic explosives, detonator, booby trap, letter bomb, parcel bomb, special device (**Example: walkman)
Killed by exposure	EXPOSURE	Person dies after being subjected to extremes such as **heat, cold, weather, exercise, forced labor.**
Killed by multiple causes	MULTIPLE	The person is killed in a variety of ways (use the appropriate definitions from other categories).
Necklacing	NECKLACING	Burnt with **petrol** and **tire.** Necklacing is coded separately from Burning because it featured heavily in the past, so it is useful to distinguish between burning with petrol and a tire and burning in a house, for example.
Petrol bomb	PETROLBOMB	Killed by a **burning bottle of petrol**. Petrol burning falls in between burning and bombing, so, like Necklacing, it is useful to code it separately. It was also called **Molotov cocktail**.
Shot dead	SHOOTING	Person is shot and killed by **live bullet, gunshot, bird shot, buck shot, pellets, and rubber bullet.**
Stabbed to death	STABBING	Killed with a sharp object such as a **knife, panga, axe, scissors, spear (including assegai).**

Suspicious suicide or accident	STAGED	Person dies in **suspicious suicide** or **fatal accident**. This should only be used if it is not clear whether it was really an accident or not, otherwise use the appropriate category and explain in the description that there was a cover-up. Examples: **slipped on soap, jumped out of window, fell down stairs, hanged himself, car accident, booby trapped hand grenades or explosives, shot himself**.
Stoned to death	STONING	Person is killed with **bricks, stones** other **missiles** thrown at them.
Tortured to death	TORTURE	Person is tortured to death.
Killing involving a vehicle	VEHICLE	**Dragged behind, thrown out, driven over, put in boot** but NOT car bomb. (See Bombing). Specify what type of vehicle was involved (for example: **car, train, truck, van, bakkie, hippo, casspir)**.
Other type of killing	OTHER	All other methods of killing including **buried alive, strangling, tear-gas, decapitation, disembowelment**. Make sure that it is clear in the description of the act exactly how they died.
Unknown cause of death	UNKNOWN	Person is dead but there is no further information

2. ATT KILLING	**Code**	**Definition**
Attempted killing by beating	BEATING	Attempt to beat a person to death by being **hit, kicked, punched**. State that part of the body was assaulted if known. Example: **feet, face, head, genitals, and breast.** If an object was used in the beating, specify the object; e.g. Sjambok, baton, gun, rifle, stick, whip, plank, beat against the wall.
Attempted killing by burning	BURNING	Attempt to kill victim in a fire or by using petrol**, chemical, fire, scalding, and arson** but does NOT include Necklacing or Petrol Bomb. The last two are separate codes.
Attempted killing by poisoning, drugs or chemical	CHEMICALS	Attempt to kill person by use of **poison, drugs** or **household substance** such as bleach or drain cleaner.
Attempted killing by drowning	DROWNING	Attempt to kill the person by drowning in a **river, swimming pool** or even in a **bucket** of water.
Attempted killing by electrocution	ELECTRIC	Attempt to kill by an electric shock.
Attempted killing by execution	EXECUTE	Attempt to **kill by hanging** or **shooting** as decided by a formal body (court or tribunal) such as the state, homeland state or a political party. It is the consequence of a death sentence.
Attempted killing in an explosion	EXPLOSION	Attempt to kill a person by a manufactured explosive or bomb but NOT a petrol bomb (see below). Explosives include **dynamite, land-mine, limpet mine, car bomb, hand grenade, plastic explosives, detonator, booby trap, letter bomb, parcel bomb, special device (**Example: walkman)

Attempted killing by exposure	EXPOSURE	Attempt to kill person by subjecting him/her to extremes such as **heat, cold, weather, exercise, and forced labor.**
Attempted killing by multiple causes	MULTIPLE	Attempt to kill the person in a variety of ways (use the appropriate definitions from other categories).
Attempted killing by Necklacing	NECKLACING	Attempt to kill by burning with **petrol** and **tire.** Necklacing is coded separately from Burning because it featured heavily in the past, so it is useful to distinguish between burning with petrol and a tire and burning in a house, for example.
Attempted killing by petrol bomb	PETROLBOMB	Attempt killing by a **burning bottle of petrol**. Petrol burning falls in between burning and bombing, so, like Necklacing, it is useful to code it separately. It was also called **Molotov cocktail**.
Attempted killing by shooting	SHOOTING	Person is shot and injured by **live bullet, gunshot, bird shot, buck shot, pellets, rubber bullet,** or possibly shot at close range or with deliberate intent to kill but **not injured.**
Attempted killing by stabbing	STABBING	Attempted killing with a sharp object such as a **knife, panga, axe, scissors, and spear (including assegai).**
Attempted killing by suspicious suicide or accident	STAGED	Attempt to kill a person by staging a **suspicious suicide** or **fatal accident**. This should only be used if it is not clear whether it was really an accident or not, otherwise use the appropriate category and explain in the description that there was a cover-up. Examples: **slipped on soap, jumped out of window, fell down stairs, hanged himself, car accident, booby trapped hand grenades or explosives, shot himself**.
Attempted killing by stoning	STONING	Attempt to kill a person by throwing **bricks, stones** or other **missiles** at them.
Attempted killing by torturing	TORTURE	Attempt made to kill a person by torturing to death.
Attempted killing involving a vehicle	VEHICLE	**Dragged behind, thrown out, driven over, put in boot** but NOT car bomb. (See Bombing). Specify what type of vehicle was involved (for example: **car, train, truck, van, bakkie, hippo, casspir)**.
Other type of attempted killing	OTHER	All other methods of attempted killing including **buried alive, strangling, tear-gas, decapitation, disembowelment**. Make sure that it is clear in the description of the act exactly how they died.
Attempted killing by unknown cause	UNKNOWN	Unspecified attempt to kill a person.

3. TORTURE	**Code**	**Description**
Torture by beating	BEATING	Person is tortured by being beaten severely or for a long time (example: **hit, kick, and punch**). State which part of the body was assaulted e.g. **feet, face, head, genitals, breast**). If an object was used in the beating, specify

		the object (example: **Sjambok, baton, gun, rifle, stick, rope, whip, and plank, beat against the wall**). Specify if victim is **pregnant** or **miscarried.**
Torture by burning	BURNING	Person is burnt with **cigarettes** or **fire,** for example.
Torture with poison	CHEMICALS	Tortured with **poison, drugs** or **household substance** such as bleach or drain cleaner.
Torture by deprivation	DEPRIVE	Person is tortured by withholding essentials, such as **sleep, food**, or **medical attention with serious injury** or **need** (this does NOT refer to general lack of medical care while in custody). See Associated violations).
Electric shock torture	ELECTRIC	Electric shock to the body. Specify which body part was shocked (for example: **genitals, breasts, fingers, toes, ears**, etc).
Torture by exposure to extremes	EXPOSURE	Person is tortured by subjecting them to extremes such as **heat, cold, weather, exercise**, **labor, noise, darkness, light** (including flashing lights, blinding by light), **blindfolding, and confinement to small space, smells, and immobilization.**
Psychological or mental torture	MENTAL	Person is tortured psychologically, mentally or emotionally for example: by **simulated execution** (includes Russian roulette,) **solitary confinement**, **degradation** (includes use of excrement, urine, spit), **insults, disinformation** (telling a person that a loved one is dead), **threats, witnessing torture, forced participation in torture, exposure when washing or on toilet, threat of torture**.
Torture by bodily mutilation	MUTILATE	Torture involving injuries to the body where parts of the body are partly or wholly cut, severed or broken. Specify body part, for example: **genitals, ears, fingernails, hair, etc.** It includes **amputation of the body parts, breaking of bones, pulling out nails, hair or teeth, scalping.**
Torture by forced posture	POSTURE	Person is tortured by forcing the body into painful positions, for example: **suspension, helicopter, tied up, handcuffed, stretching of body parts, prolonged standing, standing on bricks, uncomfortable position** (including squatting, imaginary chair, standing on one leg, pebbles in shoes), **forced exercise, forced labor, blindfolding and gagging.**
Torture by sexual assault or abuse	SEXUAL	Person is torture by attacking them using their gender or genitals as a weak point. This does NOT include electric shock, mutilation or beating (instead, use those categories and specify genitals as the body part abused). It includes: **slamming genital or breast in a drawer or other device, suspension of weights on genitals, squeezing genitals or breasts, rape by opposite sex, rape by same sex, gang rape, forced sexual acts (e.g. oral sex, simulating intercourse), introduction of objects into the vagina or rectum, sexual abuse using animals, threats of rape, touching, nakedness, sexual comments or insults, sexual enticement, deprivation of sanitary facilities for menstruation.**
Torture by suffocation	SUFFOCATE	Torture by stopping someone from breathing, for example by: **bag, towel, tube over head (wet or dry), drowning (head, whole body submerged), choke, strangle, stifle, throttle, teargas, bury alive.**
Other type of torture	OTHER	All other methods of torture. Make sure that it is clear in the description of the act exactly how the person was tortured. It includes **use of animals** (specify animal e.g. snake, tortoise, baboon), **use of vehicle.**
Unknown type of torture	UNKNOWN	Person is tortured but the method is not known.

4. SEVERE ILL-TREATMENT	**Code**	**Definition**
Burning of a building	ARSON	Burning of a home, building or other fixed property; indicates substantial **destruction** of **property**.
Severely beaten	BEATING	Person is badly beaten, or beaten for a long period of time. They may be **hit, kicked, punched, twisted**. State which part of the body was assaulted if known. Example: **feet, face, head, genitals, and breast.** If the person was beaten with an object, specify object (for example: **sjambok, baton, gun, rifle, stick, rope, whip, plank, wall)**. Specify if victim is **pregnant** or **miscarried.**
Injured by burning	BURNING	Person is injured by burning with **fire, petrol, chemical, scalding** but does NOT include necklacing or Petrol Bomb. The last two are separate codes.
Injured by poison, drugs or chemical	CHEMICALS	Person was poisoned or injured by **poison, drugs** or **household substance** such as bleach or drain cleaner.
Deprivation	DEPRIVE	This usually relates to treatment while incarcerated and would include deprivation of **food, medical treatment, sleep, and clothing.**
Injured in an explosion	EXPLOSION	Person is injured by a bomb or explosives but NOT petrol bomb (this is coded separately). See below).Explosives include **dynamite, land-mine, limpet mine, car bomb, hand grenade, plastic explosives, detonator, booby trap, letter bomb, parcel bomb, special device** (e.g. booby trapped Walkman)
Incarceration, imprisonment or detention	INCARCERATE	Person is **detained,** possibly under the **State of Emergency** legislation, or held as an **awaiting-trial** prisoner for an unusually long period or **imprisoned** as a sentenced prisoner.
Psychological or mental ill-treatment	MENTAL	Person is severely psychologically, mentally or emotionally ill treated, for example by **simulated execution** (includes Russian roulette), **degradation** (includes use of excrement, urine, spit), **death threats, threat of torture.**
Bodily mutilation	MUTILATE	Person is injured by having parts of their body mutilated or damaged. Specify body part, for example **genitals, fingernails, ears, hair**, etc.
Necklacing	NECKLACING	The person is injured in an attempted necklacing.
Severely injured by a petrol bomb.	PETROLBOMB	Person is injured in an attempted petrol bombing.
Sexually assaulted or abused	SEXUAL	All forms of attack on a person using their gender or genitals as a weak point, for example: **rape by opposite sex**, **rape by same sex, gang rape, forced sexual acts (e.g. oral sex, simulating intercourse), introduction of objects or substances into vagina or rectum, sexual abuse using animals**.
Injured in a shooting	SHOOTING	Person is injured by being shot by **live bullets, gunshot, birdshot, buckshot, pellets, rubber bullet.** Specify body part injured, if known.
Stabbed or hacked with a sharp object	STABBING	Injured with a sharp object such as a **knife, panga, axe, scissors, spear (including assegai).**

Injured in stoning	STONING	Person is injured with **bricks, stones** other **missiles** thrown at them.
Tear-gassed	TEARGAS	Severe injury caused by tear-gassing in a confined space (for example; teargas in a prison, van or packed hall).
Injury involving a vehicle	VEHICLE	Injury caused by being **dragged behind, thrown out, driven over, put in boot** of a vehicle. Specify what type of vehicle was involved (for example: **car, train, truck, van, bakkie, hippo, casspir)**.
Suffocated	SUFFOCATE	Injury or ill treatment by stopping someone from breathing, for example by **drowning (head, whole body submerged), choke, stifle, throttle, teargas, bury alive.**
Other type of ill-treatment	OTHER	All other methods of ill treatment. Make sure that it is clear in the description of the act exactly how they ill-treated.
Unknown type of ill-treatment	UNKNOWN	Person is ill-treated but the method is not known.

5. ABDUCTION	**Code**	**Definition**
Illegal and forcible abduction	ABDUCTION	Victim is forcibly and illegally taken away (for example, kidnapping), but the person is found again, returned or released. It does NOT mean detention or arrest. It is not a gross violation of human rights to be arrested (see Associated Violation).
Disappearance	DISAPPEAR	Victim is forcibly and illegally taken away and is never seen again. It does NOT include cases where somebody goes into exile and never returns. It must be done by force. This DOES include people who have disappeared but it is not clear why they have gone (instead of abduction, they might have just run away or were shot and buried). In this case, a finding will be made and the code will be left as it is or changed to Killing if the person was killed or changed to be out of the mandate of the TRC.

6. ASSOCIATED VIOLATIONS	**Code**	**Definition**
Beating	BEATING	Person is beaten, but it is not a severe or prolonged beating. It includes **a once-off mild beating**. Specify if **in custody** or if victim is **pregnant** or **miscarried.**
Violation after death	CORPSE	Body of victim was violated after death, for example by: **improper burial, body mutilated or burnt or blown up, funeral restrictions, funeral disruption, anonymous burial, mass grave**.
Deprivation	DEPRIVE	Deprivation of facilities or essentials, for example: **medical attention, food, water, sanitary facilities, and privacy.**
Destruction of property	DESTROY	Includes violations such as **arson, destruction, vandalism, theft, forced removal and eviction.**
Person disappeared and	DISAPPEAR	This is for unresolved disappearance (not abductions and not killings).

has not been seen since		The person may have disappeared while intending to go into **exile**, or while in exile from a liberation movement **camp**, or while as a **combatant** in an operation within the country.
Financial impropriety	FINANCIAL	Person was subjected to **bribery, extortion, pay-off, ransom, blackmail and ruin of business**.
Framing	FRAMING	Person is labeled as an **informer (impimpi), collaborator or criminal**. Also **false information** is spread about the person or a **smear campaign** against the person is started.
Incarceration or imprisonment	INCARCERATE	Includes **police custody, detention, house arrest, restrictions, banning, banishment, prison, and informal prison**.
Intimidate or harassment	INTIMIDATE	Victim is intimidated or harassed by **dismissal from work, threats, animals killed, visits, telephone calls, surveillance, boycott enforcement, pointing of firearms** (NOT is custody) and **threat of violence**. It does NOT include vandalism or arson. This comes under Destruction of Property.
Sexual harassment	SEXUAL	Person is sexually harassed. It includes **threats of rape, touching, nakedness, sexual comments or insults, sexual enticement, deprivation of sanitary facilities for menstruation.**
Professional misconduct	PROFESS	Person was subjected to professional misconduct by one of the following: **Doctors** (district surgeon, private doctor) who neglect or ignore injuries, collaborate in torture or conceal the cause of death or injuries. **Judiciary** (magistrates, judges, etc.) who ignore torture allegations, for example. **Lawyers** who neglect the case, ignore or tamper with evidence, misappropriation of funds or failure to hand over damages. **Businesses** who collaborate with perpetrators.
Tear-gassed	TEARGAS	Victim was tear-gassed but NOT while in custody (see Torture).
Theft or stealing	THEFT	Money or possessions were stolen from the victim.
Other type of associated violation	OTHER	All other types of associated violations, including **released into hostile environment, released into unknown place, left for dead, rough ride, detention of family or loved ones**. Give full details in the description of the violation.

Unknown type of violation	UNKNOWN	Not clear from the statement what type of associated violation the person suffered.

7. OTHER	**Code**	**Definition**
Other type of violation	OTHER	Other violations are specified by the victim, which do not fall into any of the above classifications.

8. UNKNOWN	**Code**	**Definition**
Unknown type of violation	UNKNOWN	Not clear from the statement what type of violation the person suffered.

Appendix 3

Examples of Coded Killings

This appendix shows how the different killing methods are coded for individual acts of killings.

KILLING	BEATING
KILLING	BURNING
KILLING	CHEMICALS
KILLING	DROWNING
KILLING	ELECTRIC
KILLING	EXECUTE
KILLING	EXPLOSION
KILLING	EXPOSURE
KILLING	NECKLACING
KILLING	PETROL BOMB
KILLING	SHOOTING
KILLING	STABBING
KILLING	STAGED
KILLING	STONING
KILLING	TORTURE
KILLING	VEHICLE
KILLING	OTHER
KILLING	UNKNOWN

Appendix 4

Acts of Violence

Specific acts of violence and their synonyms where relevant are shown in this appendix. As described in the section, Chronology of Events, these acts are the result of three months discussion and brainstorming.

Asphyxiation (Synonym-Choke)

Strangling (Synonym-Throttle)
Suffocation (death)
Bag overhead
Wet towel over head
Tear-gassing
Buried alive
Drowning
Submerge in water
Gagging

Assault (Synonym-Strike with an object)

Batoning
Beat (Synonyms-hit/batter) with a sharp object concealed in a cloth
Hacking
Sjambok
Stab (Synonym-cut/wound/gore)
Stoning
Cane
Flog (Synonym-whip/thrash/lash)
Beating with a rifle
Pistol whipping

Assault on Specific Parts of the Body

Beating on the soles of the feet
Beating pregnant women on the stomach
Clapping (Synonyms-whack/bang) on ears with both hands
Kaffir Klap (cheek)
Banging the head against a wall
Scalping (removal of hair from scalp with knife)
Removal of nails

Beating

Slapping (Synonyms-spank/thump/bump/strike/knock)
Kicking (Synonyms-boot/stomp)
Punching
Breaking (Synonyms-fracturing/crack/shattering/snapping) of bones

Assault Using vehicles

Dragged (Synonym-pull) behind a vehicle
Attached (Synonym-fastened) onto a moving vehicle
Thrown (Synonym-chuck) out of moving trains / taxis
Driven over
Rough ride
Put in boot

Abduction (Synonym-Kidnapping/ Apprehend/ Capture/Seize/Catch)

Disappearance

Bombing (Explode)

Land mine

Grenade
Mortar / shell
Hand grenade
Explosive / bomb
Booby trap bombs
Letter bombs
Car bomb

Burns (Synonym-scorch)

Chemicals
Cigarettes
Boiling water
Live fire / burning sticks
Necklacing
Arson

Deliberate (Synonyms-Premeditated/Planned spreading of disease) Psychological -Torture (Synonym/Torment/Pain/Anguish /Suffering/Agony/Tribulation/and Ill-treatment) - excludes Threats.

Verbal abuse (Synonym-Mistreatment/Indignity/Violation/Insult/Offence/
Malign/Denounce/Defame/Misuse/Deceive/Subvert/Mishandle/Betray/
Unjust/Crime/Condemnation/Censure/Defamation)
Simulated execution
False and alarming information / disinformation
Detention of children and family members to extract information
Russian Roulette (Gun against the head with one bullet left)
Suspension (Synonyms-Hang/Dangle) from a great height/moving vehicle
Members of family forced to watch or participate in torture
Solitary confinement
Surveillance (Synonym-Watch)
Threatening acts e.g. brandishing guns
Dismissal from employment as a result of political affiliation
Harassment

Threats (Synonyms-Coercion/Intimidation/Warning)

Against the targeted person
Against a family member of the targeted person
Against a colleague or work associate of the targeted person
Against a friend of the targeted person
Against someone working on behalf of the targeted person e.g. lawyer, human rights worker
Threats against children
Verbal threats

Deprivation (Synonym-Loss)

Deprivation of medical attention, treatment
Deprivation of food and/or water
Deprivation of sleep
Deprivation of sanitary facilities
Denial of privacy
Overcrowding (Synonyms-Packed/Strafed/Crammed/Filled)
Placed in isolation
(Synonyms-Seclusion/Solitude/Isolation/Aloneness/Separation
Confinement (Synonyms-Detention/Incarceration) in a small space

Degradation (Synonym-Shame/Embarrassment/Abasement/Humiliation)

Deprivation (Synonym-Loss of personal hygiene)
Denial (Synonyms-Refusal/Reject) of toilet facilities
Nakedness
Abuse with excrement

Denial of privacy
Derogatory (Synonym-Disparaging/Rude) language

Destruction (Synonyms-Damage/Ruin/Vandalize/Smash/Devastate/Wreck/Raze) of property

Destruction of homes/offices/schools/buildings/vehicles/personal property/arson

Extortion (Synonyms-Blackmail/Coercion/Ransom/Bribe/Pay-off

Theft (Synonyms-Pillage/Plunder/Rob/Root)

Poisoning (Synonyms-Contaminate/Pollute/Infect)

Poisoning of food
Poisoning of clothing
Intravenous poisoning

Murder (Synonyms-Liquidation/Permanent removal/Annihilation/Carnage/Manslaughter/Slay/Homicide

Assassination
Extra-judicial/illegal unlawful execution
Hanging
Electrocution
Ritual murder
Witchcraft
Use of animals

Sexual Molestation (Synonyms-Mistreatment/Violation/Abuse) and Rape

Forced performance of sexual acts other than rape
Introduction of objects into the rectum/vagina
Rape by someone of the opposite sex
Rape by someone of the same sex
Gang rape
Physical assault and touching
Body searching by members of the opposite sex
Pumping water into the uterus
Abuse with body fluids
Abuse with animals
Assault on genitals
Suspension of weights from the testicles

Imprisonment (Synonyms-Detention/Locking up/Confinement/Captivity/ Arrest/Incarceration)

Banning
Banishment
House arrest

Forced (Synonyms-Bound/Compelled/Obliged/Postures) position -

Physical Stress (Duress/Pressure/Force/Strain)

Suspension: hanging the victim by arms, legs, etc.
Forced exercise
Excessive exercise
Forced stationary posture - standing, kneeling, sitting, standing on two bricks
Forced labour
Stretching of limbs and trunk
Helicopter? - hanging the victim from the stick between knees and arms bound tightly together
Stopping of blood flow
Forced carrying of heavy weights
Buried alive

Stress to the Senses

Loud noises or music
Screams and voices

Powerful lights
Blindfolding
Exposure to extreme heat or cold
Bound or tied up
Complete immobilization
Overcrowding
Confined to small space
Bad smells

Staged accidents / suicide

Forced jumping or being thrown from heights
Car sabotage

Use of drugs

to effect psychological damage
to effect physical damage

Torture as a witness

Victim is forced to watch or listen to the torture of others
Victim is forced to participate in the torture / assault of others

Electric Shock

Electric shock to the genitals
Electric shock body - toes and fingers, etc

Shooting

Random shooting
Rubber bullets
Live ammunition
Birdshot
Buckshot

Capital Punishment

Post Mortem - Violation after death

Mutilation
Decapitation
Disembowelment
Improper burial - burial in a shallow grave
Blowing up bodies or body parts
Burn or braai a body
Removal of body parts

Appendix 5

Completed Data Processors' Coder's Sheet

As described in the section, Coding Using the Data Processors' Coder's Sheet, the data processors coded acts of violence in chronological order.

Act No.	Person ID	Last Name	First Names	Day	Month	Year
1	10278	Molokoane	Barney	30	11	85
2	10278	Molokoane	Barney	30	11	85

continues in the next block

Town	Description of place	Description of violation	Outcome & consequences
Piet Retief	Houtkop; next to the Swaziland border on the road	Shot in an ambush	Death
Piet Retief	At the township cemetery in Piet Retief	Improper burial	

continues in the next block

HRV Category	HRV Type	Reason for Violation
KILLING	**SHOOTING**	Victim - with his two other comrades - were from a MK mission in South Africa.
ASSOCIATED	**CORPSE**	*(repeat the above reason)*

continues in the next block

Political context	Age	Org No.	Organization name
It was at the time when MK cadres infiltrated the country on sabotage and other missions.	30	22	Umkhonto weSizwe
(*repeat the above political context)*	30	22	Umkhonto weSizwe

continues in the next block

Event Number	Event Name
	(both fields used by researchers for their research coding)

Appendix 6

A TRC Statement

Below is the full complete statement made by an HRV victim.[2]

The aim of a Gross Violation of Human Rights Statement is to try and gather as much information as possible about the gross violations of human rights suffered by South Africans between 1 March 1960 and 5 May 1994. The questions that form the basis of the STATEMENT are designed to make explicit the *circumstances* (broader context), the *nature* (type) and the *consequences* of the violations.

What are "gross human rights violations"?

These are serious human rights violation like the killing of people, the kidnapping of people, torture, or the severe ill treatment of people.

Who are victims of gross human rights violations?

Victims of gross human rights violations are people who are killed, abducted, tortured or severely ill-treated; and family members or dependants of a person who was killed or who disappeared.

What happens to your statement?

Your statement will be recorded on the computer and you will be given a reference number (JB04500/01GTSOW). The HRC Committee will carefully consider your statement. You might be asked to come to a public hearing to talk about your case. The Committee will then decide if you qualify as a victim in terms of the law that set up the Truth and Reconciliation Commission. It will send you a letter telling you whether or not you qualify.

If the Committee on Human Rights Violations finds that you are a victim, it will include your case in the report it sends to the Committee on Reparation and Rehabilitation. The Committee on Reparation and Rehabilitation will look at all the cases sent to it.

Ms. Dudu Chili voluntarily gave the following statement to me and can be contacted at 27 11 331-3719 (W) and 27 11 462-7240 (H).

Ms. Dudu Chili's statement

I, Dudu Chili, declare under oath in English that I am a female aged 54 years, ID number 411028 0191 084, and residing at number 7556 Maseko Street, Orlando West, P.O. Orlando 1804, Soweto in the district of Johannesburg.

I wish to state that on the 28th February 1989 my house, at Orlando West in Soweto, was bombed by the Mandela United Football Club (MUFC) and that I lost everything in it. My family and I were left with what we were wearing.

I lost my niece – Finkie Msomi - who was thirteen years old. Finki, who was in my bedroom, was shot in the head with an AK47 and died on the spot. Thereafter petrol bombs were hurled into my house and it was burnt down. My cousin Barbara Chili was also burnt while trying to save Finki from the fire. Barbara suffered third degrees burns on her waist. Finki's sister, Ntombenhle Msomi, was slightly burnt on the foot.

Sometime in 1986, Winnie Madikizela-Mandela formed the Mandela United Football Club. She demanded that all the youth in our area, Orlando West, should join her club. Those who refused were labeled sell-outs and hunted down to be killed. Since my son, Sibusiso Chili, refused membership of the club, he became a target and I tried to intervene to protect my son. I approached my cousin, Matilda Dlamini, to plead with the MUFC to spare my son's life. Matilda, a long-standing best friend of Winnie Madikizela-Mandela, temporarily succeeded. Matilda was married to Mosetlha. Mosetlha's daughter was married to President Mandela's son, Makgatho.

Two years later, in 1989, the hunting down of Sibusiso started again. A former member of the MUFC, Lerotodi Ikaneng, had deserted the club. No one was allowed to leave the team. Lerotodi was later caught and had his throat cut with garden shears by Jerry Richardson – the former MUFC coach. Lerotodi survived. Some months later after this incident, Lerotodi pointed out one of his

---

[2] In some cases of multiple similar entities (e.g., perpetrators, witnesses), where it does not affect understanding, we have omitted one or more entities.

assailants to Sibusiso. Lerotodi informed Sibusiso that this man had held him (Lerotodi) down while Jerry Richardson cut his throat. Sibusiso then suggested to Lerotodi that they approach this man and asked him to accompany them to my house to explain why they tried to kill Lerotodi. The man agreed. At that time, I was highly involved with the youth in Orlando West. Since it was late at night, I promised to attend to the matter the following day and asked this man to spend the night at my place. He agreed and slept with Sibusiso and the other boys. That night, I phoned Mrs. Sisulu to come and help to solve this problem. Mrs. Sisulu agreed and contacted the other leaders in the area.

The next day I phoned a Mr. Ndo, who was the co-president with Mrs. Sisulu, to attend the meeting. I also phoned a Mr. Steward Ngwenya who was a member of the Soweto Civic and he promised to attend.

Whilst waiting for the above civic leaders to come, the young assailant requested to go home to wash and changed into fresh clothing. He came back and was questioned on the motive to kill Lerotodi and on the harassment of other youths that were not affiliated to the MUFC. He was also asked why he was not attending school. The young man regretted his acts in the attempted murder of Lerotodi and left.

I heard that some youths that were members of the MUFC reported to Winnie Madikizela-Mandela that they saw this young man in the company of Sibusiso at my house. Lerotodi's assailant was summoned to appear before Winnie Madikizela-Mandela and her daughter, Zinzi Mandela, to explain his visit to my house. In that meeting a decision was taken to eliminate i.e. to kill Lerotodi and Sibusiso because they have become "too problematic". Some MUFC members were mandated to "carry out the order". The late Maxwell Madondo and the self-exiled Katiza Cebekulu were part of the group entrusted with the task to kill Sibusiso and Lerotodi. Katiza Cebekulu was also asked to point out Sibusiso to the other members because they did not know him.

Immediately after the meeting, Dodo, a member of the MUFC club, rushed to both Lerotodi's place and my house to warn us of the impending attack. On hearing this, I immediately called Alfred Msomi – Finki's father – who lived, at the back house opposite to mine. Dodo immediately left the township fearing for his life for alerting both the Lerotodi's and I about the decision to kill our sons.

The following day I was surprised to see my house being strategically guarded by people wearing scarves and balaclavas. I informed Finki that these people were armed and apparently their mission was to attack the house and kill Sibusiso. Sibusiso and his brothers had all gone into hiding after being alerted by Dodo. This guarding of the house continued for several days – 24 hours a day. These MUFC members apparently were not aware that we already knew of the attack.

I wish to point out that when the hunting down of both boys started, I had just arrived from London. I had gone there to attend an anti-apartheid movement conference at Sherfield. There was a concern shown by Winnie Madikizela-Mandela on my trip. I heard that she thought I had gone to London to report her about the Stompie Sepei affair to the ANC leadership and other anti-apartheid movements (i.e. the UDF, the Civic, the youth and the church leaders). Stompie Sepei – a young activist from the Free State - had been kidnapped and killed the previous December in 1988. Stompie and three other youth – Kenny Kgase, Gabriel Megoe and another – had been kidnapped from the Methodist manse under Rev. Paul Verryn and taken to Winnie Madikizela-Mandela's house in Diepkloof. The remaining youth at the manse reported the matter to me since Rev. Paul Verryn was on holiday. I was the first person to hear of the kidnapping. This trip annoyed Winnie Madikizela-Mandela and I also became her target.

During the change of guards, my sons would sneak home to wash, change clothing and rush back to their different hideouts. We too, had our spies watching the changing of shifts and would immediately notify Sibusiso and others. One day Sisusiso was on his way home when he met three of the MUFC members and a fight ensued. Immediately the word went out in the township that some MUFC members had caught up with Sibusiso. The township youth ran to Sibusiso's rescue. One the three MUFC members, Maxwell Madondo, was clubbed and stoned to death. The other two escaped and reported the killing of Maxwell to Winnie Madikizela-Mandela. Dempsey of the South African police arrested me. First Dempsey said they were going to question me about my trip to London. Dempsey wanted to know which ANC members did I meet and talked to. When they could not extract this information from me, I was charged with the murder of Maxwell Madondo. When Maxwell was killed, I was in my house. I was detained for a week and my letters were confiscated. My house was bombed the same day I was arrested. The following day after my arrest,

Dempsey took me home in a police car. On our way, I read a poster stating in bold “Thirteen-year old girl dies”. It never occurred to me that this girl was my niece, Finki. On arrival at my place, I found my house destroyed by fire. Everything was completely gutted. All our belongings – furniture, clothing, etc. – were burnt. Nothing was left except for the clothes we were wearing.

The police did not allow my neighbors to speak to me. My sisters informed me that my boys were safe but that my niece Finki had died and that my sister Barbara had burnt her foot and was in hospital. She hurt herself while trying to drag the body of Finki from the fire. I was taken to Kliptown police station. During the court proceeding I was informed by the prosecutor that the charge against me was withdrawn.

In conclusion I wish to state that Winnie Madikizela-Mandela was behind all the unfortunate happenings both in Orlando West and at my home. She was in charge of the MUFC and the members of this club took orders from her. She controlled the issuing of guns and ammunition. One of the MUFC members – Charles “Bobo” Zwane - is serving a life sentence. Most of the MUFC members refused to implicate her since they feared for their lives.

Processing a TRC statement:

Extract all personal details from the statement for capturing and check to see whether these people are already on the database. We should avoid registering people more than once. We must also check that they are accurately captured.

DEPONENT / VICTIM

Reference Number	JB04500/01GTSOW
Person ID number	3
Surname	CHILI
First Names	Dudu Olive
Aliases / Nicknames	
ID/Passport number:	411028 0191 084
Date of birth:	28-10-1942
Home language:	isiZulu
Sex:	Female
Citizenship:	South African
Race	African
Occupation	Logistics Officer
Employed:	Yes
Street Address:	7556 Maseko Street, Orlando West, Soweto, Gauteng.
Postal Address	P.O. Box 925, Johannesburg, 2000 Gauteng
Home phone:	(011) 936-7278
Work phone:	(011) 333-6330
Contact name:	
Contact Address:	
Prison:	
Contact phone:	
Prison number:	

VICTIMS/WITNESSES:[3]

	1.	**2.**	**3.**	**4.**
Reference Number:	JB04500/01GTSOW	JB04500/01GTSOW	JB04500/01GTSOW	JB04500/01GTSOW
Person ID number:	56557	56562	56566	56567
Surname:	CHILI	MSOMI	MSOMI	MSOMI
First Names:	Barbara	Ntombenhle Mantanta	Ntukayaboni Alfred	Ntombana
Aliases / Nicknames:				
ID/Passport number:				
Date of birth:				
Home language:	isiZulu	isiZulu	isiZulu	isiZulu
Sex:	Female	Female	Male	Female
Citizenship:	South African	South African	South African	South African
Race:	African	African	African	African
Occupation:				
Employed:				
Street Address:	7556 Maseko Street,	7556 Maseko Street,	7556 Maseko Street,	7556 Maseko Street,
	Orlando West,	Orlando West,	Orlando West,	Orlando West,
	Soweto,	Soweto	Soweto,	Soweto,
	Gauteng.	Gauteng	Gauteng.	Gauteng.
Postal Address:	Gauteng	Gauteng	Gauteng	Gauteng.
Home phone:	(011) 936-7278	(011) 936-7278	(011) 936-7278	(011) 936-7278
Work phone:				
Contact name:				
Contact Address:				
Contact phone:				
Prison:				
Prison number:				

[3] Number 5 omitted for space reasons.

VICTIM

	1.
Reference Number	JB04500/01GTSOW
Person ID number	56551
Surname	MSOMI (DECEASED)
First Names	Finkie Maria
Aliases / Nicknames	
ID/Passport number:	
Date of birth:	
Home language:	isiZulu
Sex:	Female
Citizenship:	South African
Race	African
Occupation:	
Employed:	
Street Address:	7556 Maseko Street,
	Orlando West,
	Soweto,
	Gauteng.
Postal Address	Gauteng
Home phone:	(011) 936-7278
Work phone:	
Contact name:	
Contact Address:	
Contact phone:	
Prison:	
Prison number:	

WITNESSES:

	1.	**2.**	**3.**
Reference Number:	JB04500/01GTSOW	JB04500/01GTSOW	B04500/01GTSOW
Person ID number:	78516	60611	9530
Surname:	KHUMALO	MASINGA	VERRYN
First Names:	Sifiso	Benjamin	Reverend Paul
Aliases / Nicknames:		Slash	
ID/Passport number:			
Date of birth:			
Home language:	isiZulu	isiZulu	English
Sex:	Male	Male	Male
Citizenship:	South African	South African	South African
Race:	African	African	White
Occupation:			
Employed:			
Street Address:	Soweto,	Soweto,	Orlando West,
	Gauteng.	Gauteng.	Soweto,
Postal Address	Gauteng	Gauteng.	Gauteng.
Home phone:			Gauteng
Work phone:			
Contact name:			
Contact Address:			
Contact phone:			
Prison:			
Prison number:			

PERPETRATORS[4]

	1.	2.	3.	4.
Reference Number:	JB04500/01GTSOW	JB04500/01GTSOW	JB04500/01GTSOW	JB04500/01GTSOW
Person ID number:	56550	29508	4759	4760
Surname:	DEMPSEY	VAN ZYL	MADIKIZELA-MANDELA	MANDELA
First Names:	Abram	Andries Rossouw	Winnie	Zinzi
Aliases / Nicknames:		Slang		
ID/Passport number:				
Date of birth:				
Home language:	English	Afrikaans	isiXhosa	isiXhosa
Sex:	Male	Male	Female	Female
Citizenship:	South African	South African	South African	South African
Race:	White	White	African	African
Occupation:	Officer	Officer		
Employed:	Yes	Yes		
Street Address:	Protea/Norwood Police Station	Protea/Norwood Police Stations	Orlando West	Orlando West
	Soweto/Johannesburg,	Soweto/Johannesburg,	Soweto,	Soweto,
	Gauteng	Gauteng.	Gauteng.	Gauteng.
Postal Address	Gauteng	Gauteng.	Gauteng	Gauteng.
Home phone:				
Work phone:				
Contact name:				
Contact Address:				
Contact phone:				
Prison:				
Prison number:				

[4] Perpetrators numbers 5, 6, and 7 omitted to save space.

Then capture the following:

Full name of statement taker
Date of interview
Place of interview
Language of interview
TRC office; JHB, DBN, CT or EL
Victim / Deponent prepared to attend a hearing
Name of data processor
Date registered in the computer
Date captured in the computer
Documentation status: R = registered, P = processed and C = captured.
Total number of pages of the statement.
Notes – any observations which we may have made.

Katiza Cebekhulu, a former Mandela United Football Club member who is now in London, is alleged to have left the country before the **Winnie Madikizela-Mandela** trial in 1991, in which he was a co-accused in the **Stompie Sepei** trial.

Maxwell Madondo, a cook at the Winnie Madikizela-Mandela house and a member of the Mandela Football Club, was killed when **Sibusiso Chili** dropped a rock on his head in February 1992.

Chili's defense was that he has acted in self-defense and that Madondo was part of a hit-squad of three Football Club members who had instructions to kill him. In court two of the three were named as Madondo and **"Killer"**. The third was not named.

However, a British Broadcasting Corporation (BBC) program later named the third person as Cebekhulu and interviewed him. He said that at a meeting at the offices of Winnie Mandela, it had been decided that Sibusiso Chili and another Football Club member, **Lerotodi Ikaneng**, should be killed.

The hit-squad was to have killed five youths who were accused of selling out to the police, but instead Madondo was killed and six youth stood trial. Police later found the hit-list with five names at the home of Winnie Madikizela-Mandela, where the Football Club members were living.

According to the BBC, "the most extraordinary development came near the end of the trial – an incident that surprisingly went unreported by the South African media. The defense and prosecution advocates stepped outside the courtroom to confer. The defense said they would call as witness the third unnamed youth who had been with Madondo just before he was killed." The BBC said they had learnt that this youth was Katiza Cebekhulu and he had made a statement for the defense confirming there had been a meeting in Winnie Madikizel-Mandela's office in Orlando West, at which it had been decided that Chili and Ikaneng would be killed.

He told the lawyers that the meeting had been chaired by Winnie Madikizel-Mandela and that Zinzi Mandela and Jerry Richardson were present. He named others who were there.

After conferring with the defense, the State read the following statement into the court record:

"The admission the State will make is that the deceased Maxwell Madondo was a member of the Mandela Football Club and that a decision was made by Mrs. Mandela and the football club to kill accused no. 1 (Ikaneng) and no. 6 (Chili). But the witness, m'lord, whose name I will not mention

now, together with "Killer" and the deceased, was instructed and went out to kill accused no. 1 and no. 6. That the person known as "Killer" was in possession of a firearm was to carry out the mandated decision."

Chili was the only one found guilty and he was sentenced to one year's jail. However, Jerry Richardson, who was sentenced to death – later commuted to life – has made statement to the TRC and he will be able to confirm or deny Katiza's allegations.

Chili's mother, Dudu Chili, told the BBC that she had worked with Albertina Sisulu to assist boys to escape from Winnie Madikizela-Mandela's home. She said she had been warned that a decision had been taken to kill her son and she had warned him. Dudu Chili was one of seven originally charged with Madondo's murder. She was released on bail on condition that she stayed away from Soweto for her own safety, and was discharged before the trial began. Her house was, however, burnt down allegedly by Football Club members and her 11-year-old niece was shot and burnt to death.

The summaries - read the entire case through highlighting the names of people mentioned and make a short summary of the statement. It should include '**WHO** did **WHAT** to **WHOM, WHEN, WHERE** and **WHY**'. Use names of victims and perpetrators.

The deponent, Dudu Chili, claims that her niece, Finkie Msomi was killed by a bomb and a bullet shot on the 28th of February 1989 at Maseko Street in Orlando West. Winnie Madikizela-Mandela, her daughter Zinzi Mandela and other members of the Mandela Football United Club are implicated in this act of human rights violation. The deponent further claims that Madikizela-Mandela and the MUFB members had begrudged her sons for refusing to join the club. They were labeled sellouts. The person who had been targeted for murder was her son, Sibusiso Chili, who narrowly escaped death after a mob, including Sibusiso, killed one of the assailants in the name of Madondo in self-defense. Prior to these attacks, Lerotodi Ikaneng, another targeted youth, had also escaped death after an attempt to murder him by cutting his throat. Ikaneng had sinned by pulling out of the club. Chili asserts that they were thought to be dangerous because they had all the information about the activities of the club.

(Give all known reference numbers of statements – HRV and amnesty - related to this case).

Refer: JB04520/01GTSOW, JB04637/01GTSOW, JB04519/01GTSOW, JB05408/01GTSO, JB05194/01GTSOW, JB05714/01GTSOW, JB03657/02PS, JB05407/01MPNEL, JB05262/03NW, JB05845/01GTSOW, JB05846/01GTSOW, AM2422/86, AM3690/96, AM6400/97, AM6401/97, AM6402/97, AM7351/97, AM7511/97, KZN/MP/017/BL.

Extract as many *acts, victims, witnesses* and *perpetrators* as possible.

ACTS – Ensure that you use the controlled language when describing an event. For every description consult the controlled language and ensure that **a word in bold** is used. When multiple injuries led to a death i.e. a person was bombed, shot and burnt, it is unclear which act was the cause of death. State all the above three acts under Severe Ill-treatment and add a fourth under Killing, thus Killing / Unknown.

VICTIMS – Write 'DECEASED' or 'DISAPPEARED' in brackets for all victims killed or disappeared respectively.

PERPETRATORS – The person who performed the act, people who gave orders or people who were involved in the conceptualization of the act.

WITNESSES – Two categories of witnesses. Those who actually saw the event and those who may not have seen it but can corroborate it or give more information.

ACTS FROM THE ABOVE STATEMENT[5]

Act 1

Victim	CHILI, Dudu Olive
Age	45
Victim Number	3
Organization	Federation of the Transvaal Women
Date	28-02-1989
Place	Protea
Specific place	Protea Police Station and Kliptown Police Station
Details	Detained.
HRV Type	Associated Violation, Incarceration or Imprisonment.
Outcome	
Reason	The police wanted Chili to give them the names of the ANC people she met in London.
Political Context	Because of the state repression at the time, an opportunity was created for gangs like the Mandela Football Club to emerge. The club terrorized the community around Soweto. The club was under the leadership of Winnie Madikizela-Mandela. Anybody not cooperating with the club was branded as a sell-out and liable to be killed.

PERPETRATORS

Name	Number	Organization
Dempsey	56550	South African Police
Van Zyl 'Slang' Andries Rossouw	29508	South African Police

[5] Thirteen acts were defined based on this statement. We show only Acts 1, 6, and 13.

Act 6

Victim	MSOMI (DECEASED), Finkie Maria
Age	13
Victim Number	56551
Organization	
Date	28-02-1989
Place	Orlando West
Specific Place	Inside the bedroom at Dudu Chili's house
Details	She was burnt during the bombing of Chili's house.
HRV type	**Severe ill-treatment, Injured by burning**
Outcome	Injury
Reason	Madikizela-Mandela and Zinzi Mandela wanted to revenge the death of Madondo who had died on his assignment to kill Sibusiso Chili.
Political Context	as in (1) above

PERPETRATORS:

Name	Number	Organization
Madikizela-Mandela, Winnie	4759	Mandela Football Club
Mandela, Zinzi	4760	Mandela Football Club
Richardson, Jerry Vusumuzi	51965	Mandela Football Club
Sonwabo (DECEASED), Vuyo	26263	Mandela Football Club
Zwane, Charles	25027	Mandela Football Club

WITNESSES:

Name	Number	Eye witness
Chili, Barbara	56557	Yes – she was in the house when it was petrol-bombed.
Msomi, Ntombenhle	56562	Yes – was also in the house
Msomi, Alfred	9530	He watched helplessly as the house burnt with Finkie his daughter.

Msomi Ntombana (DECEASED),	56567	She also watched helplessly as the house burnt with Finkie her daughter inside.

Act 13

Victim	CHILI, Sibusiso
Age	
Victim Number	56564
Organization	
Date	February 1989
Place	Johannesburg
Specific Place	At the Norwood Police Station
Details	Kept in isolation for almost a year in a dirty and filthy cell.
HRV type	**Associated Violation, Incarceration or imprisonment.**
Outcome	Detained.
Reason	Punished for the Maxwell Madondo killing. Also the police wanted to know the whereabouts of his other brothers: Mbuso, Nhlanhla and Kelly.
Political Context	as in (1) above

PERPETRATORS:

Name	**Number**	**Organization**
Van Zyl, 'Slang'	29508	South African Police Special Branch –

WITNESSES:

Name	**Number**	**Eye witness**
Khumalo, Sifiso	78516	Yes – Detained together.
Masinga, Benjamin	60611	Yes – Detained together.

Documentation and Statements.

Clearly indicate all attached documents. Further, indicate if the whereabouts of other documents are mentioned, e.g. if the death certificate is at home or with a lawyer. This could assist with the low level corroboration.

Type	Attached?	Details
Medical Report	No	Dr. Mhinga – and ANC doctor
Medical Report	No	Chris Hani Baragwanath Hospital

Consequences of the experience for the victim & expectations.

Briefly state whether the violation caused any permanent physical injuries or left any emotional suffering of pain. Describe the nature of the injury, what treatment was received for the injury and state the present physical health.

Mention the exact expectation the person would like from the TRC e.g. school bursaries, pension, gravestone erected, national monument with all the names, accommodation, etc. Are these at the person or national level – capture both.

Relations

Relate from the deponent's side i.e. from Dudu Chili.

Biographies

Capture the persons' political party, union, community involvement, etc.

A letter of acknowledgement.

The following letter is sent to each deponent / victim immediately after his / her statement has been processed.

(letter head)
Reference No.: JB04500/01GTSOW
Ms. Dudu Chili
P.O. Box 925,
Johannesburg,
2000
Gauteng

16th September 1998

Dear Ms. Chili,

Re: Statement submitted to the Human Rights Violation Committee

We would like to thank you for making a statement to the Truth and Reconciliation Commission. We apologize for the long delay in responding and ask for your understanding in this regard.

The Human Rights Violation Committee of the Commission is in the process of determining whether or not you or the persons mentioned in your statement are victims of gross violations of human rights as defined by its mandate. You will be notified of our finding by no later than 31st March 1998.

When a finding has been made, those who were found to be victims will be referred to the Reparation and Rehabilitation Committee. This committee will send these victims a Reparation Application form in due course. The Reparation and Rehabilitation Committee will the make recommendations to the State President on how the government should help those victims found to have suffered gross violations of human rights.

Your willingness to trust the Commission with your memories will assist us to find out the truth about South Africa's past and will help bring about the healing that you and our country need.

Thank you very much for volunteering to be part of the process of healing and reconciliation in our country.

May you be blessed.

Yours sincerely,

The Most Reverend Desmond Mpilo Tutu

Chairperson

Structure of the Reference Number

JB04500/01GTSOW

JB	Johannesburg regional office
04500	sequential number to count number of statements
team	01
GT	– Johannesburg commissioners were divided into three teams thus: Johannesburg city, suburbs and townships.
SOW	Statement taken from Soweto.

Other Examples

JB00099/01ERKWA

JB	as above
00099	as above
01	as above
ER	East Rand – i.e. Springs, Brakpan, Benoni, Heidelburg, etc.
KWA	Kwa-Thema, a black township in Springs

JB01238/03VT

03	team 03
VT	Vaal Triangle – Boipatong, Sebokeng, Evaton Vereeniging

JB04211/03WR

WR	West Rand – Randfontein, Krugersdorp, Carltonville, Mohlakeng

JB03331/02PS

02	team 02
PS	Pretoria/Soshanguve area, Mamelodi, Atteridgeville

JB04100/02NW

NW	North West – Mafikeng, Zeerust, Potchestroom

Appendix 7

Data Processing User Guide

This is the text of the Data Processing User Guide discussed in the section, TRC System User Guide. It is complete except for certain omissions and changes made to save space. This guide was used primarily in the training of data processors.

TRC Main Menu - After logging into the TRC database you will enter the Main Menu screen. From this menu you can run any of TRC modules available to you.

Source Menu

Accesses all modules relating to GHRV statements, amnesty applications and any other sources of information.

Note: To save space, we list all field names appearing on screens in the left to right order in which they appear on the actual screen.

Reference Number	Violation Type	Document Date	Place Taken	TRC office	Person ID Number	Surname	First Names
Attend Hearing	Summary	Summary Print	Details Print	Status	Date Registered	Date Captured	Captured by
Date Processed	Processed by	Taken by	Notes				

Source Details

Once a source document has been registered and all necessary processing has taken place, details regarding the HRV, amnesty application are entered. All the information entered at registration time will appear on the first page.

The second page of the Source Details module allows for the capturing of details of specific acts of human rights violations, which appear in the source document (see Figure 1, above). Any number of perpetrators or witnesses can be captured for a single act.

Person Details

Accesses all modules relating to people referenced within the system, including their relations and biographies.

Person ID	Residential	Surname	First Names
Aliases	Postal Code	Title	Province
ID Type	ID Number	Work	Date of Birth
Sex	Phone Home	Nationality	Race
Phone Work	Occupation	Contact Person	TRC Employee
Language	Postal Code	Prison	Phone Home
Prison Number	Phone Work		

Person Details Screen

Any person referenced anywhere within the system need only be captured once. This module allows for the capturing and querying of detailed information concerning a person. The Person ID is system assigned when the record is saved in the database. The only other required field is the

person's surname. This field may also be used as descriptor if nothing specific is known about the person. The more detailed the information entered here, the more powerful the analysis and research that may be done on it at a later stage.

The system also provides for the capturing of aliases and nicknames. It also allows for the logging of multiple biographical episodes or periods for each person stored in the database. This then serves for the building up of a political curriculum vita or any other biographical image of interest for each individual person referenced within the system. If a person is deceased or disappeared, indicate by writing (DECEASED) or (DISAPPEARED) in brackets after the surname.

Registration Menu

Reference Number	Violation Type	Document Date	Place Taken
TRC Office	Taken by	Surname	First Names
Language	Total Pages	Status	Date Registered

Register Source Document

Once an interview has been conducted or an amnesty application has been received, this module is used to register the source information within the TRC system. The reference number assigned to this source must be entered in the appropriate field and all other necessary information about the source provided.

Starting off:

Check that registration is correct.

Read entire case through from beginning to end before you start capturing.

Highlight the names of people mentioned in the case. Determine whether they are victims, perpetrators or witnesses. People who do not seem to fit into any category should be classified as witnesses (district surgeons, lawyers, etc.).

Capture the names of the people mentioned in the case, first checking whether they exist on the database. However, be careful not to deduce links that are not 100% clear. That is, do not assume that two identical names are for the same person. Put as many details as possible into the **Person Details** screen for each person and check that they are accurately captured. Corrections should be made and highlighted, especially with ID numbers. If you have no details, then at the very least put a general value in the address, i.e., the province field. Check for ages, especially of the victims. Every key person mentioned in the statement must be given a number - **Person ID.**

Write the person numbers of all victims/perpetrators/witnesses next to their names in the statement. Use the coder's sheet.

Call up the case number - **Reference Number** - on the **Source Details** screen.

Check that the right deponent is linked to the case number. The Documentation Officer does (this linkage at registration time).

Summaries

Observe the following two mnemonics in your summaries:

What, **W**here, **W**hen, **W**ho, **W**hy & **H**ow - the five wives and a husband.

Accurate, **C**lear & **C**oncise.

The summary should be as objective as possible and contain facts found in the statement. This is necessary because each statement submitted needs to be listed in the Final Report.

Three key elements which should appear in the summaries in order of writing:

What happened (Two people died and two were injured in an attack on a house in Soweto in 1990).

Contextual remarks (The day before, there had been an Inkatha rally).

What happened afterwards (People were taken to Chris Hani Baragwanath hospital by ambulance and private car. The deponent laid a charge with the Jabulani police).

In the summary, include the names of deponent and victim(s) and their relationship to each other and what happened, when, where, why and how it happened. Include the age of victim(s) if available. Give an indication of the affiliation of victim(s) and alleged perpetrator(s). Be sensitive when mentioning accusations made by anybody. Use phrases such as "alleged perpetrator or perpetrators," and "are believed to be." If the perpetrator is identified, then refer to him as "named," or

"identified." Construct sentences carefully: "they's" and "he's" in a sentence without names can be confusing. The summary must give some indication of the political motive (if this is clear from the organization of the victim and perpetrator and that would be sufficient). Be brief.

Fill in the "***Date captured"*** and put your name, i.e., your person number in the fields for "***Processed by"*** and "***Captured by***".

The "***Notes"*** field is for the following type of comments: (a) reference to other statements, (b) victim has appeared in a hearing, (c) an indication that the statement was not clear which perpetrator was linked to which violation and (d) an indication that the statement was confusing or there were discrepancies with dates, etc. This is where we need to capture any observation which we may have made, such as the fact that it is linked to Boipatong massacre, Trustfeed massacre, etc. It is NOT for the data processor to comment on whether or not you think this is a gross violation of human rights - that is up to the HRV Committee to decide.

Save your work.

Move to the next screen for the violations.

Capturing the violations:

Refer to Appendix 5 (Completed Coder's Sheet).

Call up the victim - by "***person number***"- under the acts.

Log the "***date***" of the violation. You need at least the year. If no date is given, you might be able to work it out from checking the TRC chronologies and/or checking related statements. If you do this, then state it in the "***Notes"*** field. If you cannot work out the date, log the year as **00** and put an explanation in the "***Notes"*** field.

Log the "***town"*** where the violation happened. If this is not specified but from the context seems to be the hometown of the victim, then use this. You can usually work this out from the context of the statement. Often the victim's hometown is also the deponent's hometown. If you cannot work it out, put it as "**Unknown"**. Put a note in the "***Notes"*** field if the town was not specified.

Add the "***Description of place***." Give whatever detail possible, e.g., "At home, White City, Soweto" or "at John Vorster Police station" or "open ground next to the main road through Duncan Village." This is essential in order for the researchers to code properly for the "location."

Add the "***Description of violation.***" This is the detailed free text area that was left out of the summary.

You do not have to repeat any information captured in other areas, such as who did it or where it was done. This area is crucial for capturing any information concerning the act not captured elsewhere. In essence you need to capture what happened, what was used, where on their body they were injured, how many times, etc. An example is: "**jumped out of window.**" Remember to specify if a woman was **pregnant or miscarried or if a killing was an assassination.**

Use the coding sheet definitions as a guide and include the necessary keywords. You need to ensure that you are all using the controlled language for description of events. For every description consult the controlled language and ensure that a word in **bold face** is used, i.e., use the "catch phrases." This is to ensure that at a later stage searches can be made on the free text searching through key words from the controlled language. This should not just be one word but it must be a description of what happened. .

Add the ***Outcome and consequences***. Here you capture all mental and physical injuries, such as:

If they are not able to work as a result of their violation. This is not to be confused with if they lost their job because of discrimination, or absenteeism whilst in detention, etc. These two fall under Associated Violation.

If they lost any benefits which they should have received. Double check with the R & R" (Reparation and Rehabilitation) form and "Further Question" for this information as well.

If any friends or relatives, including the deponent have been affected, this needs to be captured here as well.

State anything that happened immediately after the event. Examples: "fled the area," "death," "released after seven (7) weeks detention," or "dies four (4) days later as a result of injuries," or "permanently disabled and died of complications five (5) years later," or "signed a confession after two (2) hours torture and subsequently convicted on arson charges." If outcome is not specified under that section of the statement you can often pick it up somewhere else in the statement. In cases of incarceration, state the length of time involved, not just the date of arrest.

This needs to be captured in a short, concise way.

Add the "***HRV category***." See coding sheet as a guide for this. Here you select one of the following: Killing, Attempted Killing, Torture, Severe Ill-treatment, Disappearance, Associated Violation, Other, and Unknown.

Add the "***HRV type"***. See the coding sheet. Here you select one of the 90 codes for acts. These codes were kept close to words so that they are easier to remember but at the same time they can only have eight characters. Sometimes the acts have come out with shortened words.

Add the ***Reason for violation***." This is the immediate reason for that specific violation - thus the "**State of Emergency**" is inappropriate.

Here you need to state why the person was violated. You are trying to capture what motivated the perpetrator to commit that specific act. For example: "the victim belonged to a particular political party," "he was a shop steward," "shot after throwing stones at the police," "beaten because he was regarded as a police informer," "tear-gassed because he was part of the march," "tortured to force her to make a statement implicating someone else," etc.

It is very important that you do not presume anything here. You must only capture reasons which are stated in the deponent's statement. Do not assume. If it is left blank, assume that no reason was stated.

You do not capture here that they were innocent bystanders, for instance. This is not part of the reason why they were violated. This is captured under circumstances.

Add the "***Political context"*** if you are able to do this. The political context has two aspects to it. It can include both the national political context, e.g., there was political conflict, a national stay-away, etc. The second aspect which should be included is **if** what the person was doing at the time is relevant, e.g., if the person had recently returned from exile, if they were attending a night vigil, etc. Sometimes none of this information is available, just indicate that it is not stated. Be careful that if the context is different for different acts in the same statement that you clearly state two different contexts. For example: a person may have been in a march, beaten up and imprisoned and whilst in prison they are tortured. These two acts have different direct contexts and this needs to be indicated. You must never assume a political context even if you are aware that this act was part of the June 16 Soweto riots. If the deponent does not mention this, you cannot put it into the coding but add a comment on the "***Note"*** field. Usually the deponent provides the political context either under that question in the statement or in some other part of the statement.

Add the "***Victim age"*** if possible. This is the **age at the time of the violation**. Please check the dates if the deponent has supplied the age at the time. You can do this by checking the date of birth and the date of the incident where both are supplied. Please note that deponents often make mistakes on the date of the time of the incident. If there is a discrepancy between your addition - based on the date of birth and the date of the incident - and the deponent's version, then go with your addition but put a note in the "***Notes"*** field about the discrepancy.

Organization number and name. If the deponent refers to more than one organization, then use this to refer to the main organization or to the organizations which resulted in him/her being attacked. Remember that some organizations – resident's associations - were affiliates to others like the UDF. Use the main organization – UDF—in this case. We need the organization which the victim belonged to at the time of the violation.

Event number and name. This is used by researchers for the research coding. You can fill it in from the main list of events if it is clearly part of that event. Do not add events without consulting with the researchers. It is up to the researchers to create new events.

Save your work.

Victims:

In the Surname field, please write "DECEASED" or "DISAPPEARED" in brackets for all victims killed or disappeared respectively.

Witnesses and Perpetrators:

In order to classify someone as a perpetrator, you do not take only the person who actually pulled the trigger, for instance. If the names of people who gave the order or who were involved in the conceptualization of the act are given, they are all captured as perpetrators.

The witnesses and the perpetrators are linked to the specific violation. So, for each violation you must add appropriate witnesses and perpetrators in the fields below the acts. Be careful not to put wrong witnesses and perpetrators onto acts.

Use the perpetrator "***notes"*** field for any additional information.

Remember that there are two categories of witnesses: those who actually saw the event and those who may not have seen it but can corroborate it, give more information, etc. You need to indicate (Yes/No) whether or not they saw it. Examples are: "saw the victim shot" or "did not witness the shooting but saw the victim's wounds the day after" or "was the doctor who attended to the victim" or "acted as the deponent's lawyer," etc.

Ensure that it is possible to understand the context around the witnesses and perpetrators - it is not sufficient to just log the names onto the acts without explaining.

If a person fills in a statement himself/herself, that person is registered as the deponent. If someone assists a deponent to fill in a statement the person who assisted should be mentioned in ***Documentation*** below. At the same time not every single person named in the statement needs to be extracted, e.g., if the deponent went to the morgue and found the deceased victim, and while there bumped into a friend "X," "X" does not have to be captured as a witness.

If there is not much information on the perpetrators, just put what you have, e.g., SAP.

Save your work

Move to the next screen for details of documentation.

Documentation

Record documentation in the field marked "Statements made & other documents or items provided."

Only record documentation which the TRC actually has in possession and NOT documentation which we would like to have.

Mark whether or not the documents are attached.

If the items are too bulky to attach to the file, e.g., X-rays or a large file annexure, DO NOT indicate that the items are attached. State clearly - in the details field - that the TRC is in possession of the items but that they are filed elsewhere.

Include dates of documents

Use the attached list as a guide for how to list documents.

Save your work.

Move to the next screen for capturing **Expectation & Consequences**.

Examples of type of documentation

Type	Details
TRC research report	Annexure A - Background on Uitenhage conflict 1976-1989
Witness statement	Annexure B - Andile Xaba; interviewed 19/5/97 at Uitenhage
Medical record	Annexure C - Post mortem report
Court record	Annexure D - Civil claim - Johnson vs. Ministry of Defense, case no108/89, Uitenhage court
Court record	Annexure E - Criminal trial - State vs. Johnson, case 52/89, PE Supreme Court
Legal documents	Annexure F - Particulars of claim, case 106/89 Uitenhage court
Press clippings	Annexure G - Soldiers shoot marchers, EP Herald, 29/10/89
NGO records	Annexure H - extract from IDAF list of detainees, October 1989
NGO records	Annexure I – Black Sash report on shooting incident, October 1989
Police records	Annexure J – Photocopy of p172 of cell register, Jeffreys Bay police station, May-Dec 1989
Police records	Annexure K - Police docket, CC95/89 - public violence charges against S. Johnson.
Police record	Annexure L - Letter from Station Commander, 27/5/97, re: destruction of records.

Death certificate	Annexure M - Samuel Johnson, 20/10/89
Medical record	Annexure N - Records from Frere Hospital, EL, for S. Johnson, 19/10/89
Birth certificate	Annexure O - Birth certificate for S. Johnson, 16/2/1970
Photos	Annexure P - Photographs of injuries to S. Johnson, 19/10/89

Expectation & Consequences

Type the number of the question on the TRC statement onto the first column ("question"), i.e., 7.1 or 7.2 or 7.3

Then move to "Reply" and write in the answer in the free text (for example, "Bursaries for two school-age children").

You cannot have more that one line per question number (i.e., the database won't accept the same question number being used more than once). This means you have to put all the replies to one question in one line (e.g., "Bursaries for two school-age children and information about incident"). Be brief.

Remember that the database will be set to a current form of the TRC statement - if you are capturing a different version of the statement, then the answers will be in different fields.

Do the same with consequences

If there are no answers given to the Expectations & Consequences questions, then do not fill them in on the database - just fill in the ones that have information.

Save your work.

Reparation & Rehabilitation (R & R) coding:

The R & R coding was treated differently by the four regions. In our region one data processor was appointed to access the R & R screens of the database and capture the relevant code for that information.

Click on ***R & R sub-menu*** at the top of the screen and then choose ***R & R coding***.

Click on the query button and then type the number of the case you were working on; press F8 to call it up.

Go to **code** field and call up the list of codes (off the list of icons on the left side). Choose the code which suits the answer for 7.10

Go to the **question** field and scroll down to the next question number (7.20). Then move to the **code** field again, call up the list of codes and choose the appropriate code.

Repeat this for question 7.30

Remember that different protocols will have different question numbers.

Save your work.

Vetting – Quality control

Checking of coding errors. This was done in different ways. At the end of each day, each data processor would hand over the statements s/he processed that day. I would take one or two statements from s/he pile and re-process them comparing the victims, the acts of violations, the witnesses, perpetrators, etc., that have been extracted. I would mark all the errors in red and send the statement for re-processing. The other way is to use peer evaluation, where data processors would exchange statements and check each other's errors. Also the corroborators, researchers, investigators and commissioners would in a way help to check errors when doing their work.

STANDARDIZATION

To maintain consistency within the regions and among the data processors themselves, a document on standardization was produced and circulated to all people concerned. One of the most important issues is the importance of extracting all the acts from the statement, including the Associated Violations.

General issues:

Use small caps:, e.g., summary, description of violation, outcomes, notes, etc.

I.D. numbers - divide them as they are recorded, e.g., 19860214 5449 088

Phone numbers - write the codes in brackets, i.e., (011) 333-6330

Titles - if not stated, call all men "Mr." and all women "Ms."

General Comment:

A concise, short summary/overview of the full statement.

If the case is linked to any other statement list those reference numbers.

If the case is linked to any major events such as Boiphatong Massacre, mention this.

Comments from different units - data processors, HRV committee, researchers, investigators, etc. First state what unit the comment is from and then make the comment.

Further Comments:

State any particulars which are only pertinent to that particular act within the statement. It is not general to the whole statement but rather specific to that act, and therefore, has not been captured by the general comment or in any other place.

Circumstances:

Here you paint the broad picture of what was happening in the community at the time of the event. You paint the broad community context, such as violence between particular political groups, school boycotts, stay-aways, etc.

This also where you capture the individual activities, for instance on their way home from school, participating in a march, only doing her/his washing, watching from his doorway, walking to the shops, etc.

Biographies:

This is where we need to be capturing peoples' political/union/community involvement, etc. This needs to be used for all people captured, if we have the information.

Relations:

We need to be relating **from** the deponent, i.e., the victim is a young boy, the deponent is his mother and the witness his sister, etc.

References

Ball, Patrick, 1996. *Who Did What to Whom? Planning and Implementing a Large Scale Human Rights Data Project*. Washington: American Association for the Advancement of Science.

Ball, Patrick, Ricardo Cifuentes, Judith Dueck, Romilly Gregory, Daniel Salcedo, and Carlos Saldarriaga. 1994. *A Definition of Database Design Standards for Human Rights Agencies*. Washington, DC: American Association for the Advancement of Science and Human Rights Information and Documentation Systems International.

Chapter 4

The South African Truth and Reconciliation Commission: *Database Representation*

Gerald O'Sullivan

Introduction

The work of the Truth and Reconciliation Commission (TRC) was dominated by information processing. By the time the Human Rights Violations Committee of the TRC had completed its work, it had gathered 21,298 statements, containing 37,672 gross violations of human rights. The Amnesty Committee of the TRC received a total of 7,127 applications for amnesty. At this time (mid-1999), the work of the Amnesty Committee is not complete, so the total number of violations gathered by the amnesty process is not known, but could ultimately be in excess of 10,000.

The anticipated volume and complexity of the information was such that the Commission decided to set up a wide-area network and develop its own database to process the data. As it turned out, the network and database comprised the backbone of the organization, structuring its work in a systematic way. The end result is a rich, complex, logically disaggregated set of corroborated data which enables researchers to make powerful statements about human rights violations.

Information technology in South Africa is sophisticated despite South Africa's violent past, under-developed economy and years of sanctions. It has become more so in the years since the ban on liberation movements in 1990 was removed. With the necessary hardware, software and skills available, the TRC was able to rapidly build a powerful electronic infrastructure.

In this paper, I describe the TRC's experience of putting together this electronic infrastructure. I will describe 1) the basic network structure, 2) the organizational structure of the TRC, 3) the information flow by which the data was loaded onto the database, 4) the logical model of the database and finally 5) give some examples of the analytical results that such a database model provides. In the appendices, I give the complete statement used to gather data and the coding frame.

The editors excerpted and summarized lessons learned for this chapter and for Chapter 3. This section appears as Appendix 3.

The Basic Network Structure

The local area networks

The TRC had four regional offices, based in Cape Town, Johannesburg, Durban and East London. Each of the offices had a local area network (LAN) consisting of a number of workstations and heavy-duty printers connected together by an Ethernet network with a Windows NT server at the center of each LAN. The communication protocol was TCP/IP. There were about 250 workstations in the Commission's four offices.

The workstations ran Windows 95, and the Microsoft Office suite was used to carry out the administrative work in the offices. The e-mail facility in particular proved to be a valuable internal communications tool. In addition to Microsoft Word, which was used for word processing, the researchers used the Excel spreadsheet to analyze trends in the data on human rights violations and to graph the results of their analyses.

These products were easy to install and use, readily available and well suited for a network of this scale. However, the domain structure of Microsoft NT complicated the management of the network, and the stability of the servers was often compromised by the shortcomings of the network operating system.

In addition to the off-the-shelf software, all users involved in the TRC Information Flow (data processors, corroborators, researchers and commissioners) had access to the TRC Database. The database was a client-server design, so users had screens loaded on their workstations enabling them to query and update the information stored on the file servers.

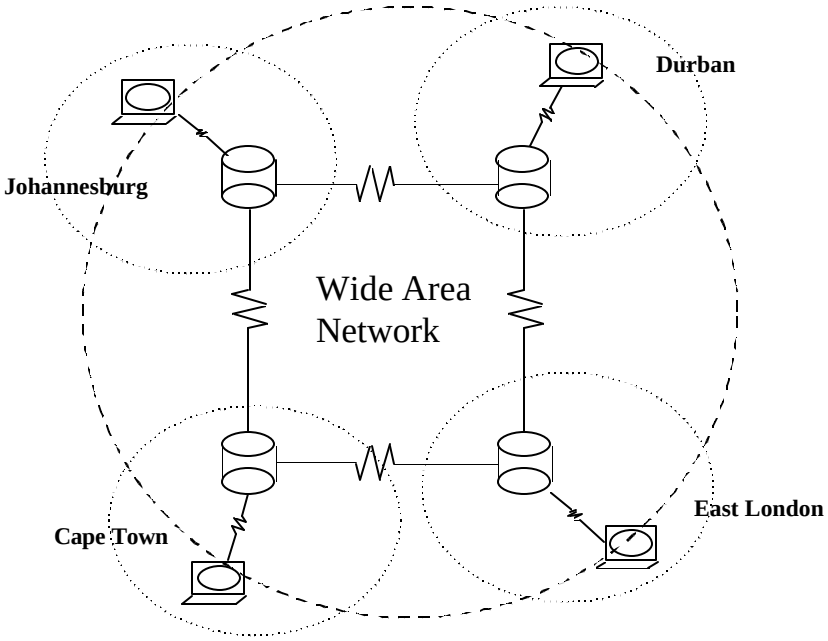
Most of the TRC staff had had little or no prior experience with computers, so each office had a Computer Officer who fixed the hardware, installed and maintained the software and provided the

users with support. A commercial network service provider supported the Computer Officers by performing the more complex hardware and networking tasks.

The Wide Area Network

A wide-area network (WAN) connected the four regional networks together, using 64k digital lines rented from Telkom, the national telecommunications provider, as shown in the schematic diagram of Figure 1.

Figure 1. Schematic diagram of the Wide Area Network



The WAN allowed users to send e-mail from one office to another, transfer word-processed documents between regions and share database information between the offices.

The commission network was not connected to the Internet for security reasons. Instead, each office had one or more freestanding computers (i.e., without a connection to the network) with dial-up access to an Internet Service Provider. There was no physical connection between the TRC network and the Internet. This was the simplest, most reliable, least expensive way of isolating the network from potential intruders, although more computer-literate users were frustrated by the lack of e-mail connections to the outside world.

The Organizational Structure of the TRC

The TRC consisted of three sub-committees:

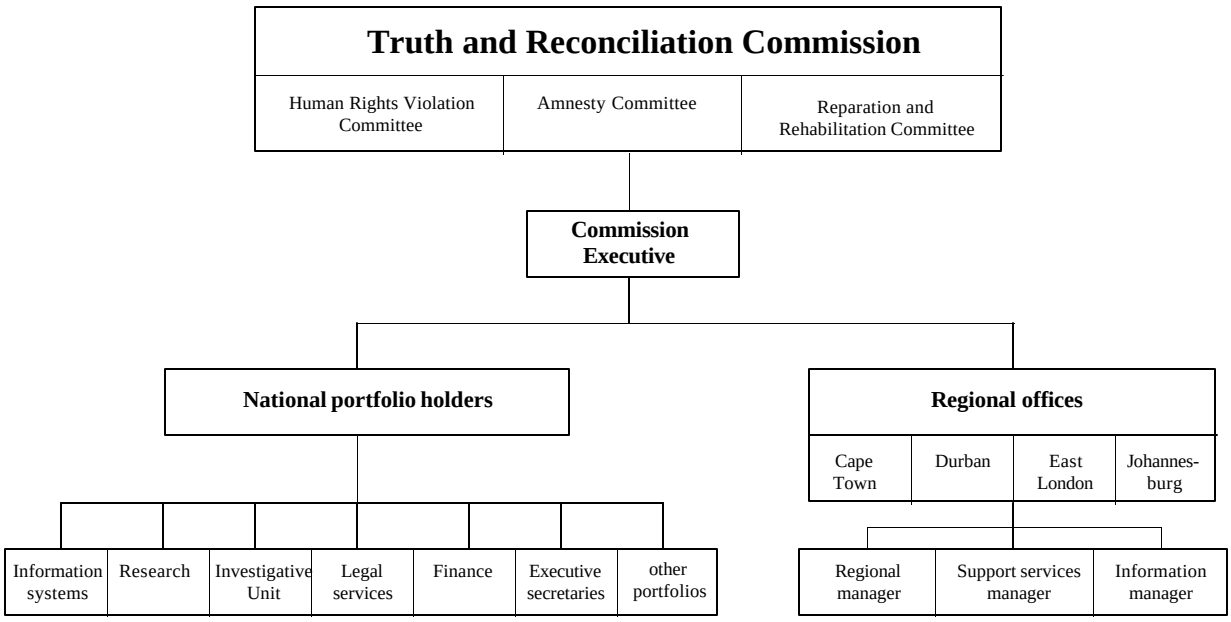
Committee	Responsibility
Human Rights Violation Committee (HRV Committee)	Collecting statements of human rights violations from victims or their surviving relatives
Amnesty Committee	Dealing with applications for amnesty from perpetrators of politically-motivated violations
Reparation and Rehabilitation Committee (R&R committee)	Making recommendations for reparation and the rehabilitation of victims identified by the TRC

The executive arm of the commission consisted of national portfolio holders reporting to the chief executive officer (CEO). They worked with the managers of the four regional offices to carry out the operational functions of the TRC and gathered and processed the HRV statements and amnesty applications on which the commissioners made findings.

Responsibility for the database and network fell under my charge as the Information Systems Manager. I worked closely with the Information Managers in each of the regional offices to ensure that the database functioned as expected, making enhancements to the functionality as more processes in the information flow came on stream. The Information Managers kept the information flow moving and ensured that the data gathered by each office was loaded onto the database efficiently and accurately.

The structure of the commission was as shown in Figure 2.

Figure 2. Structure of the South African Truth and Reconciliation Commission



By vesting the responsibility for the electronic information systems in a position reporting directly to the CEO, the database was assured of a high profile in the organization, thereby avoiding contests of ownership. It was not relegated to a purely "research" function or subsumed in the work of the investigative unit.

Indeed, the reverse was a greater problem. It was difficult to get the Research department, Investigative Unit and Commissioners to take ownership of the data that fed their own processes. The main focus of the work of the Commission was on the public hearings, rather than on gathering statements. Thus, for nearly two years, the attention of the researchers, investigators and commissioners was directed away from the database, towards the logistics of preparing for hearings.

In the absence of involvement from other portfolios and committees, the perception emerged that the contents of the database (quality, volume, and integrity) were the responsibility of the Information Systems portfolio and the Information Managers in the regional offices only. This had a substantial negative impact on the quality of the data since none of the principal users added value through active use of the data, until the findings process began in earnest and the writing of the final report was started.

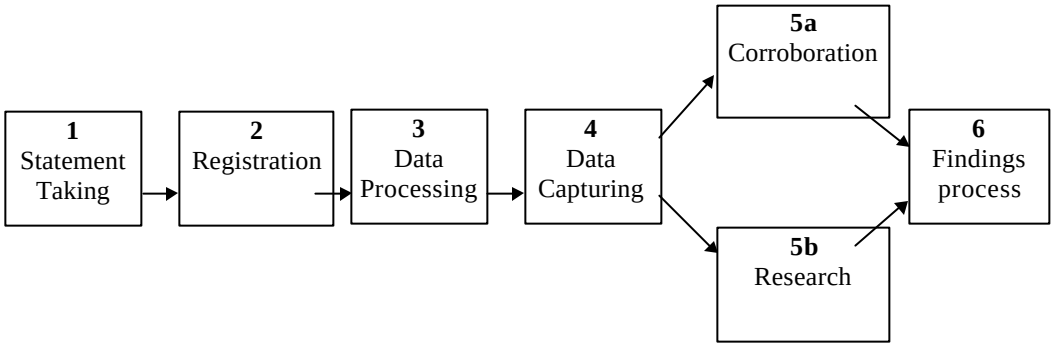
The Information Flow for Loading Data onto the Database

The TRC database was originally designed to be a research tool, but subsequently become integrated into the operations of the commission. It was at the heart of the Information Flow in which

the violations reported in statements made by deponents were analyzed, captured onto the database, corroborated by investigators and finally passed to commissioners who made findings on whether the violations constituted gross violations of human rights as defined by law.

The information flow was as shown in Figure 3.

Figure 3. TRC database information flow



The first four stages of the information flow were implemented early in the life of the commission. There was enormous pressure to get the database up and running and filled with data. As soon as the first phase of the database development was completed (database engines installed on the servers, input screens developed and installed on workstations), the registration, processing and data capture began.

At the same time as the database development was underway, the commission started its process of holding public hearings. These hearings generated enormous coverage for the work of the commission and the statement-takers were able to harness the energy of the hearings to gather statements. Unfortunately, the hearings diverted the focus of the commission from stages 5 and 6 and the crucial processes of corroboration, research and the making findings were put on hold. The data processors and data capturers worked in isolation during this time, and received no feedback on the quality or quantity of their work.

Once the process of corroboration began, and researchers began to rely on the primary data from the statement-takers to prepare for hearings, rather than using mainly secondary source material. As before, the quality of the data improved dramatically. The corroborative material (death certificates, press clippings, medical files, photographs) added enormous value to the database. Late in the life of the commission, the findings process started and the data were authenticated.

Although laborious, the process of corroboration proved invaluable and gave the findings a legitimacy they might otherwise have lacked. Before this, the data gathered often represented the data-processor's understanding of a hastily written statement, translated into English during an interview with a possibly traumatized deponent, recounting events which may have happened several years previously. Under these conditions, the probability of, and the scope for error were enormous.

The HRV statement

The data gathered by the statement-takers were written down on a standard protocol sheet. The format of the statement changed in the early stages of the commission's life. It was initially a free-flowing narrative, with sheets attached for the statement-takers to record the specific details, and to remind them to ask certain questions. Statement-takers complained that it was too repetitive, and because at that stage the commission's rate of statement-taking was very slow, the format of the statement was questioned. Some members of the HRV committee then argued for a format in the style of a motor-vehicle license application, so that these forms could be distributed by public bodies (the post office, churches, non-governmental organizations, and so on). With such a format, deponents could fill in the forms themselves, or be assisted by family members, official functionaries or religious leaders.

A form was drawn up and put into use. The quality of the data declined dramatically, and I was able to show this by doing word-count comparisons of the two styles. I did this by counting the minimum, average and maximum numbers of words entered on the database by the data-capturers

to describe the nature of the violation, the outcome of the violation and the description of where the violation took place. By comparing the counts from the free-text statements to those from the "license application form," it was easy to see that we had lost almost all context and gained nothing in the process. This form was dropped and the HRV Committee eventually compromised on a semi-structured statement (see Appendix 1).

This semi-structured statement had advantages and disadvantages when compared to the free-text statements, as I discovered by doing word-counts and by comparing the number of violations, victims and perpetrators per statement and the number of violations per victim. The results were mixed but interesting.

In those offices where data processing was known to be weak, the numbers of violations, victims and perpetrators improved, but in those where data processing was known be better, the rates dropped. The structure helped weak data processors to identify the relevant violations, victims and perpetrators. Previously, they had been lost in the narrative, but better data-processors had less narrative from which to draw, and the structure of the statement only allowed for one victim per violation type, such as killing, torture, severe ill-treatment, etc. (See Appendix 1). Thus, they ended up with fewer violations per victim and fewer victims and perpetrators per statement.

The word-counts showed little change in the amount of detail captured to describe each violation and the consequences of each violation (whether the office increased the number of violations per statement or not). However, they did show a definite improvement in all offices where details about the perpetrators, political context and the place of violation were concerned. It was clear that the semi-structured statement focussed the attention of statement-takers on questions that had been previously neglected. A deponent's testimony is understandably centered on the trauma of the violation itself, so less detail was gathered about the context in which the violation took place.

The Logical Model of the Database

The database was developed in great haste, as there was enormous pressure get the information flow functional, and with very little idea of how it would operate in practice, since the commission itself was in the process of defining its role. Drawing on the experience of Patrick Ball, as well as the work of the Human Rights Information and Documentation Systems International (HURIDOCS) team, and my knowledge of corporate databases, we created a functioning database in record time.

The design process began with the work of the Database Development Group. They reviewed the Promotion of National Unity and Reconciliation Act[1] to identify which sources of data were legally available to the TRC and which print-outs and analyses would be needed from the database to fulfil the reporting objectives of the Commission. This review was then used by a small team consisting of myself, a researcher, a systems analyst and a programmer, to design the database model and its interfaces.

The process development time was reduced by using industry-standard software. The end product was a client-server relational database using Oracle software, which is the database technology of choice for a project of this scale and nature. We chose Oracle partly for technical reasons, but also because SyBase, a database platform with similar functionality, was distributed and supported locally by a South African company with close ties to the arms procurement agency of the old regime. Commissioners feared that to use SyBase could have been seen as a conflict of interest.

We used Oracle's Computer Aided Software Engineering (CASE) tool, Designer 2000, to store the logical model, data flow functions, validation rules and the data integrity rules. By investing more effort early in the design process, we were able to use the CASE tool to generate the bulk of the code needed to create the screens, as well as the SQL scripts needed to create the database tables. Once the CASE tool had generated the code, it was a simple matter to fine-tune the screen designs and add indexes to the tables to enhance performance.

The database consisted of four database engines (one on each server in each regional office) which were kept synchronized across the wide-area network, using the standard Oracle replication functionality. This automatically integrated the data from all the offices, providing one coherent national picture.

[1] This is the Act of Parliament which established the TRC and defined its mandate.

The users had a suite of programs on their workstations that connected them to the database servers. This arrangement allowed them to register statements and amnesty applications, capture the contents of the violations, carry out complex searches on the data, extract data into spreadsheets, and print a variety of computer-generated reports such as: the content of statements or amnesty applications, corroboration carried out, letters of acknowledgement, perpetrator details, incident reports, as well as statistics for monitoring the performance of the information flow.

The data model

The underlying logical structure of the database is relatively simple, with just six core data entities. There were a total of 50 entities in the end, most of which were code tables for lookup purposes. The others were added over time to provide additional administrative functionality.

The six central entities, with the attributes[2] of relevance for the purposes of this paper, were PERSONS, SOURCES, ACTS, PERPETRATORS, WITNESSES and EVENTS. We first describe these entities with their attributes and then show the relationships among these entities in a flow chart.

PERSONS The PERSON entity consists of current or static information about the person, whether he or she was a deponent, victim, perpetrator or witness to a violation, and details about staff members.

person_seq	sequential number to uniquely identify persons (only partial details may be known about a person, so a system-generated primary key was used)
last_name	last, or family, name
first_names	first name(s)
id_number	South Africans have a unique 13-digit identification number which can be used to determine date_of_birth or sex; this field could also be used to hold passport numbers, or the old apartheid reference book number if the ID number was not available
race	human rights violations are often about ethnicity or race, uniquely so in the South African context; this attribute was valuable when analyzing patterns in the violence
date_of_birth	the ages of victims at the time of the violation or at the time of taking the statement can be calculated from the date of birth
sex	another useful analytic variable
address_lines	for contact purposes
phone_no	for contact purposes
town	a lookup table is essential for entering the name of towns
staff_members	details about staff members

A number of other PERSON attributes were on the system, but did not prove as useful as the above, because the information was either unavailable or unreliable. These attributes included marital status, religion, employment status and language. Other attributes not included here involved administrative functions - notes about the person, date of the victim finding, etc.

[2] For clarity the names of the attributes here are not exactly the same as were used in the database.

SOURCES The SOURCE entity holds details of the source of the information about the violations in question. In the case of the TRC, violations either came from Human Rights Violation statements, or Amnesty applications. Secondary source material was only used for corroborative purposes.

reference_no	file reference number allocated to the document
protocol_type	a code to indicate whether the document was an HRV statement or amnesty application; because several different versions of the HRV statement were used, the code also identified the version
deponent	the identifier of the person who made the statement or submitted the amnesty application; this had a foreign-key constraint to person_seq in the PERSONS table
place	the town where the statement was made or amnesty application lodged
office	code of the TRC office responsible for the document
status	the status field was used to track where in the Information Flow the document was, Registered, Processed, Corroborated, or Finding
date_taken	the date the statement was taken, or amnesty application made
interviewer	the identifier of the staff member who took the statement or application
date_registered	date of registration
registrar	the identifier of the staff member who registered the document
date_processed	date the document was processed
processor	the identifier of the staff member who processed the document
date_corroborated	date the document was corroborated
corroborated (by)	the identifier of the staff member who corroborated the document
date_finding	date the document completed the findings process
commissioner	the identifier of the commissioner who made the finding

The dates and person identifiers above held valuable details of the progress of the document through the information flow. They were particularly useful for monitoring blockages in the system, finding the location of backlogs and monitoring the performance of individual staff members in terms of speed and accuracy. These fields were not normalized for ease of programming and database performance. Strictly speaking, a SOURCE_HISTORY entity should have been used.

ACTS The ACT entity was at the heart of the database. This entity held details of the What and Whom, as well as When, Where, How and Why. It has a many-to-one relation to SOURCES (one document can describe many violations) as well as to PERSONS (one person can be violated many times).

reference_no	file reference of the source document
act_seq	sequence number of the violation as narrated in the document
victim	identifier of the victim
violation_desc	narrative description of the violation
violation_type	code used to categorize the nature of the violation. In practice, the TRC conflated the category of the violation as defined in terms of the legislation with the modus operandi of the violation, so the codes were of the form KILLING/SHOOTING or TORTURE/ELECTRIC; in retrospect, we should have had two fields, one for the legislative category and one for the mode of the violence. The approach used was the result of a lack of clarity regarding the coding frame at the start of the process. (See Appendix 2 for the coding frame)
outcome	narrative description of the outcome of the violation
outcome_type	code used to categorize the outcome of the violation. Unfortunately, due to time pressures, this was not used systematically, but it does have enormous analytic capacity for assessing the human cost of gross violations of human rights
place	town where the violation took place
location_desc	narrative description of the location of the violation (in a police cell, for example, or at the training camp, at the chief's kraal)
location_type	like the outcome_type, this was not used systematically, but had it been used, it could have contributed to the recommendations chapter of the Final Report
day	the day of the month of the violation; the date of the violation was split into its three components - day, month and year - because on many occasions, only partial date details were given in the documents
month	the month of the year of the violation
year	the year in which the violation occurred
victim_org	the code of the organization to which the victim belonged. This was selected from a lookup table to ensure uniformity of spelling, etc.
circumstances	narrative description of the political circumstances prevailing at the time.

The ACT entity had a several other attributes for administrative purposes, including a "veracity" indicator. This was subsequently used to record the commissioners' finding on whether the violation constituted a gross violation of human rights, or whether amnesty was granted in respect of the offence.

PERPETRATORS The PERPETRATOR entity holds details of the individuals who carried out the violation. It has a many-to-one relation to the violation, because many perpetrators can carry out one violation.

reference_no	file reference of the source document
act_seq	sequence number of the violation as narrated in the document
perpetrator	identifier of the person who carried out the violation
perp_org	the code of the organization to which the perpetrator allegedly belonged. This was selected from the same lookup table as the list of victim organizations
weapon	narrative description of the weapon used
characteristics	identifying characteristics of the perpetrator

The perpetrator entity proved to be very useful for analytic purposes, especially with respect to the alleged organizational allegiance of the perpetrator. However, in most cases, the rest of the information was too sparse to be of much value for investigative purposes. In most cases deponents remembered little of substance other than the name of the organization involved; the other attributes, such as vehicle_used, or place_last_seen, were rarely used.

WITNESSES The WITNESS entity holds details of the individuals who witnessed the violation. It has a many-to-one relation to the violation, because many individuals can witness one violation.

reference_no	file reference of the source document
act_seq	sequence number of the violation as narrated in the document
witness	identifier of the person who witnessed the violation

The WITNESS entity proved less useful than was anticipated at the start. It was intended help the investigators follow up the details of the case, but in most cases, the deponents themselves were the best witnesses.

EVENTS The EVENT entity was used to group violations from a variety of documents into conceptually meaningful events. For example, this entity was used to group all violations pertaining to the Ratanda bus massacre in one event. The event was a recursive entity, so small events could be grouped together into larger events.

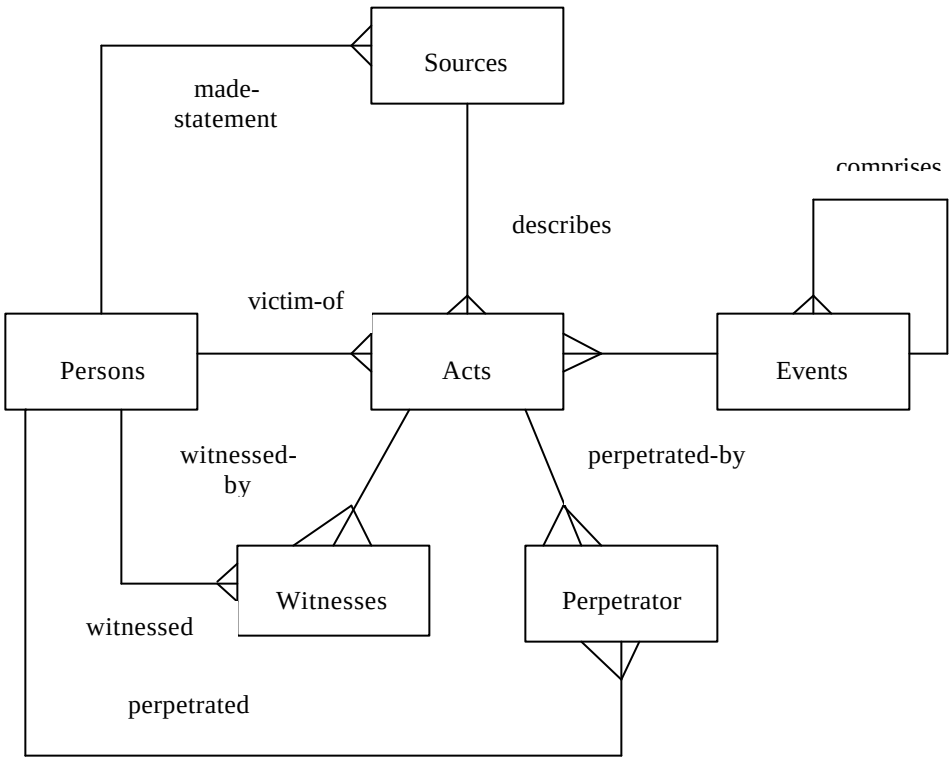
event_id	sequential number to uniquely identify events
title	the title of the event
start_date	the date when the event started
end_date	the date when the event ended
notes	notes about the event
parent_event	identifier of the larger event, of which this is a part
owner	identifier of the person who created the event

The EVENT entity had great potential, but was not used to its full capacity by the researchers who were expected to be the major users of this entity. Due to other pressures, they were unable to devote enough time to learn how to make it useful for their needs. Ultimately, it proved useful to the investigators preparing for hearings who used it to extract violations, which they then loaded into a tool, which drew diagrams of links between thousands of incidents, perpetrators and victims in a matter of seconds. The Event entity was also later used by the Amnesty Committee to plan hearings by grouping violations from various amnesty applications together.

Despite the relatively few entities in the data model, it was complex enough to model all the real world events that were brought before the commission. For example, the same person could be a victim at different times and in different places. A person could be a deponent telling about the death of a relative, and simultaneously be a victim in his or her own right. A person may be the victim of torture, and then perpetrate a gross violation of human rights in retaliation at a later date.

These entities were related as shown in Figure 4.

Figure 4. Entity relationship diagram

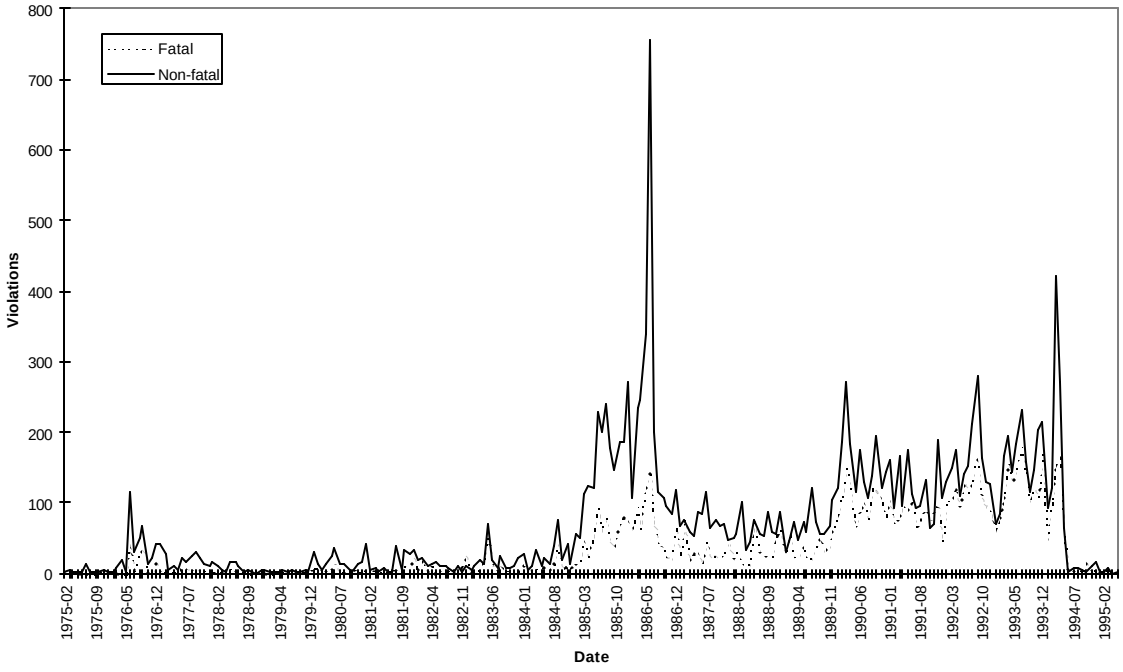


To keep the database design as straightforward as possible and to minimize the time spent on the design and build phase, no history of changes to entities was maintained. Instead, the same record was updated as new information became available or errors identified.

Given more time, it would have been of great benefit to design a database capable of holding various versions of the violation, for example, to keep the original version as told by the deponent separate from the corroborated, or "the finding" version. With such a capability, researchers could have investigated the nature of oral testimony as compared to the "official" version of history. Also, operational managers could have seen where errors were corrected and why, or if needed, to revert to an earlier version.

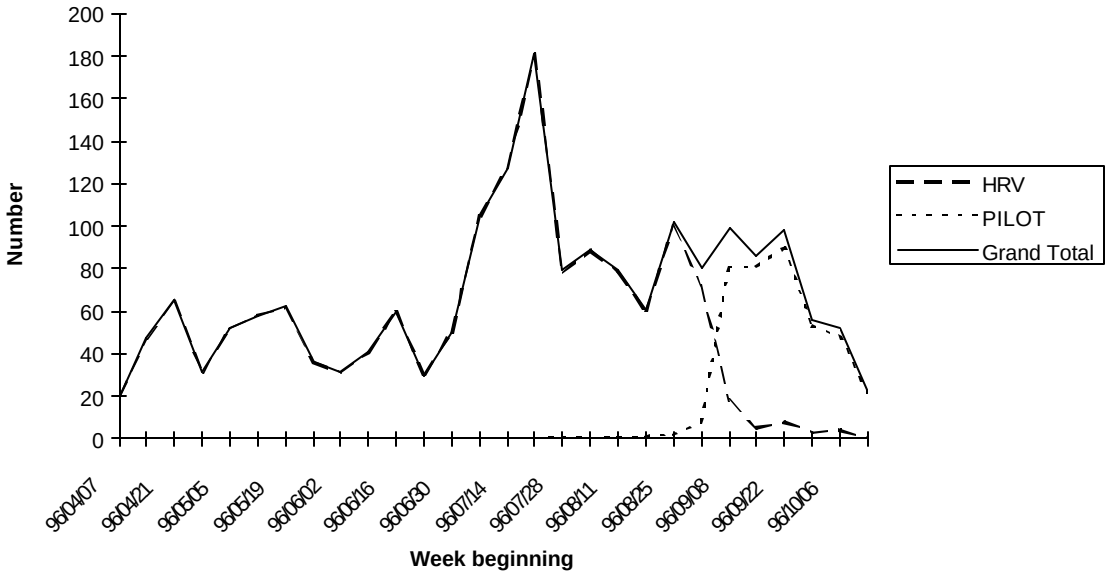
Examples of Analytical Results

It took several months before results were obtained from the database. This delay caused a great deal of anxiety on the part of the commissioners, until six months later when the first graphs of the results were produced. The following graph which compares the number of fatal to non-fatal violations reported to the commission over the mandate period is one of the first to be produced.



Other analyses were done on the ages of victims, their gender, their political affiliation, and by the type of abuse suffered. For example, graphs were drawn of the different age cohorts of deponents for each gender, which showed that the perception of statement-takers that most deponents were middle-aged women was true.

Besides its analytic value, the database was used to monitor processes in the information flow. For example, the graph below shows the progress of implementation of a pilot HRV statement in an office.



This type of analysis informed research work, as well as policy formulation for the Rehabilitation and Reparation Committee and strategic planning of the commission's work. The results contributed substantially to the final report of the commission, underpinning the narrative text in a way that dramatically highlighted the scale and extent of the violence of the past.

Appendix 1

The TRC Gross Violations of Human Rights Statement

Note: Throughout this appendix, we have reduced the spacing between lines and removed blank space for entries to reduce space and make it easier for the reader to determine the structure. Where blank spaces for entries, their presence is indicated by entry lines (" .....") of varying length.

S T A T E M E N T

concerning

GROSS VIOLATIONS OF HUMAN RIGHTS

The aim of this STATEMENT is to gather as much information as possible about the gross violations of human rights suffered as a result of the political conflict in South Africa. According to the legislation, gross human rights violations are:

Killing, torture, severe ill-treatment, abduction and disappearance
or
Any attempt, conspiracy, incitement, instigation, command or procurement to commit a gross human rights violation, defined by parliament as killing, torture, severe ill-treatment, abductions and disappearances,

that occurred in a political context

between 1 March 1960 and 10 May 1994.

Truth and Reconciliation Commission (TRC)

The aims of the Truth and Reconciliation Commission are:

- to give as complete a picture as possible of the gross human rights violations resulting from the conflicts of the past;
- to restore human and civil dignity to those who experienced violations by letting them tell their stories and recommending how they can be assisted; and
- to consider granting amnesty to perpetrators who carried out the abuses for political reasons, and who give full details of their actions to the Commission.

If you have experienced or have knowledge of *Gross Violations of Human Rights* committed between 1 March 1960 and 10 May 1994, please complete this statement. Thank you for sharing your painful experience with the Truth Commission. Your contribution will help our country come to terms with the past.

Should you run out of space when answering the questions, please use the additional pages at the back (page 23 and 24).

NOTE:

- You are entitled to legal representation at your own cost, both while completing this statement and/or when testifying in a possible public Human Rights Violation hearing. You can apply for legal aid if needed. Please contact the office.
- If you make a false statement willingly and knowingly you could be prosecuted.
- If you complete this statement by yourself, please post (or hand deliver) to any of the following Truth and Reconciliation Commission offices listed below.
- Please attach additional documents (for example, copy of ID, press clippings, doctors reports, etc.).
- Please put your initials (sign) on every page of your statement at the bottom of each page.
- By submitting this statement to the Truth and Reconciliation Commission, your name may appear in the final report of the Commission; perpetrators may be informed of any allegations you make; and your medical, legal and other records may be made available to the Commission.

JOHANNESBURG Gauteng, Mpumalanga, Northern and North West Province	**CAPE TOWN** Western Cape and Northern Cape	**EAST LONDON** Eastern Cape Province	**DURBAN** KwaZulu/Natal and Free State Provinces
Dr. Fazel Randera (Commissioner – Convenor) Mr. Patrick Kelly (Regional Manager) P.O. Box 1158 Sanlam Centre, 10th floor cnr Jeppe & Von Weilligh Str. Johannesburg 2000 Tel (011) 333-6330 Fax (011) 333-0832	Dr. Wendy Orr (Commissioner - Convenor) Ms. Ruth Lewin (Regional Manager) P.O. Box 3162 Old Mutual Building, 9th Floor 106 Adderley Street Cape Town 8000 Tel (021) 245-161 Fax (021) 245-225	Rev. Bongani Finca (Commissioner – Convenor) Rev. Vido Nyobole (Regional Manager) P.O. Box 392 NBS Building, 5th Floor 15 Terminus Street East London 5200 Tel (0431) 432-885 Fax (0432) 439-352	Mr. Richard Lyster (Commissioner - Convenor) Ms. Wendy Watson (Regional Manager) P.O. Box 62612 Metlife Building, 9th & 10th Floors 391 Smith Street Durban 4008 Tel (031) 307-6747 Fax (031) 307-6742

Declaration

I, ……….………………………………………………………. solemnly declare that the information I am about to give the Truth and Reconciliation Commission, is to the best of my knowledge, true and correct and I consider the contents of this statement binding on my conscience.

_________________________________ ________________
Signature / Finger Print / Mark *Date*

____
Witness signature

If you are called to a public hearing, will you be prepared to appear? YES NO [circle]

IMPORTANT:

- Some women testify about violations of human rights that happened to family members or friends, but they also have suffered abuses. Don't forget to tell us what happened to you yourself if you were the victim of a gross human rights abuse.

DETAILS OF THE PERSON *HELPING* TO FILL IN THE STATEMENT

Please fill in this section if somebody is HELPING you to make the statement.
Full name of person helping: ………………………………………………….
Relationship to person giving statement *(for example, neighbour, friend)*:……………………
Address:..……………………………………………………………………..
…………………………………………………………...…………………

Signature of helper: …………………………………………… Date: . ………………

1. DETAILS OF PERSON MAKING STATEMENT

PERSONAL INFORMATION

Surname: ……..… Title: ……
(*for example, Mr., Ms., Dr., Prof.*)

First Names: ……..

Other names: ……..

(for example, clan names, code names, pseudonyms, nicknames, aliases)

Type of ID document: ……..…ID or Passport number: ……

(For example, ID Book, passport, birth certificate, etc.)

Date of birth: ……Sex: Male Female *[circle]*

Citizenship: ……Race as per Apartheid legislation: ……..

Occupation: …….Are you currently employed? Yes No *[circle]*

Home Language: ……..

CONTACT ADDRESS

Where does your post go to?

Block / Street and number: …...…… P.O. Box: ……

Section/Extension: …… Suburb and City: ……

Township/Suburb/City: ……..…….. Postal Code: ……..……..

Postal Code: …… Province: …… Province: ……

Home Telephone No.: ……..…Work Telephone No.: ……...…..

What is the best and easiest way the TRC can contact **you** in future?

(Could be the same address as above or could be a friend or relative with whom there is regular contact)

Name of Contact person: (if relevant) ……

Contact address: ……

Contact telephone: ( ) ……
[area code]

2. WHOSE STORY ARE YOU GOING TO TELL THE COMMISSION?

Are you going to tell the Commission about what happened to **you**? **YES** **NO** *[circle]*

[AND / OR] *[circle]*

Are you going to tell the Commission about what happened *to* **someone else***?* **YES**

NO

(for example, your son, daughter, grandchild, mother, father, aunt, friend, etc.) *[circle]*

3. BRIEF DESCRIPTION OF THE VIOLATION OF HUMAN RIGHTS

Please briefly describe what happened to you or the person you are telling us about. Please tell us:

What happened? Who got hurt, killed or kidnapped? When did it happen? Where did it happen? Who did it?

……

4.VICTIM DETAILS

Please list **ALL** the victims you have mentioned and give details as far as you know:

Full names of person violated (i.e. victim)	**Sex and age at time of violation**	**Race as per Apartheid classification**	**Relationship of person making the statement to the victim**	**Occupation at time of violation**	**Organisational involvement (give dates and position)** *(for example, Community Council, SAP, ANC MK, APLA,, SADF, trade unions, women or youth organisation, civics, religious group)*
for example Jackie Jones	*female;21 yrs*	*White*	*myself*	*student*	*UDF supporter (1983-85) Church deacon*
for example Sam Majola	*male; 34 yrs*	*African*	*my son*	*taxi driver*	*COSAS branch chairperson (1987) MK member (since 1985)*

5. POLITICAL CONTEXT

Please describe the political situation in the community at the time of each incident.

.....

(for example, there was a mass funeral in the community that day; stay-away; boycott; march; mutiny in the camp; political rally; etc.)

6. PLEASE PROVIDE *SPECIFIC* DETAIL NEEDED BY THE TRUTH AND RECONCILIATION COMMISSION

This section of the statement is to provide all the relevant information needed by the TRC concerning the specific gross human rights violations.

Please mark the boxes below, and then turn to the appropriate section and answer the questions afterwards as far as you can.

The questions below are arranged according to the different types of gross human rights violations as defined by Parliament. You are requested to:

- please indicate which categories are relevant to your experience by marking a cross (X) in the appropriate box. If you have experienced more than one type or category of violation please indicate this by putting a cross (X) in the appropriate boxes.
- If your experience does not fit exactly into any one of the types/categories of violations listed below, please use the ADDITIONAL PAGES at the end of this form to write down your story.

	Mark with an X
Killing The person died as a result of a violation(s) (for example, shot by police at a political funeral, died as a result of torture in detention).	
Serious Injury or Severe Ill-Treatment The person does not die. Examples include bombings, shootings, stabbings, burnings, sexual abuse, attempted killings. These may have occurred in demonstrations, political conflict between groups, armed combat, etc.	
Torture Systematic and intentional abuse with a particular purpose, for example, to get information, intimidation, or punishment. This happens in captivity or custody by the state or other groups. The person, however, survived the ordeal.	
Abduction or Disappearance There is evidence that someone was taken away forcibly and illegally, or the person vanished mysteriously and was never seen again.	

The person died as a result of a violation(s) (for example, shot by police at a political funeral, died as a result of torture in detention, was killed in armed confrontation with MK soldiers).	K I L L I N G

EVENT

Name of Victim. …….....……

When was the person killed? (date and time): …….....…

Where was the person killed? (exact location, including street, name of building, area, town): …….....…….

(for example, in front of the house in Akker St.; at the taxi rank in Extension 4)

Please describe how the person was killed. Include details of what weapon was used to kill the person: .....

Why was the person killed? ….....……

Was there a post-mortem or inquest? If yes, what was the outcome?

(for example, did a doctor examine the body to find out the cause of death? Did you find out how the person was killed? Did you go to court to find out what happened? Was anybody found responsible for the death?)

…….……

PERPETRATORS

Can you identify the perpetrators in any way? Give names, rank and title, and physical description. .....

(for example, Mr. Siyanda, member of people's court; four men in balaclavas; a big man with a scar called Kallie)

How do you know who they were? ……....……

(*for example. I saw them; my neighbor told me; there was a court case)*

What organization do you think they belong to or support? ……...

(for example. SAP, UDF, witdoeke, PAC, comrades, SADF, Riot Squad, Town Council, Inkatha, ANC)

Can you specify who did what? Who was in charge? Who gave orders? Who was with them? .....

(for example, Mr. Siyanda ordered the killing, Vusi poured the petrol and Toto lit the match)

Where and when did you last see the perpetrator(s)? .....

Would you like to meet the perpetrator(s)? .....

WITNESSES

Is there anyone else who knows what happened to you or the alleged victim either **before, during** or **after** the killing? If yes; please answer the following questions as fully as possible.

Name of Witness	**Contact address and telephone number of witness**	**What did this person see or hear?**
for example, Mrs Moodley my neighbour	*13 Esau St, Lenasia tel (011) 123456*	*She saw the shooting of my son and told me about it.*

ADDITIONAL INFORMATION

…….……..……

The violation did not result in death. These may have occurred in demonstrations, political conflict between groups, armed combat etc. Examples of severe ill-treatment include shootings, stabbings, beatings, sexual abuse, burnings.	**SERIOUS INJURY OR SEVERE ILL-TREATMENT**

EVENT

Name of victim. …….....……

When did the violation occur? (date and time) ……...……

Where did the violation occur? (exact location, including street, name of building, area, town): .....

(for example, in front of the house in Akker St.; at the taxi rank in Extension 4)

Please describe in detail what was done to you and/or the person you are talking about? ..…….....……

Were you or the victim sexually assaulted? Please give details: .....

Was there a court case? If yes, what was the outcome? ……...……

PERPETRATORS

Can you identify the perpetrators in any way? Give names, rank and title, or physical description. .....

(for example, Kitskonstable Jacobs; Mrs Daba and a group of comrades; four men in balaclavas)

How do you know who they were? ……...…

(for example. I saw them; my neighbor told me, there was a court case)

What organization do you think they belong to or support? …….....

(for example. SAP, UDF, witdoeke, PAC, comrades, SADF, Riot Squad, Town Council, Inkatha, ANC)

Can you specify who did what? Who was in charge? Who gave orders? Who was with them? .....

(for example. Capt Coetzee ordered the shooting; Constable Denga shot me in the stomach)

Where and when did you last see the perpetrator(s)? ….....…….

Would you like to meet the perpetrator(s)? …….....…….

WITNESSES

Is there anyone else who knows what happened to you or the alleged victim either **before, during** or **after** the incident?

If yes; please answer the following questions as fully as possible.

Name of Witness	**Contact address and telephone number of witness**	**What did this person see or hear?**
(for example) Joe Mini	*1409 KwaMashu, Durban tel (031) 123456*	*He found me being beaten by Vusi and his friends*

ADDITIONAL INFORMATION

.....

Systematic and intentional abuse with a particular purpose, for example, to get information, intimidation, or punishment. This happened in captivity or custody by the state or other groups. The person, however, survived the ordeal.	**T O R T U R E**

EVENT

Name of victim. ……..……

When were you and/or the victim tortured? (dates, times, length of time) ……

Where did the torture occur? (exact location, including street, name of building, area, town) :……..……

(for example, Loubscher's office at the police station; in the detention centre near the camp)

Please describe in detail what was done to you or the person you talking about. In other words, describe the torture: .....

Were you sexually assaulted? Please give details: ……..….

Why were you or the person you are talking about tortured? ……

(for example, to sign a statement, to become a state witness, punishment)

Describe the conditions of the captivity …….....….

PERPETRATORS

Can you identify the perpetrators in any way? Give names, rank and title, or physical description …….....……

(for example, Kitskonstable Jacobs; Mrs Daba and a group of comrades; four men in balaclavas)

What organization do you think they belong to or support? ……....…….

(for example. SAP, Security police, Mbokodo , ANC, SADF, Town Council, Inkatha, Transkei police)

Can you specify who did what? Who was in charge? Who was with them?

(Capt Piet was in charge of my interrogation; Botha applied electric shocks; Commander 'Zizi' suspended me upside down)

Where and when did you last see the perpetrator(s)? …….……....….

Would you like to meet the perpetrator(s)? …….....

ADDITIONAL INFORMATION

Describe any visits by doctors or District Surgeons. Give names and details: .....

Describe any visits with a magistrate. Give names and details: .....

Did you see a lawyer? Was there a court case? Was the torture experience described in court? What was the outcome of the case? .....

Is there anything else you wish to tell the Commission about this experience of torture? ……..…..…..

WITNESSES

Is there anyone else who knows what happened to you or the alleged victim either **before, during** or **after** the incident? .....

If yes; please answer the following questions as fully as possible.

Name of Witness	**Contact address and telephone number of witness**	**What did this person see or hear?**
(for example) Mrs Khumalo	*14 Grange Str, Meadowlands tel (011) 123456*	*She was in the police cell with me and saw my wounds*
(for example) District Surgeon can't remember name	*Pretoria Central Prison*	*He saw my injuries and refused treatment*

Someone is taken away forcibly and illegally, or disappears and is never seen again.	**ABDUCTION OR DISAPPEARANCE**

EVENT

Name of victim . ……

When did the abduction/disappearance take place? (date and time) ……

Where did it happen? (exact location, including street, name of building, area, town)

:……..

(for example, from his house at 1711 Loerie St.; from the taxi rank in extension 5)

Please describe how it happened. ……..

Where was the person taken to? (street, building, town) ……..

Why did it happen …….....……..

What was the outcome? Did the person come back? ……..……..

(for example, They let me go after two weeks; my son's body was found the next day)

PERPETRATORS

Can you identify the perpetrators in any way? Give names, rank and title, or physical description. .....

(for example, Mr Siyanda member of people's court; Chief Ndlela , leader of Mbokodo; four men in balaclavas)

How do you know who they were? ……

(for example. I saw them; my neighbor told me, there was a court case)

What organization do you think they belong to or support? …….

(for example. Security police, vigilantes, comrades, Mbokodo, Town Council, Inkatha, ANC, SADF)

Where and when did you last see the perpetrator(s)? ...…….

Would you like to meet the perpetrator(s)? ……...…….

WITNESSES

Are there any witnesses to the violation either **before, during** or **after** the incident?

.....

If yes; please answer the following questions as fully as possible.

Name of Witness	**Contact address and telephone number of witness**	**What did this person see or hear?**
(for example) Mr Mpokeli	*629 Site C, Khayelitsha*	*He saw my son being dragged into a taxi by five men in balaclavas.*

7. EXPECTATIONS

An important part of the Truth and Reconciliation Commission's proposals to the President will be about symbolic acts which will help us remember the past, honour the dead, acknowledge the victims and their families and further the cause of reconciliation.

Please give us your opinion on what should be done:

7.1 For individuals: .....

(for example, medals, certificates, street names, memorials, grave stones, etc.)

7.2 For the Community: .....

(for example, a peace park, build a school, special ceremony, annual religious service, etc.)

7.3 For the Nation: .....

(for example, a monument, national day of remembrance, etc.)

8. CONSEQUENCES OF THE EXPERIENCE

The following questions are specific to the victim who experienced the violation.

8.1 If the violation(s) caused permanent physical injury, please describe the injury:

……..…..…….

8.2 What treatment did the victim get for the injury? Do you still need medical treatment?…….……

8.3 Please describe the present physical health of the victim: .....

8.4 If the person cannot look after themself, please explain how they live with the situation:

.....……

8.5 Did the violation cause emotional and psychological suffering or pain? .....

8.6 Please explain how the victim coped with the suffering: .....

(for example, did somebody help you deal with the pain of the event? Did you see a therapist or your priest, or a traditional healer?)

8.7 Did the violation affect relationships with friends, family, partner or children?

.....

(for example, I have lost contact with them; my marriage broke down; we do not talk together like we used to, etc.)

8.8 Has the person's behaviour changed since the violation? .....

(for example, he is depressed all the time; she feels like dying; I am always angry; I hate going near that place; etc.)

8.9 How did the violation affect the health, education, accommodation, and finances of the victim's family? .....

8.91 Health .....

(for example, since the death of my daughter, we have been suffering from depression.)

8.92 Education .....

(for example, since my husband died, my son had to leave school to earn money.)

8.93 Accommodation: .....

(for example, since my son died, we are living in this shack.)

8.94 Finances .....

(for example, before I was shot, I was working, but now I am not working and can't feed the children.)

9. DOCUMENTATION DETAILS

Have you already made one or more statements about this incident? **YES NO** *[circle]*
If yes, please specify:

To WHOM statement was made? *(for example, Foundation for Equality before the Law)*	**WHEN?** *(for example1993)*	**CONTACT details / person** *(for example, Adv. Strydom tel. (***) - *** ***)*

Do you have any documents that will help the Commission understand the situation and experience you have described? **YES** **NO** *[circle]*
(for example, Doctor's Certificate, Membership card, Diary, Newspaper clippings, Legal Documents, Post-Mortem report, Hospital records, Police records, Court records, Inquest reports etc).

Type of Document	**Where is this document at the moment?**
(for example) Inquest report	*with the lawyer Smith, Jones and Associates*
(for example) Death certificate	*at home*

What legal action did you or the victim take? Please give dates and the name of the lawyers, magistrates and judges if you can. .....
(for example, was there a court case about the violation? Did you sue the perpetrators for damages? Did you lay charges against the perpetrators?)
What was the result? .....

ADDITIONAL PAGE

Please mark clearly which question or paragraph you are answering on this page.
.....

CHECK LIST FOR THE PERSON FILLING IN THIS STATEMENT

This page is to help check that the statement has been completed as fully as possible.

CHECK LIST	**yes / no**	**Other comments**
Were all the questions either asked or considered?		
Is the DECLARATION on page 3 above signed?		
Is the RELEASE FORM on page 23 signed?		
Are all the relevant pages (including the additional pages used) initialed?		
Are all the relevant documents attached to the last page of this statement?		

FOR OFFICIAL USE ONLY

To be completed by ALL Statement Takers (including Designated Statement Takers - DST)

Full Name of Statement Taker _____

Signature of Statement Taker _____

Date of Interview ___/___ _/_____

(*day / month / year*)

Place and Town of Interview_____

Language of Interview_____

Are you a **TRC** statement taker or a **DST**? *[circle]*

TRC Office: Johannesburg / Cape Town / East London / Durban *[circle where appropriate]*

COMMENTS / OBSERVATIONS OF THE STATEMENT TAKER.

……

RELEASE FORM:
Medico-Legal Records

I, ……..……....

(name of person giving permission

hereby grant permission for the Investigative Unit of the Truth and Reconciliation Commission to obtain copies of all

medico-legal records of

…..…….……..……. who is

(name of victim)

……..…….....,

(relationship to victim, for example, myself, my son, my daughter)

for the purposes of ongoing investigation being conducted by the Truth and Reconciliation Commission.

Yours faithfully,

Signature: ..……Date: …….

Appendix 2

Coding frame for gross violations of human rights

Introduction

The task of the TRC is to identify those people who suffered gross violations of human rights, which are defined as follows: Killing, Abduction, Torture and Severe Ill-treatment. In addition to these four, there is a fifth category which is not a gross violation of human rights, but is important for understanding the context, called an Associated Violation.

Each of the five categories has several sub-headings, which explain how the violation took place (a person can be killed in different sorts of ways, so we need to identify how they were killed). By breaking the categories into sub-headings, we can then do meaningful counting for the final report.

The HRV categories

When coding, the first step is to decide under which HRV category does the violation fall. The table below shows the categories with their definitions (the words in **bold** should be used in the description of the act when they are applicable).

HRV Category	Code	Definition
Killing	KILLING	A killing is when a person dies, in one of three ways: **Assassination** - killing of a *targeted person* by a person or group who *developed a secret plan or plot* to achieve this. Person is targeted because of political position. **Execution** - capital punishment (death sentence) imposed and carried out by a legal or authorised body such as a court of law or tribunal. Victim is aware of death sentence. Perpetrators are the state, homeland governments, or security structures of political movements. **Killing** - all other deaths, including a killing by a crowd of people.
Torture	TORTURE	Torture happens in *captivity* or in *custody* of any kind, formal or informal (for example, prisons, police cells, detention camps, private houses, containers, or anywhere while tied up or bound to something). Torture is usually to get information, or to force the person to do something (for example, admit to a crime, or sign a statement), but it is also for punishment, degradation, and systematic breakdown of an individual. It includes mental or psychological torture (for example, witnessing torture, or telling the person that their family is dead).
Severe Ill-treatment	SEVERE	Severe Ill-treatment covers attempted killing and all forms of inflicted suffering which caused *extreme* bodily and/or mental harm. It tends to take place outside of custody (for example, injury by a car bomb, or beaten up at a rally), but a person could be subjected to severe ill-treatment in custody too (for example, a once-off severe beating, or teargas in the cell).
Abduction	ABDUCTION	Abduction is when a person is forcibly and illegally taken away (for example, kidnapping). It does NOT mean detention or arrest. It is not a gross violation of human rights to be arrested (see Associated violations). If the person is never found again, it is a disappearance.
Associated violation	ASSOCIATED	These are not gross violations of human rights, but are important for understanding the context of the violation (for example, detention, harassment, framing, violating a corpse after death).

The violation types

The next step is to decide how the violation took place (for example, was it a beating, or electric shock, and so on). The tables below show the HRV categories and the types of violations within each.

KILLING	Code	Definition
Beaten to death	BEATING	Person is beaten to death by being **hit, kicked, punched**. State which part of body assaulted if known e.g., **feet, face, head, genitals, breasts**. If an object was used in the beating, specify the object, e.g., **sjambok, baton, gun, rifle, stick, rope, whip, plank, beat against wall.**
Burnt to death	BURNING	Victim is killed in a fire or burnt to death using **petrol, chemical, fire, scalding, arson**, but does NOT include Necklacing or Petrol Bomb (these are separate codes).
Killed by poison, drugs or chemicals	CHEMICALS	Killed by **poison, drugs,** or **household substance**, such as bleach or drain cleaner.
Killed by drowning	DROWNING	The person is drowned in a **river**, **swimming pool,** or even in a **bucket** of water.
Killed by electrocution	ELECTRIC	Killed by an electric shock.
Killing by death sentence	EXECUTE	**Hanged** or **shot** as decided by a formal body (court or tribunal) such as the state, homeland state, or political party. It is the consequence of a death sentence.
Killed in an explosion	EXPLOSION	Killed by any manufactured explosive or bomb, but NOT petrol bomb (see below). Explosives include **dynamite, land-mine, limpet mine, car bomb, hand-grenade, plastic explosives, detonator, booby trap, letter bomb, parcel bomb, special device** (e.g., walkman).
Killed by exposure	EXPOSURE	Person dies after being subjected to extremes such as **heat, cold, weather, exercise, forced labour**.
Necklacing	NECKLACING	Burnt with **petrol** and **tire**. Necklacing is coded separately from Burning, because it featured heavily in the past, so it is useful to distinguish between burning with petrol and a tire and burning in a house, for example.
Other type of killing	OTHER	All other methods of killing including **buried alive**, **strangling, tear-gas**, **decapitation, disembowelment.** Make sure that it is clear in the description of the act exactly how they died.
Petrol bomb	PETROLBOMB	Killed by a burning bottle of petrol. Petrol bombing falls in between burning and bombing, so, like necklacing, it is useful to code it separately. Also called **molotov cocktail**.
Shot dead	SHOOTING	Person is shot and killed by **live bullet, gunshot, birdshot, buckshot, pellets, rubber bullet**.

Stabbed to death	STABBING	Killed with a sharp object, such as a **knife, panga, axe, scissors, spear** (including assegai).
Suspicious suicide or accident	STAGED	Person dies in a **suspicious suicide** or **fatal accident**. This should only be used if it is not clear whether it really was an accident or not, otherwise use the appropriate category and explain in the description that there was a cover-up. Examples are: **slipped on soap, jumped out of window, fell down stairs, hanged themself, car accident, booby trapped hand-grenades or explosives, shot themself.**
Stoned to death	STONING	Person is killed with **bricks, stones** or other **missile** thrown at them.
Tortured to death	TORTURE	Person is tortured to death.
Unknown cause of death	UNKNOWN	Person is dead, but there is no further information.
Killing involving a vehicle	VEHICLE	**Dragged behind, thrown out, driven over, put in boot**, but NOT car bomb (see Bombing). Specify what type of vehicle was involved (for example, **car, train, truck, van, bakkie, hippo, casspir).**

TORTURE	**Code**	**Definition**
Torture by beating	BEATING	Person is tortured by being beaten severely or for a long time (for example, **hit, kick, punch)**. State which part of body was assaulted e.g., **feet, face, head, genitals, breasts**. If an object was used in the beating, specify the object, e.g., **sjambok, baton, gun, rifle, stick, rope, whip, plank, beat against wall.** Specify if victim is **pregnant** or **miscarried**
Torture by burning	BURNING	Person is burnt, with **cigarettes**, or **fire**, for example.
Torture with poison, drugs or chemicals	CHEMICALS	Tortured with **poison, drugs,** or **household substance**, such as bleach or drain cleaner.
Torture by deprivation	DEPRIVE	Person is tortured by withholding essentials, such as **food,** or **medical attention with serious injury or need** (this does NOT refer to a general lack of medical care while in custody. See Associated violations).
Electric shock torture	ELECTRIC	Electric shocks to the body. Specify which body part was shocked (for example, **genitals, breasts, fingers, toes, ears,** etc.).
Torture by exposure to extremes	EXPOSURE	Person is tortured by subjecting them to extremes such as **heat, cold, weather, exercise, labour, noise, darkness, light** (including flashing lights, blinding by light), **blindfolding, confinement to small space, smells, immobilisation.**

Psychological or mental torture	MENTAL	Person is tortured psychologically, mentally or emotionally, for example by **simulated execution** (includes Russian roulette**), solitary confinement, degradation** (includes use of excrement, urine, spit), **insults**, **disinformation** (for example, telling the person that a loved one is dead), **threats**, **witnessing torture**, **forced participation in torture, exposure when washing or on toilet, threat of torture.**
Torture by bodily mutilation	MUTILATION	Torture involving injuries to the body where parts of the body are partly or wholly cut, severed or broken. Specify body part, for example, **genitals, fingernails, ears, hair,** etc. It includes **amputation of body parts, breaking of bones, pulling out nails, hair or teeth, scalping.**
Other type of torture	OTHER	All other methods of torture. Make sure that it is clear in the description of the act exactly how the person was tortured. It includes **use of animals** (specify animal e.g., snake, tortoise, baboon), **use of vehicle.**
Torture by forced posture	POSTURE	Person is tortured by forcing the body into painful positions, for example, **suspension, helicopter, tied up, handcuffed, stretching of body parts, prolonged standing, standing on bricks,** **uncomfortable position** (includes squatting, imaginary chair, standing on one leg, pebbles in shoes), **forced exercise, forced labour, blindfolding** and **gagging.**
Torture by sexual assault or abuse	SEXUAL	Person is tortured by attacking them using their gender or genitals as a weak point. This does NOT include electric shock, mutilation or beating (instead, use those categories, and specify genitals as the body part abused). It includes: **slamming genitals or breasts in drawer** or other device**, suspension of weights on genitals, squeezing genitals** or **breasts, rape by opposite sex, rape by same sex, gang rape, forced sexual acts (e.g., oral sex, simulating intercourse), introduction of objects into vagina or rectum, sexual abuse using animals, threats of rape, touching, nakedness, sexual comments or insults, sexual enticement, deprivation of sanitary facilities for menstruation**.
Torture by suffocation	SUFFOCATE	Torture by stopping someone from breathing, for example by **bag, towel, tube (wet or dry) over head, drowning (head, whole body submerged), choke, strangle, stifle, throttle, teargas, bury alive.**
Unknown type of torture	UNKNOWN	Person is tortured, but the method is not known.

SEVERE ILL-TREATMENT	**Code**	**Definition**

Severely beaten	BEATING	Person is badly or severely beaten, or beaten for a long period of time. They may be **hit, kicked, punched**, **twisted.** State which part of the body was assaulted (e.g., **feet, face, head, genitals, breasts)**. If the person was beaten with an object, specify object (for example, **sjambok, baton, gun/rifle, stick, rope, whip, plank, wall).** Specify if victim is **pregnant.**
Injured by burning	BURNING	Person is injured by burning with **fire, petrol, chemical, scalding,** but NOT Necklacing or Petrol Bomb (these are separate. See below). Specify body part, if localised burning.
Injured by poison, drugs or chemicals	CHEMICALS	Person was poisoned or injured by **poison, drugs, household substance** (for example, bleach or drain cleaner).
Injured in an explosion	EXPLOSION	Person is injured by a bomb or explosives, but NOT petrol bomb (this is coded separately. See below). Explosives include **dynamite, land-mine, limpet mine, car bomb, hand-grenade, plastic explosives, detonator, booby trap, letter bomb, parcel bomb, special device** (e.g., booby-trapped Walkman).
Psychological or mental ill-treatment	MENTAL	Person is severely psychologically, mentally or emotionally ill-treated, for example by **simulated execution** (includes Russian roulette), **degradation** (includes use of excrement, urine, spit), **death threats**, **threat of torture.**
Bodily mutilation	MUTILATE	Person is injured by having parts of their body mutilated or damaged. Specify body part, for example, **genitals, fingernails, ears, hair,** etc. It includes **amputation of body parts, breaking of bones, pulling out nails, hair or teeth, scalping.**
Necklacing	NECKLACING	Person is injured in an attempted necklacing.
Other type of severe ill-treatment	OTHER	All other types of severe ill-treatment. Make sure that it is clear in the description of the act exactly how they were ill-treated. It includes **strangling, drowning, spreading of disease.**
Sexually assaulted or abused	SEXUAL	All forms of attack on a person using their gender or genitals as a weak point, for example **rape by opposite sex, rape by same sex, gang rape, forced sexual acts (e.g., oral sex, simulating intercourse), introduction of objects or substances into vagina or rectum, sexual abuse using animals.**
Injured in a shooting	SHOOTING	Person is injured by being shot with **live bullets, gunshot, birdshot, buckshot, pellets, rubber bullet.** Specify body part injured, if known.
Stabbed or hacked with a sharp object	STABBING	Injured with a sharp object, such as a **knife, panga, axe, scissors, spear** (including assegai).

Injured in a stoning	STONING	Person is injured with **bricks** or **stones** thrown at them.
Teargassed	TEARGAS	Severe injury caused by teargassing in a confined space (for example, teargas in a prison van or packed hall).
Suffocated	SUFFOCATE	Injury or ill-treatment by stopping someone from breathing, for example by **drowning (head, whole body submerged), choke, strangle, stifle, throttle, teargas, bury alive.**
Unknown type of severe ill-treatment	UNKNOWN	Person was severely ill-treated, but it is not clear how.
Injury involving a vehicle	VEHICLE	Injuries caused by being **dragged behind, thrown out, driven over, put in boot** of a vehicle. Specify the vehicle (for example, **car, train, truck, van, bakkie, hippo, casspir)**.

ABDUCTION	**Code**	**Definition**
Illegal and forcible abduction	ABDUCTION	Victim is forcibly and illegally taken away (for example, kidnapping), but the person is found again, returned or released. It does NOT mean detention or arrest. It is not a gross violation of human rights to be arrested (see Associated violations).
Disappearance	DISAPPEAR	Victim is forcibly and illegally taken away and is never seen again. It does NOT include cases where somebody goes into exile and never returns. It must be done by force. This DOES include people have disappeared but it is not clear why they have gone (instead of abduction, they might have just run away or were shot and buried). In this case, a finding will be made and the code will be left as it is, or changed to Killing if the person was killed, or found to be out of the mandate of the TRC.

ASSOCIATED VIOLATIONS	**Code**	**Definition**
Beating	BEATING	Person is beaten, but it is not a severe or prolonged beating. It includes **once-off mild beating.** Specify if **in custody** or if victim is **pregnant** or **miscarried.**
Violation after death	CORPSE	Body of victim was violated after death, for example by **improper burial, body mutilated or burnt or blown up, funeral restrictions, funeral disruption, anonymous burial, mass grave**.
Deprivation	DEPRIVE	Deprivation of facilities or essentials, for example **medical attention, food, water, sanitary facilities, privacy, family visits.**
Destruction of property	DESTROY	Includes violations such as **arson, destruction, vandalism, theft, forced removal, eviction**.
Financial impropriety	FINANCIAL	Person was subjected to **bribery, extortion, pay-off, ransom, blackmail, ruin of business**
Framing	FRAMING	Person is labeled as an **informer**, **collaborator** (impimpi) or **criminal, false information** is spread about the person, or a **smear campaign** against the person is started.

Incarceration or imprisonment	INCARCERAT	Includes **police custody, detention, house arrest, restrictions, banning, banishment, prison, informal prison**
Intimidation or harassment	INTIMIDATE	Victim is intimidated or harassed by **dismissal from work**, **threats**, **animals killed**, **visits**, **telephone calls**, **surveillance**, **boycott enforcement, pointing of firearms** (NOT in custody), **threat of violence.** It does NOT include vandalism or arson. This comes under Destruction of Property.
Other type of associated violation	OTHER	All other types of associated violations, including **released into hostile environment, released into unknown place, left for dead, rough ride, detention of family or loved ones.** Give full details in the description of the violation.
Sexual harassment	SEXUAL	Person is sexually harassed. It includes: **threats of rape, touching, nakedness, sexual comments or insults, sexual enticement, deprivation of sanitary facilities for menstruation**
Petrol bombing	PETROLBOMB	Severely injured by a burning bottle of petrol. Also called **molotov cocktail**.
Professional misconduct	PROFESS	Person was subjected to professional misconduct by one of the following: Health professionals (including **doctors, nurses, orderlies, clinicians, district surgeons, psychiatrists, psychologists** and others) who neglect or ignore injuries, collaborate in torture, or conceal the cause of death or injuries. Judiciary (**magistrates, judges** etc.) who ignore torture allegations, for example. **Police** who neglect the case, ignore or tamper with evidence. **Lawyers** who neglect the case, ignore or tamper with evidence, misappropriation of funds or failure to hand over damages. **Businesses** who collaborate with perpetrators.
Teargassed	TEARGAS	Victim was teargassed, but NOT while in custody (see Torture).
Theft or stealing	THEFT	Money or possessions were stolen from the victim.

Appendix 3

Lessons Learned

By the editors

Entity, Function	Lesson	Recommendation	Issues
Volume and complexity of information	Wide Area Network and development of own database facilitated work.	Don't even think of working without a network. Don't use "standard" human rights software	Scope and nature of networking. Whether to outsource software development, network, or database design
E-mail	Absolutely essential	Incorporate into system	Whether to have a closed system for e-mail, cutting off access to internet services and external e-mail
Network software and hardware	Domain structure of Microsoft NT complicated network management; stability of the servers compromised by shortcomings of OS	Choice of OS calls for intense study	Having individuals with sufficient experience and skills to make good judgements; getting sufficient time and funds to make a considered decision
Security of system	Security and Internet access can be achieved	Free-standing computers connected by dial-up to Internet is simple, reliable, inexpensive way to provide Internet access	Computer-literate users will be frustrated by the lack of outside e-mail connections
Ownership of information system	Contests of ownership and a high profile can be assured.	Have the persons with responsibility for the electronic information systems in a position reporting directly to the CEO	Having supporters of this recommendation in a position to make it happen
Ownership of data and information	Users may not take ownership of data they use until late in process	Get the users involved early in the project	Getting the message across to users
Corroboratio, research, getting findings	If the system serves several purposes with higher political profiles, corroboration, research, getting findings will be delayed	Work to maintain these activities despite distractions	Easy to say, hard to do. Stakeholders in the system are in conflict and highest political priority may take over
Data collection	Free-flowing narrative may be too slow, rigidly structured form may lose context	Balance these two requirements to produce a form appropriate to the job mission, conditions, and resources	Prior to some initial data collection, it may be impossible to make a good compromise

Process development time	Using industry-standard software reduce development time.	Use industry standard software when possible to do so while achieving goals	Evaluation of software in system context
Tracking information through the system to its successive sources	Details of the information flow progress of the documents useful for monitoring blockages in the system, finding the location of backlogs and monitoring the performance of staff members	Use a SOURCE_HISTORY entity, store various versions of the violation	Time to design
Acts of violation	Must be kept to a reasonable number	Reduce to a reasonable number by appropriate method	Finding “appropriate method.” At TRC, head processors and researchers could not reach consensus until top management mandated that consensus be achieved in a finite time. This approach may not work in all situations
Reparations	To support a reparations function, the system must record the consequences of violation to the individual and the resulting individual needs	Determine whether a national policy on reparations exists or is likely to be instituted	Getting a clear commitment on reparations from the start
Deponent statements	If a group of statements is analyzed and captured as a group they may be confused	Analyze statements individually and finish each before moving to next. Code statement without evaluating	Discipline

Chapter 5

The United Nations Mission for the Verification of Human Rights in Guatemala: *Database Representation*

Ken Ward

Introduction

The United Nations Mission for the Verification of Human Rights in Guatemala (MINUGUA) was created within the framework of the peace negotiations between the government of Guatemala and the National Revolutionary Union of Guatemala (URNG). In the Comprehensive Agreement on Human Rights signed on March 29, 1994, the parties asked the Secretary General of the United Nations to establish a mission for the verification of the status of human rights and compliance with the commitments of the agreement.

On September 20, 1994, one day after the UN General Assembly approved the establishment of MINUGUA, a technical team was sent to Guatemala to work out the logistical arrangements for the mission's installment. This included drafting a handbook on verification methods and the designing training seminars for the international monitors who were to verify the human rights situation in the country. MINUGUA was formally installed on November 20 and its first regional office was opened three days later in Guatemala City.

MINUGUA's mandate was to cooperate with national institutions and entities for the effective protection and promotion of human rights, sponsor technical cooperation programs, carry out institution building, and support the judiciary, the prosecutor's office and governmental human rights offices. Thus, its central role was monitoring and reporting on human rights violations.

By the time its first report was issued in March 1995, MINUGUA had eight regional offices and five sub-regional offices and a staff of 211 international members, including 72 UN Volunteers and 30 civilian police observers almost exclusively involved in human rights monitoring. By time the peace accords were signed and the mandate of the mission expanded to include other aspects of the accords, approximately 150 members of the mission were involved directly in monitoring human rights.

In addition to simply monitoring human rights violations, officials in the field offices worked to prevent human rights violations or intervened to prevent additional violations.

Methods for Recording and Processing Violations Data

The first step for monitoring human rights involved interviewing deponents either in the field or in a field office and recording the information gathered in the interview on a standard format sheet. This format included information on the human rights violation event (called a *case*), such as unique case number, when and where the violation occurred, name of the deponent or deponents, and description of the event.[1] A second sheet allowed for the recording of personal information of the victim such as name, birth date, home address, occupation, and ethnicity. A third sheet allowed for the coding of the violations alleged to have occurred and one perpetrating institution. The coding sheet allowed for up to three violations to be recorded. Thus, it was left to the office coordinator to determine which violation would be considered the primary or "most serious" violation.

If there were more then one victim in a given event, the office would determine which victims suffered serious violations and then open a case file for each victim. Victims of lesser violations in an event would be noted in the description of the case. For example, if an event involved the arbitrary execution of four individuals, the office would open four separate cases. If the event involved the attempted execution of one victim and threats against another, it is possible that only one case would be opened for the attempted execution and the threats would simply be mentioned in the case summary.

[1] The terms *case* and *event* are not used synonymously. Several cases might be generated by one event. This was 1992, early in the development of the AAAS methodology and definitions were not in a state of development. Some of this growth in understanding is evident in this paper and others.

Each office used a list control sheet to monitor the status of their cases. It consisted of a basic table with each row containing the event number, event location, victim name, primary violation, perpetrator, and the status of the verification.

Methodological Problems

By limiting recording of an event to its "primary" violation (that is, the violation deemed to be most serious), only one violation will be recorded for a victim suffering several. This is a gross understatement of the nature of the victimization of the individual and leads to a false view of the events and distortion of trends. To illustrate this latter problem, consider Table 1, below, representing the recording of counts of violations in this "one victim-one most serious violation" schema.

Table 1. Example of the recording of counts and violations in the "one victim-one most serious violation" schema

Violation	June	July
Arbitrary Detention	2	0
Torture	1	0
Arbitrary Execution	1	3

From this table, it would seem that the number of victims of Arbitrary Detention declined from two in June to zero in July. But given the three cases of arbitrary execution that happened in July, we cannot be sure that this decrease is real. The executed people may have been detained and tortured before they were killed, in which case detention and torture went up in July. Once data have been coded and represented in this way there is no way to find out what happened in each event.

Also, by separating victims of the same event into different case files, the relationship of the victims to the same event can be confused or lost. If an event involves many victims, many different violations and/or multiple perpetrators, important information on individuals and acts will be disassociated, hidden in text, or lost altogether.

Thus, when trying to analyze what happened, there is confusion as to what exactly was recorded. When the mission talked about human rights "cases," it was not clear whether a case referred to a single human rights violation against one victim, an event with many violations with only one violation having been recorded or one victim in an event where there were many victims. All of these interpretations are equally possible.

A second problem with recording only a primary violation involves information management. Since the functions of the field offices include prevention and intervention, in depth knowledge of the human rights situation in a particular region is essential for effective results. However, for any one person to understand the case history of an office that person would need either personal knowledge of the caseload or knowledge gained by extensive reading of individual files. For a new member of the human rights team to determine if a perpetrator had a history of committing violations or to determine if an individual had suffered previous attacks, it would be necessary to consult individual office members or review each case file from memory or individual notes, a time-consuming and arduous task.

Thus, the primary source of information about the connections between cases and events was the *individual employee*, who depended on memory or personal notes. At best, this is a poor solution to the problem. However, at MINUGUA, it was compounded by the continuing rotation of personnel in the regional offices. Police observers were usually assigned to the mission for only six months. UN volunteers rotated from one office to another after six months to a year.

Analysis of trends and patterns of violations were equally difficult without personal knowledge of each case. This problem was even more pronounced in the main office where verification officers worked from case summaries and lists sent from the regional offices. The consolidation of cases from several regional offices increased the workload for the individual verification officers at MINUGUA's headquarters and made it harder for them to extract hidden details of cases.

A Report Example

In March 1995 MINUGUA presented its first report to the Secretary General of the United Nations on human rights in Guatemala, including anecdotal cases of human rights violations and a table representing 288 cases of reported human rights violations admitted for verification, classified by violation.

A footnote in the table explained that when there was more then one violation per case, only the "most serious" was considered although "most serious" was not defined. See for example, Appendix 1, Table 1.

The table of the number of cases gave a possibly misleading impression of relative importance of each violation (as measured by rate of occurrence). For example: cases where the violation against the right to life (extra judicial killings, tentative killings, and death threats) was the primary violation represented 37% of all cases accepted. Cases where the primary violation was reported as violations against physical integrity represented only 23% and cases of personal liberty only 12%. There may have been a great many cases of personal liberty violations that were not deemed "primary violations" occurring in cases where right to life was the only recorded violation. A ranking based on the primary violation of the case might then lead to distorted understanding of the human rights situation.[2]

Implementation of the Database System

During the initial setup mission when the verification manual was designed and for its first year of use, no systematic methodology to structure all information relating to violation events was used. Nor did the project management make any plans for such systematization. Some team members present during that first year reported that the concept of creating a database for systematic processing, analyzing, and monitoring of the human rights violations was discussed but there was no follow-up.

With the support of the head of the human rights division, many division members, and apparently the head of the mission, preliminary steps were taken to hire a database designer to design such a system.[3] However, the head of administration of MINUGUA opposed the hiring of a database designer, arguing that MINUGUA's Electronic Processing Department (EDP) was capable of implementing the necessary database system as an administrative function, rather than as a central issue for substantive work. Unfortunately, at that time the staff of EDP was fully loaded in the work of maintaining the mission's computer systems in headquarters and in the field, and lacked experienced programmers.

In October 1995, almost one year into the mission, the first steps were taken to design and implement a human rights violation database. These steps came about casually after it was learned that I had a computer programming background, and had expressed interest in creating a violations database for my regional office. I was then a United Nations volunteer working as a human rights observer,

At the insistence of the Human Rights Division, the volunteer (myself) was transferred by the Volunteer's Office from the field to the headquarters office to create a violations database for the mission. Since the mission had never seriously considered implementing a database, they gave me full responsibility for the structural design of the database. The Human Rights Division wanted to do more accurate and sophisticated analyses and made the rapid creation of the database a priority. Working 16-18 hour days seven days a week, I designed, coded and tested the system in less than one month.

The final product was a database system based on the following two principles (Ball, et. al., 1994):

1. A human rights event is a collection of violations, victims, and perpetrators, all of which are related.
2. A person's role in an event is independent of his/her identity.

---

[2] The mission's periodic human rights report continued this format until its November 1996 report. In May 1995, Patrick Ball was employed as a consultant to MINUGUA and helped to change the ways in which these data were recorded and reported.

[3] I was not fully informed of the managerial decision-making process.

The simplified model used for the first year and a half could not represent this complexity; it reduced a case to one violation, one victim, and one perpetrator. However, most human rights cases or events are complex collections of one or more violations or acts, suffered by one or more victims, possibly at the hands of one or more perpetrators. In addition, it is possible that in each event not every victim suffers the same series of violations and not every perpetrator commits each violent act. I designed the new database to represent this complex structure of human rights cases and preserve information relating to the number of victims, acts and perpetrators. By using this structure, it would later be possible to recreate exactly who suffered what violation and who committed that violation.

In addition, a person's role at the time of a violation (victim, witness, or perpetrator) is not part of who that person is; rather, it reflects his/her place in a violation at a specific time. This was possible even though a person could be a victim in one human rights violation event, a witness in a subsequent event and a perpetrator in another. Therefore, the database represented individuals not as victims or perpetrators but rather as members of the list of all people who are in some way associated with human rights cases. Personal information on each individual was stored in the person's record, such as name, date of birth, ethnicity, etc. References linked the individuals to the roles played in each event. This structure allowed for accurately counting exactly how many victims of violations there were and permitted the analysis of patterns of behavior, for example, of a public prosecutor that is repeatedly involved in obstruction of justice cases.

User Interface Considerations

The lack of a formal database team, the frequent movement of personnel, and the wide distribution of the regional offices required that the system be designed with a strong emphasis towards the end user.

In truth commissions and tribunals the function of a database is to store large amounts of information on violations to be processed at some later date, usually near or at the end of the commission. In contrast, MINUGUA's primary functions were reporting on human rights violations and prevention and intervention. Hence, it was essential that the system provide quick access to extensive information on events. Such information included the actual state of verification, number and types of violations, the names of victims and perpetrators and their association with other events, relations among events, and results of interventions on behalf of the victims.

I did the initial training of field office members in the use of the system but in the long run, the system had to be as intuitive as possible so that subsequent training of new members could be delegated to the offices. Also, since there was minimal pre-coding of the interview formats prior to their entry into the system, the interface relied heavily on input controls, some custom made, to assure that input was quick and restricted to predefined parameters.

Although the unit of analysis for the system was the violation or the act, the interface followed the established logic of recording acts as being grouped into cases. Thus, the main entry point for the user to the system was a case list. Selecting a control for a new case would open up a *Case Window*.

- After the user entered initial event information (i.e., event location, date of event, text summary of case, etc.) the user could specify the deponent or deponents by selecting a button control for deponents. The mission could investigate cases on its own initiative without someone reporting it. In such cases no deponents were specified.
- If the deponent button was selected, a *Person Window* was laid over the Case Window with the title "deponent." Personal information about an individual (first and last name, birth date, occupation, ethnicity, etc.) was added on this form. A lookup list consisting of all persons who had previously been associated with cases in some role was displayed as part of the window. This feature allowed users to first search the list of names and compare individual information to determine if the person to be added was already part of the system. It they were already in the system, they could then be selected without having to re-enter their information. This would also assure that the same person would not be counted twice in later analysis.
- In keeping with our differentiation of *individual* versus *role* identity, there was not a field in the person window to mark this person as a deponent. The connection between this individual and their role as deponent in this specific event was automatically created when

the user finished entering the personal information and selected an *Accept* button closing this window and revealing the overlaid Case Window.

- The user selected a button for acts to add victims, their associated violations and alleged perpetrators.
- The same Person Window was then superimposed on the Case Window, but this time the title would specify that a victim was being added. Again, individuals could first be searched for in the persons lookup list or the user entered their information if they were new to the system. Selecting Accept created the relationship between this individual and their role as victim in this event.
- Since an act requires a victim, a violation, and a perpetrator, the following step would not return to the Case window but lead the user to a third window where a perpetrator (or perpetrators) and a violation (or violations) could be selected. Adding perpetrator(s) followed the same process as before and an individual or an institution (only possible in this case) is defined or selected. Once again, when the user accepted the perpetrator(s), their role in the event is established. Violations were selected from a control lookup list of possible violations. After specifying all pertinent information, the user selected the Accept button and the system created the relationships among victims, violations and perpetrators.
- Additional fields on the main case window allowed the users to add text for qualitative case follow up and analysis.

We completed the design of the database in November 1995, and users started installation and training in each of the 13 field offices. By the end of January 1996, every office had incorporated its prior caseload (created since the beginning of the mission) and added new cases as they arrived. Once a month at first, and later, every two weeks, the information was transmitted to the head office using electronic mail. There it would be consolidated with that of the other regional offices. To assure confidentiality the information was encrypted prior to transmission using Pretty Good Privacy (PGP) public key encryption software with keys of 1024 bits.

In December 1996, the mission hired a UN volunteer to work full time on maintaining and modifying the database system. This person was also in charge of producing statistical tables and lists used by the verification officers in the Human Rights Division and other areas for analysis, creating a standard list of statistical reports and performing *ad hoc* queries for data. These results were produced as hard copy and given to the requesting party.

Introducing a New Methodology

The effect of the ability to quantify human rights violations information was first seen in MINUGUA's fourth report to the Secretary General March 1996. In the fifth report released in November 1996, the change in the statistical table was explained as follows:

> In the Comprehensive Agreement, the parties requested the Mission to receive, consider and verify complaints of human rights violations and to determine whether or not such violations had occurred. In previous reports, the Mission included statistics on the complaints admitted, classified by the right affected in each case, with a proviso as to the extent which they formed the basis for its conclusions. (United Nations, 1996)

Taking only one complaint per case as a statistical reference made it hard to shed light on the full range of victims and human rights violations covered. To overcome that limitation and to provide a more representative assessment, the Mission modified its methodology for recording and systematizing the data from the verification process. The Mission designed and set up a database for recording complete information on all the victims and human rights affected in each case reported. Thereafter, when the report refers to violations, these were understood to mean each violation of each victim. This new methodology made it difficult to draw comparisons with previous periods because of the differences in the data structures.

The fifth report introduced a statistical table based on MINUGUA's new methodology. While maintaining the numbers of cases admitted which were classified by primary violation as in previous reports, it also included new columns for the number of individual reported violations, verified violations and confirmed violations (violations in which the state was determined to be responsible), all classified by type. In this table *admitted cases* of extra-judicial killings account for 61 cases where it is the primary or "most serious" violation, as opposed to arbitrary detention which ac-

counts for only 14 cases, a greater than four to one ratio of killings to detention. But if we look at the number of *reported violations* of extra-judicial killings compared to the number of reported violations of arbitrary detention we see the ratio is almost one to one (69 to 66). If we look at *confirmed violations*, arbitrary detention out ranks extra-judicial killings by three to one (18 to 6).

From my preceding review of the developmental process, it is clear how different database representations may lead to different views of reality. I feel that the original approach of counting only one violation per case presented a misleadingly simplistic view of the human rights situation in Guatemala. It is important to carry out the database structuring correctly, as the findings are dramatically affected by the nature of the system. Of course, in human rights situations it is hard to know exactly what the nature of reality is until data collection has been in process. Ideally, the database designers will create a design that is flexible and robust so that it can deal with changes as the project proceeds.

Effects of Quantitative Analysis on Reporting

During its first year and a half of operation, MINUGUA was not able to report information on trends and patterns of human rights violations. However, after the design and implementation of the systematized database, MINUGUA could use its easy access to a rich dataset to perform sophisticated analyses.

In its fourth report to the UN Secretary-General, MINUGUA concluded that during the period of the report, 44.8% of the violations of ill treatment occurred while an individual was being arbitrarily detained and that the National Police were responsible for the ill treatment 44.2% of the time (United Nations, 1996). There is no such statement in any of the three previous reports.

In the sixth report (United Nations, 1997), MINUGUA concluded that in 1995, nearly 45% of violations were attributed to the main institutional protagonists of the armed conflict (the armed forces, including military commissioners and Voluntary Civilian Defense Committees and the URNG—the coalition guerrilla organization).[4] This percentage declined steadily in 1996 to about 34%. Accordingly, violations attributed to the National Police, the Public Prosecutor's Office and the judiciary become a larger proportion; these three institutions, which in 1995 accounted for 32% of the violations reported, rose to 44% in 1996. Hence, the analysis of the data revealed a shift away from the attribution of responsibility for violations committed in the context of the armed conflict by institutions under the direct control of the parties to State institutions over which the Government has no direct control, except in the case of the National Police.

Using the new database, MINUGUA now regularly cites the percentage of violations committed by the different state institutions as part of its quantitative analysis of the human rights situation.

In addition to providing reliable statistics, the creation of a database has permitted quick access to case information allowing the mission to perform timely evaluation of compliance. The Peace Accords call for the creation of a new police force of a civilian nature, to be created from new recruits and "recycled" members of the old police force who have been evaluated to weed out those previously involved in the repressive actions of the past. Comparing a list of National Police members enrolled in the National Civilian Police Academy with confirmed cases involving police agents as perpetrators, MINUGUA was able to establish that the government had allowed the incorporation of almost 20 members of the National Police to the Academy who had previously been involved in human rights violations including several involved in extra-judicial killings.

The systematization of information has also allowed for the cross-referencing of information between offices. For example, where case histories of perpetrators from one office can be requested by another office in the event that the individual is transferred into its region as is often the case with police agents, public prosecutors and judges.

[4] When Ríos Montt took power, he expanded the civic action aspects of the counterinsurgency efforts, including the peasant militias, under the name "civilian self-defense patrols." During the subsequent transition to civilian government, the army changed the name to "Voluntary Civilian Defense Committees" and renamed local *comandantes* as "committee presidents."

Lessons Learned

Table 2. Summary of lessons learned and recommendations

Entity, Function	Lesson	Recommendation	Issues
Carrying out full analysis of large-scale human rights violations	A relational database is needed.	The establishment of such a database into the mission should be an essential part of the commission's activities.	Will the initial planning, often dominated by legal and political parties, have the knowledge and understanding of the need for and requirements of such a database?
Design and implementation of database	Without the self-directed proposal of a volunteer, it is uncertain whether or when the database would have been ready.	Incorporate database needs into the initial planning for the project. Do not depend on chance events, such as the possibility that someone on the staff will have the skills and volunteer to do the work.	Does achieving this recommendation depend on the presence on the commission of an advocate for such a database? If no knowledgeable persons are part of the managerial and administrative structure, can this recommendation be achieved?
Data structure and unit of analysis	Don't use the structure, "one victim-one most serious violation."	Follow the guidance in Ball (1994) for the data structures based on who did what to whom.	Database designers need to be familiar with the rationale discussed in these proceedings. Will they? Law enforcement often uses the "one victim-one most serious violation" method; users may not realize the implications in a human rights situation.

Conclusions

For the year and a half prior to the implementation of a violations database MINUGUA had only the capacity to draw broad conclusions about the human rights situation in Guatemala. Reports to the Secretary General of the United Nations — MINUGUA's official evaluation of the human rights situation in the country —relied almost exclusively on anecdotal evidence. The design and implementation of a large-scale relational database has changed that situation.

The implementation of a database allowed the mission to present a more profound analysis of trends and patterns of violations. The violations database has also allowed the mission to concretely signal government noncompliance of its commitments as in the case with the National Civil Police Academy and has allowed the fluid interchange of information between previously isolated regional offices.

A final note: As shown in this paper, the implementation of MINUGUA's violation database was *ad hoc*. Such a database was not incorporated into initial planning and apparently its importance was not understood by decision-makers until after a year and a half of operational experience. Even then, but for the availability and willingness of a skilled volunteer on the staff, we can only guess how much longer it would have taken to undertake a design and implementation project. MINUGUA could have made better-supported, stronger arguments at a much earlier time, exploiting the wealth of information collected by a large team if a relational database system had been planned and implemented from the start of the project.

Appendix I

Table 1. Second Report to the Secretary General of the United Nations, August 1995, Complaints admitted by category of presumed violations*

Right to Life	
Extrajudicial execution or death in violation of judicial guarantees	54
Tentative extrajudicial execution	25
Death threat	146
Total	**225**
Right to Physical Integrity Security	
Torture	10
Cruel, inhuman or degrading treatment	3
Physical abuse	4
Excessive use of force	6
Other threats	117
Total	**140**
Right to Personal Liberty	
Arbitrary detention	14
Detention in violation of judicial guarantees	20
Kidnapping	3
Taking of hostages	-
Forced disappearance	6
Forced, unjust or racist military recruitment	7
Total	**50**
Legal Due Process	
Procedural guarantees	20
Right to habeas corpus	2
Right to access to justice	54
Total	**76**
Political Rights	**2**
Right to Free Expression	**0**
Right to Free Association	**35**
Right to Freedom of Movement	**8**
Other Violations Relating to the Internal Armed Conflict	
Harm or suffering caused to civilians	20
Attacks against civilian property	7
Attacks against goods and services indispensable for the survival of the civilian population	1
Terrorist acts	4
Failing to protect health or religious workers	-
Participation of minors 15 years and under in the internal armed conflict	2
Total	**34**
Total	**570**

(*) The number of complaints by right violated may change during the verification process

Table 2. Fifth Report to the Secretary General of the United Nations, September 1996

	Complaints admitted*	Number of violations	Violations verified	Violations proven
Right to Life				
Extrajudicial execution or death in violation of judicial guarantees	61	69	13	6
Tentative extrajudicial execution	19	54	42	39
Death threat	101	267	91	53
Total	**181**	**390**	**146**	**98**
Right to Physical Integrity Security				
Torture	4	8	2	0
Cruel, inhuman or degrading treatment	7	10	9	5
Physical abuse	39	73	27	21
Excessive use of force	12	116	103	96
Other threats	78	1060	1010	973
Total	**140**	**1267**	**1151**	**1095**
Right to Personal Liberty				
Arbitrary detention	14	66	23	18
Detention in violation of judicial guarantees	7	21	14	13
Kidnapping	9	16	7	2
Taking of hostages	0	0	0	0
Forced disappearance	1	4	4	0
Forced, unjust or racist military recruitment	2	6	2	1
Total	**33**	**113**	**50**	**34**
Legal Due Process				
Procedural guarantees				
Presumption of innocence	0	0	**	**
Right to judgement by an independent, impartial and competent authority	1	1	**	**
Right to be tried within a reasonable period	2	7	**	**
Right to a defense and assistance by a lawyer	2	5	**	**
Right to an interpreter	2	3	**	**
Right against self incrimination	0	0	**	**
Right to all legal recourses	0	0	**	**
Right of habeas corpus	0	0	**	**
Access to justice			**	**
Interfering with the National Police, Prosecutors office, of Judiciary	7	512	**	**
Obligation of the State to investigate and punish	55	590	**	**
Right to compensation	0	2	**	**
Total	**69**	**1120**	****	****
Political Rights	**3**	**4**	**3**	**2**
Right to Free Expression	**1**	**4**	**2**	**0**
Right to Free Association	**7**	**64**	**43**	**36**
Right to Freedom of Movement	**5**	**882**	**879**	**873**

Other Violations Relating to the Internal Armed Conflict				
Harm or suffering caused to civilians	28	648	634	100
Attacks against civilian property	1	2	1	1
Attacks against goods and services indispensable for the survival of the civilian population	0	0	0	0
Terrorist acts	0	1	1	0
Failing to protect health or religious workers	0	0	0	0
Participation of minors 15 years and under in the internal armed conflict	0	0	0	0
Failure to protect and assist wounded and captured	0	0	0	0
Total	**29**	**651**	**636**	**101**
Total	**468**	**4495**	**2910**	**2239**

(*) The number of complaints by right violated may change during the verification process

(**) Verified and confirmed violations of due process are not quantified while the verification of the legal process continues

References

Ball, Patrick, Ricardo Cifuentes, Judith Dueck, Romilly Gregory, Daniel Salcedo, and Carlos Saldarriaga. 1994. *A Definition of Database Design Standards for Human Rights Agencies*. Washington, DC: American Association for the Advancement of Science and Human Rights Information and Documentation Systems International.

United Nations. 1996 "Report of the Director Of The United Nations Mission for the Verification of Human Rights and of Compliance with the Commitments of the Comprehensive Agreement on Human Rights in Guatemala (MINUGUA)." 4th Report: UN Doc. A/50/878, 24 Feb. 1996

United Nations. 1996.--. 5th Report: UN Doc. A/51/1006, 19 Jul. 1996

United Nations, 1997.--. 6th Report: UN Doc. A/51/790, 31 Jan. 1997

Chapter 6

The Recovery of Historical Memory Project of the Human Rights Office of the Archbishop of Guatemala: *Data Processing, Database Representation*

Oliver Mazariegos

Introduction

The REMHI (Recovery of Historical Memory) project in Guatemala originated at the Human Rights Office of the Archbishop of Guatemala (ODHAG), when the peace agreement negotiated by the Guatemalan government and the Guatemalan National Revolutionary Union (URNG) approved the creation of the Historical Clarification Commission (CEH). The mission of the CEH was to investigate crimes of the 36-year history of armed conflict.

The draft agreement allotted a working time of six months to one year for the CEH investigation. ODHAG was concerned about this limited amount of time for the CEH to operate. Familiar with the experience of El Salvador, ODHAG knew the difficulty of gathering evidence in such a limited time. They recognized the need for an in-depth investigation and preparation of a database that could be transferred to the CEH, and set up REMHI. The REMHI project was to provide a reconstruction of the country's history from the victims' perspective, not just supply a series of unprocessed lists and statistics to transfer to the CEH.

The concept of the task is what differentiates REMHI's work from other, similar organizations. REMHI's purpose was not to attempt to reveal or interpret the history, but to arrange and describe it through the voices of the very victims who, after all, had the best knowledge of the truth.

This project was conceived and initiated by Bishop Juan Gerardi Conedera at the end of 1984 and was communicated to the rest of the bishops in the country with the intent that it would be adopted by the Episcopal Conference *in toto*. The Episcopal Conference of Guatemala decided that each bishop should individually choose whether to carry through the proposed work in his own diocese. Accordingly, work on the project started on April 1, 1995, as the coordinated effort of ten of the eleven dioceses in Guatemala.

REMHI's work is defined as "interdiocesian" because it was the result of the dioceses' coordination and it is precisely from their involvement, commitment and especially their "taking ownership" of the project, that the project developed and enhanced its activities.

The project was therefore conceived not only as a contribution to the peace process, but also as a factor in the reconciliation and reconstruction of the social fabric. This is why a fourth phase known as "the return" was added to the initial three phases of the project (preparation, collection of testimonies, and analysis).

This fourth phase is the principal contribution that the project can give to assist in the reconstruction of the Guatemalan social fabric, for it started its work by listening to the demands and proposals of the people interviewed. The return phase continues at the time of this writing (mid-1999).

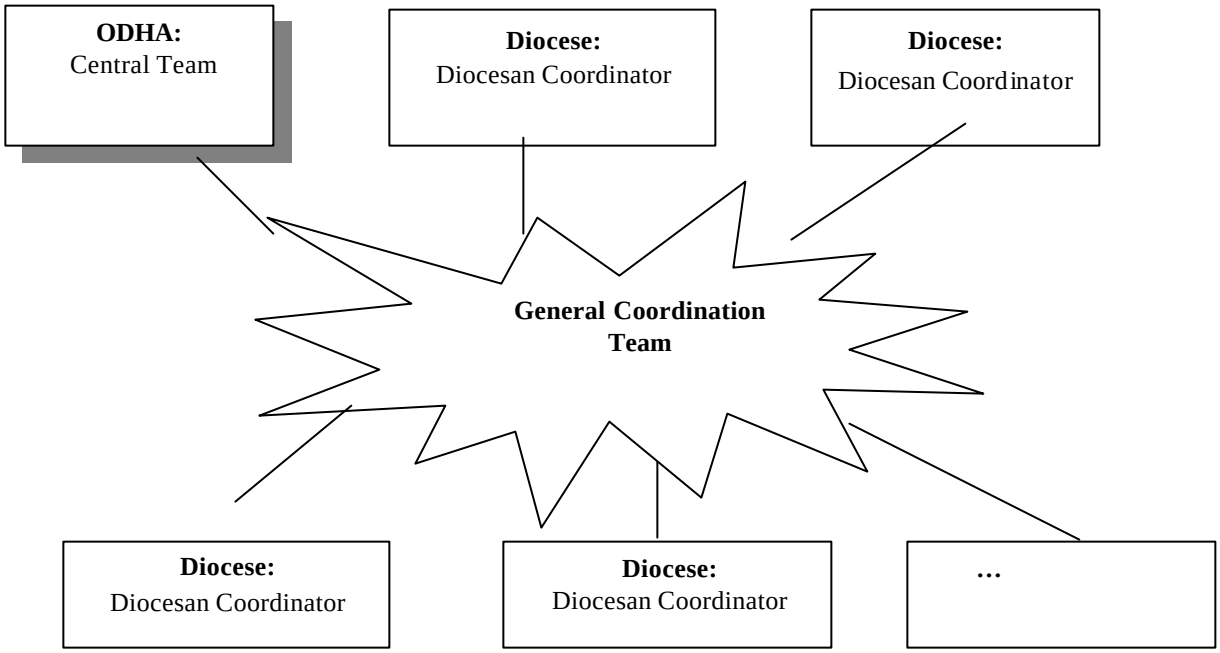
Work Methodology

The Human Rights Office of the Archbishop of Guatemala set up a work team whose function was to establish the necessary foundations to complete the proposed work.

This team — known as the Central Team — drafted an outline of the work methodology and completed the first project phase: preparation. The diocesan bishops designated trustworthy people to coordinate the work in their respective dioceses; they were the counterpart of the Central Team for work in the countryside.

Throughout the preparation phase, the Central Team outlined the work methodology. The diocesian coordinators, expanding the proposals presented by the Central Team completed these plans.

The functional diagram of the project follows:



Sources of Information

Direct interviews are the basis of the information used by the REMHI project. In addition, data were obtained by analyzing the print media ("journalistic monitoring") from 1960 to 1996, case studies (civilian defense patrols, women, etc.), interviews with key informants (perpetrators and experts on related subjects), declassified information provided by the National Security Agency (NSA), and a series of studies known as monographs. The latter were documents covering investigations the leading actors of the internal armed conflict (the church, guerrillas, etc.).

To complete the interview information, the experts used monographs as a starting point, and used journalistic analyses to obtain information on context; informants filled in any gaps.

Databases in the REMHI Project

To systematize the information on which the project would rely, REMHI noted the need for three types of information:

- **Human Rights Violations** derived from the interviews.
- **Journalistic Monitoring** based on reviewing the major cases published in print media during 36 years of war. (The design of the journalistic monitoring database was simple, almost to the point of being a flat database.)
- **Information on Military Structures** that included names, posts and positions of members of the Guatemalan armed forces.

It was possible to create databases for the first two types. Due to lack of information, it was difficult to complete the data for military structures.

Due to the straightforward nature of the journalistic monitoring database and the impossibility of building Military Structures database these two subjects are not discussed any further in this document. In the balance of this paper, I discuss the Interview Database, my central theme.

Interview Database

One of the most important and difficult steps at the start of the project was defining what was expected of the information system, since the response to this question had implications that would influence the total development of the project.

The main definitional difficulty was articulating the project objective. According to the original conception of the project, we were to assemble a database with statistical and documentary aims, which could quickly transfer information to the CEH.

I explain the design of the database as having three principal phases, in chronological order:

1. Specification of the interview form
2. Creation of the database
3. Analysis of the first interviews

Specification of the Interview Form

Our greatest analysis and design problem was the creation of the interview form. The creation of the form was a four-month effort during which a series of proposed forms were presented to the project's General Coordinating Team that had final approval of the design of the forms.

It was this interaction that allowed the preparation of suitable forms for the case. Although the experience of the Truth Commission of El Salvador was useful, the Guatemalan reality is completely different because the mechanisms of terror were different. Guatemala has a different geography and 23 linguistic groups. Most importantly, those who conducted the interviews —*reconciliation facilitators* — had a different background from the interviewers on whom the El Salvador Truth Commission relied. The interviewers in El Salvador were foreign, salaried professionals who did not know the country and its history. In the REMHI project in Guatemala, on the other hand, interviewers were volunteers, indigenous peoples or residents of the area where they conducted the interviews. Also, most were religious with low levels of education and in many cases, spoke Spanish as their second language.

Implementation of the Interview Form

As the project evolved with respect to its original conception, we saw that the interviewers' role was much more than a simple interaction with the interviewee. They were individuals of the same region, volunteers whose roles transcended the initial interview. The organizers, whose training included a diversity of functions, besides interviews, were mostly drawn to the reconciliation ministry that had developed in some dioceses as a result of REMHI's return phase.

It was precisely these volunteers, with their deep personal commitment, who changed REMHI's vision, and it was because of them that the interview form changed drastically during its development. Initially it was a technical form, specific and limited. The early version of the form was confusing and impractical for many and instead of serving as a helpful tool to conduct interviews, it created additional problems in the interview process.

The main obstacle to implementing the interview form was that our culture does not have a written tradition. Rather, it is a narrative culture. This tradition limited the interviewer's ability to collect and transfer narrative information since the continuity of the story was lost in filling out a series of forms. Another obstacle — which now appears obvious — was that the interviewers would have to develop an initial coding system for the interview and that structure was reflected in the interview form. The coding system entailed specifying the code of the responsible force (perpetrator), the classification of the crime, and descriptive information about the actors as social groups for those who were victims (unions, NGOs, ethnic groups, etc). This had further negative implications that were discovered later. It proved impossible to assure that 700 team members working separately throughout the whole country would exercise uniform criteria and make similar decisions.

To correct these obstacles and other relatively minor obstacles such as the length of the form, the size of the paper, etc., we developed a less technical and more practical new set of forms. These forms helped guide the thread of the interview and allowed a more complete collection of information in a more orderly narrative manner. This important development was achieved through a continuing series of corrections that often seemed to be interminable.

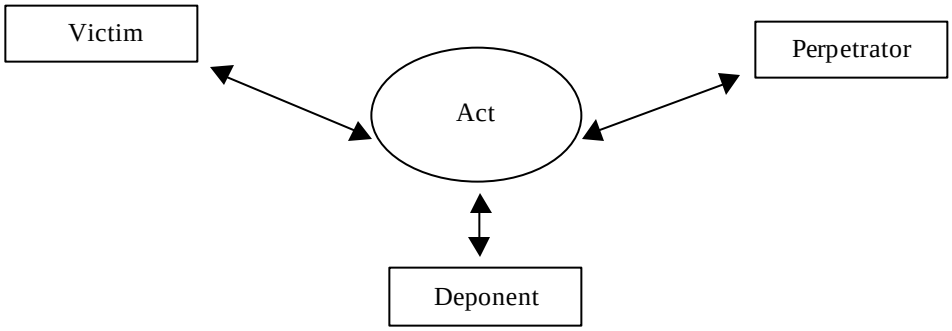
In addition, we made a series of practical changes to improve the interview form. These included using larger letters and more readable fonts, providing ample writing spaces for organizers with little aptitude for writing, numbering the forms, identifying the different sections included in the form (deponent, victim, cover page, etc.) with different colors, etc. With prior authorization of the deponent, a tape recorder could be used to record the complete interview.

We also included specific spaces in the form to record information relating to coding. The purpose of this provision was to facilitate data processing.

The final version of the interview form is shown in Appendix 1.

Creating the Database

As already indicated, the starting point for the design of the database was the preparation of the interview forms. However, it was not until the time we designed the database that we considered issues relating to treatment of exact information on violations of human rights.



We recognized that the primary goal of the database was statistical documentation. To serve this goal, we attempted to classify the greatest possible amount of information: sociodemographic data, individual data, information relating to time and space, etc. The main challenge in this process was to break down the information to a level that would make possible the reconstruction of the facts.

Logical schema

Our concept of the appropriate information methodology was based on the following logical principles: There are three actors: one **victim**, one **perpetrator**, and one **deponent** (on whom we rely). These three people are related to each other by one act, the **violation**.

These *parts* (or *roles*) that individuals play cannot be fixed nor are they exclusive. The deponent can be the victim in another violation, or the victim or the perpetrator can be the same deponent, etc. Besides, the result of counting of these four units of information can be zero in the perpetrator's case, or multiple, since in a violent act there can be various victims, various perpetrators, different deponents or various abuses.

This large number of possible combinations was the main complication in the design. It led to a series of questions that were difficult to resolve. At first we required that the database tell us **who** did **what** to **whom**, and in addition, **who** reported this information. This requirement greatly complicated counting the actors, since the greater the breakdown we tried to achieve, the more complicated it was to maintain a structure (*links*) that would permit us to reconstruct the facts from the systematized information.

We confronted such dilemmas as how to create a database that in addition to showing the victims and cases would tell us exactly what the deponents reported. Thus, we tried to create a database that could relate what a certain deponent stated and who the deponent identified. In case another deponent mentioned another victim or other victims later, it was necessary to know the level of overlap that the interviews presented in order to affirm that deponent *x* mentioned victim *m* while deponent *z* mentioned victim *m*, and also deponent *n*.

At the level of database design this situation could have been easily resolved. However, such a solution would have complicated inputting the information to a database and in the long run

would have been impractical. In view of this situation, we settled on only maintaining the link by case. Hence, we would know who were the perpetrators, the victims and the deponents.

As a result of this reasoning, the structure represented above has the case at its center. This is the linking mechanism for the three actors (victim, perpetrator, and deponent) to achieve the goal of indicating who did what to whom, and furthermore who told us. From this point we can choose the most convenient unit of analysis, which could be the interview, victim, violation, victim per violation, etc. The important thing was that the database should not limit this choice so that we could make a final decision later, since we were not set on any of the three choices from the start.

Data structure

For the definition of a fact, we considered that a fact could contain different violations, each one with its own respective data (date, place, responsible force, etc.) that bear a close relation such as causality, context, etc. This definition, similar to what Patrick Ball defines as *context*, is what permits us to differentiate a series of violations committed together against one or various victims from another series of violations committed independently one by one. It is what permits us to maintain the relation in a disappearance-torture-murder *modus operandi* and differentiate an act of torture and murder performed on one victim but carried out in a different context.

Below is a schematic representation of the data structure. We explain it, working from left to right:

Interview Number	Collective Case Number	Fact Number	Violation Number	Order

- The *Interview Number* is the unique identifier of the case.
- The *Collective Case Number* is information compiled where there was a group of victims in which some people suffered certain violations different from the rest of the group. Our need for this became apparent when the first tests were made with the database. The most common case was that in which a group of people were detained, subsequently tortured and the women sexually abused, and some people did not survive the torture and died. In this case, three numbers of different patterns were specified: one for the victims that were disappeared and tortured; one for the victims that were disappeared, tortured and sexually abused; and another for victims who were disappeared, tortured and who died (extrajudicial execution). However, in order to maintain the unit in these cases and subsequently regroup them, a number was assigned to the common, collective case for these three patterns.
- The *Fact Number* is information that permits us to group those violations that bear a direct relationship, such as torture and extrajudicial execution, which occur in succession. These violations were counted with one common fact number so they could be linked.
- The *Violation Number* is a foreign key that directs us to a table of violations in which pertinent data, such as the date, place, type, etc., are specified.
- The *Order* indicates the sequence of cases since there could be more than one violation in the pattern. Thus, if a person was detained and tortured, the detention would have Order number 1 and torture, Order number 2.

  A proposed alternative to this division would have been to create a disappearance/torture pattern for all of the victims, and add sexual abuse or death as a second and third pattern. It is important to note that this structure of collective patterns can become confusing since a detached look at it reveals an artificial division of a concrete case that should be kept as a single unit.
- The utility of the interview number was that it retained the relation of the database information with the written documents.

For performance considerations, other data were added, for example, the type of violation. Although this variable appears in the table of violations, this field allowed us to make general calculations without needing to relate it to the rest of the tables in the database.

It is important to know how relations were handled in the database:

- The deponents were related by their interview numbers. Since a number of deponents could have spoken about the same case, the interview number would tell us the indirect relation.
- The victims were related by the fact numbers. Note that we assume that the victim suffered all of the violations shared by the same pattern number.
- The individual perpetrators were related by way of the violation number, since they are different from the victims in that one cannot assume that the perpetrator participated in all the specific violations of the pattern.

The treatment of massacres

Most of these inconveniences were resolved, but there was one issue we discovered in the form preparation that could not be resolved until the creation of the database. This issue related to the treatment of the massacres: How were we to handle the massacres? The answer to this question depends on the answer to a larger question: How do we define a “massacre”? To define a criterion by which to label a case as a massacre is more complicated than it appears to be at first sight. It could be according to the number of victims, the number of violations, the brutality with which the acts were committed, etc. Finally, as we often did for other decisions, we opted for an unconventional, but functional definition. We defined *massacre* as that confusing act (from the deponent’s/victim’s point of view) whose final objective was **communal** destruction.

This involved considerations of intention, outcome, etc. Clearly, no one criterion exists to label or not label an act as a massacre; it was a **policy** that permitted analysts of the interviews to have a parameter of evaluation that was sufficiently broad and yet as precise in measure as possible.

We designed the interview database to meet these criteria and followed them closely in the majority of design decisions that we made.

For practical effects and with the intent of facilitating database input, cases that concerned one victim were differentiated from those that concerned more than one victim. Thus, it was at the interface level of the database and the forms that the differentiation between *individual cases* and *collective cases* took place. The flow charts for the three cases -- individual violations, collective violations and massacres -- are shown in Appendixes 2, 3, and 4, respectively.

It was anticipated that for massacres we would have to use a slightly different structure, since the count for actors and cases was highly variable. We realized that the compiled information would be complex, confusing and above all, incomplete. For that reason, our first database design did not consider the systematization of information relating to the massacres. We decided to wait on the first interviews to make the necessary decisions with real cases and detailed information.

However, massacres were coded simply as massacres, and did not use the detailed codes of kinds of violations. The omission of the detailed violation types from the massacre coding -- a representational error -- created an accidental bias. Many people in massacres were raped, tortured, and disappeared, but they were not coded as having suffered these violations -- they were coded only as having suffered "massacre." The other violations were not coded. (Victims of "massacre" were treated in the analysis as having been killed but not counted as having been raped or another other violation) As a result, after the non-massacre and massacre data were mixed together, the statistics reported for all violations (except death) actually meant 'violations excluding massacres' because if those violations did happen in the context of a massacre, they were not recorded in the database. But this was not how the statistics were interpreted.

A higher proportion of massacres were committed by some perpetrators than by others, and this proportional difference among perpetrators was greater for massacres than for some other violations. The representation error biased the proportion of responsibility attributed to some perpetrators relative to other perpetrators for some non-death types of violations. Perpetrators who committed more massacres were artificially counted as having committed proportionally slightly fewer non-fatal violations.

Data Processing

With the final design of the interview forms and the first database design completed, we started our work on the processing of data. This function was supposed to last approximately three months. It depended on a work team of five to eight people whose task was to input the data from the interviews.

Coding

For this work, two major tasks were identified, *coding* and *data-input*. Coding was the task of assigning codes to diverse classifications on which we relied, such as the place of the events, sectoral classifications, responsible forces, etc. Data-input was the task of transcribing the forms on paper to the database system.

Due to (1) the nature of coding, the mechanism designed for data entry, (2) the short period of time needed for the team to accomplish the work, (3) the status of computer technology at the time (the beginning of 1995), and (4) the systems analyst's experience, it was urgent to start the work as soon as possible. We decided that inputting the information to the database would be done with a text-based interface and that subsequently we would create a system using a graphic user interface for data query.

The Human Rights Office of the Archbishop of Guatemala at the time relied on a Novell Netware 3.1 Network Operating System with an Ethernet protocol using star topology installations. A small computing center was established with four workstations with a topological bus for cost reasons. To avoid overloading traffic on the network, an additional network card was installed in the server exclusively for the computing center. The database was developed on FoxPro 2.6 for DOS, the same as the journalistic database.

Analysis of the First Interviews

With the start of information compilation, the first interviews from in the countryside came in. They were the input to the first tests and evaluations of the designed system. The objective of these tests was to determine the efficacy of both the manual and the automatic procedures designed for the information system.

In addition, we had a series of detailed situations about which we had not had sufficient information to make decisions in the design phase. We hoped to get information from the incoming interviews that would give us more hard facts to work with. For example, the treatment of massacres was among the main problems. Prior to receipt of interviews from the field, information on what was expected was often incomplete and sometimes confusing.

Although some case information was specific enough, most interviews gave inexact references that complicated quantifying the violations in a massacre ("They separated the women and raped the youngest ones"). For this reason we opted to quantify the number of dead and disappeared victims within a list of victims that did not specify if the victims listed were dead or disappeared ("We never heard of them again"). Since this decision left out cases of torture, rape, threats, attempts, etc., we covered this gap with qualitative explanations, such as the *modus operandi* of the massacres. (In the end it turned out to be more valuable to indicate that in most massacres women were raped, rather than indicate the number of rape victims which would in any case be an approximate figure.)

Thus, at the level of database structure, the massacre became a new type of violation with a violation number and its own pattern number that would make use of the existing data structure.

Working with the massacre interviews was more complicated than working with the individual and collective violations because the data came from a much larger number of interviews, which created conflicting versions. There were similar names, and we did not know if these represented the same people (e.g., José Antonio Velázquez versus José Velasco), contradictory data (especially the date), and above all, differing versions because of each deponent's statement as to what he or she could observe.

To have a level of specialization that would allow us to determine whether an interview was related to what was mentioned in another interview, we had the analysts distribute the work by geographical areas. This helped to determine the exact date of a massacre, for example.

Consequently, the list of victims was maintained separately from the main list, but kept exactly the same structure. However, from the information on which we could draw we managed to obtain the name, sex, age, and at times, the ethnic group identification of the victim.

Among the new scenarios encountered upon receiving the interviews and entering them into the database, we noticed that when there was more than one deponent for the same case, we would come across data that could be either complementary or contradictory. For example, one deponent might report a number of victims and another deponent gave us a different number. Even worse were cases where one deponent informed us of a disappeared victim and another deponent mentioned the death of the same victim.

Since the project did not investigate or dig deeper into the interviews we received, in many instances we lacked sufficient resources to disqualify an interview. The answer to this dilemma was that we would have to adapt the database so it was able to store different versions of the same case.

This decision implied a potential artificial inflation of the statistics. Therefore, at the time of calculating the statistics we had to make decisions to resolve this problem to avoid biasing the results. At the level of the database structure we resolved this problem in the following manner:

1. The information was complementary. For example, one deponent is specific about the date of the violation, but the other deponent is not. We would then modify the violation previously stored in the database, use the same pattern number, violation number, collective number and order number, but specify a different interview number.
2. The information was contradictory. We recorded everything anew as if it were its own case so that in the end, we could group the patterns by victim and decide which of the different versions we would use in the final analysis.
3. The information was neither complementary nor contradictory. Duplication was taken into account in the creation of statistics and final lists with the aim of not artificially inflating the statistics.

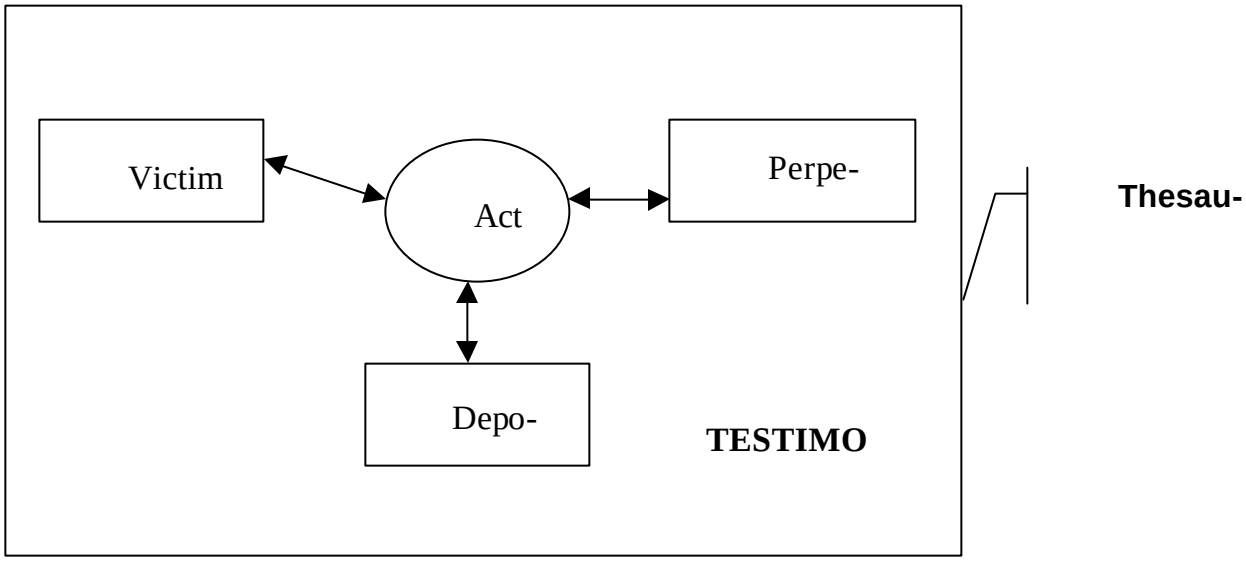
It was during the analysis of the first interviews that the analysts and investigators discovered the great potential of interviews as investigative material. However, until that time we had not taken measures at the level of the database so that we could recover this information.

The Thesaurus

Since this material was mainly qualitative information, the cost of incorporating it in the database made it an almost impossible task. We therefore created the *Thesaurus*, which was a list of keywords identified by project investigators. The words dealt with subjects such as the *modus operandi*, effects on victims and their families and communities, demands, proposals, cultural questions, ethnic issues, etc.

The Thesaurus was initially proposed by investigators according to subject —religion, perpetrators, effects, demands, etc. — and throughout its use was enhanced by the information processing team. The Thesaurus is summarized in Appendix 5.

The Thesaurus-based system was the tool on which investigators depended when maximizing the narrative capacity of those interviewed. In this way we hoped to conduct a detailed investigation (the individual effects on women in a certain region in the western part of the country, for example) that would cross the base information regarding violations with Thesaurus keys to obtain a list of interviews mentioning the subject. Thus, our conceptualization of information would remain as is specified in the following figure:



This new complexity and revised use of the database created the need for an interface for the database that should be easy to use. The new interface would allow investigators to perform reference and cross checks in the database. We developed this new interface with FoxPro 2.6, in a Windows 3.11 graphic environment.

Data Input

The data processing team (*coders*) had to carry out tasks and develop methods that had not at first been contemplated. Among the most important were transcribing the interviews, in some cases, six hours in length. This called for analyzing the Thesaurus, interacting with analysts, and discussing the parameters and policies that guided how decisions were made (such as the case of the difference between a disappearance and a forced disappearance). Inputting information to the database was a process that ultimately involved 18 people and took 20 months.

Once information input to the database was complete, we created cleaning processes to reduce duplications in the database. We did this even though from the beginning, the computerized system indicated the actors whose first name and surname coincided with data that was specified at the time the information was inputted.

To calculate descriptive statistics, we exported the database to Excel and through pivot tables (dynamic crosstabulations) we were able to perform most calculations and create desired charts.

Lessons Learned

Problem	Solution	Issues
Lack of uniformity. Everyone did not always understand the policies and take similar actions in similar situations.	The decisions were made by the coding team, which took into consideration the opinions of all project personnel. Important decisions concerned the violation type, use of thesaurus, classifications. Internal workshops to structure the discussion, training in different aspects (gender, ethnic affairs, etc.) and sharing of experiences.	Sometimes, the discussions seemed annoying and tedious, but in the end were perceived as helpful. The quality and profile of the coding team is an important factor in success of the discussions. The coding team was the key source for every detail.

Finding qualitative information. It was hard to find qualitative information integrated into 6,000 interviews.	Set up and used keywords (Thesaurus).	The elaboration of the thesaurus is sensitive work; anything not specified in it will be untraceable.
Mixed violations.	Don't do it again! Treat massacres just like all other violations.	
Lack of Graphics.	Next time, include more graphs	
Horror of codifying.	Any code used in the database should have a zero value option.	
Control of existing work.	Even though the input of a whole interview can take a long time, the input of general information about the interview itself is a task that consumes little time. Thus, every time interviews were received, the coding team inputted into the database the id # of the interview, and some general information (date, place of interview, and so forth)	This practice proved to be helpful for other purposes such as control of flow, distribution of work, interviews tracking, etc. When controlling the development of the activities of the coding team, it's easy to know how much has the team done, but hard to know how much is left. All you can do is to make an estimate.
Fatigue, emotional issues.	Workshops to discuss these issues. Be creative. Don't ignore this issue!	Working in data entry in a database that deals with human rights violations means more than keypunching. The "key- punchers" are people who must deal with atrocities and horrors, the pain of others, etc.
Where to start in database design.	Read Ball, *Who did What to Whom,* Washington: AAAS (1996).	Don't try to re-invent the wheel, find out what has already been done.
Incomplete information.	Build a system capable of managing incomplete data. We made printed forms for the victims of massacre. Since the original forms used one sheet per victim and most of the data was missing, we made a special form for listing the victims, their names, gender, date of birth and ethnic group.	Try always for the highest level of completeness of data. However, when working with this type of information (from a period of 5-35 years), it is certain that much of the data will be incomplete and imprecise, especially dates.
Dispersion of decision-making.	Log decisions, so you can gather all the decisions in the data analysis phase.	It is impractical and not advisable to centralize the decision-making process in one person. The process of decision-making, is carried on throughout the course of the project and is distributed in space, time and throughout the organization. An inevitable risk that must be dealt with.
Lack of ready access to, or availability of information. .	Easy-to-use interfaces so anyone can sit at a computer and search information.	
Success of the project	Those individuals who worked in more operational tasks (interviewers, encoders, etc.) are the best source of evaluations, ideas and understanding of how to make the project a success.	Facing the fact that preparation, capacity and experience of the people who design, structure and direct the project is necessary, but not sufficient.

Appendix 1

Interview forms

Cover page

Interview number	five digits
Date of interview	dd/mm/yy
Location of interview	
Type of violation (mark those that pertain)	Killing (extrajudicial execution Massacre (more than five dead) Assault Detention (forcible disappearance) Torture Threat Other
Property Loss	Yes, no.
Location of violations	
Date of violations	dd/mm/yy
Was the interview recorded?	Yes, no
How many cassettes were used?	
Additional pages	Victim Summary Perpetrator Deponent Other documents

Victim

Interview number	five digits
Victim number	eight digits
Surname(s)	
Given name(s)	
Other names used	
Name not known	
Gender	Male, Female
Pregnant?	Yes, No

Documentation	Official identification papers Birth certificate Baptismal certificate Refugee or displaced person document Passport Other
Document number	
Expires in	
Birth date	dd/mm/yy
Age	count
(Year)	
Place of birth	Canton Village *Caserio*[1] Plot *Finca*
Place name	
Place identification	eight digit code
Municipality	
Department	
Country	
Mother tongue	plus two digit code
Profession or office	plus two digit code
Where were you living when the act occurred?	
Father's given name	
Mother's given name	
Marital status	Single Married (living together) Widow (er) Divorced or separated
Given name of spouse	
Total number of children that you have (live and dead)	Given name, status
Of which groups a member? (Political, military, social, community, trade union, refugee, displaced person, etc.)	Category of group, name of group, dates, duties
Comments	

[1] *Caserio* is a smaller division than a village. Several *Caserios* comprise a village.

Summary

In narrative form, answers to the following questions:

1. Who was the victim?
2. What followed, where and when?
3. Who were the perpetrators of what followed?
4. Why did the acts follow?
5. What motives provoked the acts?
6. What was done to confront the situation?
7. What must be done to avoid a repetition of these acts of violence?

Perpetrator

Interview number	five digits
Perpetrator number	eight digits
Surname(s)	
Given name(s)	
Other names used	
Gender	Male, female
Documentation	Official identification papers Birth certificate Baptismal certificate Refugee or displaced person document Passport Other
Document number	
Expires in	
Birth date	dd/mm/yy
Age	count
(Year)	
Place of birth	Canton (district) Village *Caserio* Plot House
Place name	

Place identification	eight digit code
Municipality	
Department	
Country	
Mother tongue	plus two digit code
To what force(s) did the perpetrator belong when the acts were committed?	Name of the force Duties plus five digit code
Who was responsible for these violations?	plus two digit code plus one digit code
How is it known that this person was responsible	
Is it known if this person participated in violations in other, different cases?	
Is it known where this person is now?	
Comments	

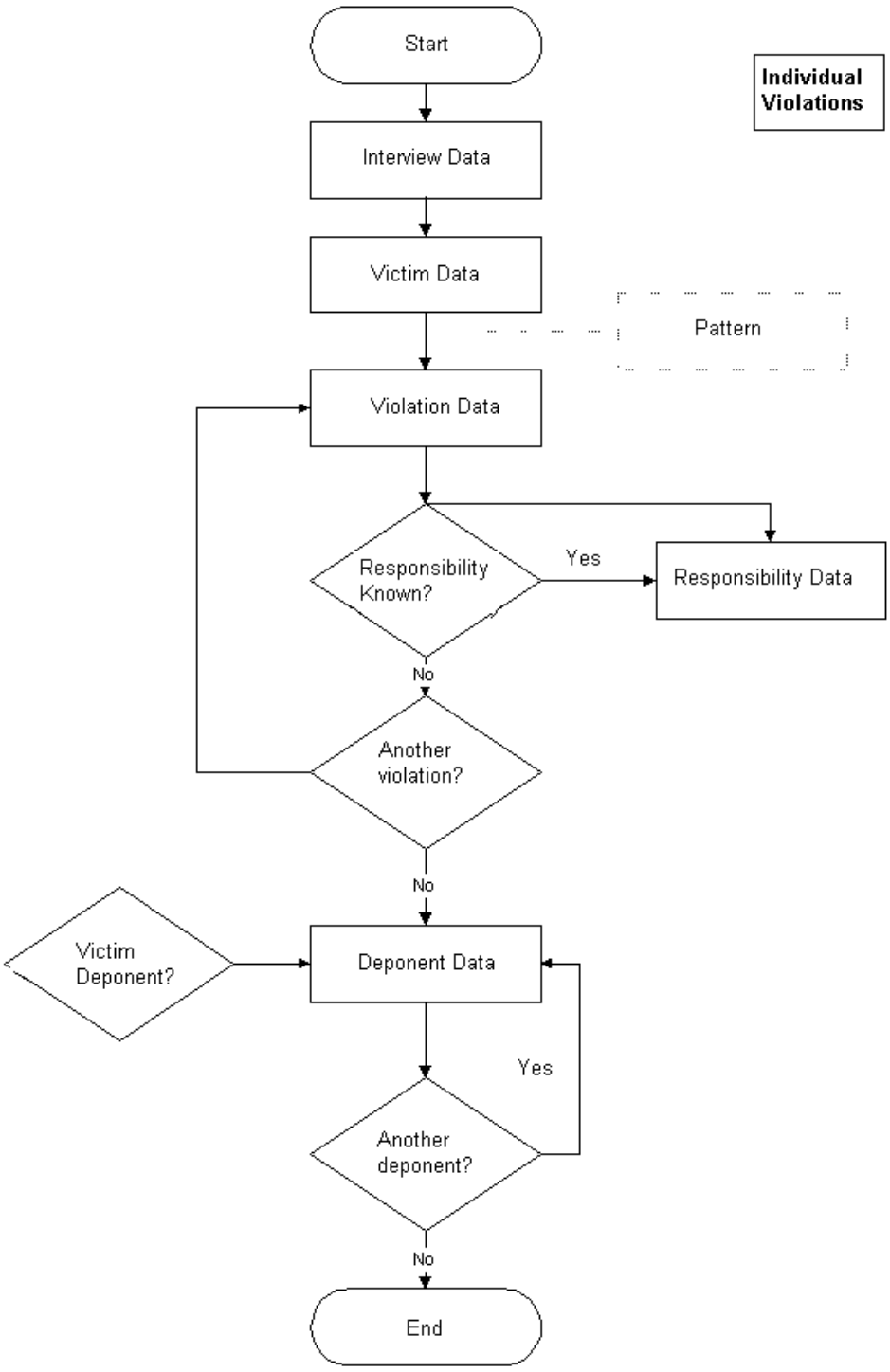
Deponent

Interview number	five digits
Deponent number	eight digits
Are you the victim?	Yes, No
Were you present during the acts?	Yes, No
Surname(s)	
Given name(s)	
Gender	Male, Female
Birth date	dd/mm/yy
Documentation	Official identification papers Birth certificate Baptismal certificate Refugee or displaced person document Passport Other
Document number	
Expires in	
Mother tongue	two digit code
Age	count
What is the relation between you and the victim?	
Do you know how to read and write?	Canton (district) Village *Caserio* Plot House
Are there other persons who know something about the acts?	Yes, No
If it were necessary, can we meet with you another time?	Yes, No how?
Do you give your permission to present your testimony to the Truth Commission, including your identity?	Yes, No
Do you give your permission to present your testimony to the Truth Commission, without your identity?	Yes, No
Signature (mark) of the deponent.	
Date	dd/mm/yy

Language in which the interview was conducted.	
Comments	

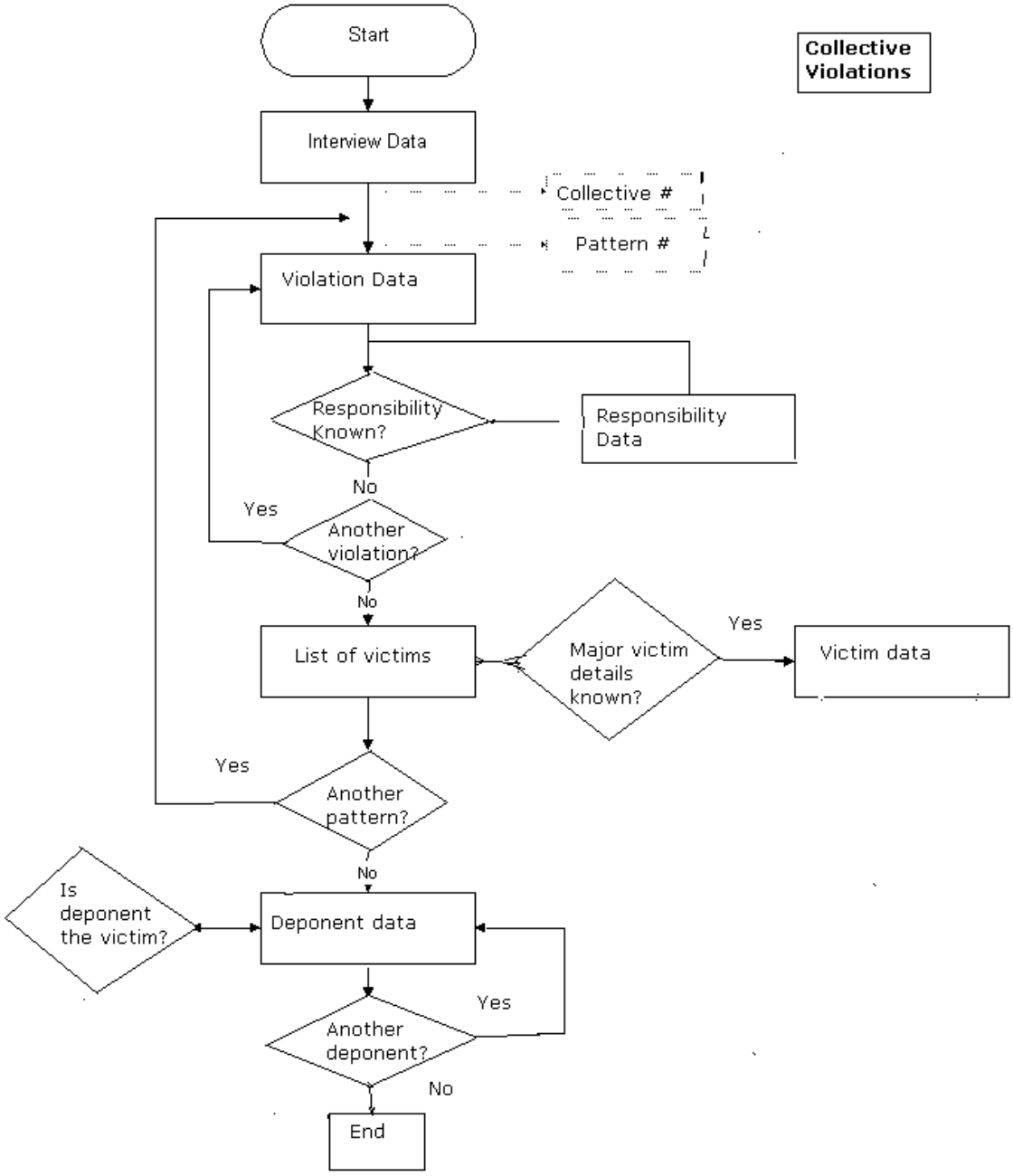
Appendix 2

Flow chart for individual violations



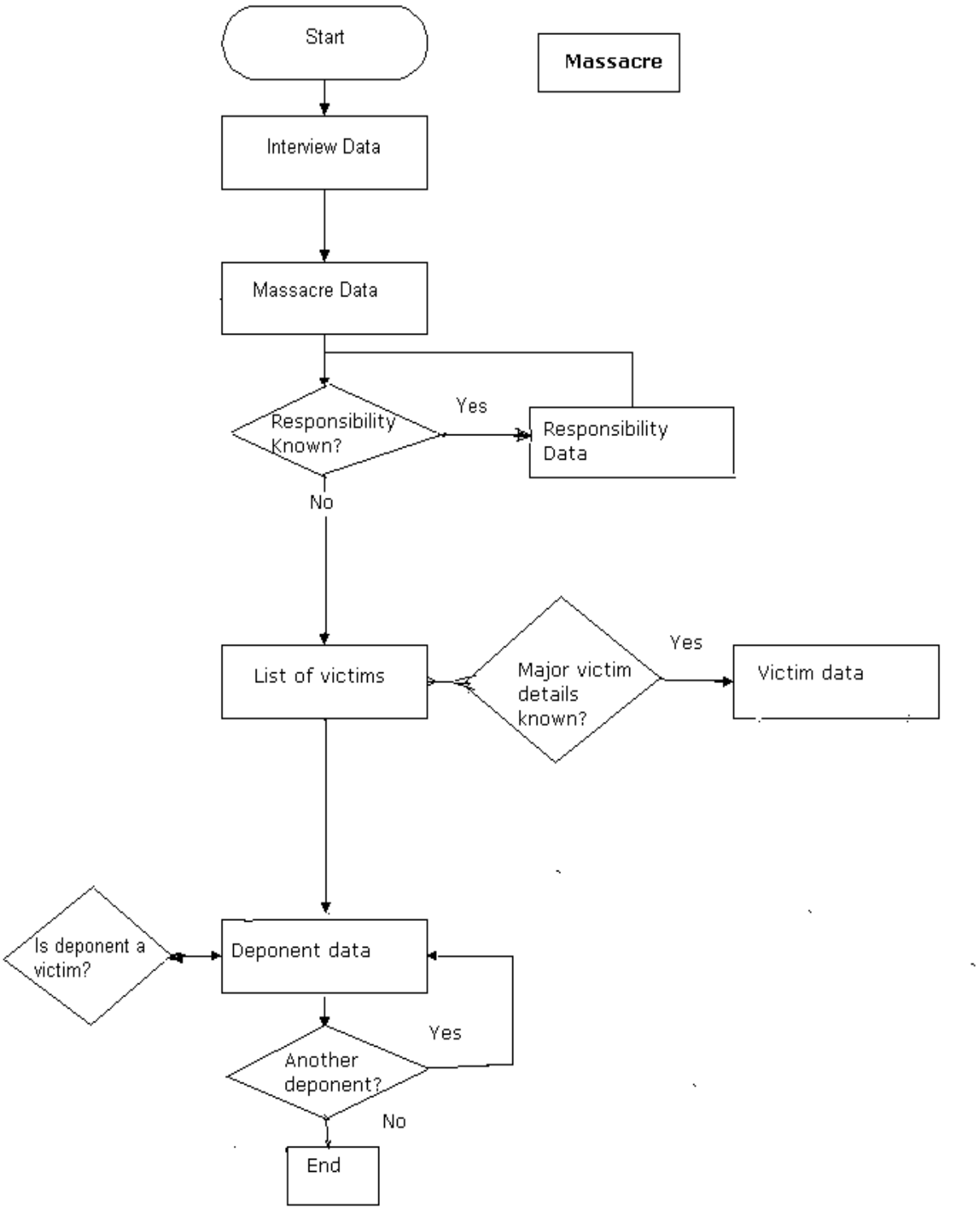
Appendix 3

Flow chart for collective violations



Appendix 4

Flow chart for massacres



Appendix 5

Thesaurus

Types of Violations

1. Death caused by:
    - 1.1 Extrajudicial execution
    - 1.2 Indiscriminate attack
    - 1.3 Bomb
    - 1.4 Artillery
    - 1.5 Explosives
    - 1.6 Mines
    - 1.7 Crossfire
    - 1.8 Other
2. Death resulting from Persecution:
    - 2.1 Suicide
    - 2.2 Hunger
    - 2.3 Illness
    - 2.4 Accident
    - 2.5 Other
3. Forced Disappearance:
    - 3.1 No Reappearance
    - 3.2 Reappeared Alive
    - 3.3 Reappeared Dead Date of Reappearance: ___/___/___
    - 3.4 Unknown
4. Disappearance:
    - 4.1 No Reappearance
    - 4.2 Reappeared Alive
    - 4.3 Reappeared Dead Date of Reappearance: ___/___/___
    - 4.4 Unknown
5. Forced Detention
6. Torture:
    - 6.1 Cruel and inhumane treatment
    - 6.2 Torture
7. Sexual Violation
8. Attack against personal integrity with injury:
    - 8.1 Knives, etc.
    - 8.2 Firearm
    - 8.3 Bomb
    - 8.4 Artillery
    - 8.5 Explosives/Mines
    - 8.6 Other
9. Attack against an institution or group with damage:
    - 9.1 Firearms
    - 9.2 Bomb
    - 9.3 Artillery
    - 9.4 Explosives/Mines
    - 9.5 Other
10. Threats against people:
    - 10.1 Bomb Alarm
    - 10.2 Death Threat
    - 10.3 Intimidation
    - 10.4 Other
11. Threats against an institution or group:
    - 11.1 Bomb Alarm

11.2 Death Threat
11.3 Intimidation
11.4 Other
12. Irregular Detention
13. Other Violations

Responsible Forces

1. Army
   1.1 EMP: General Presidential Staff (or Estado Mayor Presidencial)
   1.2 DSP: Office of Presidential Security (or Dirección de Seguridad Presidencial) (Archive)
   1.3 Presidential Guard
   1.4 G-2 Place
   1.5 S-5 Place
   1.6 *Kaibiles* [2] Place
   1.7 Traveling Military Police
   1.8 Specialists Place
   1.9 Ministry of Defense
   1.10 General Defense Staff
   1.11 Air Force
   1.12 Brigade
   1.13 Military Zone
   1.14 Military Base
   1.15 Special Command
   1.16 Outpost
   1.17 Other
2. Police
   2.1 National Police
   Section
   Station
   Substation
   2.2 Special Command
   2.3 National Guard
   2.4 Municipal Police Place
   2.5 Judicial Police Place
   2.6 Other
3. Combined Forces
4. Irregular Forces
   4.1 Commissioned Soldiers Place
   4.2 PAC: Civilian Self-Defense Patrols Place
5. Death Squads
   5.1 Mano Blanco
   5.2 ESA: Secret Anticommunist Army (Ejército Secreto Anticomunista)
   5.3 NOA: New Anticommunist Organization (Nueva Organización Anticomunista)
   5.4 JJ: Avenging Jaguar (Jaguar Justiciero)
   5.5 Other
6. Insurgent Forces
   6.1 EGP: Guerrilla Army of the Poor
   6.2 ORPA: Organization of the People in Arms
   6.3 FAR: Armed Rebel Forces
   6.4 PGT: Guatemalan Workers' Party
   6.5 Unitary Front
   6.6 URNG
   6.7 Other

[2] Special Task Force

7. Unidentified
    - 7.1 Civilian clothed
    - 7.2 Uniformed
    - 7.3 Unknown
8. Mayors, Farmers, Private security forces, etc.
9. Others

Types of Responsibility

1. Lone vigil (patrolling or similar)
2. Participation in violation of Physical Integrity (executing, torturing, etc.)
3. Participation in violation against Property (burning houses, destroying crops, etc.)
4. Intellectual Responsibility (commanding)
5. Collaborator

Chapter 7

The International Center for Human Rights Investigations: *Generating Analytical Reports*

Herbert F. Spirer

Introduction

In this case study, I review my work in conducting the analysis of the data and generating the graphs and tables for the joint International Center for Human Rights Investigations (CIIDH) and American Association for the Advancement of Science (AAAS) report on Guatemala (Ball, Kobrak and Spirer, 1999). The purpose of the report was to use statistics in conjunction with historical analysis to tell the story of state violence in Guatemala from 1960 to 1996. The published report of 154 pages contains 42 graphs, 9 tables, and numerous statistics appearing in a text that largely reflects the information in the figures. Despite the small number of graphs and tables in the final report, it was informed by many hundreds of figures, analyses, and statistics, created over a nine-month period.

I give a summary of the lessons learned from this work, and make recommendations to help others working on similar projects. The project organization, analytical tools, and working relationships used on this project are generally related to those used by industrial analysts. In view of the growing use of large-scale datasets in human rights, I expect that with time the human rights field will develop its own approaches to data analysis. This paper is intended to be a contribution to that developmental process.

The statistical methodology (described below) used in the CIIDH/AAAS Guatemala project work is straightforward and well established; neither sophisticated nor novel methods were used. Because of the need to maintain the highest standards of credibility, the dominant issue in the statistical analysis was the avoidance of error and control of the process of generation and use of analyses. For that reason, my focus in this case study is to show how we met that need.

I believe that we were effective in meeting the standard of credibility necessary for a human rights report that established state responsibility for political violence. Other workers in this field should be able to use knowledge of this case study to achieve the same standard, and may do so more efficiently. In the section **Lessons Learned**, I review the lessons learned on this project, make recommendations, and discuss how those lessons could be applied in future projects.

There was considerable similarity between the process of generating analytical reports for the CIIDH and CEH projects as carried out by the analysts. Accordingly, the analysts for these two projects, Eva Scheibreithner and myself, jointly prepared recommendations for future large-scale human rights data analysis. These recommendations appear in Appendix 1 of Chapter 10, Data Analysis Recommendations.

Preliminaries

Data Analytical Methods in the CIIDH Report

Descriptive statistics

Descriptive statistics are measures that summarize and describe the overall characteristics of a dataset. In general, this refers to a set of well-known statistics derived from the data set that describe one or more variables. For this project, the descriptive statistics are the number of observations, mean, variance or standard deviation, minimum, maximum, median, and the sum for each of the variables in the dataset. I used descriptive statistics in this project largely as a means of detecting, correcting, and avoiding error.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is "about looking at data to see what it seems to say. It concentrates on simple arithmetic and easy-to-draw pictures. It regards whatever appearances we have recognized as partial descriptions, and tries to look beneath them for new insights." (Tukey, 1977:

p. v.) EDA uses the statistical measures of descriptive statistics, and in addition a number of other methods that involve creative analysis and interpretation. These methods include tabulations and crosstabulations, time series plots, scatterplots, transformations of variables, autocorrelation analysis, regression analysis, difference analyses, and others. Generalizations drawn from EDA may be extended to a larger universe, but cannot be given meaning in terms of mathematical probability.

Inferential statistics

In statistical inference, the analyst generalizes from sample data to make probability-based statements about the larger universe from which the data were obtained. These probability statements are usually expressed in hypothesis test results or as confidence intervals. For example, in the CIIDH report, Appendix 5, Monthly Seasonal Variation Analysis, I use a hypothesis test to infer that the observed monthly seasonal pattern of killings and disappearances is essentially certain to have been caused by an organized plan.

The Data

The CIIDH database is a relational database consisting of cases culled from press sources, documentary, and direct testimonies. CIIDH team members collected over 10,000 cases from newspapers, by reading every newspaper published during the 36-year period of armed conflict in Guatemala. Four thousand additional cases came from documentary sources such as the archives of several Guatemalan non-governmental organizations and the publications of the Justice and Peace Committee of the Guatemalan Church in Exile. Members of the CIIDH team directly collected over 5,000 testimonies for inclusion in the database.

We define a case as the information given by a single source (press report, or interview, or document) concerning violations that are reported as having happened at a particular time and place. "Violations" are instances of violence, including killings, disappearances, torture, kidnapping, and injury. "Victims" are people who suffer violations. A case may be simple (one victim who suffered one violation) or complex (many victims, each of whom suffered many different violations). In the CIIDH analyses, the unit of analysis is almost always the violation.

The basic data with which I worked were contained in four flat datasets (two-dimensional tables of information without established relationships to other tables), each with variables chosen from a common set of variables. Complex Structured Query Language[1] (SQL) queries and extensive programming produced these datasets with variables selectively chosen from the listing of variables shown in the data dictionary of Appendix 1. Unfortunately, the variables did not keep the same definitions in all data sets.

The four basic datasets were denoted by ctanon, ctnmd, rtanon, and rtnmd as indicated in Appendix 1. In this terminology, the prefix "ct" denotes complete, in that these are the data net of overlaps among data sources (interviews, documents, and periodicals). The prefix "rt" denotes reduced, in which the source categories "other" and "non-CIIDH interviews" were folded into the "documents" category.

The suffix "anon" indicates that the dataset consists of both anonymous and named violations for which victim identification exists, and the suffix "nmd" indicates that the dataset consists only of precisely named violations.

I also worked with four additional datasets in which only killings appear with additional variables to describe the nature of the killings and the size of the group in which they occurred. These datasets carry the additional suffix "k". All datasets were followed by "v" with an integer (1, 2, 3,…) suffix to indicate the version of the dataset. By the completion of the report, the version number had reached 16. Table 1 is a summary of the datasets.

Table 1. Datasets.

Data set name	Description
ctanon	Complete, anonymous plus named

[1] Structured Query Language is a computer language used to retrieve, update, and manage data.

rtanon	Reduced, anonymous plus named
ctnmd	Complete, named
rtnmd	Reduced, named
ctanonk	Complete, anonymous plus named, killings
rtanonk	Reduced, anonymous plus named, killings
ctnmdk	Complete, named, killings
rtnmdk	Reduced, named, killings

Data Control Documents

Data control document refers to the data dictionaries, dataset descriptions, variable position dictionaries, and derived dataset descriptions.

By *data dictionary*, I mean a tabulation of the names of each field, the values that can appear in each field, and a verbal description of the meaning of each field variable. Some data dictionaries include the dataset list. However, for my purposes, I separated the descriptions of the data from the description of the files. As mentioned earlier, Appendix 1 is the data dictionary for this project.

The *dataset description* includes as a minimum the name of the dataset, the number of records, and in this case, the number of violations. In some cases, I also included the number of killings. Appendix 2 is an example of such a description.

By *variable position dictionary* I refer to a summary by dataset of the columnar position of variables, which implicitly shows whether a variable is in a particular dataset. The need for this control document was because of the use of Excel for statistical analysis. Appendix 3 shows a variable position dictionary.

The *derived dataset description* includes as a minimum the names and brief description of datasets derived from the source datasets discussed under Background, The Data. It may also include Excel versions of the underlying basic data that were received as a file in xBase format, with the extension dbf. The derived dataset description also includes the source data, where relevant, and comments. Appendix 4 shows part of a typical derived dataset description, with my original footnoted comments.

I updated all of these documents with successive versions.

I used control documents on this project for two purposes. First, I needed them to keep track of the rapidly growing number of files and versions and the field names and values, which were changed during the analysis phase. Second, they played a role in checking for error. As will be discussed in detail later, every new version and revised configuration for a working file was tested with respect to its predecessor. When I could directly predict the expected effect of a change, I used these documents to verify that the expected changes occurred or to explain their absence.

Checking for Data Integrity

Statement of the Problem

There are many challenges to the integrity of the data. I considered every transition involving a dataset a potential source of harmful alteration of the data. Transitions are events in the transfer, conversion, and use of the data. Most of these events occur in the use of data in any form of analysis. For example, the analysis of data can reveal inconsistencies, outliers or suspicious results that result from errors in the working data set. These errors must be corrected and thereby result in new versions of the working datasets.

However, as described in detail in this section, most of the transitions are the result of the overall methodology of this project. Throughout the analysis period, this project was a work in progress, with a strong research component. The results of a particular analysis could reveal fea-

tures that were not anticipated, calling for a recoding or revised query of the dataset. Consequently, transitions were frequent.

For this project, the general sequence of transitions was as is shown below in steps 1-12:

1. Patrick Ball (PB) creates a file.
2. PB transmits the file to me (HS) in a .dbf or .xls format, whichever was convenient for one or both of us as an e-mail attachment. We had to use both formats because of initial conflicts in Excel versions. My actions would then be to:
3. Download the file in native form and archive it.
4. Convert the file to Excel format and archive it.
5. Make a working copy of the file.
6. Filter, reorder, consolidate, summarize, and otherwise manipulate the data to facilitate a desired analysis.
7. Perform the desired analyses.
8. Transmit the results to PB.
9. Create the graphs.
10. Transmit the graphs to PB.
11. Revise the graphs in accordance with format and analytical needs through joint exchanges with PB and Paul Kobrak (PK).
12. Transmit the graphs as attachments by e-mail.

The likelihood and form of the data integrity challenge at a transition is dependent on the transition and the circumstances. For example, I cannot recall an instance in which we found an error resulting from download transmission or format conversion (2-5, 8, and 12, above). However, I only know that these transitions were error-free because I was checking the results. I had many errors – often minor -- develop in the other transitions, which were detected and corrected. Our concern for even minor errors was to avoid the possibility of any negative effect on our credibility.

There were also challenges to the integrity of our results that relate to handling the data. For example, Excel apparently has internal instabilities, or as yet undocumented capacity limitations. On a number of occasions I returned to a workbook several days or weeks after creating graphs and found that the graph had disappeared or that formatting features were altered. I never had this problem in a small worksheet. Archiving the original data and any revised datasets that entered into analysis is essential. However, this action is another transition where the integrity of the dataset itself is in jeopardy from the failure to archive the latest version, or the inadvertent deletion of a file.

Throughout the process described above, I carried out different levels of checking, as I judged appropriate, as discussed in the next section.

Verification Methods

My approach to verification is based on applying descriptive statistical methods to the dataset or pair of datasets (before a transition and after a transition). By definition, summary statistics reduce the information content of the data to facilitate an understanding of the whole set. I show the descriptive statistical measures used for numerical and categorical variables listed in Table 2.

Table 2. Descriptive statistics used for verification

Numerical	**Categorical**
One-way tabulation	One-way tabulation
Record count	Record count
Crosstabulation	Crosstabulation
Extremes (high, low)	
Mean	
Median	
Sum	

For a single dataset, I look for reasonableness in the values. For example, if a dataset contains the variable SEXO for gender, a one-way tabulation should show some number of males and females, which may be coded "m" and "f". What other value might reasonably appear in the tabulation? If there has been agreement on the representation of unknown gender values as d (for *desconicido*), then we expect some number of d's to appear in a one-way tabulation. If I find no d values but a number of –1's, then I suspect that there may have been a change in the assignment of unknown values in this dataset. Of course, this would have to be reconciled.

But if **both** –1's and d's appear, then something is seriously wrong. It may be miscoding or a more fundamental problem. Or perhaps, the tabulation includes blanks. What might be signified by a blank, a missing value that was not properly coded or entered, an input error, or a blank record (which may reflect a serious error)?

With two datasets – one before and one after – I look for a reasonable comparison in the values. If there are two datasets, and the second is one in which records have been removed from the first dataset described above, then only m, f, or the missing data value should appear, and in no case with a higher count than in the first set.

Extreme values of numerical variables (maximum, minimum) can be a symptom of a problem. If there are a large number of numerical values, a one-way tabulation is usually more confusing than revealing. Extreme values may be outliers in the sense that they either are unreasonable or differ greatly from the normal range of deviations. For example, although –1 might be used as a missing value indicator for ages, what do we make of a –2 also appearing in the dataset? Is a maximum age of observation of 95 an error?

Comparison of the median and mean values is a quick way to determine skewness of the distribution of numerical data. To carry out this comparison, the analyst needs a sense of what the distribution of the data is, or should be, or how it would be changed by some transitional step using before and after comparisons.

The sum of columns is a simple check and it is easy in Excel to maintain sums of numerical fields at the bottom of the dataset. I monitored sums and record counts on a continual basis while working with a particular dataset. Using the sum on a continuing basis is a process that has its own problems, because of the automatic selection of data by Excel for certain procedures, and my own errors.

Many of the desired analyses are crosstabulations, and in themselves provide a basis for checking the dataset integrity. While I infrequently made crosstabulations as a check on a dataset, I almost invariably compared marginal totals in crosstabulations to the values produced by independent one-way tabulations.

It is tempting to think of automating these checks and verifications to reduce the dependency on human intervention. Without automation, some person has to make a conscious effort to carry out the check. But with automation, you may have another source of errors and lose the judgmental insight that can only come from knowledge of the data and what its attributes should be, or are

most likely to be. Accordingly, during the project, I used only the Excel built-in functions (where appropriate) to obtain values for the verifications discussed above. However, as will be discussed in Lessons Learned, I also used an intermediate approach.

In the final analysis, human intervention is critical. In one case, the routine checks suggested a possible incorrect coding. To track this down, I visually scanned approximately 10,000 records by observing the "play" of the patterns on the screen as I scrolled rapidly through the dataset. Using this method, Patrick Ball found a coding error, traceable to key entry at the source.

There is no substitute for vigilance and scrutiny.

Examples of errors

The following are a few examples of errors that were detected in the process of analysis. Some related to problems existing in the database or the query process. These have significance for project personnel other than data analysts and therefore have general interest and applicability in the management and implementation of the information system. However, the overwhelming majority of errors were the results of my own actions, occurring on a continual basis and which, by and large, can only be generalized to the need for each individual analyst to work constantly at avoidance, detection, and correction of error.

Early in the project, time series plots showed a midyear peak in violations with a clear, pronounced peak of violations in the sixth month, June. At first we were concerned only with revealing this pattern, but attempting to find out why such a pattern should exist led to the investigation of the coding process by which violations were assigned to a particular month. When the precision of the date of the violation was one year (that is, the violation could only be placed somewhere or at some time within a particular year), the violation was arbitrarily assigned to June. This resolved the problem of giving it a date, and would not affect any analysis of annual patterns. However, when the data were summarized by month across all years, the number of violations in June was improperly inflated by violations that could have happened at any time during the year.

When analyzing the patterns of collective and individual killings that required the use of named datasets, I routinely summed killings by individual and obtained the maximum and minimum values in the column of sums. A minimum of zero would indicate the presence of a zero due to one or more entry errors, corruption of a cell, or records that should not have been in the data set. A maximum above 1 would indicate miscoding, entry, or corruption errors. Two different cases were uncovered by this check:

1. In the early phases of analysis, I found instances where an individual was reported as suffering more than one death. This anomaly resulted from more than one source of data reporting an individual's death. This problem was traced back to duplicate reporting leading to miscoding.
2. In another case, the same individual was reported as killed by the same source at different dates. This is a genuine error, but I found only one.

In the data description associated with a dataset and the data block associated with an analysis, the number of violations is reported. The dataset rtanonkv7 contains only killings and hence, its violations total should have been the same as the count of killings in its source dataset, rtanonv7. Observation revealed that it was not the same, 34,747 compared to 34,659, a difference of 85! While this is an error of only 0.2%, we could not overlook it for reasons of credibility and because it might reflect larger compensating errors. On examination, Patrick Ball found that those 85 death records were reported as more than one of the three death killing categories -- cadavers, individual, and collective. His new program brought the two totals into agreement.

Early in the analysis, a one-way tabulation of ages in the named dataset showed ages of 0 and -1. Both values had been used to represent missing values of age. Conflicts in the number of missing values found at the same time were traced back to a revision of the coding process that caused the loss of the ages of 540 people (out of about 10,000, depending on the dataset).

Performing the Analyses

Statement of the Problem

In the data analytic aspects of this project, our goal was to describe, summarize, and explore the data. By and large, our mission was not to infer some parameter from a sample but to reveal the facts inherent in our data. The broader interpretation of these facts derived from an incomplete

coverage of the actual events must – as it did – come from the interplay and conjunction of the quantitative knowledge gained from the data and the equally relevant anecdotal and qualitative knowledge of subject matter experts.

Our data could not be obtained by probability sampling, which would have enabled the use of inferential statistics and its related disciplines of statistical hypothesis testing and confidence interval estimation. However, we did use probabilistic approaches to evaluate the apparent monthly pattern (Ball, et. al, 1999: Appendix 1).

Thus, most of the tools that are usually called "statistical methods" in the educational process and in much research did not apply to the analyses used in the body of the CIIDH report. The challenge in this project was to apply simple methods to complex, large-scale datasets in such a way that the voice of the data is heard and understood by both the knowledgeable members of the project team and the lay audience of CEH, researchers, and the interested public.

Methods of data analysis

Accordingly, we used **summary statistics, tables** and **graphs** as our primary tools of analysis. In Excel, graphs are called "charts," reflecting the orientation of Excel to business applications. Most statistical programs (e.g., Stata) call them graphs, as we do in this case study.

Our use of tables did not extend beyond the two-way crosstabulation. In our analyses, we used graph formats (e.g., logarithmic axes) and types (e.g., scatterplots) that we did not present in the final report. In fact, with few exceptions, the tables and graphs appearing in the report are fully described in the AAAS/HURIDOCS handbook, *Data Analysis for Monitoring Human Rights* (Spirer and Spirer, 1993). One exception is the **comparative histogram** that relates absolute and relative rates of killing by age (Ball, et. al., 1999: Figure 16.2), another is the time series plot of percent of victims by age that uses stacked line plots (Ball, et. al., 1999: Figure 11.4).

Since readers of this paper may want to relate our approach to the formal discipline of statistics, we reiterate that we have used the tools of *descriptive statistics* -- describing, presenting, and summarizing data to reveal or gain a better understanding about the processes that created the data. Exploratory Data Analysis (EDA) is a related set of techniques for understanding, analyzing, and presenting data, its structure and systematic patterns (Tukey, 1977). Easily understood by non-professionals, these methods have much to offer in the adversarial human rights environment. Their effectiveness has been demonstrated, as in (Hoaglin and Velleman, 1995: 277):

> Our examination shows that approaches commonly identified with Exploratory Data Analysis are substantially more effective [than a long list of advanced model-fitting methods] at revealing the underlying patterns in the data and at building parsimonious, *understandable* [my emphasis] models that fit the data well.

Graphs: The Visual Display of Information

Our approach to the visual display of our analyses conforms to Tufte's standards for Excellence in graphical representation (Tufte, 1983: p. 13):

…. Graphical displays should

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else
- avoid distorting what the data have to say
- present many numbers in a small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from a broad overview to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation …
- be closely integrated with the statistical and verbal description of a data set

Some of the goals above can be achieved through formatting. Accordingly, in the final version of the report, the three authors agreed to a spare, uniform format. There are no gridlines and no frames on any of the graphs, as these embellishments add nothing to comprehension.

In time series plots, I omitted point markers, used ticks sparingly, and only major values are labeled and associated with ticks. Appendix 5 shows a typical time series plot. Appendix 6 shows that a spare style does not mean that complex relationships are not portrayed.

Not all of Tufte's recommendations can be achieved through formatting. Graphical presentation and analysis are interdependent. Table 3 shows the frequency distribution for the types of graphs used in the final report. This is by no means a summary of the graphs used during the analysis, but indicates the types of graphs that are likely to be used in an analytical report on human rights violations for this type of audience.

Table 3. Frequency distribution of graph types

Type	Number of dependent variables	Number of graphs
Time series plot	1	18
Time series plot	2	6
Vertical bar chart	1	6
Time series plot	3	2
Horizontal bar chart	1	2
Horizontal bar chart	3	2
Vertical bar chart	2	2
Stacked line plot	3	1
Stacked bar chart	2	1
Histogram	1	1
Comparative histogram	2	1
TOTAL	—	42

Analysis: The What, How, and Who

Analysis for this project was an iterative process, inseparable from the creative interaction of two and sometimes, three persons. For any consideration of the analysis process, keep in mind that the figures and many textual statistical references of the report are a small fraction of the total number of tables, summary statistics, and graphs that were produced during the analysis process.

As the preface of the report itself states, Patrick Ball designed the analyses, and I carried them out. However, this was not a rigid hierarchical process. As is often the case in analysis, my instructions might be as vague as "see what the relationship is between X and Y," or as precise as "make a bar chart for X and Y for A and B, with A on the left and B on the right." While the general flow of instructions was indeed from him to me, there was interaction in both directions, one analysis leading to another in an ongoing, often iterative process. I did not always confirm expected relationships, and unexpected results were frequent occurrences.

What kinds of decisions did we have to make? Given that an analysis led to a significant result (in the sense that it was worth passing on to the reader), how can we most effectively present it to the reader? Sometimes, issues of modeling were involved. Modeling is integrated with presentation – presentation for us and for the reader. For example, the annual number of killings of women is a highly skewed distribution. When analyzing such a distribution, an analyst's first instinct is to

transform to the logarithm of the variable. This transformation makes it possible to view all values without an indecipherable cluster at the low end of the axis. If the transformed distribution is normal, have we learned anything about the process?

In this case we know that the number of killings has a skewed distribution because of factors pertaining to the actions of the state and the skewed high end results from the actions of the Laugerud García and Ríos Montt regimes. If no comparison is being made (for example, between the number of killings of males and females, where the high ends differ by a large factor), there is no good reason for using the logarithmic transformation. A logarithmic transformation is often used to normalize a skewed data distribution in order to use the methods associated with normal distributions; in this case there is no need for such a transformation except to make the scale of values visually tractable.

The relevant model for the time series of violations is analogous to the standard model of industrial quality control. In this model, many sources of variation common to all data points (called "common causes") accumulate to give a background level of random variation. In terms of this model, the time sequence of killing in Guatemala has a "background" level due to common causes. One or more significantly large deviations from that level would be denoted as due to "systematic causes." The analyst then searches for the systematic cause, which in this case usually is the imposition of state policies.

Control and Traceability

Which analyses have been or are to be carried out? What scientific questions do we want to answer? What is the status of analyses in process? Who is responsible for particular data or results? What is the reference identification for a particular analysis? These are the questions that pertain to control of the analysis.

My analysis control documents followed the progression of the project from exploration to preparation of a final report. In the first phase, Patrick Ball was proposing exploratory analyses and I was producing them. This was an interactive process as is reflected in the control document. By and large, his proposals were received in text e-mail messages and it was my problem to keep track of them and respond either with the result or an additional query. Appendix 7 shows the control document used in this phase, the *Reconciliation of Instructions.* I include a fragment of this document to reveal both the format and the interactive nature of both the process of analysis and of tracking instructions and results, both intermediate and final. The instructions shown in this document are taken from e-mail or verbal instructions.

The second phase was after we incorporated Paul Kobrak into the process and we determined the final structure of the report. We now had to maintain tight tracking of the progress and in particular, tracking the ongoing revisions of both chart and figure numbers. The control document for this phase was the *Figure List*, a typical version of which is shown in Appendix 8.

An analysis results in a table or graph that we would ultimately identify by a figure number in the final report. However, many hundreds of analyses were performed. We recognized early in the project that there was a need to associate a minimal set of data with each analysis. These factors were that:

- Figure numbers were context sensitive and in a state of flux until completion of the final report.
- The creation of a table or graph and assignment of a figure number was another transition that might produce error.
- The dataset version used in the analysis was a moving target due to ongoing revisions.
- Excel in itself produced no record of the source dataset (workbook, worksheet) or the process of analysis.
- Certain summary data was relevant to every analysis (e.g., number of violations included in the analysis). For example, the number of violations or other count of units entering into the analysis is essential to our evaluation of the results as we passed from draft to final copy.
- The analyst identification is needed to determine source of analysis for questions or revisions.

- The date of the analysis provides a control for error. For example, the date of the analysis must be consistent with the date of the source dataset. Also, the date reveals whether the latest desired version of the analysis has been performed by reference to the Figure List.

Thus, Patrick Ball proposed, and I agreed that I would associate an *informational data block* with each analysis. This block was attached to every analysis until the final report was prepared. It took several iterations before we fixed on a standard format, shown below:

Date of analysis [*date*]
Analyzed by: [*analyst*]
Records included: [*count of the records used in the analysis*]
Violations included: [*count of violations used the analysis*]
File Reference: [*workbook(s), worksheet(s)*]

The data block count of the records used in the analysis was not necessarily the total included either in the source dataset, the workbook derived from it, or the particular worksheet. It was the number used to perform the particular analysis. Of course, I made errors in the data blocks and had to apply the same constant verification as in the case of the datasets themselves. However, in the long run, these data blocks proved to be invaluable in verification and in finding a way among the many dozens of subsidiary workbooks and sheets when I needed to revise or verify and analyze.

Backup

I started with a simple backup strategy. I backed up my work locally on removable disks and each week mailed a complete compressed copy of my project files on a 100MB ZIP disk to the AAAS for archival storage. Since we finished the project and have been able to create full electronic archives at the end, this aspect of the project could be considered a success. We can trace any analysis in the final report to a figure including a data block, and hence, reconstruct the original analysis.

However, my inconsistent directory structure and file naming conventions made this more difficult than it should have been, as will be discussed in the following section, Lessons Learned. These problems came in part from the fluid nature of the project, which was essential to a creative process.

Lessons Learned

In a successful project such as this one, the retrospective issue is to set the stage to carry out successful projects in the future. By showing what we did in the preceding sections, I hope that others will get guidance for their own future work. In this section, I specifically target functions and methods that worked well, and those that did not work well, in order to make recommendations that can be applied both by others and us in similar large-scale human rights data analysis projects.

Large-scale analysis of human rights data rarely occurs in the same environment twice; it is much closer to social science research than industrial statistical analysis. A common issue in applying lessons learned to recommendations is a tendency to introduce central control, uniformity, standard procedures, and conformance to rules as a way to improve efficiency and effectiveness. This is a valid approach in situations where **control** is important. On the other hand, freedom of action and tolerance of diversity is vital to **creativity**. I regard the establishment of the appropriate balance between these two poles is the major administrative and personal challenge that we face. My own preference at this stage is to lean toward promoting creativity. As a minimum, each contributor should have a unique individual approach to resolving the common problems – but in such a way that other team members can access and comprehend his or her work.

Another common and general issue is self-discipline. If you set up a rule for naming files or a procedure for backups, and so forth, stick to it. This is not easy when trying to get new answers to new problems under time pressure, but it is precisely those circumstances when lack of discipline will hurt the work the most.

Our lessons learned and related recommendations are summarized in Table 4, following. A more comprehensive jointly authored set of recommendations for data analysis, based on both the CIIDH and CEH experiences appears in Appendix 1 of The Guatemalan Commission for Historical Clarification: *Generating Analytical Reports*, by Eva Scheibreithner.

Table 4. Summary of lessons learned and recommendations

Entity, Function	Lesson	Recommendation	Issues
Data dictionary	Valuable to analyst	Ideal would be a common data dictionary, used and updated by all who create variables	If common, who is allowed to make entries? Who is required to make entries? Should this be networked (private web site)?
Directory structure	A rational project structure would help everyone. Backups from other team members would be comprehensible on sight.	Agree on a project directory structure for common use.	Will a common structure serve all? Can a single structure be used throughout the duration of the project?
File name system	Patrick Ball's dataset naming rules worked well. Mine quickly became unsatisfactory. It was good only for a small-scale project, here no better than sequential serial numbers.	Use appropriate file naming rules that will be understandable to all.	Can satisfactory rules be set at start of project?
Field names	Ambiguity in field names is treacherous	Don't use the same field name for different variables even if appearing in different datasets.	Self-discipline.
Control documents	I can't work without the control documents described in this report	Some people need control and some don't. Do what fits you.	Finding approaches to shared documents that are mutually satisfactory to team members working together.
Update of control documents	If you don't keep your records updated, you may be sorry.	Don't end up being sorry.	Self-discipline.
Errors, transitions	Human, machine, program, transmission errors happen	At every stage, be vigilant and scrutinize.	Self-discipline. See also, Facilitating error checking and verification, below.
Backup	Backup of files is a Good Thing	Have individual and project backup system.	Present system seems satisfactory; is it good enough for the next project?
Software	Different software, different versions lead to unnecessary inefficiencies and errors. I had to switch both computers and software versions to match his. These transitions caused a number of problems.	Have team members working together use the same programs and versions from the start.	Agreeing on software and versions at project start. Cost of upgrading team members' resources. Site licenses and project-owned hardware: reasonable approaches?

Audit trail of procedures	No way to know what series of edits and operations have been performed in Excel	Don't use Excel for analysis in large-scale data projects.	Which programs to use for statistical analysis? Is Stata, for which AAAS has license, the statistical software of choice? Can we get adequate graphical output from Stata?
Graph objects	Encapsulated Postscript Graphics would have been a lot easier to work with in final report than Excel pictures.	Make sure that the statistical software produces graphical files that facilitate report production.	Is Stata good enough? Do we have to consider other alternatives?
Analysis	It is a nuisance to have to integrate a revised version of a dataset in Excel.	It would be a Good Thing to be able to re-run a set of analyses on a new data set without concern for changes in number of variables or records.	Choose and agree on a statistical program that will do this.
Variable formats	It is not a Good Thing to define numerical variables to have textual values.	Don't do it!	Be careful with variable definitions.

Facilitating Error Checking, Maintaining an Audit Trail, and Updating

Stata and other statistical programs (e.g., BMDP, SPSS) have functions that make error checking and verification of datasets semi-automatic and more reliable. In addition, they produce logs of the edits and operations, outputs, and instructions used to achieve results. In Stata, the commands, *codebook, describe, list,* and *inspect,* have value in error checking and verification. The Stata commands *log* and *edit* enable the analyst to maintain a log of the steps in producing a given result. I will give a brief summary of what these Stata functions can do for the analyst. This is not the place to give a tutorial on Stata, since we do not have agreement that it will be the software of choice.

codebook examines the variable names, labels and data to produce a codebook describing the data. You can determine the pattern of missing values, automatically obtain summary numerical statistics for continuous variables and tabulations for categorical variables. (Stata, 1997: v. 1, p. 151-4)

describe produces a summary of the contents of the dataset. You can list the variable names in a compact format (Stata, 1997: v. 1, p. 206-10).

list displays the values of the variables (Stata, 1997: v. 2, p. 335-7).

inspect produces a summary of numeric variables. It reports the number of negative, zero, and positive values; the number of integers and non-integers; the number of unique and missing values; and a miniature histogram (Stata, 1997: v. 2, p. 271-3).

log enables the user to maintain a log of both commands and results (Stata, 1997: v. 2, p. 341-2).

edit adds to the log changes made to the data in the Stata editor (Stata, 1997: v. 1, p. 251).

A Stata log can be converted to a procedure that can be applied to new versions of a dataset without modification, facilitating upgrading to a new version.

No software is without problems, and I am not proposing that Stata (or any other program) is a panacea. If the user does not have a commitment to vigilance and scrutiny to the extent described above, these or similar procedures will be unused or poorly used. If the user does not exercise judgement, intuition, and creativity in dealing with the data, the choice of software will mean little. If logs are poorly edited and maintained or cannot be properly accessed and associated with results because of poor control and naming conventions, they will be of little value.

For example, when the user is engaged in the disorderly but creative process of searching for the appropriate data configuration and analysis, the Stata log can become large and cluttered with false starts. To reduce this log to the final set of procedures can be time-consuming. For tracking logs by name, the Stata team itself has nothing better to offer than date stamping. This is not good enough for our work. At the minimum, we would have to produce logs that are subsets of the daily log. For the individual and possibly for the project, a systematic organization of final procedures will be needed.

Stata has problems and bugs, and there is an active list server in which users share experiences. The descriptive commands discussed above are sensitive to the ways in which data are defined in the data set, and can give incomplete and confusing results. As always, the functions, advantages, and disadvantages of using different software must be evaluated to find the balance appropriate to the job at hand. It may well be that for all the problems and inefficiencies, the software and methods used on the AAAS/CIIDH project were somewhere near the optimum, given that the project goals were met in a timely fashion.

Conclusion

In this report I reviewed my work on the CIIDH/AAAS report, summarized the most important lessons learned, and made recommendations for work of this nature on future projects.

I know that this summary will help me to do a better job on the next similar analysis project. I hope that it will also help others, and in that spirit, close with this quote:

> After spending many years with the Estonian Cancer Registry [I] now think more intensively about data quality than about the application of refined statistical and cartographic methods to data analysis.[2]

[2] From a book review of *Global Geocancerology: A World Geography of Human Cancers*. *The Scientific American*, Feb. 1987, pp. 27-31.

Appendix 1

AAAS/CIIDH Data Dictionary

Data Dictionary			**Appears in file:**			
Field name	**Meaning(s)**	**Values –meaning**	***ctanon***	***ctnmd***	***rtanon***	***rtnmd***
AGE	Age of victim	integer >= 0 -1 – missing value		x		x
C_NMD	Number of victims having age and gender	integer = 1		x		
C_TOT	Number of victims including not named and without age or gender values	integer = > 1		x		
CERTFECH	Precision of dates	1 – day 2 – month 3 – quarter 4 – semester 5 – year 6 – decade 7 – season, no year 8 – no idea of date	x	x	x	x
CIV_CIV	Civilian perpetrators present	0 – not reported	x	x	x	x
EST_EJR	Army involved	0 – no 1 – yes		x		x
EST_EJR	Number of violations in which Army involved	integer = > 0	x		x	
EST_PAC	Civil patrollers involved	0 – no 1 – yes		x		x
EST_PAC	Civil patrollers involved	integer = > 0	x		x	

EST_POL	Police involved	0 – no 1 – yes		x	x	x
EST_POL	Number of violations in which Police involved	integer = > 0	x			
FW	Number of violations for a given case	integer >= 1	x		x	
MON_VLCN	Month in which violation occurred	1 – January … 12 – December	x	x	x	x
MONYEAR	Month and year of violation	59-03 to 95-12	x	x	x	x
OVERKILL	Presence of additional indignities to a victim either killed or being killed	0 – not reported 1 – reported present	x	x	x	
P94_NAC	1994 national census population for municipio of birth	integer > 0 -1 = missing		x		x
P94_VLN	1994 national census population for municipio of violation	integer > 0 -1 = missing		x		x
PAR_PAR	Paramilitary present at violation(s)	0 – not reported 1 – reported present	x	x	x	x
PRES_URN	Guatemala National Revolutionary Union present at violation(s)	0 – not reported 1 – reported present	x	x	x	x

REGIME_N	Regime in which violation occurred as identified by President	01 – Ydígoras Fuentes 02 – Peralta Azurdia 03 - Méndez Montenegro 04 – Arana Osorio 05 – Laugerud García 06 – Lucas García 07 - Ríos Montt 08 - Mejía Víctores 09 – Cerezo Arévalo 10 – Serrano Elias 11 - de León Carpio	x	x	x	x
REGION	Homogeneous geographical region	01 – Occidente 02 – Costa Sur 03 – Verapaces 04 - Petén 05 – Oriente 06 – Meseta Central	x	x	x	x
SEX	Gender of victim	F – female M – male d – unknown	x	x		x
SVNUM	Serial number of named victim	sv concatenated with 7 digit integer except: SV0050217 (input error)		x		x
TYPE_SOU	Type of source of case	DOC - documentary ENT – CIIDH interview PER – periodical			x	x
U_R	Type of area	r – rural u – urban d – unknown	x	x	x	x

URN_URN	Guatemala National Revolutionary Union involved	integer = > 0	x		x	
URN_URN	Number of violations in which Guatemala National Revolutionary Union involved	0 – not reported 1 – reported present		x		x
V_DEPTVL	Department of violation	department no se sabe - unknown	x			
V_DPMU	Department and municipo of violation	department, municipio otro – no department, other municipio no se sabe, no se sabe – department and municipio unknown	x			
V_IND	Ethnic category of victim	Desconocido – unknown Indigenous Ladino	x	x	x	
V_MUNINA	Birthplace of victim	municipio name no se sabe - don't know		x		x
V_MUNIVL	Place of violation	municipio name no se sabe - don't know otro – other	x	x	x	x

V_ORG	Victim's organizational affiliation	civ-camp – civilian, peasant civ- ddh – civilian, human rights civ – emp – civilian, employee civ-ind – civilian-indigenous organization civ-otr –civilian, other civ-rel – civilian, religious civ-sin – civilian, labor civ-uni – civilian, university otr-otr – other other pol-mil – political, military pol-pol – political, political	x	x		
V_SEXO	Gender of victim	F – female M – male d – unknown			x	
V_TRAB	Occupation of victim	occupation name no tenia trabaj – unemployed otro – other blank – missing		x		
VLCN	Type of violation	Mu – killing Ds – disappeared Se – kidnapping or illegal detention Hr – injury To – torture	x	x	x	x
YR_VLCN	Year in which violation occurred	1959, …, 1995	x	x	x	x

Appendix 2

AAAS/CIIDH Dataset Description

Primary Data Sets[3]

Filename	No. Records	No. Violations	Killings
ftanonkv8	5,601	34,660	34,660
ftnmdkv8	8,968	8,968	8,968
rtnmdkv8	8,964	8,964	8,964
rtanonkv8	5,585	5,585	34,656
ftanonv8	8,242	43,547	34,660
ftnmdv8	13,917	13,917	8,969
rtnmdv8	13,906	13,906	8,964
rtanonv8	8,205	43,535	34,656

[3] Data sets with the prefix "f" for full were not used in my analysis.

Appendix 3

AAAS/CIIDH variable position dictionary

Variable number gives order in data set[4]

Filename	No. of records	svnum	vlcn	mon_vlcn	yr_vlcn	certfech	muni_vlcn	type_sou	c_nmd	c_tot	age	sex	munitot	hompct	v_ape1	cnt	ape2cnt
ftf2t2v1.dta	17,941	1	2	3	4	5	6	7		8							
ftf2t1v1.dta	14,025	1	2	3	4	5	6	7			8	9					
fintaba.txt	13,821		1	2	3		4	5	8	9	6	7					
ft11.txt	14,025	1	2	3	4	5	6	7			8	9					
ft1.txt	14,025	1	2	3	4	5	6	7			8	9					
ft2.txt	17,941	1	2	3	4	5	6	7		8							
mn4.txt,.xls, .dta	34		1				2						3	4			
mnap.txt, .xls, .dta	410		2				3	1					7		4	5	6

[4] Data sets with the prefix "f" for full were not used in my analysis.

Appendix 4

AAAS/CIIDH derived dataset description

(Fragment shown only for illustration)

Filename	Description	Source	
age distribution.xls	Demographic vertical two-sided plot of male, female age distributions for Guate based on 1994 census.		
ageBOX.gph	Primitive box plot of age distributions by violation	ageBYvln.dta	
ageBOX.wmf	Primitive box plot of age distributions by violation		DEL
ageBOXvln simple.doc	Primitive box plot of age distributions by violation		
ageBYvln.dta	ages in columns, one column for each violation type	rtnmdv3AG	
AgeBYvln.txt	ages in columns, one column for each violation type		
Book1.xls	can't recall		DEL
Book2.xls	dummy block		DEL
censo_a full for source.xls	working copy	censo_a	
censo_a.xls	original	censo_a	
Ch05AgeHistOpen with rtnmdv3A.xls	age histogram for Chapter 05 linked to rtnmdvA.xls	rtnmdvA.xls, histogram template	
D1_VctmDemo.xls	PB South Africa automated graph generator		Look at
Dummy for Time Series.xls	dummy block		
fintaba.dta	working copy	fintaba	
fintaba.xls	working copy	fintaba	
Formatted Sections.xls	Contains standard sections. For example, names of presidentes in chronological order, region names, etc.		

Notes:

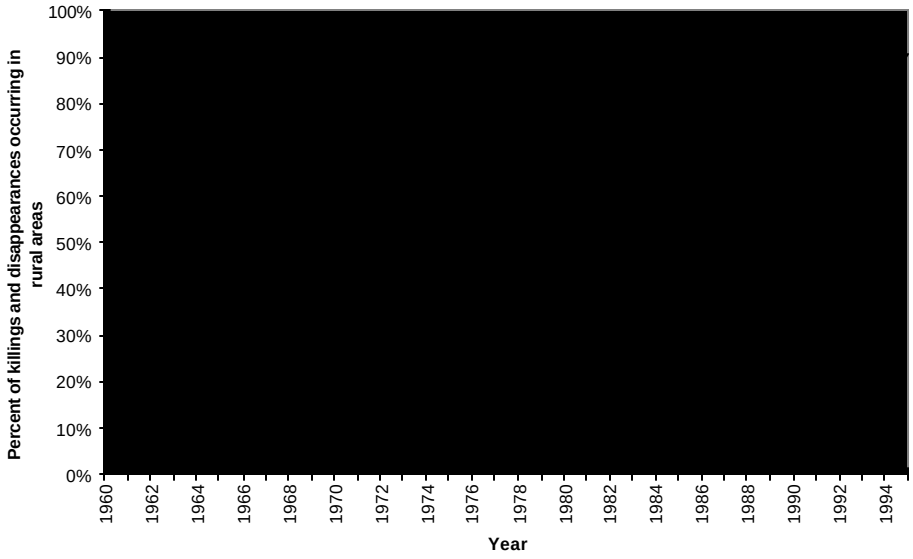
1. If an entry is given in "Source," then the file is derived from an original data set described in Data Set Descriptions.doc.

2. Dummy blocks contain blocks of entries to add to data sets. With these blocks included, Pivot Table will have something to chew on for each year. This makes the year variable continuous and complete.
3. Proliferation of .xls files with same prefix on name was to avoid excessively large files. If I had to do it over again, I would simply number them sequentially.

Appendix 5

Typical AAAS/CIIDH time series plot

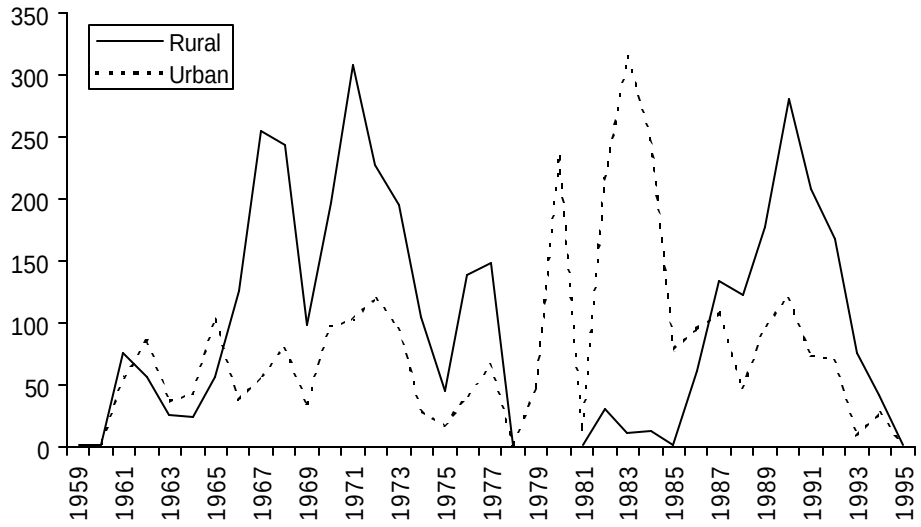
Figure 8.3 Percent of killings and disappearances occurring in rural areas by year, 1960-1995



Appendix 6

Complex AAAS/CIIDH time series plot

Figure 9.1: Number of killings and disappearances reported in the press by area (rural vs. urban) and by year



Appendix 7

AAAS/CIIDH Reconciliation of instructions

(Fragment shown only for illustration)

ID	Date	Chapter	Instruction	Action
01	?	03	Monthly 1st diff sources and violations	not needed
02	?	03 and ?	Time can be broken into segments, 61-78, 79-85, 86-96	Done when appropriate. Use of logarithmic reexpression allows use of all at once when relevant.
03	?	04	monthly first differences for departments, limited to ENT/DOC, for all killing, then COL/IND	Done for annual. Too many months with no violations, too much detail.
04	5 May 98	?	Order of departments by number of killings among three sources and relative proportions	Done in Chapter 4
05	5 May 98	03	Analysis by regime as a kind of time	Done in Chapter 4
06	5 May 98	04	Urban rural differences by regime, separately by violation	Done in Chapter 4
07	5 May 98	04	Killings overall, disappearances overall	Killings done, very few disappearances, so not done
08	5 May 98	03	Compare killings by IND, COL and ENT to PER	Done
09	5 May 98	03	Do regime comparisons at end	Done
10	9 Apr 98	All	Break into numbered sections	Done, needs watching and agreement on style
11	9 Apr 98	03	Numerous position changes of sections marked on draft	Done
12	9 Apr 98	03	Compare one kind of violation by ENT PER DOC sources, a figure for each type of violation	Done

Appendix 8

AAAS/CIIDH Figure List

Figure list as of 24 October 1998

Fig# 24 Oct	Fig# 14 Oct	Fig# 13 Oct	Figure title	Derived from figure, notes	Filters	Datase t	Resp .
01.1	1.1	1.1	Number of killings and disappearances by year, 1960-1995		Net of URNG, certfech<=5	CTanon	HFS
02.1	2.1	2.1	Number of killings and disappearances by year, 1960-1969		Net of URNG, certfech<=5	CTanon	HFS
03.1	3.1	3.1	Number of killings and disappearances by year, 1970-1979		Net of URNG, certfech<=5	CTanon	HFS
04.1	4.1	4.1	Number of killings and disappearances by year, 1980-1989		Net of URNG, certfech<=5	CTanon	HFS
05.1	5.1	5.1	Number of killings and disappearances by year, 1990-1995		Net of URNG, certfech<=5	CTanon	HFS
06.1	6.1	6.1	Number of disappearances and killings, by regime	Figure 3.26, killings only, without means	Net of URNG, certfech<=2	CTanon	HFS
06.2	6.2	6.2	Average monthly number of deaths and disappearances, by regime	this is the other half of Fig 3.26; ordered by regime	Net of URNG, certfech<=2	CTanon	HFS
06.3	6.3	7.1	Number of killings and disappearances by month, July 1979-June 1984		Net of URNG, certfech<=2	CTanon	HFS
07.1	7.1	8.1 & 8.2	Number of killings and disappearances by year and source (press vs. documentary vs. interview)	Fig. 3.7; cut off vertical axis at 700, leaving the DOC and ENT peaks off the graph	Net of URNG, certfech<=5; PER and DOC	RTanon	HFS
07.2	7.2	8.4	Number of killings and disappearances by regime and data source	Fig. 3.31, PER/DOC/ENT	Net of URNG, certfech<=25; PER, DOC and ENT	RTanon	HFS

References

Ball, Patrick, Kobrak, Paul, and Spirer, Herbert, 1999. *State Violence in Guatemala, 1960-1996: A Quantitative Reflection.* Washington: American Association for the Advancement of Science and Centro Internacional por Investigaciones en Derechos Humanos.

Hoaglin, David, and Velleman, Paul, 1995. "A Critical Look at Some Analyses of Major League Baseball Salaries." *The American Statistician* (August 1995), pp. 277-285.

Spirer, Herbert, and Spirer, Louise, 1993. *Data Analysis for Monitoring Human Rights.* Washington, DC: American Association for the Advancement of Science and HURIDOCS.

Stata, 1997. *Stata Reference Manual.* Release 5. College Station, TX: Stata Press.

Tufte, Edward, 1983. *The Visual Display of Quantitative Information.* Cheshire, CT: Graphics Press.

Tukey, John W., 1977. *Exploratory Data Analysis.* Reading: Addison-Wesley Publishing Co.

Chapter 8

The Guatemalan Commission for Historical Clarification: *Data Processing*

Rocio Mezquita

Collection of Data: The Case

The field collection of the human rights violation data for the Guatemalan Historical Clarification Commission (CEH) was done by the *case*. A case was defined by one or more violations against one or more persons, perpetrated by one or more perpetrators, in the same act or operation.

The case was worked in the field by the interviewers who collected the testimonies, identified the cases and summarized each one. If several deponents spoke of the same case, then the interviewers had to put these testimonies together, rewriting the summary, reconstructing the facts of the case and resolving contradictions. A printed summary was then sent to the CEH database, located at the Central Offices of the CEH, in Guatemala City.

Original testimonies were rarely recorded or written. The interviewers wrote a summary of what the people said at the interview. The printed summary was attached to a set of five *forms*. These were; the case summary form; the victim(s) form(s), the violations pattern form; the perpetrator(s) form(s); and the deponent form.[1] The information collected in the field by interviewers is described in detail in Appendix 1.

Different deponents gave testimony about the same case in different CEH field offices. Thus, if the case could not be reconciled and reconstructed into a single case summary in the field, this task was left to the database processing team.

When could a violation be considered a CEH case?

To be accepted by the interviewer as a case, one ore more of the violations had to be a defined CEH violation (see the next section). The initial list of CEH violations was discussed, developed and accepted by the commissioners of the CEH, the Central Team, and the Database Coordinator. During the period of data processing the list was expanded to account for knowledge gained during the collection of the data.

To accept one of the CEH violations as valid for a case, the violation had to have a relation to Guatemala's war between the years of 1960 and 1996[2], no matter the perpetrator. A listing of the CEH violations, with their definitions, appears in Appendix 2.

How Violations Were Recorded

Violations were recorded in a *pattern form*. A *pattern of violations* is a sequence of violations that were inflicted on one or more victims in the same place and date by the same perpetrator(s). A case must have at least one pattern of violations, but may have an unlimited number of patterns. By using these pattern forms, repetition of each sequence of violations for victims was avoided. This was especially convenient in the case of a massacre, where there could be more than 200 identified victims.

The following is an example of a pattern in which information was given for the sequence of violations in a case.

1) **Violation which occurred first**
   Geographic location

[1] Editors' note. The reader of other papers in this volume will notice that this structuring is defined elsewhere. We retained these redundancies so that each paper is self-explanatory.

[2] The original mandate of the CEH, specified in the Final Peace Accords was that the period of interest was 1960 to 1996. After a subsequent historical analysis, the CEH Commissioners decided that the "internal conflict" started in 1962 and ended in 1996, with the signing of the Final Peace Accords.

Specific location (home, street, etc)
Date
Perpetrator's group name
Individual perpetrator's information if known
Victim(s) - (identified, anonymous or collective)[3]

2) Violation which occurred second
Geographic location
Specific location
Date
Perpetrator's group name
Individual perpetrator's information if known
Victim(s) – (identified, anonymous or collective)

... And so forth for 3...*n* times, as appropriate.

Certainty of the violation and the perpetrator

The data processing team identified the certainty that the violation had happened, the type of responsibility of the perpetrating forces (direct, intellectual, collaboration...) and the certainty that this was the perpetrator responsible for the violation.

Violation:

1. Deponent is a witness
2. Deponent is not a witness
3. Deponent is not sure about the occurrence of the violation

Perpetrator's responsibility:

1. Material author
2. Collaborator
3. Intellectual author
4. Informant
5. Does not apply This was used for specific violations such as "the person disappeared for unknown reasons", in which there was no perpetrator, and therefore responsibility could not be assigned.

Perpetrator's identity.

1. Deponent is a witness
2. There are other witnesses
3. Deponent suspects
4. It is publicly known
5. Documented evidence exists
6. Does not apply This was used for specific violations such as "the person disappeared for unknown reasons", in which there was no perpetrator, and therefore responsibility could not be assigned.

The data processing team had therefore to identify the following elements in a case:
who the victims were
what happened to them
who was responsible
where
when
violation's certainty
perpetrator's responsibility
certainty about the perpetrator's identity

[3] The victim was directly related to the pattern, but victim information was recorded separately.

The perpetrators could be identified in a collective way by the group to which they belonged. The listing of such groups is given in Appendix 4. If the perpetrators' names were known, they could be related directly to the violation.

The victims

The victims were directly related to the pattern. Accordingly, there were three types of "countable" victims:

identified
anonymous
collective

Identified victims

The identified victims were those victims of whom we knew at least two of the three fields used to identify the victim in the database, one for the name and two for the last names (father's and mother's).

Example #1:

NAME: *Francisco*
1st LAST NAME: *Pop*
2nd LAST NAME: X *(unknown)*

Example #2:

NAME: *Juana*
1st LAST NAME: *X (unknown)*
2nd LAST NAME: *Ramirez*

Example #3:

NAME: *X (unknown)*
1st LAST NAME: *Cu*
2nd LAST NAME: *Caal*

Anonymous victims

Anonymous victims were victims for whom there was no personal information. Until almost the end of the project, the program also counted the "xx" (individual victims whose names we did not know even though we knew their sex, age or ethnicity), as anonymous victims. Initially the program did not count this valuable information in its statistics, and it was lost. Finally, this was changed and the "xx" victim was automatically converted into a "collective" of one person. Thus, in the end, this information was counted in the statistics.

Collective victims

This term denoted two or more victims for whom we had some information, such as sex, group identity (e.g., catechists, or peasants from a specific village, or an ethnic group).[4]

Difficulties Encountered

The definition of a massacre

How to define a massacre was an issue of ongoing concern throughout most of the CEH project. As the term "massacre" was never a violation in itself, the identification was made through key words at the beginning and at the end, through the title.

Initially, a key word code was used when the when the case testimony mentioned a massacre. At that time it was the **testimony** that identified a massacre, and not the **case**, as there was a many-to-many relationship between cases and testimonies.[5] A massacre was at that time defined as the

[4] A discussion of the nature of the definitions of collective victims and the relationships inherent in these definitions appears in Chapter 9.

[5] One testimony could relate information about many massacres; each massacre might be described in many testimonies.

killing of a "significant" number of people, but that number was not given in advance. Thus, whatever number the data processor thought to be significant determined whether the event was denoted a massacre. Unfortunately, a specific key word for massacre cases was not used. The case could be a massacre or its testimony could – among other things – relate a massacre, which was not the violation of the case itself. This was a confusing situation.

Later in the project, when almost all of the data processing of the cases had been completed, a final definition was agreed to. This final definition was: "A massacre shall be considered the execution of five or more people, in the same place, as part of the same operation and whose victims were in an indefensible state."

A code was created to identify those cases, which fell under the new definition. All the cases that were already in the database and which had five or more victims of executions were revised and re-coded. A problem was that some cases which were massacres did not appear in the list of "five or more victims." Thus, they could not be re-coded under the new definition. They were not in the list because the victim type was collective and when we did not know how many victims there were we counted the collective victim category as including only two victims.

When this problem was found, it was decided to automatically apply the first massacre definition to the second one. It was assumed that the first coding was more limited and that all the cases that had this code should also fall under the new more specific definition of massacre. Unfortunately, it was forgotten that the first massacre code has been used not only for massacre cases, but for all the cases in which the testimony mentioned a massacre whether it belonged to the case or not.

As a consequence, the new code for massacre lost its relation to the new massacre definition because it included testimonies which spoke of a massacre, or where the case itself was not a massacre, or from massacre cases under the old definition and massacre cases under the new definition

There was no way to identify all the massacres in the database that fell under the new definition, as the code had been altered. To solve this newly created problem, it was decided to search for massacres by the "title" section, which appeared on the summary case form. This approach to a solution had its difficulties. Unfortunately, not all the massacres were identified correctly in the title because some interviewers used the wording "indiscriminate attack." Another source of problems was a spelling error, which made it impossible for the program to correctly identify the code. Final resolution was achieved after a number of revisions of the entire database and all the massacres were identified and listed.

Names of the categories

At the time of data processing, the database did not distinguish different violations for different perpetrators. For example, *extrajudicial execution* for state actors (as stated in international human rights law) and *assassination* or *killing* for the guerrilla forces. Extrajudicial execution was accepted and coded equally for any perpetrator. At the time of the final report, the category name was changed and the killing of a person was defined as "arbitrary execution."

All such violations, which were denoted homicides in the sense that the reason for the killing was personal and not political, were considered extrajudicial execution whether committed by state agents or the guerrilla forces. The reasoning for this designation was that the perpetrators committed these violations under the impunity that the context of war offered.

The ambiguous category, *wounded or killed*, was created to keep a record of the combatants who were mentioned in the different cases and were either wounded or killed. This category was often used to identify those people who joined the guerrilla forces, and never came back. In such a case, it was assigned with a certainty of "it is suspected." There was no certainty that they were killed or wounded. They may even have become refugees in another country. This information could have been useful if it were decided to look for those combatants who never returned, and whose families continue to search for them.

The identification of forced disappearances also had many problems. In some exceptional instances, there was no specific information on whether a witness observed the kidnapping or the testimony clearly stated that no one witnessed it. However, the context strongly suggests a forced disappearance. For example, the victim may have previously been threatened, or belonged to a group likely to suffer political violence in the Guatemalan context. In these special instances, the violation was classified as forced disappearance.

Other special cases arose when the body of the person was never found. In all of the following cases, the violation was denoted a forced disappearance:

- A person disappeared and people suspected s/he was dead.
- The victim's body was seen in a non-public place (such as military base) and it was clear that there was intent to disappear the victim.
- When one of the perpetrators confessed (usually under the influence of alcohol) that he had killed the person.
- When the person was taken away, but there is no more information in the testimony to determine whether the person reappeared, never appeared, or reappeared dead.

People with the same name/repeated victims

When entering the name of a victim whose name was already in the system, the program alerted the operator entering the information. At that time, whether the person was the same was not checked. This was done later, in the filtering of repeated names.

To verify if the two names belonged to the same person or not, it was necessary to check and compare the person's information as well as the context and the violation itself, and decide whether the person is or not the same person. If a person was discovered to be the same as another one in a different case, the cases were joined into one. This was a slow and time-consuming process.

Unifying a case

When by some circumstances (for example, two or more cases have the same victims or two or more cases contained the same violation) the data processing team had to unify the cases, cross the information of the violations, the victims, the place, date and perpetrators. If contradictions were found (which almost always occurred), the data processing personnel had to make a decision based on the information of both testimonies checking and taking into consideration which of the deponents was more reliable. Among the criteria for reliability was a closer relationship to the victim, or whether the deponent was a witness or not, etc.

Massacres were frequently described by many different deponents in the different field offices of the CEH so many cases of the same massacre arrived at the database without unification. To avoid going through the same cases again and again, as new cases for a same massacre were inputted, it was decided to accumulate all the cases which spoke of a specific place, and reconcile these cases out at the end of the reconciliation of testimonies. Massacre cases were unified in the same way as any other case. The major problem was that there often were many testimonies.

Deponent

Every case was constructed of one or more interviews. The relation of the person who gave the testimony was directly related to the case itself, never to the specific information that appeared in it. If more cases appeared which mentioned the same act of violation as another one, the additional deponent information was also added to the case.

Frequently the interviews were *collective interviews*. Sometimes entire communities would assemble and collectively give their testimony. This *collective deponent* was seldom identified by the interviewer, and therefore, was not recorded at the database.

For security reasons it is important to decide how and where the deponent's form is going to be filed. At the CEH, the deponents' forms (which contained their names and signatures) were filed separately from the case, for security reasons. Every testimony had to have at least one deponent sheet. This sheet either had a name, or if the person did not want to give their name, a note saying "deponent is afraid of giving the name" or stating whatever reason the name is suppressed. Only in this way, could we use the database to count all the people who gave testimony.

Training the Team

Data processing consisted of two main steps: 1) reading the testimony, identifying the information (violations and other qualitative information), entering it on forms in code or text, and 2) entering the codes and text into the database.

The interviewers and data processors did the first step. Then, the data entry personnel who *captured* the codes and text on the forms did the second step. No one person did both of these tasks. This specialization was good because at best it allowed team members to develop a high level of skill at their task with a greater level of accuracy and higher speed. On the other hand, there were some disadvantages to specialization.

The disadvantages included the difficulty of reading other people's handwriting, which was a continuing problem. Team personnel would have to contact the persons who carried out prior tasks to get clarification. This caused some lost time and may have resulted in undetected errors by persons who did not realize they should seek help. Also, most of the interviewers, data processors, and data entry personnel were not Guatemalan and were not always sure of the spelling of names. In a few cases, the data processor might fail to enter some information that was required. Errors due to these causes were avoided, but time and effort was expended correcting the errors.

However, specialization had its advantages. Data entry personnel were able to increase the speed of entry as they became more experienced (further along the learning curve). Data processors could concentrate on reading the testimonies and dealing with the specific problems that appeared when determining and interpreting the acts comprising a violation or a statement.

The database coordinator and her assistant trained the data processors. The topics covered in the training were the following:

- Information to be obtained from testimonies (violation, date, perpetrator's group name, perpetrator's individual name, etc.)
- CEH violations
- Key words
- How to properly fill in the forms
- Use of codes and textual entries
- Coding categories, for example, how to determine an identified, collective or anonymous victim or perpetrator
- Explanation of how to properly complete the forms, specifying which information should be coded and which should be text

Lessons Learned

Problem	Lesson learned	Issues
Codes and definitions were not unique throughout the project.	A unique code should be kept for each definition. If a new definition is to be coded, do not code it under an existing code. Give that definition its own unique code.	Enable the system to keep track of all the information in a separate way. Then, the system will be robust with respect to changes introduced by heads of the investigation, interviewers or analysts.
In some cases, original information was coded but the details not retained.	Record testimonies if authorized by the deponent. Always keep the original testimony.	
Inefficiencies resulting from the lack of knowledge of Guatemalan languages, history and geography.	Assure that the database team personnel have a good knowledge of the language, history and geography of the country. This can be achieved through training and education.	It would be good to have country nationals "seeded" throughout the teams if the proportion of foreign personnel is high.
The same problems were not always resolved in the same way.	Make consistent rules for resolution of problems and distribute to team personnel.	
Different rules for determining duplicate names were used by different team members at different times.	Make consistent rules for resolution of duplicates and distribute to team personnel.	This is the most important special case of the prior lesson learned.

Problem of apparently anonymous victim who bears a known relationship to an identified victim.	Consider on an individual project basis.	For example, Juan Perez is an identified victim and it is known that another victim is his son, but cannot be fully identified by name.
Not all violations had a controlled vocabulary at all times.	Do not allow violation coding without a controlled vocabulary.	
Duplication of names because deponent testifies in several cases.	Make clear on the forms that this deponent is duplicated, and on which forms the name is repeated.	
Lack of information about number of deponents participating in a collective interview.	In future, have a space on deponent forms to indicate number of participants in a collective interview.	
Filtering for duplicate names was inefficient and consumed too much time at end of project (two months!).	Filter suspected duplicate names early in the process of data entry. Provide an entry on forms for indicating that an interviewer, data processor or other person believes that a form contains a duplicated name. Store this notation with the record in the database.	When duplicate names are not cleansed early in the process, each duplicate entry is processed. Thus, many data processing operations are duplicated unnecessarily.
Inefficient re-coding because it is done by a different person.	Identify who does coding so that person can re-code if needed.	
Balancing too little information with too much.	Education of data processing team on objectives of project, especially if they are changed.	It is not always clear in advance what information will be useful. Too much is better than too little, except that resources, especially time, limit what can be done.
Continuous communication between interviewers and database team to facilitate correct reconstruction of cases.	Encourage and assure continuing communication between interviewers and database team.	
Inefficiencies at the start of the project because of incomplete organization, confused ideas about information to be collected, and lack of understanding of principles of case/violation structure.	Establish principles and rules at start of project; inform data processing team members, making sure that they are aware of changes as they occur.	Unfortunately, there was still disagreement over some of these principles going on after data collection had started.

Appendix 1

Information Collected in the Field by the Interviewers

Victim:

1st last name	
2nd last name	
Name	
Other names	nicknames, war names, pseudonyms
Sex of the victim	M/F
Identity document.	number and date of issue; one of the following was accepted: Identity Card, Birth Certificate, Refugee Card, Demobilization Card, Passport. This information was almost never completed
Date of birth or age at the moment of the violation. Certainty of this information	there were several levels of certainty options: 1) total 2) 1-2 years 3) 3-5 years 4) 6-10 years 5) none
Nationality	text, not coded
Place of birth	department, municipality, town, village, with a code number from a coded geographical dictionary of Guatemala
Mother's language	A coded list of languages spoken in Guatemala, as well as other languages, was used to answer this question
Type of victim	multiple and non exclusive options were allowed here
Where did the victim live at the moment of the violation?	text, not coded
Was the victim forced to leave the place where she/he lived because of war?	Y/N
Name of the father	
Name of the mother	
Marital status	at the moment of the violation, options were: 1) single, 2) married 3) widowed 4) divorced
Name of the wife/husband	
Number of daughters and sons	dead or alive
Name and age of the daughters and sons.	for the age, the deponent had the option to tell the age at the moment of the violation, or the age when the testimony was given
Additional comments on the victim	text, not coded

Perpetrator (individual):

1st last name	
2nd last name	
Name	
Other names	nicknames, war names, etc.
Sex	M/F
Identity document	number and date of issue. One of the following was accepted: Identity card, Birth Certificate, Refugee card, Demobilization card, Passport
Date of birth or age at the moment of the violation. Certainty on this information	there were several levels of certainty options: 1) total 2) 1-2 years 3) 3-5 years 4) 6-10 years 5) none
Place of birth	department, municipality, village. With a code number from a coded geographical dictionary of Guatemala
Mother tongue	a coded list of languages spoken in Guatemala, as well as other languages, was used to answer this question.
Groups)/organizations) to which the perpetrator belonged	multiple and non-exclusive options were allowed here. A coded list of groups of perpetrators was used.
Post of responsibility in this organization	a coded list of posts of responsibilities was used.
Violations for which the perpetrator is responsible	The violation had to be related to the pattern and to the order in the sequence of violations. This section included the kind of responsibility: perpetrator, intellectual responsibility or informant, and the type of evidence -- deponent is witness, other people saw him, deponent suspects, “everybody knows,” there are documented proofs.
Other violations in which the perpetrator participated	Text, not coded
Where is the perpetrator living now and what is he doing?	Text, not coded
Other comments on this person	Text, not coded

Deponent

1st last name	
2nd last name	
Name	
Where does he/she live now?	Text, not coded
Sex	M/F
Identity document	number and date of issue. One of the following was accepted: Identity card, Birth Certificate, Refugee card, Demobilization card, Passport. This information was almost never completed.
Date of birth or present age.	
Mother tongue	a coded list of languages spoken in Guatemala, as well as other languages, was used to answer this question.
Is the deponent also perpetrator?	Y/N
Other characteristics of the deponent	one of following options should be selected if it corresponds: a) refugee b) displaced c) reinserted ex-combatant d) returned refugee, e) survivor of a massacre f) victim of non typified violations g) witness or survivor of the political violence h) other text
Relation of the deponent with the victims	as the deponent could be the same in several cases, this section specified the victim, the victim's number, and the case in which that victim appears. Then, the relation was specified. There was a coded list of the most common relationships.
Other people who the deponent knows were also witnesses of violations	Text: name of the witness, how to find her/him, and any security problems there may be for this witness if s/he is looked for by a CEH interviewer
Expectations that the deponent has from the CEH	Text, not coded
Date of the interview	
Place of the interview	Text, not coded.
Language in which the interview was taken	a coded list of languages spoken in Guatemala, as well as other languages, was used to answer this question.
If necessary, can the CEH contact the deponent again?	Y/N How? Text.
Signature of the interviewer	
Signature or fingerprint of the translator	
Signature or fingerprint of the deponent	

Appendix 2

CEH Violations

The CEH used definitions for most violations that were derived from international humanitarian law, except for some violations that had CEH definitions. The following were the CEH accepted violation categories with the approved definitions.

a) Human rights violations and cases of violence resulting in death	**This is a general category. However, the database processing team used this category to classify all conflict-related deaths that did not correspond to any of the approved specific violations under this general category. For example, people who died as a consequence of torture -- not immediately, but several months or years later – were included in this category, as well as related suicides.**
Extrajudicial execution	Legal definition, including incidents where a guerrilla was the perpetrator. Also, when a person died immediately after or as a clear consequence of torture.
Death by forced displacement	People who died as a consequence of the displacement that persecution, fear, and massacres produced. Includes deaths due to hunger, sickness, depression, lack of medical attention due to displacement (pregnant women who died in labor, etc.).
Civilian death during hostilities	According to the international humanitarian law definition for “hostilities.”
Civilian death by indiscriminate attack	According to the international humanitarian law definition for “indiscriminate attack.”
Civilian death by the use of mines	Death as a consequence of touching or walking over an anti-personnel mine.
Death resulting from use as a human shield	Deaths in events in which civilians or paramilitary state-approved forces (e.g., PACs) were used by military forces on patrol to protect themselves from guerrilla forces. Not easy to identify, as not all deponents would have known that these people were being used as human shields at the moment they were killed.
b) Human rights violations and cases of violence resulting in grave injuries	**This is a general category. However, the database processing team used this category to classify all conflict-related wounds and injuries, which did not correspond to any of the approved specific violations under this general category.**
Wounded during an attempt on one’s life	Victims who survived an attempt at extrajudicial execution, but were wounded in the attempt.
Wounded during forced displacement	People who were wounded as a consequence of the displacement that persecution, fear, and massacres produced. Includes wounds due to hunger, sickness, depression, lack of medical attention due to displacement (pregnant women injured in labor, etc.).
Civilian wounded during hostilities	According to the international humanitarian law definition for “hostilities.”
Civilian wounded by indiscriminate attack	According to the international humanitarian law definition for “indiscriminate attack.”
Civilian wounded by the use of mines	Injury as a consequence of touching or walking over an anti-personnel mine.

Wounded while being used as a human shield	Injuries or wounds in events in which civilians or paramilitary state approved forces (e.g., PACs) were used by military forces on patrol to protect themselves from guerrilla forces. Not easy to identify, as not all deponents would have known that these people were being used as human shields at the moment they were injured or wounded.
c) Torture and cruel, inhuman and degrading treatment	**As defined in international humanitarian law, but applied to state agents, guerrillas or any other group. This category includes ill treatment.**
d) Sexual violations	**Sexual abuse by any perpetrator. If a person was raped by more than one perpetrator at the same place and date, the database team counted only one act of sexual violation.**
e) Disappearances	**This general category was never assigned.**
Forced disappearance	Used in general only when the victim was seen being disappeared by a perpetrator. In cases where the context (e.g., prior threats, membership in a targeted group) strongly suggested a forced disappearance, this category was also assigned.
Disappearance by unknown cause	This category was used in the absence of information about the circumstances of the disappearance (e.g., a person left his/her house or was last seen somewhere, and after that, never seen again.). If there was a suspicion that a specific perpetrator disappeared the person, then the violation should be “suspected forced disappearance”.
f) Kidnapping	**This category was used only where the kidnapping was from guerrilla actions, when extortion was involved. If the person died as a result of the kidnapping, the violation should be classified as kidnapping and arbitrary execution.**
g) Others	**This category is for violations not included in the original list, but are violations or other events that the team considered valuable for future analysis or investigation.**
Threats, intimidation	Originally to be used only when there was no other CEH violation which made it possible for the violation to be recorded as a case. Later, also used when this violation was important to a case.
Burning crops	Rarely used. Defined to allow recording of this act as part of a context.
Deprivation of one’s liberty	Any action which aims to deprive the victim of physical liberty. This category was mainly used to classify those violations in which a victim was kept as a captive for a specific period of time, and then freed, tortured, killed or disappeared. This violation could happen several times to a same person, if she or he was transported from one place to another. For example, often a victim would be kept in a specific military building, and then moved to another one where s/he could have been tortured, and then to another location, and so on until s/he was freed, killed, or never appeared again.
Forced recruitment	Not easy to distinguish from other deprivations of liberty. Only when the testimony gave specific information that the victim had been forced into military service was this violation assigned.
Homicide	*Not a CEH violation.* This violation was used when the testimony gave clear elements to conclude that the death of a person was not the result of political violence.
Dead or injured combatants	*Not a CEH violation.* For historical record only.

Prisoners of war	*Not a CEH violation.* Recorded combatants who were taken prisoner by the military. If other CEH violations (i.e., torture) occurred during the detention, these were recorded.

Appendix 3

Key Words

Key words are categories that make it possible to classify information according to qualitative criteria. These key words were classified into the following 12 major groups. A summary of the key words follows:

Human rights violations and cases of violence	violations that were not registered as types; i.e., forcing a person to commit acts of violence against one's family or community; forcing a person to witness acts of violence; publicly displaying cadavers; committing extremely cruel actions; destroying housing; violating other rights, such as civil or political rights; persecuting populations; etc.
Violations to cultural integrity	violations committed in relation to the ethnic background of the victims (indigenous people)
Strategies of parties to the peace accords Specific military strategies Specific guerrilla strategies	strategies and actions employed by armed actors against the population: among others, forced recruitment; massive oppression; sociological actions; prisoners of war; infiltration; scorched earth policies; use of paramilitary groups; use of informants; accusing a person of being a guerrilla or collaborator with the military, etc.
Modus operandi	themes related to the way in which parties to the peace accords acted: the use of disguises; acting like other parties to the peace accords; wearing hoods; use of arms; use of specific vehicles, etc.
General context	information on local conflicts; local power structures (social, economic, political, religious, and the state); impunity, etc.
Community reaction mechanisms	alternatives sought by the civilian population to protect themselves from violence: forming popular organizations; forming communities of people in resistance; fleeing; displacement; hiding in the mountains, etc.
Consequences of armed conflict	the effects of war: poverty; displacement; dispossession of land; physical and psychological illnesses; marginalization; returnees, etc.
Women victims of violence Child victims of violence	specific violence against women and children respectively: sexual assault; exploitation and forced work against women; persecution; abandonment; trade and exploitation of children, among others
Cemeteries	hidden graves, communal graves and exhumations, and characteristics of violence in Guatemala

Appendix 4

Perpetrator groups

The perpetrators could be identified in a collective way by the group to which they belonged. The following is the listing of the perpetrator groups that were used in the project.

Military	41 identified military sections
National police	17 identified National Police sections
National Guard (*guardia de hacienda*)	
Security forces	
Military Commissioners	
Self-defense Civil Patrols (PACs, *Patrullas de Autodefensa Civil*)	
Paramilitary groups ("Death Squads")	at least 10 different death squads were identified
Guatemalan Revolutionary National Unity (URNG, the guerilla umbrella organization)	*Ejercito Guerrillero de los Pobres* *Fuerzas Armadas Revolucionarias* *Organizacion del Pueblo en Armas* *Partido Guatemalteco del Trabajo*
Armed groups	
Unidentified	
Civilians	plantation owners, businesspersons, etc.
Public employees	
Mexican military	
Mexican police	
Ex-military	
Ex-guerrilla	

Chapter 9

The Guatemalan Commission for Historical Clarification: *Database Representation*

Humberto Sequeira

Introduction

The purpose of the database for the Guatemalan Commission for Historical Clarification (CEH) was to make it possible to process the human rights violations reports collected by the investigators who did the field work in the areas most affected by the violence in Guatemala from 1960 to 1996. The design was straightforward and simple to implement, and allowed for improvement and expansion of the database to other areas of the project.

When I joined the CEH, my experience in Information Systems was in traditional commercial areas: Point of Sales, Inventory, Export/Import, Container Control, etc. The concept of a human rights information system was new to me, and my first reaction was negative. My initial concern was that the case capturing system and database designs were almost complete. In addition, Visual FoxPro, the database application program in use, was not what I would have expected to use for a large-scale project. But I focused on the project as a whole and decided to do the best work I could with the available tools.

The choice of Visual FoxPro was a good one, despite some initial problems. Its ease of design and programming proved invaluable. In addition, we were not limited in the number of users accessing the database. Hence, in-line SQL statements proved to be excellent for testing and searching the database for errors and basic information, and for producing the tables needed for the CEH team's analyses and studies. While the selection of Windows NT Server as our server software was not optimal from the standpoint of speed, it was easy to administer. In fact, the server went down only twice in a year's work. Speed was the major problem with our server and if the CEH had been network-ready, we would have had difficulty accessing, manipulating and distributing the information. Our server was too large for the database, considering it took less than 500 MB to run the database and its software. With distributed systems we could have done much more than only serve the database team. Distributed systems would have supported more extensive collaboration on various subjects, and the ability to share resources would have been an asset.

One major disappointment was that the CEH was not network-ready in their central offices. This was the most troublesome aspect of the otherwise excellent work environment. Without a network we could not give some teams the timely information they needed to do their jobs. Also, we could not track the information the database was producing and some members lacked the technical and statistical knowledge needed to understand some of the graphs, statistics and reports they received. This created an uncomfortable relationship between the database team and other CEH workers.

In this paper, I discuss the information we used to design, implement and develop a database system that addresses human rights violations, which was our principal interest. I describe the current system with proposed improvements, its design, development, testing, training, implementing, contingency proposals and time/cost estimation, with proposed improvements. In addition, I offer recommendations for future human rights database systems in Appendix 1.

Database Representation

At first, an in-house programmer was hired to design the system to collect information from testimonies received by the CEH during their fieldwork. However, the pressure of deadlines, working with new software and intense work environment caused him to resign. Subsequently, Assist, a Guatemalan software firm was hired to carry out the design.

From a user's point of view, the current system was divided into the following parts:

1. Case Summary
2. Violation(s) Pattern
3. Victim(s) General Information

4. Person(s) Denounced for the Violation(s), that is, the Perpetrator(s), the person(s) responsible for the violation(s)
5. Person(s) who Denounced the Perpetrator (Deponent)

This information comprises a *Case*. Using this structure, we obtained valuable information. I explain and discuss each of the above parts in detail in the following sections.[1]

Case Summary.

The case summary is where we kept general information about the violation(s) being reported. This is general information, and was not be used for statistical analysis.

In this summary we had information such as:

- researcher name
- interview date and place
- if the interview was taped and how many cassettes were used
- documentation given to the CEH
- given case name (from the analysts' team)
- general date of the violation or the period in which the violations occurred (also from the analysts' team)
- case summary (general overview of the case)
- comments and questions about the case raised by the database analyst
- general geographic information
- key words in the violation(s) pattern
- whether this case was reported to other organizations
- number of victims mentioned in the case

We added *markers* to validate the cases and violations. A team of CEH lawyers made legal judgements about the cases and violations to provide the validations.

Some information collected in the Case Summary is redundant, but it was useful for an overall view of the case. Among the redundancies were date and place where the violations(s) took place and the number of victims mentioned in the case. This information was also stored in Violations Pattern (2) and the victims were also counted in 3. Victim(s) General Information (3), discussed below.

These redundancies were a primary cause of discrepancies in the accounts of victims, but were an aid to quickly showing the number of victims in cases. Thus, massacres were identified more easily. In my opinion, these redundancies should not be eliminated from the Case Summary. They should appear in the Case Summary and the system should provide an account of them, but this should not be a user-editable field.

Keywords gave us the ability to store information that otherwise would have been lost. This was where data such as *Modus Operandi*, Military Strategy, Violence Against Children and Women, Cruel Actions, Destruction of Goods, and Religious Violations, etc., was stored.

However, researchers should not make conclusions based on statistical analysis using the keywords for two reasons. First, the key words qualified **cases**, not victims or violations, and cases do not have clear substantive boundaries that allow them to be meaningful quantitative units. Second, the key words were not always applied with the care and precision that were customarily employed in other classifications. The purpose of the key words was to allow analysts to group cases that share some idea such that the cases could be found again and revisited for qualitative in-depth analysis.

We were able to group Cases in important categories using the Name Assigned to the Case. We used it to differentiate between *Normal* Cases and *Massacre* Cases. We were originally using the keyword Massacre to identify the Cases that had more than five Arbitrary Executions. However, problems in the definition of massacre caused some cases to be identified as massacres and others to be left out. So in the final stage of the CEH, we agreed to use the first part of the case name as Massacre to identify massacres. This was important, because we then could differentiate between Massacres and Normal Cases. A keyword validation is a more useful method, because the

[1] Editors' note. The reader of other papers in these proceedings will notice that this structuring is defined elsewhere. We retained these redundancies so that each paper is self-explanatory.

name can change and all references to a case being considered as massacre will be lost. However, with the use of a key word to identify a case called a massacre we have an historical analysis of the changing definition of massacre during the project.

Violation(s) Pattern.

The Violation(s) Pattern is where we kept track of **what, where, when** and **to whom**. As discussed in (Ball, 1996), this structure is essential when the case has more than one violation. Clarifying the structure of this information component was the most discussed aspect of how the data were represented in the database.

Here we stored the following information:

- Violation order
- Type of violation
- Certainty level for the violation
- Place of the violation
- Date of the violation
- Group responsible for the violation
- Type of responsibility for the group responsible
- Type of evidence about the responsible group
- Number of identified victims
- Number of collective and anonymous victims affected by the violation(s)
- Number of victims who died
- Disposition of body
- Location of the burial site
- Information about victim death
- Was body identified
- Was body found
- Was the victim tortured
- What kind of torture the victim received
- How the person reporting the violation knew who committed the violation.

As you can see, the Violation(s) Pattern was the category where we stored most of the information given to members of the CEH. It includes some redundant information, which concerned the victim's account according to the violation pattern. The violation pattern was the easiest way for us to represent violations to one or more persons. Some information here, like torture information, was also presented in graphical form.

Due to this design the representation of **multiple** victims receiving one or more violations was excellent, but it was not the best way to represent **one** person receiving one or more violations. The reason that this was not a good way to deal with one person receiving one or more violations was that is that the basis of our system design was to use violations as the unit of analysis rather than victims. We could give our co-workers breakdowns based on victims, but we wanted them to think in terms of violations. Some of the officials at the CEH could not grasp this approach. This is understandable, because the information they wanted to share with others were personal accounts, such as how many people were affected, not how many times they were affected.[2]

For example, some violations, like Arbitrary Execution (or any other in which the victim died or disappeared), were straightforward mappings of persons to violations in a one-one relation. But others like Freedom Deprivation were not so clear. In such a case, one person can be deprived of his/her liberty on more than one occasion, not only in the same case, but also if the person appeared in another case and was again deprived of liberty. Thus, we would have a one-to-many relation of one person to multiple violations.

It is important to remember not to have more than one death-causing violation in one pattern. This may seem pretty clear, but it happened to us occasionally. We took the necessary steps to fix this problem. We used a listing on the person's name, place and time of violations, along with age and parents' name. We also got some help in dealing with this problem when we stored violations. For instance, we have the Violation Order field, as we capture violations in an orderly fashion. So if

[2] For an explanation of the importance of this approach, see (Ball, 1996). Patrick Ball's concepts and guidance were of great assistance to me in my work on this project.

a Freedom violation occurred first and then a Sexual Violation, then the Freedom violation will be given a lower order number. The order field was also useful in pinpointing the responsible forces participating in the violation, so in the same order number we could have the Army participating in the same violation with the Military Commissioners.[3] The restriction for this use is that the violation must occur on the same place and date.

The following is an example in which the Army, Military Police and Civilians participate in one Sexual Violation.

ORDER	VIOLATION	GROUP RESPONSIBLE
1	Sexual Violation	Army
1		Military Police
1		Civilians

Some information was infrequently used, either for information analysis or questioning by the investigators making the interview. An example is information about the disposition of the body and whether the burial site of the victim was known. Also, in this category was information as to the death of the victim, and if the body was identified, or not found at all.

As we discussed, during the course of the project we realized that we had to update the Violation Pattern, and we came up with a solution, which will be discussed in detail in the section on database design.

Victim(s) General Information.

Here we store the information about the person suffering the violation. This "person" can be an individual or a group. If it is an individual person this part must include the minimum required information. Or this information may describe a group of persons who have some characteristics in common. In particular, we store here the following information:

- Last Name
- Second Last Name (Mother's Last Name)
- First Name
- Alias (es)
- Sex
- Documentation for the person (if given)
- Date of Birth (DOB)
- Age at the Time of the Violation (if DOB not given)
- Year of age (if DOB not given)
- Certainty on the DOB
- Nationality
- Place of Birth
- Language
- Organizational Sector of Victim
- Residence at the Time of the Violation
- If the Victim Had to Move Because of the War
- Name of the Father
- Mother
- Wife (if married)
- Sons/Daughters (all these in the form of Last Name, Second Last Name and Name and Ages at the time of the violations or in present time)
- Civil Status
- Comments

[3] *Ejercito* and *Comisionados Miltares*¸ respectively.

As mentioned earlier, the victims were categorized as Individual or Group. The information stored for both categories was the same except for some information that was unnecessary for Group victims such as age, date of birth, relatives' names, etc.

In the Group information we entered information regarding a group of persons, such as Sex, Birthplace, Location Where Found, Age Range, and some other general terms. We used this information in the name we gave to the group. For example, if a group of children was found dead near the river Rio Negro, we used the Last Name, Second Last Name and Name to call them "Children of Rio Negro."

We tried to determine an approximate number for how many people were affected by the violation using these classes; 2, 4, 5, 10, 11-20, 20-50, more than 50.

Sometimes it was not possible to use the classes, and terms such as "many," "a few," and "a lot" would appear, indicating that we had no approximate number to use. As time went on we needed a number to assign to these victims. We were conservative and assigned an approximate value of two. We also were faced with the anonymous victim assignment problem and gave them the same approximate number of two. The Anonymous victims were people about whom we knew almost nothing, and for that reason we stored their information in the Violation(s) Pattern.

Perpetrator (s) General Information

The information stored here closely resembled that stored in the individual victim component. This included the following information:

- Last Name
- Second Last Name (Mother's Last Name)
- Name
- Alias (es)
- Sex
- Documentation for the person (if given)
- Date of Birth
- Age at the Time of the Violation (if DOB was not given)
- Year of age (if DOB was not given)
- Certainty of the DOB
- Nationality
- Place of Birth
- Language
- To which group the person belonged and his role in that organization
- With what violations and cases this person was charged
- Type of Responsibility and Evidence for this Accusation
- If the person participated in other violations at any time or any place
- If the person reporting the violation knew where this person is now going and where he might be located
- Comments.

The most secret materials kept at the CEH were the perpetrator's name along with the information about the person reporting the violation, because it was essential that the information given to the CEH should not be used for revenge. **In fact, the most important aspect of the Database is to produce information that is without bias and not manipulated in any way to benefit any group or person in particular.**

The data stored here was information about the individual perpetrators identified in a violation or violations. We soon found out that an individual perpetrator could appear in more than one case. This is the same as for victims or persons reporting the violation to the CEH, and we made the necessary adjustment to reflect this relation. A victim can also be a person denounced for a violation or/and a person reporting a violation in some other case. The lack of accurate information was responsible for the problem that a person might have been mistakenly referenced as a perpetrator; however the analysis team made its best effort to avoid this situation.

There can be many persons responsible for a case, but only one can be represented as the Person Denounced for the Violation(s), the Perpetrator who is the person responsible for the violation(s). Also, we did not accept a general description such as "Person living in the village of Barraza in 1991." We demanded a name or at least, an alias.

Person(s) Who Denounced the Perpetrator

Here we stored the information for the persons. (*deponents*) reporting the violations to the CEH. As with the victim(s) general information part, it can represent an individual or a group of persons. We stored here the following information:

- Last Name
- Second Last Name (Mother's Last Name)
- Name
- Sex
- Documentation for the person (if given)
- Date of birth (DOB)
- Actual age (if DOB not given or known)
- Year of age
- (If DOB not given)
- Certainty of the DOB
- Nationality
- Place of birth
- Language
- If the person making the report is an individual accused of a violation (from an analyst's point of view) and if that individual is accused
- The perpetrator case and identification number
- If the person making the report is a victim
- The victim's case and identification number.
- If the person making the report was a refugee
- Displaced person
- Resettled person
- Returnee
- Survivor of a massacre
- Victim of some other violation
- Witness or violence survivor
- The relationship of this person to the victims of this case. The information required here is the Case number, victim's identification number and relation to the victim

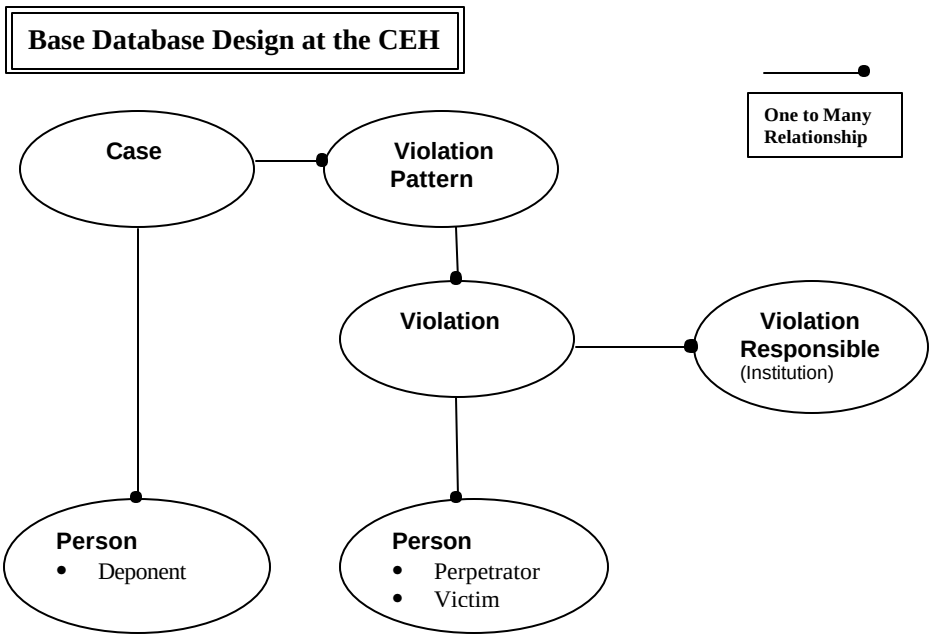
Once again it is important to note that we should always protect the identity of the persons who have trusted the CEH with the information stored by this screen.[4] It should remain only in the archives of the CEH.

It is essential not to change the stored information in some tables. For example, David Sequeira is a victim (victim #2), his mother is Magda Sequeira and his father is Walter Sequeira. We save this information on the victims screen. In this case Magda is also a victim (victim #6). Since David is also the person reporting the violation, some users entered incorrect information stating that the relationship was son/daughter, thus changing the relationship between Magda and David.

Database Design

The following flow chart is an overview of database design at the CEH.

[4] When we refer to a screen, we are referring both to the screen as shown by the computer for data entry and the physical form that may have been used to record the information.



As mentioned previously, Patrick Ball designed the CEH relational database so that we could use SQL syntax to code a wide range of information. The design not only counted violations as the primary unit of analysis, but was flexible enough to support other units of analysis.

The first table used is the CASE table, [5] which is the main table in the database. In this table, we stored the basic case information, such as the information stored in the Case Summary. The field investigators assigned most of the case numbers according to a limited set of numbers given for each area, so that an area such as Guatemala City could range from 1 to 1000, for Zacapa, 1001 to 2000, for Quiché, 2001-3000. In a similar fashion, ranges of numbers covered all the sites where CEH offices operated.

We used a technique called *automatic number generation*, in which case #9 will not necessarily be #9 internally in the database. Instead, case #9 could be internally represented by CA0000257. The cases were numbered, as entered, one by one by the database and given a sequential number. Although you might normally expect that case #9 would be entered before case #1125, with automatic number generation, this is not necessarily true.

In addition to the information in the Case Summary, we stored here the Case Creation Date, Case Modification Date, User Who Entered the Case and User who Modified the Case.

The master key for this table was the CASE_ID, which was used throughout the system.

Following with the case information we also had a table to store the keywords used to qualify a case, CASE_CLV. A case can have more than one qualifying keyword. The information was CASE_ID, PALA_COD and AUTHOR. Using this table, it is possible to record some essential information that was not recorded in the VIOL table and otherwise would have been impossible to obtain. PALA_COD is the keyword code.

If a case was reported to another institution besides the CEH we used a table, CASE_DEN. Here we stored the CASE_ID, the person who reported it, the institution that received the report and the date.

The PATTERN table is the glue that holds together the information in the database, as here we store the Pattern number along with the number of identified, collective and anonymous victims, and the estimated magnitude of the number of persons mentioned ("many," "a few," "a lot," etc.). This table is directly linked to the CASE table in such a way that a case can have several PATTERNS. The key for this table is CASE_ID+PATTERN_ID.

The PATTERN table is the parent table of the VIOL table. In the VIOL table, we stored the pattern number, order number, the violation, the date, the place, and some geographic description to

[5] In the discussions that follow, I follow conventional practice in information system design and implementation of capitalizing the names of tables, keys, codes, and field names when appropriate.

help us identify the place and violation certainty level. The full key for this table is PATTERN_ID+ORDEN_NUM+VIOL_ID, which we can link to the CASE table using the PATTERN table. Thus, if we had only the VIOL table, some of the information would be available, such as **what, when** and **where** but we would still be missing **who** were the perpetrator(s), which is the table I describe next.

We stored information about the organization responsible for the violation in the table VIOL_RSP. We stored the VIOL_ID code for the perpetrator group, type of responsibility, evidence, and place of the group assigned to the violation. In this table we can show if more than one group shared the violation, up to *n* groups. The key for this table is the VIOL_ID.

The table PATR_VICT stores the victim's id number and the pattern in which the victim's violation occurred, so we only need the PATR_ID, PERSON_ID and victim number. Thus, we have all the violations that a person suffered in a given case or in more than one case. It doesn't matter if the victim is individual or collective; we only need the PERSON_ID number. We also use the VICT table as a backup for PATR_VICT; the only difference here was that we also stored the CASE_ID number.

The PERS table stored all the information regarding a person, any kind of person. The key for this table is PERS_ID. Victims (Individual or Collective), Persons Accused of Violations, Persons Reporting a Violation (Individual or Collective), Brothers, Parents, Son/Daughters, Wives, Husbands, etc. Every name of a known or unknown Individual or Collective person is here. By July 25, 1998, we had about 30,000 persons in the PERS table. We had three fields to mark if a person was a Victim, a Perpetrator of Violations and a Person Reporting a Violation. We also had a field to show if the victim was an Individual or a Collective person. Important information regarding the person such as: Full Name, Alias (es), Sex, Age, DOB, Nationality, Language, Documents, Place of Birth, Civil Status, Comments, Age for Deponent.

Because we wanted full information about a person we were faced with a problem in time representation. For example, a person who doesn't remember his DOB (a frequent occurrence) reports a violation he suffered in May 1982 when he states his age as 21. He then reports a violation he suffered in July 1993, when he states his age as 34. But this is 11 years later, and if his 1982 age were correct, then he should be 33. The report was filed in 1996 and in it he says he is 36 years old. If the 1982 age were correct, he should be 35. On the other hand, if the 1993 age was correct he should be 37. Which age do you accept as the age of the person? We represented the last age entered in the database, because the person can't be entered in the database again to show a different age. If this person was reporting a violation, we stored his age in a different field, and when performing calculations we used the best age possible. Still, it is not a guarantee that we had the person's correct age.

We also had problems with the Comment field. Some comments about the person were about when he was a victim. Others were about when he was reporting a violation and others when he was accused of a violation. This was because the victims, persons reporting a violation and perpetrator all share the same information; the individual can have more than one role. This was a programming error and was fixed as soon as I found out about it.

For any dates used in the CEH database, we used the Russian date format, yyyymmdd (year, month, day). We used this format so that we could use zero (00) in representing the month and day. We had to, because on some dates we could not get any certainty at all, for example:

- In the winter of 1986
- Some day in January 1981
- At the end of Easter 1983
- I don't remember when

The IHCH table is where we kept the information about the individual accused of the violation (the perpetrator), such as CASE_ID, PERS_ID and IHCH number. This is basic information about this person and useful if we knew that a person participated in a violation, but were not sure what the violation was. More complete information on the person's participation can be found in IHCH_VIOL which included the CASE_ID, PATR_ID,VIOL_ID and evidence and responsibility.

The PERS_CLS table stores information on the activities of the victim, or to what organizations the person belonged (such as union, religious organization, etc.). The information needed is the PERS_ID and VICTCLS_COD (Type of victim according to our catalog).

Still working with the PERS table, we used the PERS_HIJO to store relations between people. Here we show the relationships of the victims or persons reporting the violations to the CEH. This is a good approach and worked for a while, but catalog limitations (imposed by ourselves) made us misplace some information. For example, if a person only had one parent, we could not know if it was the person's father or mother, because we used Mother/Father as one category.

We used the group of tables PATR_DS, PATR_MU and PATR_TO to store information on whether the victim was disappeared, dead or tortured, respectively. This information is basic and the only thing required here is the victim's form or screen number along with the PATR_ID.

The USERS table stored the information about users who entered information on the database. This information included user code, name, security level and password.

Some of the other tables in the database are the catalogs, which we used to complete the information, such as the Institutions, Type of Victim, Key word, Relation, Language, Nationality and others. We hope to see them when the CEH authorizes the use of the information stored in the database.

Application Programming: Development, Testing, Training, Implementing

As mentioned previously, Assist, a Guatemalan company, did the initial programming and I did the balance. The system was network-ready and with minor adjustments after it was implemented, worked quite well. In this section, I give some details of the process.

Once the database design was provided, Assist delivered their product in about five weeks. I needed completion quickly, so we agreed that I would finish the application and they would correct some residual problems. This was of value to me and I became familiar with the final product. At first I wasn't happy with Visual FoxPro as the development tool, as I formerly worked with UNIX database systems and used Visual Basic for a visual interface. However, Visual FoxPro had an edge: no per-user cost to access the database, inline SQL access (not possible with Visual Basic), no need to back me up in UNIX administration and most important, no delay in startup. So I gracefully accepted Visual FoxPro as the tool of choice. I have not regretted that decision, and would only complain about the slow speed of operation.

The first version of the system was a FoxPro Application, which meant it ran inside the Visual FoxPro Environment and each machine had to have FoxPro installed. We had a number of problems with this version. The application kept producing Error Protection Faults (EPF) either from the Kernel or from Visual FoxPro, two or three times a day per machine. Fortunately, the tables were not damaged (as in previous versions of FoxPro) but too much time was lost, as we had six machines. So we built an executable file from FoxPro and reduced system outages to only two or three times per month resulting from problems affecting the Kernel.

The FoxPro executable is not a true Windows executable, but behaves as one. Using the FoxPro executable, we did not need to install FoxPro in the machines, only several DLL files necessary for some functions. Everything else is in the executable file.

As mentioned, we did not have the proper time to test the systems (about one half-hour per change no matter what size) but it worked out well. The system was solid, so we did not lose a lot of time. Training was also informal, but we tried to give the people working with the system at least one full day to become familiar with it.

The CEH system can still be improved. Some procedures can be automated to give the user entering the case more speed and less overhead. These are:

Problem	**Solution**
Beginning and Ending date of the Case	This can be taken from the Pattern form automatically (scanning all violations)
Places where the case took place	This can be taken from the Pattern form automatically (scanning all violations)
Keywords for the cases	Keywords cannot be repeated

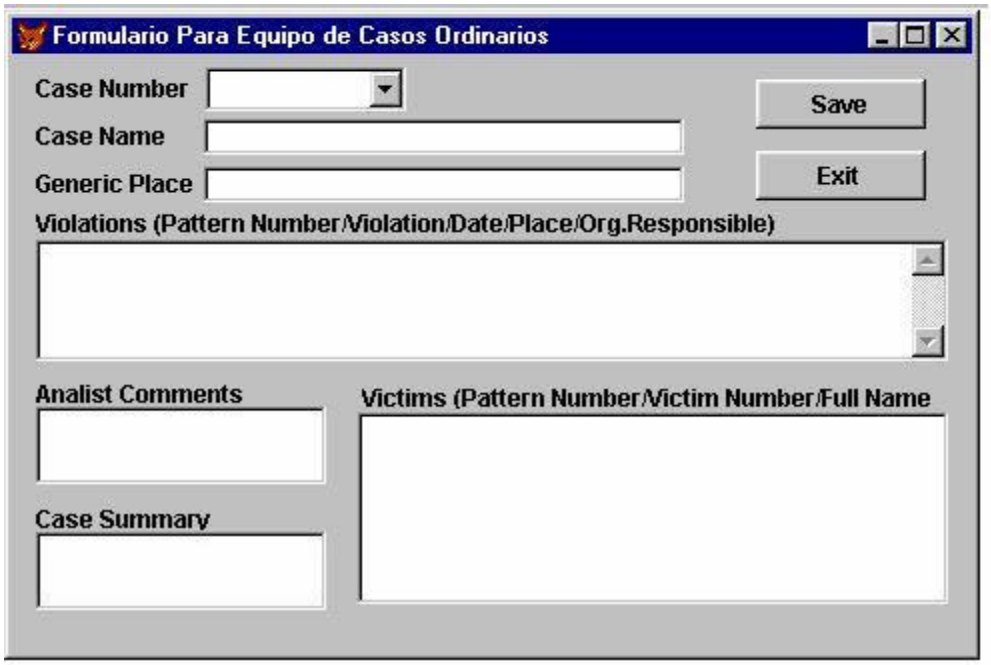
Other institutions where a violation was reported	A catalog of Human Rights or other institutions can be made, including military institutions
Only one interviewer can record the case	More than one interviewer can participate in recording the interview or case
Identified, Collective and Anonymous victims in the Case form	This should be eliminated from the case form, and can be obtained automatically from the Pattern and Victims form
Identified and Collective victims from the Pattern Form	This should be eliminated from the pattern form, and can be obtained automatically from the Victims form
Where the victim lived when the violation occurred	This should be coded from the Places Catalog
Language of the persons	This should be the ethnic group to which they belonged; their language should be stored in a separate location
Type of victim	Type of victim can't be repeated. If this happens, more violations would be counted in the type of victim analysis
Personal relations	They all should be independent. Do not use something like "Father/Mother." Instead, separate them
Comments of the victims	Should be shared in all the forms. Should be common to the person
Group to which the person belongs	Group cannot be repeated. The only case is when the charge inside the group changes; thus we can keep a history of the person (mostly related to perpetrators)

As to the application programming, it doesn't matter in what language or on what platform it is developed. It should be fast for the users, and aid them in any way possible in their daily routine. This is dull work, with almost no changes in the daily routine. Every step should be taken to make the application flawless.

As time went by, a need for more information was evident. The CEH database was network-ready, but the CEH organization was not. Some teams needed to work on the cases to make quick studies (they also needed to write a short summaries), and needed many varied and unpredictable types of information. One of these teams was the Ordinary Cases Team, which needed ALL the information for the case on one page (except for perpetrators and deponents, not given for security reasons).

Accordingly, I decided to build an off-line system with requested updates, so that when the user finished with the cases I could provide them with an update, keeping the summary they have written. This took a lot of my time, because I had to use diskettes to transfer the program, and there were eight persons working with this software.

This subsystem appeared like this to the user:



I created another sub-system for the Recommendations Team. In it, I showed the comments of the deponents and allowed building keywords on each comment. We could produce statistics on the most common keywords to represent the concepts appearing in the information being collected.

In the Appendixes to this paper, I give recommendations for system automation in similar projects (Appendix 1), detailed information on our SQL queries (Appendix 2), and my recommendations for information integrity and security (Appendix 3).

Lessons Learned

In Appendixes 1 and 3 I give the body of my recommendations for future similar projects. In this section, I present other, more specific lessons that were learned in the course of my work on the CEH project.

Entity, Function	Problem	Lesson	Issues
Case reports	Started giving general date and place, Members of the CEH increasingly wanted more exact information in their reports	Show all the places and dates of the violations in a concise manner, take information from the violation (i.e., where it occurred and when). Team must decide how to handle use of information in case reports.	Must not be a user-editable field. Only so much information can be shown in a listing.
Violation pattern screen	Sometimes users would open screen and then stop.	Know who opens the violation pattern and then stops	
Personal information	Inconsistent ages reported by individuals in different cases	Make table with PATR_ID, PERS_ID and AGE, to store the information of the person's age in more than one case.	This table is a modification of the PATR_VICT table in use at CEH, without the AGE field.
Screens	Inconsistent information	Make all the information on the form or screen consistent	Discipline in the design process
Y2K problems	Unpredictable	Store dates in the Russian format using 00 for month and days	Easy to do, was done for CEH
Personal information	Confusion about whether the parent was mother or father.	Use two separate categories	
Hardware & software	Some machines had problems with certain installed virtual drivers	Remove the offending drivers in advance.	Knowing in advance which drivers are likely to cause problems
Workplace	Pressure on users resulting from daily quotas and other factors, limited individual capabilities for pressure and for work	Database administrators should review with data entry personnel to get speed without compromise of accuracy, individual progress reports rather than competition Select and hire users with necessary knowledge and ability to work under pressure.	May call for change in management style of database administrators Verification of experience Getting users with similar capabilities

Appendix 1

Recommendations on System Automation.[6]

Introduction

In this Appendix, I make recommendations for system automation in any similar project. In my work at the CEH, I became concerned about the CEH structure's weakness with relation to the group work that was about to be performed for the CEH report. Also, I make observations on the CEH computing structure and compare its structure with other more general projects and databases. I believe that if followed on similar human rights projects, the recommendations in this Appendix will enable future systems to avoid these problems.

Today's technology makes our work easy: it helps us to plan, organize and administer work. In collaborative interfaces such as that found in the CEH, the lack of planning in topology for computer networks and data servers for document administration had negative effects, causing setbacks and problems in the administrative process.

I document the need for software (locally/internally produced programs or packages from established companies) and hardware (equipment and accessories). Without these two components, it would be impossible to achieve these tasks. Today, the need for access to, and administration of, information mandates these programs communicate with each other. Hence, the organization can be seen as a sensitive, responsive, and interconnected system. Some of the necessary programs for working on all aspects of the documents are so simple and standardized that they can be internally or locally produced. However, others are so complex that the best investment might be to buy an external package that produces the desired results and meets our information and communication needs. The decision-makers in future projects should make these decisions explicitly to fit their particular needs.

All of the programs should be multipurpose and have the same user-friendly and intuitive interface. And of course, in the final analysis, they should meet both the needs of the user team and the organization.

As to hardware, the shared use of available resources (laser printers, color printers, modems, scanners, and hard discs) is one of the most pressing issues. By sharing resources, the organization can focus on acquiring the correct peripheral equipment and reduce costs. For example, a printer for every computer is no longer necessary. Information could be stored in a centralized manner in archive servers, and these servers could be divided into work groups.

Our purposes are special and human rights organizations are almost invariably under-funded. One solution is to contact the vendors of software in the beginning of a project, and possibly the Business Software Alliance (BSA), to obtain licenses for each package. By explaining the purposes and objectives of the organization, it is likely that the organization could obtain savings or discounts in acquiring software. The software vendors may be responsive to knowing that the use of their product will become public knowledge. This has in fact occurred frequently in similar situations. Often the vendors are willing to donate or provide at a large discount the version prior to the current version they are marketing.

Systems Automation for the Human Rights Project

I recommend four major approaches for the system automation plan:

1. Magnetic identification system for team members.
2. Bar code identification system for documents, supplies, and office furniture.
3. System physical facility including wiring, equipment, networking, communications, etc., for the project systems.
4. Workflow system for automating the office.

Magnetic identification system for project members

With the use of a unique identity card for all project members the physical and logical administration cost could be reduced. Personnel could become members of one or more different work

[6] I thank Walter Sequeira and Eduardo Meyer for their participation in this paper.

groups in which they would be allowed access to different documents generated and/or used in the project. This card would be valid only during the employee's physical participation in the project.

Physical security could be centralized with the personal identification number on the user's card; both the personnel and security offices would administer access to offices.

This magnetic code is generally the same as that seen on credit cards. The following benefits would occur with the use of the magnetic identification system for project members:

- Any team member would be easily located.
- There would be oversight of the project team members' time.
- Access to documents would appear more dynamic and beneficial (in combination with point three) since every document would have an owner at all times (whether its place of origin or the person who requested it) and would be easily located, ready for its return or subsequent processing.

The project member's photograph should be included on the identity card in order to make visual confirmation simpler and faster. The identity card should be non-transferable and should be destroyed once the employee terminates participation in the project.

Bar code identification system for documents, supplies, and office equipment

The bar code is a simple and economical way to label multiple classes of physical objects. In a project of this magnitude, management of documents is a priority. Documents should be maintained in good condition while still being easy to access. Their security and sensitivity should also be taken into account. Documents can be one page, lists, graphics or complete books; this makes their complete identification necessary.

The use of the bar code is a viable and, at the same time, economical option. All of the documents that are accessible to project members should have a unique code created locally in the project.

Frequent problems that occur in locating, receiving and delivering documents would be drastically reduced. With the combined use of project member identity cards, the following information on documents could be known:

- to whom it was delivered
- who has it
- where it is
- how many times it has been requested and by whom
- what time it has been used and by whom

The use of bar codes for the office team cuts administrative costs and time. Thus, office team control will be transparent and, above all, organized. The organization will have knowledge of the location of teams and to whom they were assigned.

The administration of supplies in a project of this magnitude is of vital importance since it offers:

- Projections that can be made monthly, tri-annually, biannually and annually.
- Reports on different teams' and departments' demands.
- Help in estimating the project budget using different available projections.

The use of bar codes together with the fourth theme of workflow would form a formidable combination in relation to the access and classification of documents when the option of data entry is used.

Structured wiring and necessary equipment for the project systems

None of the proposed systems in this paper would be feasible if the project did not have an adequate infrastructure. The network is the spinal cord of the whole system. The design should plan for such extra capacity in relation to the physical capacity of the location at the time of planning, to maximize workstations, and to estimate the different types of users that can occupy these places and their volume of demands and production.

Different types of project teams and their members generate different demands, such as:

- Groups that work exclusively as a team with multiple documents.
- Groups of members modifying the same documents.
- Team members with pressing demands for manipulating graphs.
- Team members with high levels of demand for printing, some in color.
- Specific employees from the different databases used in the project.
- Mobile uses, such as portable or cellular PC's.

All of these demands can be centralized in equipment (computers). The number of users will determine these demands and their specific needs.

Once a locale is chosen, requests for networking should be solicited, always keeping in mind an available route in case of growth.

One important consideration is the physical security of the wiring, since it cannot in any way jeopardize the information that will flow in it. This is the main purpose the wiring fills for a company or professional person, since the wiring necessary for linking more than one floor in a building is laborious and the limitations of distance and security require different hardware and software selections and communication protocols.

Many kinds of available networks must be taken into consideration, and the types, models, and features are constantly changing. At the current time (mid-1999), for an organization with fewer than 200 members and individual teams with fewer than 30 members, I recommend a local area Ethernet type network, the most popular, easiest to administer and most economical in a PC interface.

Workflow system for automating the office

The organization of information, both electronic information and traditional, is a project of vital importance. A new term in this realm is the *workflow*. With the combined use of hardware and software in the form of computer programs, scanners, faxes, modems and others, documents can be administered in an orderly manner.

With these new technologies documents can be:

- Indexed with a unique key
- Stored in a specific location
- Consulted on screen
- Printed on some device
- Sent by fax
- Accessed by different search methods
- Modified for various users

Often, a document needs to be revised or modified by more than one person until it arrives at its final destination. Here is a typical workflow for working with such a document; it goes from Workstation 0 (origination) to Workstation 5 (delivery of final document) after having been revised by employees 1, 2, 3 and 4. The workflow is defined in such a way that the document can be returned to the previous employee if it contains errors, or passed to the next employee if the previous revision is satisfactory. The document can be accessed for consultation at any point in the flow and other users can make modifications on it (if that is how it was defined in the flow). Maximum time periods that an employee can have the document can also be established. If the system detects that the maximum time has passed, administrative alerts are sent to the predefined users.

One must take into account that at no moment has the document been printed; all of the revisions and modifications have been done electronically. This is a small example of a document. The flow can be increased in complexity, demands, users, functions, etc. All of these functions define an electronic office. Imagine how at the moment they are generated, all of our memoranda, faxes, lists, graphs, etc., would have the ease of being sent to the necessary work posts.

One of the disadvantages of these systems is that the people who design the workflow must have a clear idea of the needs and demands of every flow, since not all require the same restrictions nor do they go to the same users. Another disadvantage is that it may be difficult initially for people to work from the screen, rather than with hard copy. However, by providing them an adequate monitor, an optimal resolution screen and an infrastructure of fast networks, these obstacles can be overcome. The workflow will not completely replace paper since some lists, reports, and other types of documents that call for revision will be printed.

There are different kinds of workflow software on the market today. Strategic project planning is what determines if only one will be used or if the configuration will be a combination in relation to the size of the work groups. That decision will determine the type of software to be used.

Server/Workstation Hardware & Software

This is an open topic. It primarily depends on the economic resources available for the project and the persons involved in the planning and their preferences. Some projects will be small;

others will have almost all the correct technology available. Availability of product and support in a given geographical location can also influence the choice of hardware and software. My choice for software is to go with Microsoft® products, except for the database, where you can choose from many vendors with more robust products.

Based on my CEH experience, I recommend consideration of the following:

Feature	Server Side	WorkStation Side
Operating System	WINDOWS NT SERVER 4.0 or better Any UNIX (Aix, Digital, SCO) or LINUX flavor	WINDOWS NT WORKSTATION 4.0 or better Windows 95/98 (If necessary access through emulation, ODBC or Middleware)
Software Network Protocol	Will depend on OS, but TCP/IP for large projects. NetBEUI for small projects.	Will depend on OS, but TCP/IP for large projects. NetBEUI for small projects.
Hardware Network Protocol	Ethernet/Fast Ethernet Fiber Optics	Ethernet/Fast Ethernet Fiber Optics
Database	Will depend on the Server platform Oracle from Oracle Corp. DB2 from IBM Corp. Informix from Informix SQL Server from Microsoft Access from Microsoft Visual FoxPro from Microsoft	N/A

Appendix 2

Information Requests and Structure Query Language

Structured Query Language (SQL) was not an area of expertise for me at the start of the CEH project, but it ultimately became of great importance to me and to the project as a whole. By the end of the project, programmers were producing up to 20 SQL queries a day, most of which were quite complex.

SQL is a specialized language by which a programmer or user can query a database that is using to the Open Database Connectivity (ODBC) standard. It uses drivers to access databases in many different formats. In our case, we used EXCEL for working with the tables produced by an SQL query, and FoxPro as the data source.

Our work with SQL took place in three relatively distinct phases in chronological sequence.

In Phase 1 we were making simple queries because the information requests we received were simple. Most of them were queries at the case level. For example, listings such as in Figure 1.[7]

Figure 1. Listings I.

Query	From Table	Comments
Case Number	CASO	
Case Name	CASO	Given by the analyst
Certainty for the Case		Based on the violations on the case
Generic Date	CASO	A generic date for when the violations occurred. The Analyst chooses the date.
Violations	VIOL	All violations in the case.
Generic Place	CASO	A generic place where the violations occurred. The Analyst chooses the place.

In Phase 2 we started producing violation-based information, such as Figures 2-4, below.

Figure 2. Violations, I.

Query	From Table	Comments
Specific Violation	VIOL	Violation Name from The Violations Catalog
Identified Victims Count **	PATR_VICT	
Collective Victims Count **	PATR_VICT	Identified + Collective
Anonymous Victims Count **	PATR	Identified + Collective + Anonymous

** This means we had three columns in the table to count the victims.

With the availability of the results from these queries, the number of requests from members of the CEH team increased rapidly. Due to the organization of our Catalog, some of the killed were

[7] Figure and table numbers in this appendix are sequential within the appendix and do not relate to the captions in the full paper.

adding to the violation count (for example, Arbitrary Execution, Death as a Result of Violence). For that reason we made the grouping shown in Figure 3, below.

Figure 3. Violations, II.

Query	From Table	Comments
Generic Violation	VIOL	Violation Name from The Violations Catalog
Specific Violation	VIOL	Violation Name from The Violations Catalog
Identified Victims Count	PATR_VICT	
Collective Victims Count	PATR_VICT	Identified + Collective
Anonymous Victims Count	PATR	Identified + Collective + Anonymous

Note: By using Generic Violation, you can obtain all the dead.

The analyses and investigations of the CEH team came to the point where they needed to include particular keywords. A typical such request would be expressed in narrative for as "a listing of cases that include the keywords 'Violence Against Children,' 'Territorial Movements and/or Religious Attacks' in the year 1982 in the province of Huehuetenango, including 'Responsible Groups.'" The response would be a level case listing such as in Figure 4, below.

Figure 4. Listings, II.

Query	From Table	Comments
Case Number	CASO	
Case Name	CASO	
Certainty for the Case		Based on the violations on the case
Generic Date	CASO	
Violations	VIOL	
Responsible Groups	VIOL_RSP	All the groups that participated in all the violations in the case.
Generic Place	CASO	

In Phase 3, we added additional information on violations. First we included information that is in a one-to-one relationship to the violations, such as the place where it happened. Thus, we added Department/Province to the table, and then information that was in more complex relationships, as shown in Figure 5.

Table 5. Additions to the tables

Option	Comments
Age (Categorized/Grouped)	Initially for only the Identified Victims, but later for Collective and Anonymous victims as well. We used the number –1 to identify the age for collective and anonymous victims, and mapped the violations to the general violation table.
Date (Month-Year/Trimester/Semester/Year)	This was to make it possible to divide the year in any way users wanted.
Place (Department and Cities)	At first only Departments, afterwards including cities where desired by the CEH teams.
Sex	Initially for only the Identified Victims, but later for Collective and Anonymous victims as well.
Forces Responsible (Institutional level Perpetrator)	This is where the violations count will be higher than that from the general (base) violations table because more than one perpetrator can participate in a violation. The first production represented all the forces responsible. The next version would group them to represent groups of interest to the CEH (i.e. URNG, ORPA, and all the other guerrilla groups will be grouped in the "Guerrilla Group." The same was done with the Government Institutions.

Of course the options were mixed, and we ended up with table analyses for many topics. These included: Age/Sex, Place/Language/Sex/Age, Place/Language, Place/Type of Victim, Place/Date, Place/Force Responsible, Place/Date/Force Responsible, and analyses for specific purposes: Massacres, Non-Massacre Analysis, Non-Guerrilla Analysis, Government Forces Analysis, Range of Victims Analysis, Non-Massacre Analysis, and so forth.

When half of the work was done, we decided to create table structures that would produce consistent information. We built structures for Identified, Collective and Anonymous Victims. These structures will lead to tables with information that will satisfy all the future requests we could conceive at the time. If we had done this earlier, it would have simplified our work. These structures were a success from both the user and programmer standpoints. With time, I became quite glad that I had learned how to use SQL, which enabled us to easily and quickly program queries to facilitate the work of the CEH teams.

Appendix 3

Security and COntingency Recommendations

In any field, at any time, we have problems of security and the need for contingency plans. Thus, we must be prepared to deal with these problems in a timely and correct manner. I make the following recommendations to achieve those results:

1. Avert data loss problems by having a tape backup on the server machine, and do automatic backups daily at midnight
2. Check backup tape every morning to see if everything was backed up properly
3. Keep a biweekly backup tape in a secure location outside the database building.
4. Do a backup and fire drill every month
5. Change users' logins and database passwords at least once every 45 days or when a database member leaves
6. Physically check the database area for unknown cables on the roof and floor
7. Close all the diskette operations on the client's machine except one, to be chosen by the DBA
8. Keep Internet access on a separate machine
9. In case of an Intranet, divide users by groups and give specific access to those groups
10. Remember that attacks to the database occur. Most of them will come from within the organization, so be selective in the personnel you choose and the members of the project you physically allow to enter the database
11. Use strong data encryption when moving information outside the database
12. Try to establish an internal audit trail data system within the database, to keep a log of updates and modifications
13. All systems should have a functional UPS, with sufficient uptime to finish the current operation and shut down the computer

References

Ball, Patrick. 1996. *Who Did What to Whom? Planning and implementing a Large Scale Human Rights Data Project.* Washington, DC: American Association for the Advancement of Science.

Chapter 10

The Guatemalan Commission for Historical Clarification: *Generating Analytical Reports*

Eva Scheibreithner

Introduction

The Initial Problems

I started working with the CEH in July 1998, when its investigation had almost come to an end, joining the database team eight months before the final report was released. Thus there were many situations which led to problems that I could have prevented had I been continuously engaged full-time in statistical analysis from the start of the project. In this report I explain the problems faced in producing the needed statistical analyses and how I found solutions. While my solutions were not always the most efficient, I found effective solutions that I could execute under the circumstances.

Before my employment, no one person had worked full time on the statistical analysis. Many different people obtained statistical output, graphs, etc., with the hardware and software that became my workstation. As every person has his/her own way of maintaining logical order, there was no orderly basis from which to work. No one had kept a permanent record, so I didn't know what had been produced, to whom it went, from where it was produced, etc. The source files and Excel worksheets had names based on no logical system that I could ascertain. Every person working with the files used his/her own system of naming and archiving files. The statistical outputs had neither uniform layout nor titles. Every graph looked different, although some of them were the same! I found it possible to identify them only by close examination. No one could find any specific result; hard copies weren't kept and there were no explanations or details describing the process, the variables, or the abbreviations used.

In the beginning I didn't know what to do with the variables. The CEH defined three different types of victims: *individual identified victims* (VICT_IND), *collective victims* (VICT_COL) and *anonymous victims* (VICT_AN). I found Excel files with three columns, corresponding to these three variables. I added another column summing the three to get the total of violations and I got different totals when repeating the process. Some days later, the programmer checked my computer and verified my suspicion that the sums were wrong because my computer had a virus. He asked me about my additional column for sums and then he told me that the variable VICT_COL already included VICT_IND, and the variable VICT_AN was the sum of all three types of victims. I never got the right totals as they had already been included. This is typical of the kind of problems I faced at the start.

The first two weeks I spent looking through the files and worksheets, trying to find a way to organize the structure. Finally I concluded that it was easier and quicker for me not to use the existing structures and to start with a completely new system of archiving, organizing and naming. I archived the old structure and started working with new updated files and structure.

I believe that the integrity, transparency, and safekeeping of the data and results are important. By transparency, I mean that if I was not available and another person was hired to do the statistical analysis, that person would be able to rapidly continue from where I stopped. I designed my system accordingly. I tried to make my system so easy-to-use that even a person who had never before worked with that structure or topic could follow my work in less than two days of study. Although this was not necessary, I know that anyone else who accesses my archives in the future will be able to follow my work without difficulty.

Importing New Data: From Programmer to Statistical Calculation

Specify New Input Files

Input files for statistical calculation were in dBase Data Format, with the extension ".dbf." Accordingly, unless it would create confusion, I use the acronym "DBF" to denote input files obtained for me by programmer queries on the source database.

To specify new DBFs, I needed to know the exact needs of the investigator: for what reason, why, how does he/she want the information? Not unreasonably, the investigators often didn't know exactly how to express themselves in their requests for information and sometimes they asked for impossible or useless information. However, we were almost always able to work it out and once the required information was identified, I passed the description of the needed blocks of data to the programmer, who provided the DBFs.

In the beginning I had some problems specifying the new datasets needed. The programmer wanted to help me and gave me more than I asked for, so that I could use the files again later (his idea). But for me it wasn't very useful. I wanted only the data I requested and had to cut out the additional, unwanted information. In fact, it's easier to work with small files; small files lead to less trouble with Excel, which performs more reliably with smaller blocks of data.

For example, I needed the disappearances for individual victims in Guatemala City between 1980 and 1982. So I asked for the data blocks for violation, region, year, and individual victims. I received, in addition, the gross violation, the collective victims, the anonymous victims and sometimes the subregions, for which I had no use.

Import New DBFs

Using the Excel function Data>ImportData, I imported the DBFs created by the programmer from his computer directly into my Excel workbook. Note that this function (Data>ImportData) appears only after an Open Database Connectivity[1] (ODBC) link to an external data source has been established. In the absence of such a link, this menu item does not appear.

Check New DBFs

The programmer always made the first check of the new DBFs. When the DBFs passed the first check I always made a rough second check after import. This second check consisted of comparing the total violations of the categories of victims (one, two or all three categories) with the up-to-date overview. The up-to-date general overview was made with every new update of the database. It contained the new totals of cases in the database (of all three categories of victims), the totals of violations (for the same categories), the five main violations (selected in the beginning by the CEH commissioners) with their totals for the same three categories. This information was used to inform -- in a short and exact way -- all the personnel working with the CEH who used the database output (commissioners, central team, investigators, database staff, etc.).

Thus, the last up-to-date general overview showed a total violation count for individual victims of 15,233. All the new DBFs that I imported from the programmer had to match with this number. Usually I checked all the three categories of victims (which meant always the individual victims - VICT_IND -, the individual victims plus collective victims - VICT_COL -, and the total of victims - VICT_AN).

Update Data, Import New Data to Existing Excel Files

To update data from the existing Excel files, I used the function Data> UpdateData within the Excel program. As mentioned earlier, this is possible only after an ODBC link to an external data source has been established.

To assure correctness, I updated existing files in a way that was a bit more complicated than necessary. I "artificially" checked every file, every worksheet and manually updated the worksheets, checking before and after the update. I could have done this much faster with a macro. However, I would still have had to check every worksheet of every file if the process succeeded. I

[1] Open Database Connectivity is a Microsoft standard using drivers to access database files in a variety of formats.

would have had to assure that the result was correct if I had made calculations or manually added columns as well. In addition, I would have needed an updated record of all files with all the sheets I wanted to update.

Check New and Old Excel Files After the Import or Update of Data

I used a process to check the files that may appear time-consuming but which I felt was necessary. I created a checking form (Figure 1, below) to standardize the process. I checked all the updated data for the main questions: what, where, when, who, to whom and a special *star-question*, a complex specific question involving the context (an example is given in the appendix).

First I compared the main violations. I required that all the totals agreed with the general overview. Then I looked for more specific details, comparing all the Excel files, one to the other, always using three different attributes.

Figure 1. Checking Form Example

EXCEL FILE	WHAT	WHERE	WHEN	WHO	TO WHOM	STAR QUESTION
Name	compare main violations to general overview, gender, age	using three regions	three different years	Military	URNG	complex question in context

Generally I spent more time checking the information than making graphs or other statistical outputs. Sometimes the process took weeks. Checking and updating about 40 different files, each with at least eight sheets, and verifying with the checking form was time-consuming. But in the end I found mistakes that had been made by people in the database chain (programmers, typists, and analysts), confirming my belief that checking is imperative and cannot be neglected. Don't automatically trust what you see on the screen!

Update Data

Before an update in the database I always discussed with the programmer the DBFs I needed updated. There were DBFs for which we had no further use, and with every step developing new possibilities, we decided not to update old DBFs (those with no further use). When the programmer got my list of still-useful DBFs, he updated them, and, when completed, passed the checked list to me. By the time the CEH report was finished, only about 20% of all the DBFs created in the whole process had been updated.

In November 1998 we updated the database. This was a busy period for statistical analysis. The investigators were finishing their reports, which created a high level of demand for graphs and statistics. I had a list of about 60 files to update. When I gave this list to the programmer asking him to update the DBFs, he was concerned about the amount of work required in view of the total workload on the system. So I took back the list and reviewed it. I eliminated another 25 files and in the end we updated only about 35 DBFs. The programmer's reaction was correct, as it is time-consuming to update files, and such time should not be spent on files that aren't going to be used in the future.

Final Update

The final update was a more extensive process to be done in a limited time. I still had many files to check, which I had not eliminated. After the normal update check I checked all the files for "white cells" (cells without information). I then passed them on to the programmers to have the cases checked individually to see if there was indeed no further information to enter in the cells or if an error had been made. If needed, the correction was made and we then completed the update to final form.

To be prepared for every possible request I still kept many of the files until the final update. If I had eliminated more of them I wouldn't have had to spend so much time checking them. After checking I did a "white-cells-check," looking for cells without any entries for age, gender, violation,

region, etc. Usually the white cells errors were typing errors, but on occasion some of programming was at fault. I still looked for the outliers, e.g., age=260, etc., and found some I had overlooked previously. As we had a special flag for "massacres," I also checked the case number files for "massacre." We flagged a case as a "massacre" if it had at least five executed victims. We could detect some "massacres" which had not been marked as "massacres" before, and others which were flagged incorrectly. I had two files with the case numbers, so I could pass the case numbers I filtered with these checks on to the data processors to check the cases again. These errors were both typing and analysis errors.

Output from the Database: Answering the Requests of the Investigators

Kinds of Analysis

Generally the information provided by the database was descriptive statistical information, easy-to-read graphs and easy-to-understand figures (examples of some graphs are given in the appendix). We primarily did calculations based on violation, but some special calculations were based on victims. In general, there was no analysis based on cases. Usually we only produced statistical output based on the information the CEH collected.

I produced some graphs based on victims for the chapter on indigenous identity, to show the percentage of individual victims identified for their ethnic characteristics. There were two exceptions. I made some graphs based on cases with the key word Massacre, in particular a time line graph for the areas of military operation. In addition, I made some graphs from the information provided by the military (e.g., how many military commissioners they had recruited in the different regions in the years of the armed conflict) that had not been collected by the CEH.

Kinds of Graphs, Figures, Numbers

The investigators used the graphs we produced to find the tendencies and confirm or disconfirm the hypothesis. Thus, graphs generally had only a few details, and because of this they were easy to read and understand.

Once the recommendation group asked me for a special bar chart. They had made a special codification with only a sample of cases (randomly selected) with instructions to create graphs to analyze and show the tendencies. I made a colored bar chart with the standard error lines. The next day another investigator from a completely different section came to our office to say that we were making useless graphs. He had found the colored bar chart and made a copy for himself. However, since the blue color lines changed to black in the copy, he couldn't identify the standard error lines, so I stopped using colored graphs and used different shades of gray for bar charts and different styles of lines for time lines.

Checking Graphs, Figures, Numbers

Every output of the database had to pass a final check on the layout and the information calculated. I used a special unified layout for all outgoing information. It contained the basic information, e.g., the date of the last update and the total number of cases from the last update, so that the investigator always knew from which set of information his output had been made (an example is given in the appendix). Occasionally, I forgot to change information within the macro that did the automatic formatting. So graphs went out with the old information. The investigators, who sometimes did not find my error, would return claiming that the graph looked different from the last one, but that the total number of cases were the same. At first I was confused and uncertain, but then I realized that I had simply forgotten to change the numbers. After this experience I began checking outgoing material even more carefully.

Lessons Learned

Problem	Alternative used	Lesson learned

Copies from graphs originally printed in color, which led to unidentifiable graphs.	Only black and white graphs used.	It is better to prevent the problem than to trust that everybody will know that copies usually are only black and white.
Output has to go out with some basic information for identification and checking purposes.	Layout of the sheet for providing the basic information which every output received.	Carefully check every outgoing graph to see if all the variables of the layout are changed and updated.
There was no check after me.	Had to check even more carefully.	It would be better to have someone else as a security check.

The Statistical Analysis Program

We used Excel for several reasons. One reason is its primary advantage, which is that it is widely used and it was easy to find a person who knew the program. Another advantage is that the interface is easy to understand and one can develop the ability to use it rapidly.

Unfortunately as we found over time, Excel has many disadvantages. Among them are: loss of graph layout (i.e., formatting would change without apparent cause); update data had to be checked carefully; large data manipulation sequences are difficult for Excel to manage; there are frequent crashes of the computer because of working with too much data, and there is no record of what was done to get the result shown in the worksheet.

Administration: Keeping a Record of the Output

Administration of the Excel Files

I made a new branch of the file directory for every new step or development within the programming process. Thus, the Excel files show their relational structure within the file directory tree, so it's easy to identify the steps of programming.

At the completion of the CEH report process, it was instructive to see the file directory tree. It revealed the whole history of the programming process. First we had different files only for massacres or only for no massacres, then we had the combined files where I could search inside the file the different violations; for example, within massacres or outside of massacres, etc.

Administration of the Outgoing Graphs

The outgoing graphs and figures were registered within two different archives. First they all had their registration numbers included in the title, which consisted of a letter and numbers. Then they were registered in the book of registration and then on the visits control sheet that I maintained. I made copies of all the outgoing graphs to record them within two different file records: one by the type of statistics, e.g., all the bar charts, and the other one for the topic or variable, e.g., all the material covering "massacres" or "department." Thus I always had several ways to find graphs and figures.

I spent almost as much time administering the output as checking it. I had a permanent horror of losing some graph while under pressure to produce results and not being able to find the graph again. So I devised a special registry system. Every type of chart got its letter: B=bar chart, L=time line, T=table, etc. The letter was followed by the indicator for the topic or main characteristic of the graph: 1=department, 2=children, 3=massacres, etc. And then I gave each output a serial number. For example, the registration number B2.32 indicated that this was the thirty-second bar chart on children. I manually kept the records in the registration book, where every type+topic combination had its own page: e.g., I had a sheet for B1, a sheet for B2, and a sheet for B3, etc. In this registration book I kept the date information, the title, where the data came from (file and worksheet name), where the output went (name of the investigator), and notes (e.g., if the graph was an updated version from a former graph).

Lessons Learned

In this section, I discuss both the lessons learned and their implementation.

There was considerable similarity between the work for the CIIDH and CEH projects. Accordingly, the analysts for these two projects, Herbert F. Spirer and myself, jointly prepared recommendations for future large-scale human rights data analysis that appear at the end of this paper in Appendix 1, Data Analysis Recommendations.

Problem	Solution	Issues
No permanent person working on statistical analysis, which led to unique outputs, and inconsistent ways of archiving and naming different layouts.	Have the same person work on statistical analysis and output from start to finish.	If not possible to have same analyst(s) throughout project , establish a uniform logical structure at the start.
No records were kept.	Because of the considerable effort that would have been necessary to recover and reorganize the materials stored under the former inconsistent structure, I started with a new recording system.	If it is too time-consuming to restructure the existing material and if you are still at the beginning of the statistical analysis, create a good new system and take the loss of former material.
There was no detailed information about how the data were processed before they were used in the statistical analysis	Immediate detection and correction of mistakes resulting from of misunderstandings concerning what was in the input data.	Start by asking for all the details you need to know for working with the data (former calculations, what the variables mean, how are they calculated, etc.).
Sometimes the DBFs provided by the programmer contained too much information.	Discussed with the programmer until we found a middle way: I received only the blocks I wanted (plus a few more)	Specify exactly the needs for producing the statistical output. This means exactly specifying the variables requested.
Data has to be checked before using for statistical analysis.	I designed a large checking system with a first rough total check on import of the data and another specific widespread check afterwards.	You can never check too much. It's not so important how, but the important thing is that there are checking steps.
Updating only the minimal number of files used to meet the needs of investigators.	There were always too many files to update. This led to a long updating process.	Eliminate the files with no further use, as there will be new ones as the archive always grows.
Data after the final update has to be as completely checked as possible and cleaned.	Extensive checking methods for the final update.	There may always be some mistakes that you will overlook, but it's always worth trying to eliminate all error.
We found mistakes when checking data.	We found mistakes from typists, analysts, and programmers.	Detecting errors is necessary and positive, but it does not mean blaming someone! Errors happen.
The investigators complained that they didn't receive what they had been told they would receive, because the investigator receiving the output didn't really know what statistical output would look like.	Alleviated by having one full-time analyst (me).	The statistical analyst making the output is the one receiving the requests, and should explain at the beginning how the investigators should make their requests and what they can expect to get.

The person receiving the requests from the investigators was not the same person as the one doing the statistical analysis.	Constantly working to stay in touch.	Only statistically skilled persons should receive requests.
The person handing the output over to the investigator was not the same person as the one doing the statistical analysis. Investigators didn't receive explanations on what the output was about, how it was calculated, where the data had come from.	Constantly working to stay in touch.	Set up system so that analyst physically gives analyses to users. Analyst should explain the meaning of outputs. May not be needed with statistically knowledgeable users.
When I started work, other persons without statistical understanding were obtaining statistical outputs. There was no control over the outgoing information. It wasn't statistically checked, so mistakes went out, and incorrect records and different layouts frustrated investigators	I was able to correct this situation, but could not undo the problems of the past.	Only one qualified person produces statistical analysis to maintain control and records. Or, if more than one person produces statistical analysis, one person has to check everything for statistical correctness and maintain the records.
Investigators haven't been educated in reading graphs and understanding statistics, no explanations were provided. Investigators deduced incorrect explanations of the figures in their chapters, even to the point of misunderstanding the meaning of the title of the graph. Also, investigators misinterpreted analytical findings and made hypotheses that did not correctly reflect the analytical findings.	I held a class and tried to inform everyone about statistics.	Periodic workshops for investigators on the use and interpretations of basic statistics, explanation of the basic graphs.
Many people working on the project had problems using statistical reasoning. This is quite common where there has been no training in statistical methods. For example, people can confuse statements such as "20 percent of the women in Rabinal were assaulted" and "20 percent of the women who were assaulted were from Rabinal."	See above.	The path from the producer to the final user of the statistics should be a short as possible to guarantee a correct result in the final version. Each added intermediary is a potential source of error or confusion, especially if they are not fully qualified. Education for everyone is important; statistical reasoning can be unfamiliar to people.
Programming develops within the process. The ability to identify the different steps must be provided.	Identified within the file directory tree, every step another branch.	Better to have everything written down and recorded in a logical structure.
Output must be easily identified.	Manually using a registration number and recording in the registration book.	Registration is necessary, but my way was very "artificial" as it was kept manually in books.

Output must be easily and quickly retrievable.	Double file record archiving system to provide the possibility to look for the output by two different criteria.	Copies of output are very useful for examples, and as proof and replacement, if needed later

Implementation of Lessons Learned

The following table reviews some of the specific actions that I took in order to put learned lessons to work during the project. The positive effects of these actions are also shown in the table.

How it was	Positive effects
I always put updated graphs and output in visible places in the team's offices to keep the team informed. As almost everybody working in the database offices had been living and working in Guatemala before the CEH started working, they were knowledgeable about the history and actual situation.	As the whole database team is involved in the process providing the data for statistics, they know the whole chain and are interested in knowing what's at the end. This led to a better identification with the group, refreshed their energy and strength, and reduced the widely held distance to statistics. They also made their own personal hypotheses and interpretations leading to interesting discussions in the team.
We produced a general updated overview with every database update, where I changed some expressions into more understandable words before handing it over to the commissioners as part of our agreement to keep them informed.	It was necessary to use common interpretations of technical terms to make the overview more understandable and easy to read. Then the commissioners, the central team and all the people working with statistical output received an easy-to-read overview periodically and were pleased that they were included in the process and could understand what they received.
The output had the same layout and basic information.	Led to a professional impression by the investigators and made any one graph more official.
I started with one three hour workshop, inviting all the investigators, commissioners, team leaders, etc. I explained the main graphs used up to this moment; the data processor explained the different variables and terms used and the programmer talked about the lists of cases provided by him.	The small audience that attended that meeting appreciated the effort and reported that they had learned a lot. From this favorable experience came my idea of periodically providing basic workshops in statistical reasoning.
I started keeping records of the visits from the investigators (unfortunately only for two months), noting their concerns and wishes.	It was easier to prevent misunderstandings and to reproduce acceptable materials later for the same person.

Appendix 1

Data Analysis Recommendations

By Eva Scheibreithner and Herbert F. Spirer

Introduction

As part of the process at the Experts' Meeting, we jointly reviewed our experiences and lessons learned and have integrated them into this set of recommendations for data analysts who will be carrying out similar missions in the future.

We make some recommendations that are explicit statements of procedures that we believe should be followed to maintain the integrity of the data while producing analytical results that faithfully report on the findings of the project. Such recommendations are those required for Verification.

We make recommendations that are general and meant as guidance to the analysts. They are for control of datasets, choice of statistical program, chart standards, an output identification system, and education. In these cases, we hope and expect that analysts will recognize the validity and value of our guidance and use it to formulate their own procedures and practices that are consistent with the context in which they are working. Such recommendations are those concerning Graph Standards.

Control of Datasets

As we have discussed, avoidance of error is critical in the analysis stage to maintain the credibility of the final results. We have found that the following requirements are the minimum needed to assure this freedom from major sources of error.

- The statistical analyst must maintain a current data dictionary. This data dictionary must contain as a minimum, the variable (field) name, the meaning of the variable, and a list or verbal description of the values that can appear in the corresponding field for the variable.
- The analyst must also maintain a cross-reference table of files and variable (field) names so that the analyst and others will know which variables appear in which datasets and which datasets contain a given variable.
- To avoid confusion among different versions of a dataset with a given name, the analyst should use a separate directory (folder) for each version, numbered in accordance with the sequence of the version. If database personnel produce these datasets through queries and store the datasets in directories, they should organize the datasets in this manner.

Choice of Statistical Program[2]

We used Excel in performing our analyses and both of us found it to have problems as described in this paper. In addition to our statistical issues with this program, it had the disadvantage of limited graphic output capability. This latter limitation caused significant problems in the production of the reports. Those problems could have been avoided by the use of Encapsulated PostScript files.

Encapsulated PostScript (EPS) is a standard format for importing and exporting graphic files in all environments. The EPS file is included as an illustration in other PostScript language page descriptions and can contain any combination of text, graphics, and images. Unfortunately, not all PostScript-enabled printers are able to print the EPS files, creating a hardware or software issue that must be resolved to facilitate the analyst's work.

In addition, Excel does not produce a log recording the actions taken by the analyst and the use of Visual Basic macros for this purpose is dangerous. Unless the analyst is diligent in keeping records, in the absence of the analyst, other personnel on the project or outside auditors may have

[2] These observations also apply to the statistical work of the TRC.-PB.

no way to recreate the analysis, except to try to repeat the process. Unfortunately, the analyst cannot recover the actions taken to produce a result from that result except by reverse engineering.

We believe that the use of a particular program should not be dictated, and the analyst needs the freedom to choose a program consistent with experience, abilities, and preferences. Balancing of costs and benefits will lead to the best choice of a program. These considerations include the skills and knowledge of the analyst. In a particular context compromise may be necessary.

Accordingly, we suggest the following as desirable goals.

- The graphical and tabular output will be in the form of Encapsulated PostScript files.
- Either the analyst or the program (preferred) will produce a detailed log of the actions taken in manipulating the dataset to produce results.
- The analyst will use standard programs to make it easier for replacement analysts to check the work rather than exotic programs or those not widely known.

Graph Standards

Choosing the appropriate graph to display information is a combination of technology and art, essentially a creative process. To give specific rules is to stifle that creativity and in the long run, will lead to results of limited value. Our approach to the visual display of our analyses conforms to Tufte's standards for Excellence in Graphical Representation, quoted in the paper The International Center for Human Rights Research Investigation, in the section, Graphs: The Visual Display of Information.

In addition to that general guidance, we recommend that:

- The purposes and needs of the data analysis be met in large part by strategic use of the following types of graph: univariate time series plot (time line), overlaid time series plot, vertical bar chart, horizontal bar chart, stacked bar chart, and histogram.
- The analysts avoid pie charts, which can be difficult to interpret and are often misleading.
- The analysts strive to avoid clutter, which means, among other things: use ticks, but don't use gridlines, don't set charts in visible frames, and don't use markers unless there is a clear need.
- Any tables be spare, and without clutter. There are a number of examples of such table layouts in the CIIDH report (Ball, Kobrak, and Spirer, 1999, pp. 70, 119, 122-3, 128-130).

Verification

The need for verification derives from both the human and machine elements at work in the process of statistical analysis. Among the sources of error are:

- programmer errors in preparing the datasets
- analyst errors in doing the analysis
- program faults inherent in the current version of the analysis software
- consequences of computer crashes
- hardware limitations, inherent in the hardware and possibly unknown to the analyst
- key-entry errors, which can occur at any stage from the initial to the final output

The ideal situation is when none of these errors occur. Analysts, programmers, and others can with experience and motivation, reduce the number of errors generated, but they can never eliminate them. No software is ever completely bug-free, and hardware is prone to both inherent flaws and degradation. Thus, to have credible analytical results, we need a verification process for detecting errors. To this end, we recommend the following to statistical analysts:

- Have programmers producing working datasets supply totals and extremes for all numerical dataset variables as a part of each version.
- Use these totals and extremes as a check on the changes from the prior version of the dataset.
- Check the dataset as received from the programmer.
- Base checks on Table 5, following. The analyst should maintain the summary described in this table. If the analyst uses a program generating a log and allowing the use of stored programs, such as Stata or SPSS, a summary will be automatically retained.

Table 5. Summary of recommended checking methods.

File name	Check	Units of analysis	
	Totals		
	Extremes		
	*-questions		
	Key tabulations of categorical variables		

Note: A *-*question* is some question about the data that will provide a check of context, such as "what proportion of women were disappeared in the month of…?"

- Check the dataset as received.
- Check the dataset at every critical transition. When in doubt, check the dataset at every change. Checking means comparing totals, extremes and *-questions for before and after values.
- Be skeptical, vigilant, and scrutinize constantly.

Output Identification System

To track graphs through the information management process, each graph or tabular output must have an informative, unique identification. With such an identification, which we call the *Graph Identification Number* (GIN), unambiguous reference to outputs can be used in communication among project members and any final or intermediate result can be tracked to its source. In this section, we propose a format for the GIN, a system for tracking the production and transit of outputs, a simple rationale for archiving outputs, and a method for tracking the subsequent changes in a given dataset.

- The GIN is structured as follows: TT-FF…FF-VV-DATE, where TT is the mnemonic output type designator (TA for table, BV for vertical bar chart, TS for time series plot, etc.), FF…FF is a variable length mnemonic for the title, VV is the two-digit version number of this particular output, and DATE is the date produced.
- Outputs are archived by GIN and by category.
- To provide an audit trail and the ability to access users, the analyst should maintain a graph tracking system as shown in Table 6, below:

Table 6. Structure of recommended graph tracking system.

GIN	Create date	Title	Source dataset	Recipient	Notes

- Since modifications of the dataset produced by the analyst to obtain particular outputs produce new internal data sets, successive versions can be tracked by the use of upper case suffixes; and versions from which outputs are produced by lower case prefixes. Hence, BRTANONV14A could represent the data set obtained by filtering out all violations except killings (RTANONV14A) and the specific subset of that data set used to create the second variation of a bar chart as BRTANONV14A.

We note that the use of **some** identification and tracking systems is the key part of our recommendation. Here we recommend a particular system based on our experiences, but other contexts may call for other approaches.

Education

The outputs of statistical analysis are a major part of the end result of a large-scale human rights data project. They represent the physical realization of the logical process of drawing meaningful conclusions about the data. To come to that point, a great deal of interaction among team members is needed. Since most team members will not have had either education or experience in statistical and analytical reasoning, we recommend that education in these topics be included in the project plan and execution.

Education of the type we discuss will have the benefit of more effective, efficient work, and better relationships among project team members. We make the following recommendations, understanding that their implementation will depend on the context and issues of resource limitations.

- Educational objectives are (1) how to interpret graphs and tables, (2) methods of descriptive and exploratory data analysis, (3) the meaning of statistical statements, (4) how to read titles and notes, (4) how to work with absolutes and percentages, and (5) how to work with conditional statements about data.
- Project management should decide on what is best for the given project; whether the educational process should involve all team members, or functional groups, workshops or classes, or continuing and periodic or episodic sessions.
- Because of the serious problems in how "statistics" is taught in schools, many people are averse to the subject and it is usually necessary to mandate attendance.[3] Team members should know that practical data analysis often bears little relationship to the content of conventional first statistics courses.

The amount of time required of team members for the educational process should be strictly limited. Because much of the education in these methods will take place in the workplace, workshop time can be limited to less than eight hours throughout the project.

References

Ball, Patrick, Kobrak, Paul, and Spirer, Herbert, 1999. *State Violence in Guatemala, 1960-1996: A Quantitative Reflection.* Washington: American Association for the Advancement of Science and Centro Internacional por Investigaciones en Derechos Humanos.

[3] Teachers of statistics know of these problems. These issues are discussed at almost every meeting of the profession. Unfortunately, there continues to be a difference between what teachers teach in the first course in statistics and what these same people do when working in the field.

Chapter 11

The Guatemalan Commission for Historical Clarification: *Generating Analytic Reports*

Inter-Sample Analysis

Patrick Ball

Introduction

This paper reports on an analytical study requested by the Commission for Historical Clarification (CEH) and carried out by the American Association for the Advancement of Science.[1] The purpose of the study was to answer the question: How many people were killed in Guatemala during the period of the CEH mandate, 1960-1996? To answer this question, we used the information in three databases of human rights violations, resulting from three projects – one conducted by the CEH, one by the CIIDH, and one by REMHI. These databases reported data from interviews with direct witnesses and victims. As a consequence of having three sources, we must first ask a) how many unduplicated killings are documented by the three projects? and then attempt to answer the second question, b) how many killings were there in all during the internal armed conflict?

Our analysis deals with these two questions. We deal first with the information collected by the three projects in light of the objectives of this analysis. We then explain the scientific methods used to estimate rates and quantities that answer the specific empirical questions derived from these objectives. We then present and interpret the results of applying the scientific methods to the information from the three databases. We subsequently analyze four regions of Guatemala in which genocide may have occurred during the period 1981-1983. Finally, using other statistical methods we show that the three projects lead to similar implications about the patterns of violence in Guatemala.

Note that in some tables, there are numbers that are not counts, but result from arithmetic operations subject to rounding error. Thus, totals shown in the table will not exactly add up to the totals of the related columns or rows. In some graphs we have retained the Spanish labels, as it is our intent to present tables and graphs as they appeared in the CEH report.[2]

The Information

The three databases were created using information gathered from interviews with witnesses and victims. Each contained a list of named victims who were killed, as well as numbers of people who were killed but who could not be identified by name. The three projects did not define "political killing" in the same ways. Therefore the measure we use in this study is *deaths*, and not the more juridically precise term "extrajudicial execution" that is used elsewhere in the CEH report. The three projects have unique definitions of murder, and to join them, it was necessary to use the broadest possible definition of the killing as a human rights violation. Thus, the totals of deaths in the AAAS study should be compared with the totals of deaths in the CEH report, and not with the totals of extrajudicial execution.

Table 1 shows the number of documented killings (victims with and without names), by time period, region, and database. Many killings were not reported to any project, and therefore, the quantities presented in Table 1 are less than the total actual number of victims who were killed in political violence in the period 1960 to 1996. Table 1 shows only those victims who were reported to one or more documentation project.

[1] At the request of the CEH, this analysis was conducted by Dr. Patrick Ball, Deputy Director of the Science and Human Rights Program of the American Association for the Advancement of Science, with the assistance and advice of Dr. Herbert F. Spirer (Adjunct Professor, Columbia University and consultant to the AAAS), Dr. Frederick Scheuren (Senior Fellow of The Urban Institute, and Adjunct Professor at George Washington University), and William Seltzer (Senior Researcher at Fordham University).

[2] The meaning of the Spanish labels will be clear to any reader.

Table 1[3]: Number of documented killings (victims with and without names), by time period, region, and project

Region	CEH			CIIDH[4]			REMHI		
	1960-1977	1978-1996	1960-1996	1960-1977	1978-1996	1960-1996	1960-1977	1978-1996	1960-1996
Region 0 (others)	271	9,916	10,187	18	2,586	2,604	84	6,888	6,972
Region I (area Ixil)	14	4,609	4,623	0	4,028	4,028	9	5,423	5,432
Region II (Cahabón)	1	532	533	0	135	135	7	453	460
Region III (Rabinal)	0	1,379	1,379	0	1,297	1,297	0	1,354	1,354
Region IV (San Martín Jilo.)	0	1,347	1,347	1	20	21	0	68	68
Region V (Nte. De Huehue)	0	1,746	1,746	0	1	1	0	1,032	1,32
Region VI (área Zacualpa)[5]	0	1,951	1,951	0	238	238	1	1,674	1,675
Region VII (Guatemala)	91	245	336	1	15	16	10	111	121
Region VIII (área Panzós)	0	169	169	11	41	52	1	1,167	1,168
Region IX (Ixcán)	3	2,421	2,424	0	48	48	5	2,751	2,756
Region X (area Costa Sur)	25	190	215	2	91	93	23	139	162
Total	405	24,505	24,910	33	8,500	8,533	140	21,060	21,200

The three projects did not equally cover all of the regions. All conducted investigations in the Ixil area (Region I) during the period 1978-1996, but only the CEH collected adequate information in San Martín Jilotepeque (Region IV)[6]. It is also clear that none of the projects adequately covered

[3] Table 1 excludes the victims for whom the year or place of death is not known.

[4]. Although the CIIDH also collected information from journalistic and documentary sources, this analysis only includes the information from direct sources supported by the witness' signature.

[5] The definition of Region VI (the Zacualpa area) includes the municipios of Chiche and Joyabaj, and therefore does not correspond exactly to the definition of the region in the section of the CEH report that examines genocide. That section includes as Region VI only the municipio of Zacualpa.

[6] The regions were defined in order to isolate areas in which there were big differences in the coverage rates among projects.

the period 1960-1977, including the massacres of the 1968-1973. Any estimate must take these limitations into account.

If no victims were reported to more than one project, the total of documented victims would be the sum of the three totals, that is 24,910+8,533+21,200 = 54,643. However, many of the same victims were reported to two or three projects. Thus, we cannot assume that the total number of victims is equal to this simple sum.

The projects were managed independently, and each victim could have been reported to more than one project. For example, assume that a victim Juan Pérez was murdered. His wife may have reported the killing to the CIIDH in 1994; his son may have given testimony to REMHI in 1996; and Peréz's neighbor might have related the story to the CEH in 1997. If we simply sum the three databases, Peréz's killing will be counted three times.

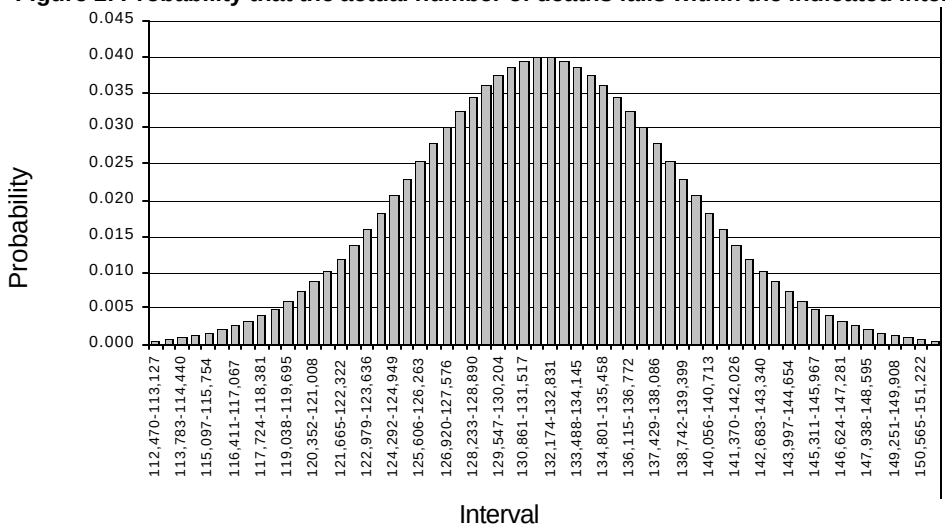
Duplicated reporting of deaths in more than one database is called "overlap." To estimate the total number of victims reported by the three databases, the overlap between databases must be estimated to reduce the contribution of each database by its overlap rate.

Two possible conditions demonstrate the limits of the overlap problem. If none of the victims in any database appear in any other database, then the total number of victims of killing and disappearance is equal to the sum of the number of victims in the three databases (54,643). This is the upper limit to the number of such victims. The lower limit can be found in the extreme case that the largest of the three databases (here, the CEH database) contains all the cases reported in the other two (REMHI and CIIDH). In this case, the total number of killings is simply the number of killings reported in the largest database (405+24,505=24,910). The sum of the total number of unique victims in the three databases must fall within these two limits, that is, between 24,910 and 54,643. The purpose of our analyses estimating the total number of documented killings is to narrow this range.

Many killings were not reported to any of the three projects. In the following section, we carry out a scientific analysis to estimate the total number of killings, [7] including those not reported to CEH, the CIIDH, nor to REMHI. The estimate from this analysis is that between 119,300 and 145,000 killings were committed, with the most likely figure being around 132,000. Figure 1 shows the probabilities that the real value falls within various ranges around the estimate of 132,000.

[7] Note that this analysis does not cover forced disappearances (they are handled separately). There was insufficient time and resources to extend this analysis to disappearances.

Figure 1: Probability that the actual number of deaths falls within the indicated interval



The number of victims based on the three independent databases

Analysis of Overlap

The information in the three databases lists victims identified or estimated by witnesses. Some, but not all, of the victims were identified by name.[8] The total number of killings in each database is referred to using the notation described below.

M_{CEH} = the total number of victims in the CEH database
M_{CIIDH} = the total number of victims in the CIIDH database
M_{REMHI} = the total number of victims in the REMHI database

None of the three databases directly estimates the total number of killings in the country during the full period of the CEH mandate. Each database is a list of victims of killings who were reported directly to the project and verified according to the methodology of that particular project. As has been mentioned, many victims were not reported to any project. The total number of victims killed in Guatemala and reported (or not) to different projects can be expressed by eight categories, defined below.

N_{000} = victims who were not reported to any of the three projects: not to the CEH, the CIIDH, nor to REMHI
N_{111} = victims who were reported to all three projects
N_{110} = victims reported to the CEH and to the CIIDH, but not to REMHI
N_{101} = victims reported to the CEH and to REMHI, but not to the CIIDH
N_{011} = victims reported to the CIIDH and to REMHI, but not to the CEH
N_{100} = victims reported only to the CEH, and not to the CIIDH nor to REMHI
N_{010} = victims reported only to the CIIDH, and not to the CEH nor to REMHI
N_{001} = victims reported only to REMHI, and not to CEH nor to the CIIDH

The total number of victims of killing in Guatemala, N, is the sum of these eight values. The total number of victims reported to one, two, or three projects, N_k, is the sum of the seven categories that are calculated directly from the databases, that is, N_{111} to N_{001} as shown in Equation 1, below.

[8] This analysis treats victims, not violations, but for killings, the two measures are identical and so this distinction is not significant. See Ball (1996).

$$\hat{N}_k = N_{111} + N_{110} + N_{101} + N_{011} + N_{100} + N_{010} + N_{001} \quad (1)$$

To get the total number of victims reported to one or more projects, estimates must be made of the number of victims reported to all the projects (N_{111}, cases that are found in all three database), and those reported to each pair of projects (N_{110}, N_{101}, *and* N_{011}), and the complements of the number reported to each project (N_{100}, N_{010}, *and* N_{001}). With this information, we can determine N_k.

Matching

It is difficult to find the same victim in any two or all three of these databases using a computer program. Victims are reported with varying information. Identical names may be spelled differently, sometimes because they were inconsistently or idiosyncratically translated from Mayan languages. Dates of birth and death can be uncertain or wrong.

Thus, it is neither practical nor accurate to match databases by automated means with computer programs. To find a person from one database (the source) in another of the databases (the target), an analyst must compare all of the data relevant to the killing, including the name, place, and date of the killing from the source with all the records in the target. This process we call "matching."[9]

Many victims are not identified by name in the databases. Often the original witnesses would mention only a group of people. Different witnesses of the same event often estimate different numbers of victims who suffered the same violations. In our analysis we assume that the match rates between unnamed victims are the same as the rates among the identified victims.

Matching databases is difficult, tedious and time-consuming. Instead of trying to match all the records of each database against **all** the records in the other databases, stratified random samples of the victims identified by name in each database were selected and matched against the records in the other two. The samples were proportionally stratified by region to assure that all regions were covered. The number of records taken in each sample is denoted by the letter m (m_{CEH}, m_{CIIDH}, m_{REMHI}). Including all the regions, the total number of records sampled and matched was 1,412, 1,351, and 1,122, respectively (see Table 2).[10]

Each person sampled (from each of the three databases) was compared against the records in the other two databases. When the same person was found in one of the other two databases, it was noted as a double-match; when the record was found in both of the other two databases, it was noted as a triple match.

Four groups of samples were chosen from the three databases. One analyst from the CEH matched one group, and a second analyst matched the other three groups. Many records were deliberately included in both samples. Only in a small number of cases were differences between the analysts' decisions found. The implication from this finding is that the error resulting from non-sampling factors was minimal.[11] The numbers of matches are shown in Table 2.

[9] Furthermore, many victims are not identified by name in the databases. When witnesses mentioned a group of victims without specifying the victims' names, different witnesses often refer to different numbers of victims. Given the already-mentioned difficulty that witnesses often confuse the exact dates of the events, it is not possible to match groups of unnamed victims. This analysis assumes that the match rates between unnamed victims are the same as the rates between named victims.

[10] Of the records mentioned in the text, 498 were resampled and matched a second time. We refer to these records in the analysis of the reliability of the matching.

[11] In the match analysis, what concerns us is that records that are true matches do not escape the analysts. Of the 498 records matched twice, 171 were true matches. Comparing these 171 records matched by two different analysts, 88% were coded identically.

Table 2: Number of matched records in three databases, by outcome

	CEH	CIIDH	REMHI
m_{111}	21	73	19
m_{110}	48	153	
m_{101}	210		226
m_{011}		121	27
m_{100}	1,133		
m_{010}		1,004	
m_{001}			850
Sample Total	1,412	1,351	1,122

Table 2 shows that of the sample of 1,412 victims selected from the CEH database, 21 were found in the CIIDH database and in the REMHI database; these 21 are triple matches. Forty-eight victims in the CEH database were found in the CIIDH database but not in the REMHI database. In addition, 210 more victims in the CEH database were found in the REMHI but not in the CIIDH database. A total of 1,133 of the victims sampled from the CEH database were not found in either of the other two databases. The interpretation of the other two columns is the same.

We obtained overlap rates shown in Table 3 by dividing each m_{xyz} in Table 2 by the total number of victims sampled in each database.

Table 3: Overlap rates for three databases

	CEH	CIIDH	REMHI
r_{111}	1.5%	5.4%	1.7%
r_{110}	3.4%	11.3%	
r_{101}	14.9%		20.1%
r_{011}		9.0%	2.4%
r_{100}	80.2%		
r_{010}		74.3%	
r_{001}			75.8%

To interpret this table, note that r_{110} on the second line indicates that 3.4% of the victims in the CEH database are also in the CIIDH database. The database of the CIIDH is smaller than the CEH database; the same estimation from the point of view of the CIIDH is that 11.3% of the victims recorded in the CIIDH database are also in the CEH database.

Note that the differences in the estimations of the rates are not exactly in proportion to the differences in size among the databases. The differences occur because of the variability that results in the process of taking a random sample, and from the errors in matching. We treat these issues in the later section on the analysis of error.

Estimation of the total number of reported victims

As discussed in "Analysis of overlap" (above), the total number of victims of killing was estimated by the sum of seven components as defined in Equation 1, repeated below.

$$\hat{N}_k = N_{111} + N_{110} + N_{101} + N_{011} + N_{100} + N_{010} + N_{001} \quad \text{Equation 1}$$

With the rates from Table 3 and the last line of Table 1 (the total number of victims 1978-1996), the components of N_k based on information in the three databases can be calculated. The results are in Table 4, below.

Table 4: Number of killings, estimated by category and by project

	CEH	CIIDH	REMHI
N_{111}	364	459	357
N_{110}	833	963	
N_{101}	3,645		4,242
N_{011}		761	507
N_{100}	19,663		
N_{010}		6,317	
N_{001}			15,955

However, to calculate N_k, the number of victims common to all three databases, the several estimates of the number of matched records (N_{111}, $N_{110,}$ $N_{101,}$ and N_{011}) must be reconciled. We used the average of each of these four components, providing the totals shown in Table 5, below.

Table 5: Estimated number of killings in all three databases (CEH, CIIDH, and REMHI).

	Mean
N_{111}	393
N_{110}	898
N_{101}	3,943
N_{011}	634
N_{100}	19,663
N_{010}	6317
N_{001}	15,955
N_k	47,803

Thus, our estimate of the unduplicated number of reported killings in the three databases is 47,803. However, we show below, this number is subject to a number of controllable biases.

Demographic Theory and The Estimation of the Total Number of Undocumented Killings

In the previous section, we explained how we estimated the total number of documented killings. In this section, we discuss how to estimate the total number of undocumented killings, denoted N_{000}. Below, in Figures 2a, 2b, and 2c, we show with the use of Venn diagrams, the three possible ways in which the databases might be related.

Figure 2a: Distribution of databases in the universe of all violations (where there is no relation)

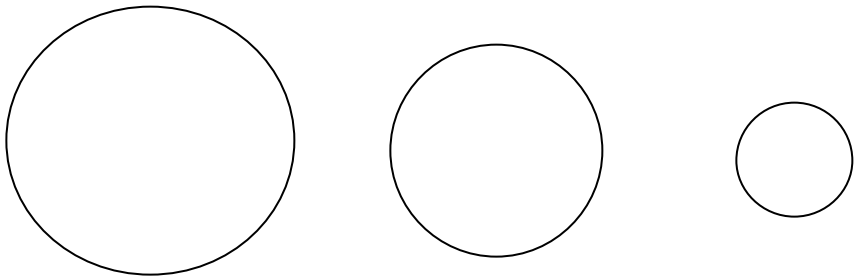


Figure 2b: Distribution of databases in the universe of all violations (total overlap)

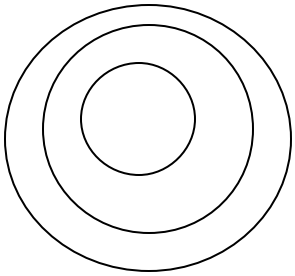
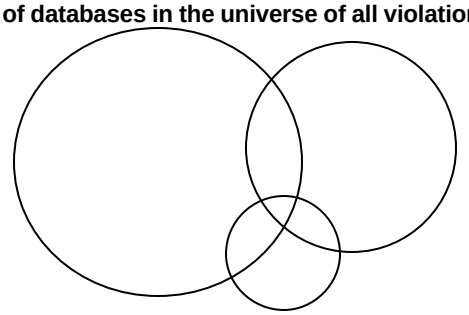


Figure 2c: Distribution of databases in the universe of all violations (partial overlap)



In Figure 2a, the databases share no violations. In Figure 2b, all the violations are contained in the largest of the three databases. In Figure 2c, some violations are shared. From the previous section, it is clear that Figure 2c is the correct representation of the three databases.

Assume that the three projects operated independently and consequently, that the probability that a project has testimony about a certain violation has no influence on whether another project has testimony about the same violation. What implication does this have for the universe of violations? In Figure 2a, the implication is that the universe of violations is large because working independently, the databases do not overlap. In Figure 2b, the implication is the inverse, that the one database is contained within the next larger, and the next larger is contained within the largest. In Figure 2c, which corresponds to our situation, the levels of overlap are partial. With the assumption of independence and the reality of overlap, the number of violations in the universe can be inferred.

Consider the case of two projects, P_A and P_B, whose databases have an overlap M in a universe of violations N.[12] Note that the probability of any given killing being documented by Project P_A is $\Pr(A)=\frac{A}{N}$ that is $N=\frac{A}{\Pr(A)}$, and the probability of any given killing being documented by Project P_B is $\Pr(B)=\frac{B}{N}$. The probability that a killing was documented by both databases, $\Pr(M)$, is equal to $\Pr(M) = M/N$, and by the definition of an event composed of two independent events, $\Pr(M) = \Pr(A|B) = \Pr(A) * \Pr(B)$.

Interchanging the terms, $\Pr(A)=\frac{\Pr(M)}{\Pr(B)}$, which reduces to $\Pr(A)=\frac{M/N}{B/N}=\frac{M}{B}$

Combining the first relation $\Pr(A)=\frac{A}{N}$ with the previous result gives us $\frac{A}{N}=\frac{M}{B}$, and therefore $N=\frac{AB}{M}$. In order to estimate only the killings that were excluded from the two projects, $N_{00}=\frac{(A-M)(B-M)}{M}$, or in the notation of the three-database system,

$$N_{00}=\frac{N_{10}*N_{01}}{N_{11}} \tag{2}$$

With the same logic, it is possible to derive an estimator for $n_{000:}$ the measure of the number of killings that were not documented by any of the three projects.[13] This estimator is presented below in Equation 3.

$$N_{000}=\frac{N_{100}N_{010}+N_{100}N_{001}+N_{010}N_{001}}{N_{110}+N_{101}+N_{011}} \tag{3}$$

Measuring the Sampling Error and the Estimator Error

The estimators of N_{000} and of the total number of killings $\hat{N}$ depend on the estimates of the overlap between the three databases.[14] To estimate of the number of killings in the categories N_{111}, N_{110},…, N_{001} that sum to $N_{k,}$ the levels presented in Table 3 above are multiplied by the total number of killings in each database. We then estimated n_{000} using Equation 3, above. Summing these two components gives an estimate of $\hat{N}$. We used the jackknife method to estimate $\hat{N}$ because that method allows us to control the ratio bias inherent in n_{000} and to estimate the variation in the three estimators necessary for this analysis (N_k, N_{000}, and $\hat{N}$). In the general explanation of the method below, the estimator $\hat{q}$ represents each of the three estimators. For example, N_k in Table 5, above, (47,803), is $\hat{q}$ for N_k taken at the national level.

The method first randomly divides the sample of matched records (total size n records) into k groups, each of which contains m records such that $n = mk$. $\hat{q}_{(a)}$ is calculated by the same method as $\hat{q}$ but with the sample reduced $m(k-1)$ obtained by omitting group a.

Define

$$\hat{q}_a = k\hat{q} - (k-1)\hat{q}_{(a)} \tag{4}$$

and

$$\bar{\hat{q}} = \frac{1}{k}\sum_{a=1}^{k}\hat{q}_a \tag{5}$$

Equation 4 gives us k values of $\hat{q}_a$ calculated from the sub-samples reduced by omitting a group k; the average of these values is $\bar{\hat{q}}$ (see equation 5), called Quenouille's estimator, and removes various biases that affect $\hat{q}$. This estimator is presented in Table 7.

[12] This explanation is taken from Marks, Seltzer, and Krótki (1974, pp. 13-17).
[13] See (Marks, Seltzer, and Krótki, 1974, equation 7.188) . Two possible estimators are given, but we chose the one preferred in cases such as ours, where there is likely to be correlation bias.
[14] This section is largely based on Wolter (1985, pp. 154-155).

The other beneficial result of the jackknife method is that the values of $\hat{q}_a$ are distributed approximately normally.[15] The standard error of the estimator (the square root of the variance) is estimated in Equation 6.

$$SE(\hat{\bar{q}}) = \sqrt{\frac{1}{k(k-1)}\sum_{a=1}^{k}(\hat{q}_a - \hat{\bar{q}})^2} \quad (6)$$

The standard errors given in Table 7 were calculated with equation 6.

Coverage in Space and Time and its Effect on N_{000}

In the discussion concerning Table 1 we noted that none of the three projects covered well violations in the period 1960-1977, and thus no estimation of $\hat{N}$ for this period is possible. The most important complication for the estimation of N_{000} is that the projects did not cover all of the regions with the same intensity. If the regions with different coverage rates among the three projects are not handled separately, the estimation could be biased.

For example, consider Region IV, in which the CEH found more than 1,300 killings, while the other two projects reported only some dozens of killings. Clearly the levels of overlap are low, but the overlap rates should not be used for an estimate of N_{000} because the concept of overlap requires that the projects actually collected data in the same areas. Therefore the estimation of n_{000} was based only in the projects that were in fact able to work in each region. The projects that contributed to the estimation of n_{000} are in Table 6.

[15] The "pseudovalues" $\hat{q}_a$ should be approximately independent and distributed identically. This assumption was tested with a normal probability plot for each set of pseudovalues, and in each case the results were consistent with this assumption.

Table 6: Projects used to estimate N_{000}, by region

Region	Projects with adequate coverage	Equation for n_{000}
Region 0 (other)	All three	Equation 3
Region I (área Ixil)	All three	Equation 3
Region II (Cahabón)	All three	Equation 3
Region III (Rabinal)	All three	Equation 3
Region IV (San Martín Jilotepeque)	Only CEH	Unable to estimate n_{000}
Region V (North of Huehuetenango)	CEH & REMHI	Equation 2
Region VI (area Zacualpa)	All three	Equation 3
Region VII (Guatemala)	CEH & REMHI	Equation 2
Region VIII (area Panzós)	CEH & REMHI	Equation 2
Region IX (Ixcán)	CEH & REMHI	Equation 2
Region X (area Costa Sur)	All three	Equation 3

Results and Some Limits on the Interpretation

We now show the results by component and by region.

Table 7: Total number of killings in Guatemala 1978-1996, by category of estimation and region

	Region											
Category	**0**	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**Total**
N_{111}	67	141	15	146	0	0	17	2	0	2	2	391
N_{110}	378	406	8	98	5	0	67	3	0	16	2	983
N_{101}	1,358	1,010	204	170	13	206	336	24	43	681	13	4059
N_{011}	133	419	16	122	0	0	0	0	0	0	0	690
N_{100}	8,260	3,187	221	1,028	1,325	1,597	1,642	226	156	1,720	182	19,545
N_{010}	2,256	2,708	85	836	16	1	195	15	41	30	91	6,274
N_{001}	5,228	3,999	295	926	59	765	1,166	77	1,099	2,054	106	15,773
N_k (without duplication)	17,679	11,870	844	3,328	1,418	2,569	3,416	347	1,339	4,501	396	**47,706**
SE (N_k)	110	135	24	79	11	44	76	9	26	77	5	228
N_{000}	38,856	17,397	466	6,467	0	5,548	5,836	561	2,265	5,052	2,019	**84,468**
SE (N_{000})	3,809	2045	105	1,152	0	1,826	1,890	350	3,062	995	1,840	6,388
$\hat{N}$	56,535	29,267	1,310	9,795	1,418	8,117	9,252	908	3,604	9,553	2,415	**132,174**
$SE(\hat{N})$	3,918	2175	127	1,218	11	1,870	1,964	357	3,087	1,072	1,844	6,568

In Table 7, it can be seen that in Region 0 there were N_k = 17,679 killings documented in all three projects. Over all regions, there were 47,706 killings documented, being the sum of the regional estimations[16]. The standard error SE(N_k) is not the simple sum of the regions, but rather the square root of the sum of the squares of the regional values (i=0, I, …, X):

$$SE(N_{kl}) = \sqrt{\sum_{i=0}^{X} SE(N_{ki})^2} \qquad \text{Equation 7}$$

Similarly, the values for N_{000} and for N are the sum of the regional values and the standard error of N_{000} and $\hat{N}$ is the square root of the sum of the squared regional values. In this way, we estimate

[16] The estimation for N_k was 47,706 murders documented between all three projects, with a standard error of 228, yielding a 95% confidence interval of 47,559 to 48,152. Note that this range includes the value estimated in Table 5, 47,803. The closeness of the value in Table 5 with the value estimated by the sum of the regions by the jackknife method implies that there is not much bias in the simple estimation. Nonetheless, the bias that required the disaggregation by regions may not have affected N_k, but yet might still affect N_{000}, and is therefore still necessary.

that there were approximately 84,468 killings that were not reported to the CEH, to the CIIDH, or to the REMHI project. Summing N_k and n_{000} to $\hat{N}$, we have as our final estimate, that there were 132,174 killings in Guatemala between 1978-1996, with a standard error of 6,568.

Possible Corrections and Limitations to the Interpretation of Table 7

There are five sources of error that cannot be quantified in this analysis due to lack of time, resources, and adequate data. In preliminary analysis of this type, the global effect of these corrections is conservative, in that the corrections tend to reduce the estimation of $\hat{N}$. The conclusion of this section is that the accumulated effect of the identified biases is does not significantly change our interpretation of Table 7.

Correlation between sources

The estimation of N_{000} depends on the assumption of independence between sources; that is, that the probability that any given respondent gives her testimony to one project has no effect on the probability that the same respondent will give her testimony to one of the other two projects. It is certain that this correlation is not zero, but is positive, for two reasons.

First, psychological research has shown that survivors of human rights violations who are able to give testimonies under supportive conditions experience improvement in their mental health. Thus, it is likely that people who give testimony in these conditions may seek additional opportunities to give their testimony, thereby increasing the overlap rate.

Second, it is known that several popular movement groups organized their social bases to present testimonies to all three projects. In this way, members of these organizations would have greater probabilities of giving testimonies more than once, thus, reporting the same violations more than once, and increasing the overlap levels. The two effects – both of which are certain – would tend to bias the estimate of n_{000} toward a smaller number.

Matching errors

If the analysts who conducted the matching failed to find victims who were in multiple databases, by accident or because there were inadequate data in the original sources, these omissions would tend to depress the level of measured overlap and in consequence bias the estimation of n_{000} upwards. In preliminary investigations, (all that are possible given the partial state of many cases) only minimal effects of this kind were found. At most, they amounted to about 12% of the final estimate of $n_{000,}$ implying about 8% of $\hat{N}$. Considering the other sources of error listed in this section, and recognizing that the data for this analysis were limited, we decided not to quantify this error (or a correction for it) in the final analysis.

Internal duplication

All of the projects that receive information from primary sources may have problems with internal duplication that results from multiple reports of the same events.[17] Internal duplication tends to artificially increase the number of killings that are reported in a single database. All three projects worked hard to clean their data to reduce internal duplications, but some always remain. In a preliminary analysis, insufficient duplication appeared to require a correction for this source of error.

Rates of overlap between kinds of victims

The amount of overlap between the three databases was based on an analysis of victims identified by name and surname. However, many victims are not identified by name as a result of large-scale violence that overwhelmed the capacity of the witnesses to remember all the victims' names. It is possible that the level of overlap between victims not identified by name could be either higher or lower than the levels measured among named victims. Given the difficulty of matching unnamed victims, it is not possible to quantify the potentially different overlap levels among victims with

[17] See, in this context REMHI (1998, pp. XXXI-XXXII), and Ball, Kobrak, and Spirer, (1999, pp. 62, note 12).

different amounts of identifying information. As mentioned earlier, we assumed that the match rates for unnamed and named victims were the same.

Geographic areas excluded from the analysis

In Table 7 we noted that there is not sufficient data to estimate n_{000} in Region IV. Given the experience of other regions, in which the ratio between N_k and n_{000} varies between 0.5 and 2, with a mode and mean close to 0.5, it is likely that the value of n_{000} for Region IV is approximately 2,500. In other regions N_k is composed of data from more than two projects. However, in Region IV information was collected only by the CEH. Thus, it is possible that the ratio between N_k and n_{000} for Region IV could be 0.25 or less, increasing the estimate by a factor of two or more. There are no other methods available to reduce the lack of certainty about this number, and therefore it is not included in the final estimate.

In addition, Region IV exemplifies a more fundamental problem: This methodology works only for areas covered by at least two of the three projects, even if the two projects only partly cover each area. In areas in which only one or none of the three projects conducted interviews, there is no basis for an estimate of the total number of excluded victims (n_{000}). Instead, in these situations only N_k (the total documented number of killings) enters the estimation process. As all three projects focussed on particular areas in Guatemala in which large scale violations were known to have occurred on the basis of journalistic and NGO accounts, it is unlikely that the excluded areas had high levels of human rights violations. However, if our analysis included any excluded area in which killings could have occurred, that inclusion would tend to increase the final estimate.

Estimation of Killing Rates, by Ethnic Group and Region

The rate of killing for a defined group is the proportion of people in that group who were killed. Quite simply, it is the number of people in the group who were killed divided by the total number in the group prior to the killings. The CEH was interested in comparing the rates of killing for defined ethnic groups during the period 1981-1983.

Six geographical regions were identified as those in which -- according to secondary sources and anecdotal evidence -- state violence was concentrated against indigenous peoples. These six regions are listed in Table 8, with the ethnic group populations according to the census of 1981.

Table 8: Populations in six regions, by ethnic group, 1981

	Indigenous	Non-indigenous
Region I: Ixil area	38,902	5,882
Region II: Cahabón	20,706	868
Region III: Rabinal	18,610	4,120
Region IV: San Martín Jilotepeque	31,690	4,876
Region V: north of Huehuetenango	53,556	11,123
Region VI: Chiché, Zacualpa, Joyabaj	51,105	10,997

To calculate the killing rate, the number of victims is estimated. We did this estimation twice, first to get the total number of documented victims, and then to get the estimated number of victims using the methods outlined above.

The following are the steps in this estimation process:

1. The number of murders that occurred 1981-1983, less those attributed to the URNG, were calculated by the ethnic group classifications indigenous, not indigenous, and unknown, for each of the six regions in the three databases. This step is analogous to Table 1 above,

but limited to killings attributed to the state during the period 1981-1983 and disaggregated by the ethnicity of the victims.

2. The number of matched victims and the corresponding rates were calculated for each of the six regions (logically following the method shown for Tables 2 and 3).[18]
3. The number of victims for each ethnic group was estimated using the regional rates of overlap and the number of victims in each database (similar to Table 4).
4. The estimates were made for each region by taking the average of each of the three database estimates (similar to Table 5).
5. The jackknife method was applied to each defined group, following equations 4, 5, and 6, in order to estimate N_k and $\hat{N}$ (and their standard errors) for each ethnicity in each region. The values of N_k are presented below in Tables 9a and 9b, and those for $\hat{N}$ in Tables 11a and 11b.
6. Taking from Table 9a the victims with known ethnicity, the victims without known ethnicity were apportioned to the categories "indigenous" or "not-indigenous" according to the proportions shown below in Table 10, creating the figures shown in Table 9b.
7. With the information from Tables 8, 9, and 10, the proportion killed of each ethnic group in each region was calculated, along with its standard error. The data, presented below in Figure 3, explain *inter alia* that according to the information documented by the CEH, CIIDH, and REMHI, more than 14% of the indigenous population in the Ixil area in 1981 were murdered by 1983, while in the same period and area, only 2% of the non-indigenous population were killed.

[18] The overlap rates were not calculated by ethnic group. Instead the regional match rates for the period 1981-1983 were applied equally to the ethnic groups in that region. This application assumes that the overlap rates did not vary significantly among ethnic groups.

Table 9a: Number of documented killings (N_k) in three databases, 1981-1983, by region and ethnicity

		Region I	Region II	Region III	Region IV	Region V	Region VI
Indigenous	N_k	1388	340	1071	1012	1020	1126
	$SE(N_k)$	25.5	13.12	33.71	0.67	17.11	16.7
Non-indigenous	N_k	32	2	13	16	8	6
	$SE(N_k)$	0.49	0.07	0.33	0.14	0.12	0.13
Unknown ethnicity	N_k	4339	186	1669	10	752	1208
	$SE(N_k)$	62.87	9.02	47.75	0.64	31.95	46.79

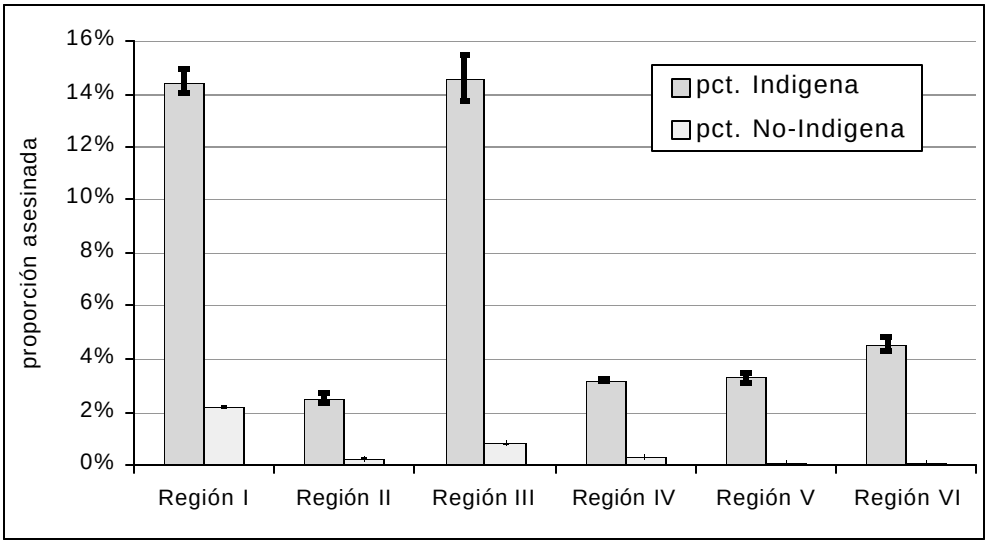
Table 9b: The number of documented killings (N_k) in three databases, by ethnic group, including victims without known ethnicity, 1981-1983

		Region I	Region II	Region III	Region IV	Region V	Region VI
Indigenous	N_k	5,632	525	2,720	1,022	1,767	2,327
	$SE(N_k)$	66.56	15.90	57.98	0.92	36.05	49.42
Non-indigenous	N_k	127	3	33	16	13	13
	$SE(N_k)$	0.49	0.07	0.33	0.14	0.12	0.13

Table 10: Percentage of indigenous of victims with known ethnicity of all victims in Table 9a

	Region I	Region II	Region III	Region IV	Region V	Region VI
Proportion indigenous	97.8%	99.6%	98.8%	98.5%	99.3%	99.4%
Proportion of victims with known ethnicity	24.7%	64.7%	39.4%	99.1%	57.7%	48.4%

Figure 3: Documented proportion of the population killed by State forces in Guatemala 1981-1983, by region and ethnic group, with the 95% confidence interval[19]



8. Note that the data presented for Region VI (Zacualpa area) in Figure 3 do not correspond exactly to the statistics presented in the genocide section of the CEH report because the definition of Region VI used here includes the *municipios* of Chiché, Joyabaj and Zacualpa. In the genocide section of the report, only the *municipio* of Zacualpa is considered part of Region VI. The statistics for Region VI (and for all the regions) in Figure 3 and in the genocide section were calculated with the same methods but with different population and violation bases.

9. The projected totals ($\hat{N}$) by ethnicity and region were calculated using the same methods described with equations 2-6, and with the same data as shown in Tables 8-10. The statistics are presented in Tables 11a and 11b and rates are shown in Figure 4.

[19] Source of the graph: 1981 census; testimonies received by the CEH, direct sources to the CIIDH, and testimonies received by the REMHI project.

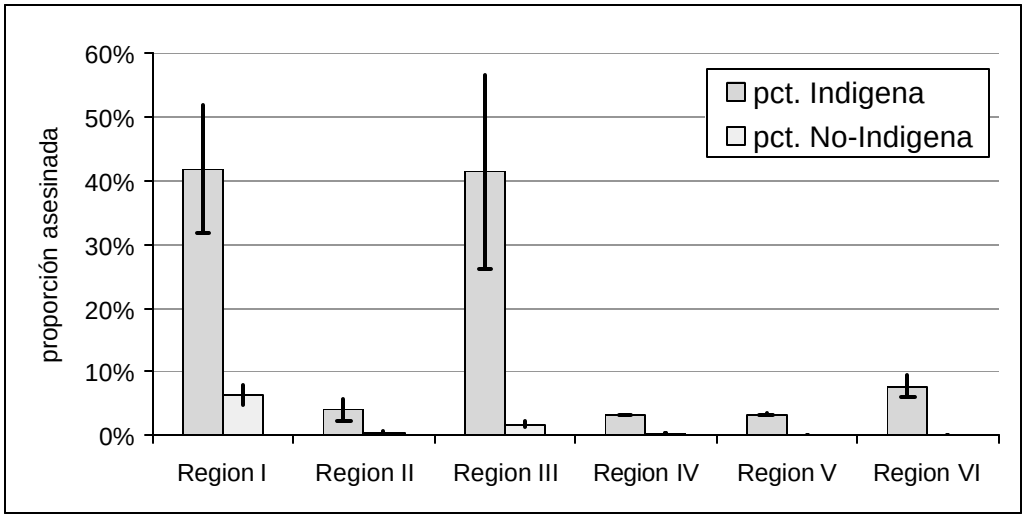
Table 11a: Number of projected killings ($\hat{N}$) in three databases, by ethnic group and region, 1981-1983

		Region I	Region II	Region III	Region IV	Region V	Region VI
Indigenous	$\hat{N}$	2578	443	2983	1012	1020	2723
	SE($\hat{N}$)	190.28	39.13	587.58	0.67	17.11	386.43
Non-indigenous	$\hat{N}$	63	2	13	15.7	7.5	6
	SE($\hat{N}$)	5.4	0.07	0.33	0.14	0.12	0.13
Unknown ethnicity	$\hat{N}$	14014.6	394	4791	10	752	1208
	SE($\hat{N}$)	1841.44	151.12	874.1	0.64	31.95	46.79

Table 11b: Number of projected killings ($\hat{N}$) in three database, by ethnic group and region, including victims without identified ethnicity, 1981-1983

		Region I	Region II	Region III	Region IV	Region V	Region VI
Indigenous	$\hat{N}$	16,284	835	7,717	1,022	1,767	3,924
	SE($\hat{N}$)	1,811.0	155.5	1,044.5	0.9	36.1	389.2
Non-indigenous	$\hat{N}$	371	4	70	16	13	13
	SE($\hat{N}$)	40.9	.6	10.5	.1	.3	.3

Figure 4: Projected proportions of ethnic groups killed by state forces in Guatemala 1981-1983, by region and ethnic group, with 95% confidence interval



Note that Figures 3 and 4 have two interpretations.[20] First, regions I and III were the most affected by state violence (based on rates). In these two regions there are clear quantitative signs that the killing was so massive that it could have been genocide. Second: in all the regions the victims were disproportionately indigenous. Note, for example that as shown in Figure 4, in Region I more than 40% of the indigenous population was killed while approximately 8% of the non-indigenous population were killed. The difference between the killing rates is a factor of five. In the structure of violence committed by the Guatemalan state, these are revealing differences.

Comparisons among Databases

In the analysis of multiple databases, the databases can be compared in order to determine the levels of temporal or geographic agreement among them, and in terms of the relative levels to which they attribute responsibility to the state or insurgent forces. In this section we compare the tendencies and statistics among the three databases.

It is clear from Table 1, that there is only a moderate level of agreement about where violence occurred among the databases. In some regions all three projects found many violations, (Regions 0, I, and III), while in other regions only two projects investigated deeply (for example, Regions V, VI, IX), while in Region IX only the CEH carried out an intensive investigation.

Although the projects covered distinct areas, the second section in this part shows that the databases agree on which months saw the peaks of the violence. The third section considers the relative proportions of responsibility attributed to the two parties to the conflict.

Coincidence in Time

If the months are ordered in terms of how many killings are reported according to each of the three databases, a relatively high level of agreement is found. In Table 12, the top ten months are shown ordered as described, presenting the percentages of the total number of killings during the entire period 1979-1984.

Table 12: **The ten most violent months in the three databases, 1979-1984**[21]

	CEH				CIIDH				REMHI		
Range	**Month**	**Total**	**Pct.**		**Month**	**Total**	**Pct.**		**Month**	**Total**	**Pct.**
1	82-01	2,256	9%		82-02	610	12%		82-03	1,330	12%
2	82-03	2,253	9%		81-06	390	7%		82-02	807	7%
3	82-02	1,880	8%		83-03	297	6%		82-07	792	7%
4	82-08	1,819	8%		82-06	279	5%		82-05	657	6%
5	82-07	1,719	7%		82-07	234	4%		81-09	629	6%
6	81-01	1,423	6%		82-01	233	4%		82-01	470	4%
7	82-06	1,146	5%		82-04	222	4%		82-04	428	4%
8	82-04	937	4%		82-05	210	4%		81-07	397	4%
9	82-05	895	4%		83-08	180	3%		80-02	364	3%
10	81-09	754	3%		81-02	174	3%		82-10	360	3%
Ten month total		**15,082**	**63%**			**2,829**	**54%**			**6,234**	**56%**
Total for 1979-1984		**23,890**	**100%**			**5,275**	**100%**			**11,065**	**100%**

The shaded five months are those for which the three databases show concordance. Within the ten worst months in each database, the three systems agree on five months: January, February,

[20] Note that in both absolute and relative terms, the standard error for each statistic in Figure 3 is greater than that for the analogous statistic in Figure 2. This is consistent, as the projections in Figure 3 incorporate more uncertainty than the estimations in Figure 2. The size of the samples on which the estimation of the unduplicated totals were based (N_k) are sufficient for those estimations which do not have such high errors as to make them unusable. The projection, however, still contains significant uncertainty, reflected in the higher error rates.

[21] In Table 12, only killings identified with dates precise to the month are included.

April, May, and July of 1982; other months of the same year (March and June) coincide in two of the three databases.

The databases are in agreement that approximately half of the killings occurred in the ten worst months (63%, 54%, and 56%). This concentration follows Pareto's Law, which states that 80% of any given phenomenon will occur in 20% of the categories. However, the closeness of these months to each other in time (all of them occur toward the first half of 1982) is strong evidence that this period is the most intense period of political violence. Furthermore, the level of agreement between the databases implies that although the projects did not investigate exactly the same regions of the country, they found the same trends in time.

Coincidence in the Attribution of Responsibility

The three sources agree in the attribution of responsibility to state forces and the insurgents: together they attribute more than 94% of the killings to the state and less than 6% to the guerrilla forces.

Table 13a: Total number of killings with identified perpetrator, by responsible entity[22]

	CEH	CIIDH	REMHI	TOTAL
State	24,121	2,916	19,177	46,214
Guerrilla	1,263	61	1,184	2,508

Table 13b: Percentage of killings with identified perpetrator, by responsible entity

	CEH	CIIDH	REMHI	TOTAL
State	95%	98%	94%	95%
Guerrilla	5%	2%	6%	5%

In an analysis of the attribution of responsibility derived from non-probability samples, such as those used for the statistics in Tables 13a and 13b, there is the possibility that the data may have been biased towards the violations committed by one entity or another because of a tendency to focus on one perpetrator. This kind of "over-focus" bias can effect the estimated proportions.

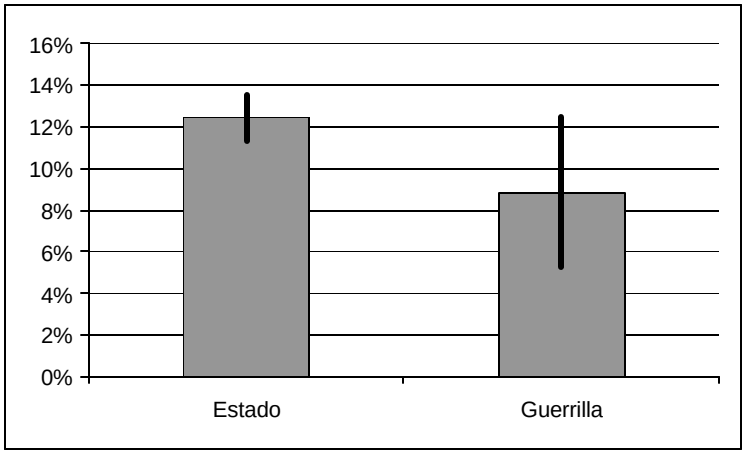
For example, if the projects looked primarily for violations committed by the guerrilla forces, while dedicating less effort to the search for violations committed by state forces, this data would reflect an inflated level of responsibility attributed to the guerrilla forces. In a probability sample, this bias is avoided by accepting testimonies that were selected at random. None of the three projects sought testimonies according to a probabilistic design and hence, there may be bias in these resulting proportions.

Taking advantage of the existence of the three projects and using the measures of overlap explained earlier, we can test the hypothesis that the projects over-focused on one or the other of the perpetrating entities. The components of the estimated documented killings must be separated into the killings attributed to the insurgents and those attributed to the state. The components that indicate overlap are summed ($N_{111}+N_{110}+N_{101}+N_{011}$) and divided by N_k; this figure gives the percent of overlap by responsible group. The results of this calculation are shown below in Figure 5[23].

[22] This only includes violations with the perpetrator identified and with a date precise to the year. It is worth reemphasizing that this analysis includes all killings, not only arbitrary executions.

[23] The table is based on Figure 5, corresponding to Table 5 divided by the perpetrating entity, shown in disaggregated form below.

Figure 5: Overlap rates for victims of killing documented by the CEH, the CIIDH, and by REMHI, for violations committed by the state and guerilla forces (with bars to indicate the 95% confidence intervals.).



With the results of Figure 5, we rejected the hypothesis that there is a significant difference between the level of coverage of violations committed by the guerrilla forces and those committed by the state forces. Although there is a small difference in the overlap rates of the state forces (12.4%) and the guerrilla forces (8.8%), the difference is within the standard error. Thus, the difference cannot be distinguished from the sampling error of the matching process.

The difference between the overlap rate for killings committed by the state and by the guerrillas is significant, neither in technical terms, nor in analytic terms. The analysis of the effect on the estimate proportions follows. The technical test is the following:

$$SE = \sqrt{\frac{p_E * (1 - p_E)}{N_E} + \frac{p_G * (1 - p_G)}{N_G}} = 0.0193,$$

which yields the confidence interval +/- 3.8%. The difference between the two rates is 12.4% - 8.8% = 3.6%; the confidence interval is more than the difference, which means that we cannot reject the hypothesis that the difference is equal to zero. This calculation confirms the intuitive interpretation from Figure 5.

The implication of Figure 5 is that all three projects investigated violations committed by the guerrillas and violations committed by the state with approximately the same level of coverage and intensity. Therefore there is no systematic disproportionality in the intensity of investigation between the two entities sufficient to change the interpretation of the proportions of responsibility attributed to each.

Category	State	Guerrilla
N_{111}	299	3
N_{110}	617	
N_{101}	3769	200
N_{011}	388	
N_{100}	19,173	1,087
N_{010}	2,165	61
N_{001}	14,430	949
N_k	40,842	2,301
Overlap Rate $\frac{N_{111} + N_{110} + N_{101} + N_{011}}{N_k}$	12.4%	8.8%
SE (N_k)	0.5%	1.9%

The standard error calculated by the conventional method for proportions derived from samples is $SE = \sqrt{\frac{p * (1 - p)}{N}}$, and the 95% confidence interval is +/- 1.96*SE.

There are two ways to consider the effect of that the overlap rates on the proportion of responsibility attributed to the state and the guerrillas. The proportions of attributed responsibility that result from N_k estimated in note 23, are presented below. The average does not come from all three databases, as implied in Table 13 in the text. Note that this analysis excludes violations for which responsibility is unclear; adding the unknown category would reduce both proportions slightly.

Figure 6. Proportions of attributed responsibility.

Estimation	State	Guerrillas
N_k	41,147	1,860
Proportion of the total N_k	95.7%	4.3%
n_{000}	73,622	3,706
$\hat{N}$	114,769	5,567
Proportion of the total $\hat{N}$	95.4%	4.6%

To see the insignificant effect of the disproportionality on coverage, n_{000} is calculated (using Equation 3 for the state, but using Equation 2 for the guerrillas because the CIIDH did not report sufficient guerrilla violations). The estimation of n_{000} includes the information about the overlap rates, and in this way n_{000} controls the effect of the disproportionality in coverage. Note that the calculated proportions of $\hat{N}$ are the same as those calculated for N_k. The conclusion is that the disproportionality in coverage of the state and the insurgents does not change the final analysis about their relative responsibility.

Appendix 1

Regional Definitions, by *municipio*

All of the *municipios* other than those listed here were classified as Region 0. Note that the definition of Region VI in this study is not the same as the definition used in the CEH report. This study included two additional *municipios* that the genocide study did not include.

REGION	*Departamento*	*Municipio*
Region I	Quiché	Chajul
Region I	Quiché	San Juan Cotzal
Region I	Quiché	Nebaj
Region II	Alta Verapaz	Cahabón
Region III	Baja Verapaz	Rabinal
Region IV	Chimaltenango	San Martín Jilotepeque
Region V	Huehuetenango	Nenton
Region V	Huehuetenango	San Mateo Ixtatán
Region V	Huehuetenango	Barillas
Region VI	Quiché	Chiche
Region VI	Quiché	Zacualpa
Region VI	Quiché	Joyabaj
Region VII	Guatemala	Guatemala
Region VII	Guatemala	Mixco
Region VIII	Alta Verapaz	Panzós
Region VIII	Alta Verapaz	San Pedro Carchá
Region IX	Quiché	Ixcán
Region X	Santa Rosa	Cuilapa
Region X	Santa Rosa	Barberena
Region X	Santa Rosa	Casillas
Region X	Santa Rosa	Santa Rosa De Lima
Region X	Santa Rosa	Oratorio
Region X	Santa Rosa	San Rafael Las Flores
Region X	Santa Rosa	Santa Maria Ixhuatan
Region X	Santa Rosa	Taxisco

REGION	*Departamento*	*Municipio*
Region X	Santa Rosa	Chiquimulilla
Region X	Santa Rosa	San Juan Tecuaco
Region X	Santa Rosa	Guazacapán
Region X	Santa Rosa	Naranjo
Region X	Santa Rosa	Pueblo Nuevo Las Viñas
Region X	Santa Rosa	Nueva Santa Rosa
Region X	Escuintla	Escuintla
Region X	Escuintla	Santa Lucía Cotzumalguapa
Region X	Escuintla	La Democracia
Region X	Escuintla	Siquinalá
Region X	Escuintla	Masagua
Region X	Escuintla	Tiquisate
Region X	Escuintla	La Gomera
Region X	Escuintla	Guanagazapa
Region X	Escuintla	San José
Region X	Escuintla	Iztapa
Region X	Escuintla	Palín
Region X	Escuintla	San Vicente Pacaya
Region X	Escuintla	Nueva Concepción
Region X	Retalhuleu	Retalhuleu
Region X	Retalhuleu	San Sebastián
Region X	Retalhuleu	Santa Cruz Mulua
Region X	Retalhuleu	San Martín Zapotitlán
Region X	Retalhuleu	San Felipe
Region X	Retalhuleu	San Andrés Villa Seca

REGION	*Departamento*	*Municipio*
Region X	Retalhuleu	Champerico
Region X	Retalhuleu	Nuevo San Carlos
Region X	Retalhuleu	El Asintal
Region X	San Marcos	Nuevo Progreso
Region X	San Marcos	El Tumbador
Region X	San Marcos	Malacatán
Region X	San Marcos	Catarina
Region X	San Marcos	Ayutla
Region X	San Marcos	Ocos
Region X	San Marcos	Pajapita

References

Ball, Patrick, 1996. *Who Did What to Whom? Designing and Implementing a Large-Scale Human Rights Data Project*. Washington: American Association for the Advancement of Science.

Ball, Patrick, Kobrak, Paul, and Spirer, Herbert, 1999. *State Violence in Guatemala, 1960-1996: A Quantitative Reflection*. Washington: American Association for the Advancement of Science and Centro Internacional por Investigaciones en Derechos Humanos.

Marks, Eli S., Seltzer, William, and Krótki, Karol J., 1974. *Population Growth Estimation: A Handbook of Vital Statistics Measurement*. New York: The Population Council.

REMHI 1998. Project Report, *Guatemala: Nunca Más Tomo IV, Victims del Conflicto*. Guatemala: Oficina de Derechos Humanos del Arzobispado de Guatemala.

Wolter, Kirk M., 1985. *Introduction to Variance Estimation*. New York: Springer-Verlag.

Chapter 12

The Guatemalan Commission for Historical Clarification: *Database Representation and Data Processing*

Sonia Zambrano

Introduction

In this report I review the processing and representation of information concerning human rights violations and other violence that occurred during armed conflict in Guatemala from 1960 to 1996. The tasks of processing and representing information were conducted by the database team of the Guatemalan Commission for Historical Clarification (CEH), which presented its final report in February, 1999.

I analyze the database work as part of an integrated process that goes beyond the representation of information and involves all parts of the organization of a truth commission responsible for reporting on large-scale violence. To achieve these aims, this report contains three parts. The first part describes the internal capacity of the database and information processing. The second part describes database functions in coordination with other CEH sectors. The third part contains my conclusions and lessons learned based on my experience as the director of the CEH database.

Information Processing in the Database

The database team of the CEH had the task of processing CEH's core information, which was testimonies presented by Guatemalans who came to the CEH and were based on cases of *human rights violations* and *other violence* that occurred during the armed conflict in Guatemala.[1]

To create a suitable database, the CEH formed a team to receive information collected by interviewers in the field, organize it, analyze it, structure it and input it into the CEH database. The goal was to systematically store both qualitative and quantitative information. This information would provide an important resource for the formulation and testing of hypotheses and analyses presented in the final report.

This process was dominated by information processing. The database was developed in several principal phases that were implemented both in series and in parallel according to the needs of the overall process.

Establishing the Database

This phase consisted of designing and implementing the database. After the design was completed, implementation – direct field collection of information by CEH and subsequent processing took place. This phase consisted of the activities described below.

Forming the database team:

A team of 21 people was formed to perform database-related tasks. Tasks were distributed as follows:

1. The *database coordinator* was responsible for the database, in charge of coordinating the entire databasing process, and in charge of coordinating the database work with respect to other CEH sectors.
2. The *database assistant* worked with the database coordinator and was in charge of coordinating internal processes by delegating tasks.
3. The *programmer* was responsible for electronic programming and designing the process for inputting information.
4. The *systems assistant* was responsible for maintaining computers and the physical infrastructure of the information system.
5. The *systems analyst* was responsible for producing statistical information.

[1] By legal definition, human rights violations are committed by state actors; other violence is committed by non-state actors such as the guerrillas.

6. The *archive assistant* was responsible for organizing the database archives, answering demands for information by interviewers, and controlling the physical movement of information to guarantee its integrity and security.
7. Nine *analysts* were responsible for analysis and preparation of information for subsequent input to the database.
8. Six *data entry specialists* were responsible for inputting information to the database using the program designed for that purpose.

The formation of the team involved a selection process (interviews and reviewing applicants' backgrounds), hiring and training the selected people, and finally a process of frequent discussions to guarantee methodological uniformity when processing information. Unfortunately, team members were hired at different times. Not having the whole team together from the start meant a loss of time, since bringing each new person up to speed entailed a new training and preparation process before starting work. This affected the workflow and efficiency of the team members who were already at work.

For example, information analysis started with five analysts. Unfortunately, a team of only five people could not process the huge quantity of information in the desired time. This meant expanding the team of analysts in the middle of the process, which called for repeating training for the new people and discussions to establish a uniform methodology.

Forming the database infrastructure

This refers to forming teams, establishing the network, setting up programs, defining security systems and protecting information, among other tasks.

Constructing the electronic database

This task involved designing and implementing the program, creating tables, defining relations, constructing the interface, testing and correcting the program, and so forth.

The consultants who started to work on creating the CEH database had to withdraw and could not complete their part of the process. When new project managers arrived to help create the database, we had to continue setting up the database at the same time field interviewers were collecting information. Due to these difficulties, neither the database infrastructure nor the program for data entry were finished when the interviewers had started collecting information in the field. This mistiming caused a setback in the data entry and analysis process. Consequently, it took longer than planned before the database entry personnel could start to integrate the information that was arriving from the field.

This delay caused a backlog of information at the input to the database and a gap between collecting and systematizing information, which affected the coordination between these two phases. This situation demonstrated to us how important it is that both the physical and electronic infrastructure of the database be completely finished before starting to collect information so that the database can start to input information as soon as collection begins.

Creating the Database Archives

This phase consisted of creating a system for receiving, classifying and filing cases, similar to designing a consulting system and service to interviewers who requested information from the database.

Creating the case archives was an activity that entailed considerably more work than was anticipated. The archival work starts at a high level from the moment the work on the database begins. Receiving, classifying and filing cases, and controlling information to guarantee its security were tasks that required care and the full-time attention of one person to perform them.

The archival work included many unanticipated activities that required a lot of time but were essential to complete. For example, the database had to respond to interviewers' requests to consult the physical archives. These requests were based on a list of cases the database prepared for interviewers according to their specific needs. This activity continued during the time it took for the analysis and preparation of the final report, and the systematized information in the database was a vital resource for the CEH.

Controlling and guarding information — that is to say, its security and integrity inside the database — depended on the organization of the archives. To achieve these goals, we had to devise a strategy for information classification and movement (lending and filing) that allows for controlling and maintaining the integrity and security of the information.

Collecting Information

Although this process was not strictly part of the database, I briefly discuss it because it is the step immediately prior to analysis and recording information and is directly related to information processing.[2]

This phase consists of information collection by CEH interviewers, who collected approximately 11,000 testimonies (collated into 7000 cases) on human rights violations and other violence that occurred in Guatemala during the armed conflict.

This was made possible by setting up 14 regional CEH offices in central locations around the country. This large number of regional offices was needed to get the widest possible coverage. Information was collected during about eight months, when the interviewers received testimonies. These testimonies account for the primary direct information below:

1. Testimonies on cases of human rights violations and cases of other violence.
2. Testimonies on the general situation or the context in which violations were committed.

In addition, CEH interviewers also collected substantial information from other sources; documents, books, official institution and NGO reports, among others. To file and systematize all of this information, we set up a documentation center, in which electronic databases and physical archives were maintained.

The database team was in charge of systematizing testimonies that were received by CEH interviewers. Interviewers wrote regional reports in which all of the information on context that interviewers collected in the field was retained. These reports were also kept in the documentation center. The information in the database and the information in the documentation center were compared and used in conjunction for making theoretical analyses, formulating hypotheses and drafting the final CEH report.

Methodology of information collection

The methodology for collecting information that was prepared for interviewers was limited to creating information collection instruments, or record forms of cases and a glossary of violation types. The classification of information was based on these tools.

Basic concepts, criteria for analysis and general categories of classification were not defined in this step but were left for a later process that was necessarily developed for the most part inside the database. There the basic parameters of information classification could be defined. This process corresponded to a previous phase of defining the general CEH methodology, and was established during the information analysis phase. This obligated the CEH to create parameters and on many occasions reformulate criteria applied in the information collection process that was already underway.

We tried to overcome such difficulties through ongoing contact between the interviewers and myself as database director. We used these meetings to work towards common standards and on the minimum necessary modifications, while trying not to affect the collection process that had already begun. For example, the violation types were created before starting the collection process and were ready when interviewers went to the field. However, there was no careful discussion by different sectors of CEH concerning the violation types. Later when they were applied in the field and interviewers were more familiar with them, the violation types were reformulated to adjust them to the reality in the field. Unfortunately there were no uniform standards for collecting and interpreting deponents' testimonies or accounts.

In the first months of data collection, quantitative information was given priority over qualitative. Thus, the work focused on filling out record forms accompanied by a short summary of events. Subsequently, the database sector started to insist on the importance of the testimonies and the need to recover qualitative information to achieve a more complete report. This suggestion led to more detailed testimonies with more information with which the database could achieve much more.

Nevertheless, the collection of testimonies continued without uniform criteria. For example, there was no clear definition of the importance of the testimony and what could be obtained from it. Every interviewer oriented his/her interviews according to his/her training and personal interests. As a result, testimonies often differed significantly and it was not easy to apply systematic criteria

[2] A detailed discussion of this process is given in Chapter 8.

to classify them. Lawyers looked for testimonies geared toward knowledge of legal instruments that would elucidate the facts, while those concerned with the sociology of the events emphasized social aspects in the context of the region and the consequences of events on the affected population. Those who had political training favored interpretations that supported their hypotheses. Since these three versions were fundamental to attain a complete overview, each was dealt with independently. As a consequence, there were frequent omissions of information in testimonies making for incomplete descriptions of reality.

Some interviewers, who were interested in specific themes such as violence against women and children, emphasized this aspect in the testimonies, while in my view others did not place enough importance on such themes. One could only count on the few testimonies that were taken by interviewers who were interested in a particular theme, and they were insufficient to quantitatively analyze the phenomenon.

Similarly, while some interviewers wrote the testimony just as they had collected it (that is, they recorded the original testimony given by the deponent), other interviewers filed their reports introducing their own interpretations of events. Thus, the database analysts could not distinguish between the deponent's version and that of the interviewer. When performing analyses, it was difficult to create adequate bases for analysis in all situations.

Another difficulty was the lack of clarity in the way forms were filled in. The most obvious case concerned the question on the victim form regarding the "mother tongue." The purpose of this question was to determine the ethnic identity of the victim, an important element in Guatemala where the indigenous population was the main victim of violence. Some interviewers correctly recorded the mother tongue spoken by to the victim or the victim's community (Mayan, Spanish, and others), but other interviewers recorded the language that the victim or the victim's community currently spoke. Since in many Guatemalan communities Mayans speak Spanish, the data was incorrect in those cases. Even if the victim speaks Spanish, s/he was indigenous and that was the primary concern.

A problem was detected when the information was analyzed, and it was no longer possible to return the cases to interviewers to recover the correct information. In this situation, the database team, with the help of several Guatemalan interviewers, backtracked case by case, cross-checking the information with data obtained from the indigenous category in the glossary of victim types and recovering the correct information. Fortunately, the process was a success, and the resulting information gave statistical and qualitative support to prove that the indigenous population represented the great majority of the victims of violence.

The above shows that the definition of methodology and clear parameters for collecting information are important, since they had direct bearing on subsequent information processing and the effectiveness of its results and the efficiency of the process. When these parameters are drawn later in the analysis phase, it necessarily affects the collection process.

Information collection instruments

To collect information, seven forms were created on which basic or necessary information was recorded to subsequently obtain statistical results.

The forms that were created are shown in Table 1, following:

Table 1. The forms used for recording basic or necessary information.

Control Form	Contains information on the case number, number of victims, number of violations, and the date and place of events
Case Summary Form	Contains information on the case number, date and place of events, number of victims, a summary of the case, and key words found in the summary
Individual Victim Form	Contains information on names and surnames, age, sex, marital status, type of victim, place and date of birth, and personal data regarding the victim
Collective Victim Form	Description of the collective victim (group, family, village, etc)

Violation Pattern Form	Contains specific data regarding the violation or violations that occurred; date and place where it occurred, the perpetrator, level of certainty that the violation occurred, level of certainty regarding responsibility for the violation, level of certainty regarding the alleged perpetrator, and total victims who suffered the violation or set of violations
Individual Case Form	Contains information on names and surnames, age, sex, institution to which the individual belongs, and his or her position
Individual Deponent Form	Contains information on names and surnames, age, sex, type of deponent, relation to the victim, and date and place of birth
Collective Deponent Form	Specific data on the group, community or village that gave the collective testimony

Although the forms were made before information was collected, they needed some modifications once they were tested in the field and discussed with interviewers. These alterations affected the collection process that was already underway.

On some forms important questions were lacking and others were not precise. For example, several questions that were important and that did not appear on the original form had to be added to both the victim and deponent forms. The question regarding the type of victim (based on the glossary of victim types that was described in detail in the section on analysis) was introduced in the victim form. Also, the names and ages of victims' children were registered. This information is basic to determining whether more than one form listing the same given name and surname relates to one or to several different persons.

A change concerning the type of deponent was introduced in the deponent form so that people who approached the CEH to testify (victims, relatives, survivors, witnesses, refugees, displaced persons, etc.) could be identified later. Likewise, the question on the relation of the deponent to the victim was modified since initially the question presented the victim with respect to the deponent. This created inaccuracies since the deponent frequently referred to several victims, and the original only had space for one relation.

On other forms questions were repeated. For example, the control form was abandoned after being used for some time since information was repeated in the summary form. This meant a loss of time for interviewers who had to write the same information several times on both forms. Furthermore, there were difficulties inputting information in the database since the recorded data on every form did not always coincide. For example, in many cases the summary form presented a different date for an event than that which appeared on the control form, even though they were both answers to the same question. Analysts could not ascertain which was the correct date since the information on which they relied was insufficient. That obliged them to ask the interviewer who had recorded the case, but the interviewer usually could not recall.

The summary and the violation pattern forms suffered from a similar problem. Both asked the same question to which there were frequently contradictory responses. For example, the summary form asked the initial and final date. Only one date for every violation could be noted on the pattern violation form, so that often the dates did not coincide. Analysts could not easily ascertain the correct date, nor could interviewers recall the correct data or the reason why different dates appeared.

Frequent contradictions were generated between the number of victims recorded on the summary form and the number recorded on the violation pattern form. To determine the count, information contained in the pattern form was considered valid, since that form referred to every violation and was more precise. The violation pattern form also created many difficulties for interviewers, since they did not have enough time to complete the form and in many cases did not understand it. The database team ended up having to complete and modify this form. These are just a few of the examples of the problems that the forms created.

Modifications that were executed resulted in different forms (the initial and the final forms) which made subsequent systematization of information difficult, since not all of the forms had the same information, and consequently all of the information could not be fully utilized in analysis.

According to CEH's experience, one can conclude that it is more convenient to reduce the number of forms, and that forms should only contain the necessary information for a statistical count and for case analyses. In my opinion, this information must be given priority over the whole story or qualitative descriptions of events. The forms should not keep information that might not be used, and above all must not repeat information. They should be previously tested in the field before applying them definitively to adjust them to the reality of the country under study.

The Systematic Classification of Information

The cases that were compiled by interviewers and recorded on forms were sent to the database where they were recorded, organized in the physical archives, and revised for later input to the electronic database. Information analysis consists of the first phase of information processing in the database and the case by case revision by analysts, applying a pre-defined methodology. Every case was comprised of record forms: case summary, violation pattern, individual or collective victim, individual case, and individual or collective deponents, and an account of the reported case.

Defining the methodology for database analysis

This activity consisted of defining the basic elements of analysis required to start the process of classifying information. It is important to distinguish this approach from the general approach in the investigation process. In this task the analyst determines the basic parameters for classifying information. As was previously explained, in view of the absence of an *a priori* methodology for the process, it was necessary to construct a methodology in the database not only for classification, but also for a general approach to defining parameters for classification.

Thus, the database methodology included the definition of categories, basic theoretical concepts, and criteria for analyzing and assessing information. This involved discussions with other CEH sectors to make the appropriate decisions. The methodology for analyzing information was developed concurrently with the process of analysis. This made it necessary to maintain a high level of flexibility to adjust criteria to the version of reality being studied. A similar level of flexibility also was applied to the collection phase, since setting the parameters for analysis involved reformulating parameters that were applied in collecting information.

We tried to overcome new problems through the coordinated efforts of the database team and interviewers, by developing the collection and analysis phases simultaneously whenever possible. Even then, we did not always meet our objective. Both gaps and successes in the process depended on coordination between the two phases.

Defining basic categories or types for classifying information

The violation categories that were used to classify information are defined in Chapter 8, and should be referenced in connection with this discussion.

Those categories were defined before starting the phase of collecting information. However, in the process of collecting information, the need to modify and adjust the categories to the reality of the situation became evident. The database team revised categories with the general rule allowing new categories when new phenomena appeared were considered important, but did not fit into existing categories. Positive results were obtained by applying this general rule.

Applying this principle, the *others* category was created. The category *deprivation of one's liberty* was created to record deprivation cases that appeared as part of a case that corresponded to the main typologies or because an interviewer had decided to collect it. This allowed the recovery of a number of significant cases to the degree that statistically, this violation constituted one of the five most frequent violations in Guatemala.

In some cases even though a category was created in the database, information could not be recovered in its entirety because those cases had not been systematically received in the field. Only information that an interviewer had decided to collect or that arrived as part of a case and that corresponded to categories could be recorded. This occurred with "dead or wounded combatants," "burning crops" and "forced recruitment."

In other cases, interviewers, commissioners and the central team were not sufficiently familiar with the enormous amount of information recorded in the database to take full advantage of it in

the final stages.[3] This was so despite the importance of this knowledge for the general analysis tasks. Such was the case for the category "disappearance by unknown cause" that illustrates and describes the phenomenon of forced disappearances in Guatemala.

Creating glossaries and tables to classify information

This task was to construct the list of categories on which we based subsequent information classification. The main glossaries were the Glossary of Perpetrators, Glossary of Victims, and the Glossary of Key Words. The database team defined every category and kept interviewers informed. However, it was impossible to create complete uniformity throughout the commission on category concepts and meanings.

Glossary of Perpetrators

This glossary exemplifies the difficulty of creating complete uniformity. Some interviewers spoke of paramilitary groups and others spoke of death squads in reference to the same type of perpetrators. There was no consensus among interviewers on the concept of an "armed group," since some used this label when they were not certain of the perpetrator and others used it for death squads lacking a specific name.

In the case of civilian self-defense patrols (PAC) and military commissioners, there was no agreement on whether to include state agents. The database team anticipated this difficulty by leaving the military, PAC and commissioner categories separate to count them independently—or, if they wished—to count them later as one set.

Every team of interviewers requested different groupings for their analysis. One example is the case of massacres. Information on massacres was requested where the general category was "federal agents" including the military, PACs, commissioners and death squads in one single group. This led to the creation of specific archives solely for analyzing massacres.

Glossary of Victims

The criterion for defining the victim's category was to consider the victim for his/her characteristics, political or social activities, or conditions facing the armed conflict. Membership in these groups represented possible causes of violence to victims. Using these categories made it possible to determine the proportion of people killed in relation with respect to membership to one or more of these groups. The last two group categories, social sector (peasant, day laborer, farm worker, student, shop owner, professor, etc.) and civilian population, were not treated as categories that are similar to the preceding ones. Rather, these categories were opened to record information on the victim whenever it did not relate to the other categories.

This glossary allowed the team to find important information in cases. However, this information could not be used to its full benefit because it was created after the collection of information was already underway and interviewers did not readily grasp its utility. This did not prevent the team from making analyses that highlighted tendencies. For example, the principal groups victimized during certain years of violence could be identified. Also, the years when there was an increase in violence against union members, students, or religious leaders could be determined. The indigenous category made it possible to determine the proportion of the indigenous population that was subject to the violence.

Glossary of Key Words

This is a list of themes or central ideas that the cases might contain, made with the purpose of classifying information according to qualitative criteria.

The glossary was of great value, since through its use important information that appeared in personal accounts was recovered. There was no other way personal accounts could have been recorded and used to classify information. Interviewers could look up classified cases by themes, which allow them to quote testimonies to support their arguments and make analyses of grouped cases to determine tendencies and strategies in the development of violence in Guatemala. For example, they could review as one set all cases in which there were massacres and cruel actions; cases in which economic or labor conflicts were perceived to have been caused by a violent event; cases in which there was violence against children or women, and so forth.

[3] The central team was in charge of coordination of the work of the CEH. It included the Executive Secretary, the Investigations Director, the Operations Manager and the Report Coordinator.

Defining Criteria for Classifying Information

This step consisted in defining criteria employed by the database team to classify information. The major criteria were the following:

Levels of Certainty

Cases recorded by the CEH are grouped according to fixed types. Within these types, groups of cases are organized by perpetrator (categories are described in this section). At the same time, for each perpetrator, cases are grouped according to the level of certainty that the event occurred.

The levels of certainty are set according to both the interviewers' and the database team's assessment of information given by the deponent. Thus, the levels of certainty or confidence in the cases are not levels of legal character insofar as they have not undergone an investigation of events. Thus, one cannot use these indicators as **proof** of confidence. They are **levels** of confidence in the deponent's certainty about the event's occurrence and the perpetrator.

Two types of certainty (event, perpetrator) and three levels for each type of certainty were used as shown in Tables 2a and 2b).

Table 2a. Certainty Regarding the Event

Level	Deponent Role
1	Direct witness
2	Deponent is not a direct witness/ there are other witnesses
3	Deponent is not certain

Table 2b. Certainty Regarding the Perpetrator

Level	Deponent Role
1	Direct witness or documented evidence exists
2	Deponent is not a direct witness or there are other witnesses
3	Deponent suspects the perpetrator or it is public knowledge

To set the three general levels of certainty of a case, the types of certainty of Tables 2a and 2b were combined as shown in Table 3. To interpret this table, note that (1) a Perpetrator Certainty of Level 1 and a Event Certainty of Level 1 gives a Combined Certainty of LEVEL 1 (entry in table), (2) a Perpetrator Certainty of Level 3 and a Event Certainty of Level 3 gives a Combined Certainty of LEVEL 3 (entry in table), etc.

Table 3. Combined Certainty Regarding the Perpetrator and Event

Certainty of Event **Certainty of Perpetrator**	Level 1	Level 2	Level 3
Level 1	LEVEL 1	LEVEL 2	LEVEL 3
Level 2	LEVEL 2	LEVEL 2	LEVEL 3

Level 3	LEVEL 3	LEVEL 3	LEVEL 3

In accordance with the above table, cases were ultimately classified in three levels of certainty. Level 1 cases consist of CEH's best-supported cases on which strong arguments could be based. Level 2 cases had a high level of doubt and Level 3 cases were cases which usually could not be confirmed by CEH.

In the same manner as other aspects, when the systematization of this information began, we could see that interviewers did not apply the same meaning to the same levels of certainty. For example, the level "it is public knowledge" for some interviewers meant that nearly the entire community assumed or had a general idea of who was responsible. For other interviewers it meant that all of the people in the community had seen who was responsible (i.e., direct witnesses of the event). Such situations generated significant difficulties when it came time to systematize the information.

Type of Responsibility

CEH authorities agreed to structure the types of responsibility as follows:

Actual perpetrator
Collaborator
Mastermind
Informer

As with the levels of certainty, although these categories were on the original forms, interviewers' interpretations were not uniform. The database team had to devise a strategy to standardize the meaning of these categories, which involved case-by-case revision to determine the correct category.

Using Secondary Sources

The criterion used for the database was that the credibility of the information entered into the database was assessed based on the received testimony. Priority of credibility was given by checking the certainty level described above. Other sources cited by interviewers, such as books, NGO printed reports, etc., helped to corroborate information in the case, but was not used as a source for assessing the certainty of the event if the testimony had minimal conditions for credibility.

Reading cases, revising and classifying information

Every case (both forms and personal accounts) was read and revised using the methodology previously defined and described. As the information analysis progressed, the analysis methodology and the definition of criteria were being perfected through the ongoing revision and reformulation that the process demanded. This process involved frequent discussions by the analysis team to unify and corroborate or modify criteria that arose in the studying the cases.

Discussions served to:

- Apply previously defined typologies.
- Apply the glossary of key words, perpetrator types and victim types.
- Apply basic concepts such as massacre, case, deponent, victim and others.
- Define and perfect the operations strategy and coordination of the database team.

Although the discussion process and unification of criteria was done by the database team, continuous contact was maintained with interviewers and the CEH central team to ensure to the extent possible the uniformity of criteria and orientation between the database and other sectors of CEH.

The methodology used in the database guaranteed a minimum rigorous standard for the systematization of information. It was consistent with the needs and objectives of the CEH and allowed them to make good use of information collected by interviewers in the field. It was also an important resource in subsequent analysis work, in formulating the hypothesis and composing the final report.

Inputting Information

At the end of the analysis, every case passed through the data entry team in which information was already prepared for input into the database with the use of a program developed in advance. As in the analysis process, the process of data entry continually gave rise to the need to set up new or revise existing patterns or standards. For example, how to input data on different people who have the same first name **and** surname? In this case the database team decided to assign a number at the end of the first name to distinguish them: Pedro Coc, Pedro1 Coc, Pedro2 Coc. As to how to handle the data that were blank on the form, the team decided that gaps should be filled with the option "not stated" or "none" so that no blank spaces would remain in the database.

These are two examples of the many decisions of a practical nature that had to be made in response to problems as the data entry process proceeded. This involved frequent discussions among the members of the data entry team to agree on how information would be entered and to establish uniform processing. Here, as elsewhere, the methodology was developed and polished concurrently with the data entry.

Cleansing and Correcting the Database

Once the team had recorded and entered information to the database, they proceeded to *cleanse* it. This involved revising the recorded information to resolve inconsistencies, duplicate information, contradictions and inaccuracies in the data, etc. The logical flow of the process was Cleansing, Detecting and Correcting Technical Errors and Detecting and Correcting Fundamental Errors.

The team had to set basic rules for cleansing. Examples include setting priority criteria for revision (the same names and surnames, the same surnames, the same secondary surnames, etc.) and technical instruments or manual tools to help speed the process (comparison tables, lists of names, places, dates, etc.). As in previous steps, many decisions were made as needs arose. Such decisions were agreed to through discussions among the database team members that were involved in the process (analysts, data entry specialists, programmers, the database assistant and the database director).

Quality Control of Database Results

Database quality control consisted of a database revision process that was the reverse of the systematization process. The team started by revising final statistical results (tables and graphs). They then checked the prior process, sequentially, until they located an error (in its order, revision of the DBF file, revision of entries, revision of the form relating to the case in doubt, revision of the analyses of cases in doubt).

This was done to control the quality, cohesion, and consistency of each process, in the context of the overall results in which errors were more evident. This approach gave the team control over the structural quality of results. This was an important phase since it allowed for large-scale errors to be identified by locating the problem at the specific point where it first arose. It was also in this phase that the scheduling problem was solved, since this method avoided the need to freeze each database team member's work in order to revise it.

Generating Analytical Reports

The database team prepared types of statistical resources for the interviewers who would prepare the final report. One resource was a set of charts, tables and statistical graphs based on the results that were obtained in the process of systematizing information. This work was conducted by the systems analyst, who regularly delivered statistical information to interviewers who requested it.

This process did not restrict the simple delivery of information since the database team had to maintain an ongoing exchange with every interviewer who requested information. The purpose of this exchange was to rationalize and optimize the use of information by informing interviewers in the use of statistics.

The second type of information prepared for interviewers was the compilation of lists of cases made by the database programmer and organized by specific classification criteria. Typical criteria

were dates, places where the events occurred, types of victims, types of violations, perpetrators, themes (list of key words), or any other criterion requested by the interviewer.

Based on lists prepared according to their needs, interviewers could consult physical archives in the database (the case documents) grouped according to one or more specific criteria. For example, they could specifically search for cases that occurred in the Chajul municipality between 1982 and 1983, each of which could have specifications about the type of victims, alleged perpetrators, key words that describe the case, number of victims, etc.

With the list, the interviewers could efficiently review case types and themes. They could select what interested them based on knowing in advance which cases would meet their needs. This resulted in a reduction in time to get results. In addition, this process facilitated an ongoing exchange between the database team and interviewers. Through this exchange, they could precisely define criteria for grouping lists of cases by searching for the best information to meet the interviewers' needs.

The production of this information continued throughout the preparation of the report. It involved more than 200 lists of cases constructed to various selection criteria. This information allowed for the efficient use of qualitative information contained in testimonies collected by CEH. The testimonies could be cited in the report to illustrate analyses that were presented and served as qualitative support for descriptions and hypotheses concerning the work.

Coordination with Other Sectors of CEH

In addition to coordinating work within the database, as database director I had to carry out various activities related to the coordination among the database team and other sectors of the commission. This work was the mechanism by which the coherence and unity of criteria were guaranteed. Work conducted in the database, the main sector for systematizing information, was coordinated with work conducted by the analysis team, who formed hypotheses and composed the final report for interviewers based on information derived from the database.

Central Team Coordination

As database director, I frequently participated in central team discussions on different methodological aspects, both theoretical and technical. The issues we worked on jointly included:

- unifying criteria
- defining indicators
- discussing the process of collecting and researching cases
- discussing CEH methodology
- coordinating interviewer teams, the documentation center, final report coordinator, and director of research

Consulting and Discussing with Interviewers

I had frequently worked with interviewers to guide them in the use of forms, and direct them in collecting testimonies, and inform them of categories used in the database, glossaries, classification tables, criteria applied in information analysis, etc.

This was done through orientation processes organized by CEH during the selection process and influx of new interviewers, through periodic visits to regional offices to bring interviewers up to date on database work, through the evolution of criteria for analysis; and through consultation services in the central office.

Participating in Composing the Final Report

The database team had the responsibility of composing the final report in the following four sections:

Creating the annex of cases presented to CEH

This annex consists of all cases that were presented to CEH and registered in the database (approximately 700) and amounts to about 1,500 pages. For each case, this information includes summary description of the reported case, followed by a list of victims of the violations that occurred.

For presentation purposes, the cases were organized under the following criteria applied in the same order as below:

- Type of violation
- Place where the events occurred
- Perpetrator of the event
- Level of certainty of the event

Drafting the statistical annex

This annex is composed of statistical graphics used in the report and referenced to different chapters.

Composing the chapter on statistical overview

This chapter consists of the statistical analysis and general interpretation of the database results.

Writing the chapter on database methodology

Conclusion and Lessons Learned

Conclusion

Despite the problems that arose during the course of the CEH project, the information system designed and implemented by the CEH database team achieved its major objective, to be the CEH's primary source of information. This information, based on the testimonies collected directly from victims of violence in the Guatemalan population, was the essential resource for analysis and preparation of the final report.

The database team assured a rigorous standard of information handling. The results demonstrated that when the work of the database team is conceived as a part of a structured and integrated process involving all sectors of the commission, problems can readily be solved and a successful outcome achieved.

Lessons Learned

Problem	Solution	Issues
Lack of uniformity in taking testimonies. Testimonies often were different, reflecting diverse backgrounds of personnel. Hence difficult to classify systematically.	Frequent discussion involving all concerned personnel.	Three viewpoints perceived: legal, social science research and political. All viewpoints essential to complete overview. All should be involved in discussions.
Initial lack of recognition of the dominant role of database.	Recognition by coordinators of need to create working database team at the initiation of the project. Coordinators to allocate sufficient physical and financial resources at start of project	Not clear how to make coordinators aware of the critical role of the database when project being defined.
Inefficiency, time delays and reduced effectiveness due to incompleteness of electronic infrastructure when data collection starts.	Get electronic infrastructure running rapidly at start of project; delay data collection until electronic infrastructure is ready.	

Delayed and changing definitions of methodology and clear parameters for collecting information. Collection process adversely affected.	Greater emphasis on preparatory work in defining methodology and parameters.	Until data collection process has produced results, it may not be clear just what factors are to be taken into account in definitions.
Non-comparability of data collected on forms. Changing items on forms makes it impossible to use some information that was already collected or to compare new entries.	Pre-test forms in field tests before finalizing the forms.	
Excessive and conflicting information on forms.	Hold number of forms and information on forms to a minimum. Give priority to information needed for statistical count and case analyses. Don't repeat information on different forms.	
Non-comparable information on forms.	Avoid changing what is on forms and forms during the project	
Interviewers, commissioners, central team not sufficiently familiar with database information to take full advantage in final stages of project.	Seminars in contents of database for all concerned parties.	
Coordination problems.	Develop collection and information analysis simultaneously with continuous feedback and a defined methodology that is flexible enough to respond to the needs of the process as they arise. Apply a well-defined system methodology from the start of the project.	
Non-verifiable data.	Do not reject out-of-hand; use to the extent that the credibility of source is assured. Create criteria for credibility in addition to verifiability. For example, set levels for assessing the information based on interviewer's assessment of sources.	
In retrospect, more effective use might have been made of collected information.	Design forms to collect both qualitative and quantitative information. Through discussions, promote clear understanding of the project's objectives, the criteria with which interviewers should collect information, the use of collection tools, the meaning of questions and the way in which they should be conducted.	