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.