The IASSIST QUARTERLY represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The QUARTERLY reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of IASSIST.

**Information for Authors**
The QUARTERLY is published four times per year. Articles and other information should be typewritten and double-spaced. Each page of the manuscript should be numbered. The first page should contain the article title, author's name, affiliation, address to which correspondence may be sent, and telephone number. Footnotes and bibliographic citations should be consistent in style, preferably following a standard authority such as the University of Chicago press *Manual of Style* or Kate L. Turabian's *Manual for Writers*. Where appropriate, machine-readable data files should be cited with bibliographic citations consistent in style with Dodd, Sue A. "Bibliographic references for numeric social science data files: suggested guidelines". *Journal of the American Society for Information Science* 30(2):77-82, March 1979.

If the contribution is an announcement of a conference, training session, or the like, the text should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event. Book notices and reviews should not exceed two double-spaced pages. Deadlines for submitting articles are six weeks before publication. Manuscripts should be sent in duplicate to the Editor: Walter Pfoevesan, Research Data Library, W.A.C. Bennett Library, Simon Fraser University, Burnaby, B.C., V5A 1S6 CANADA. (604) 291-4349 E-Mail: USERDLIB@SFU.BITNET

Book reviews should be submitted in duplicate to the Book Review Editor: Daniel Tsang, Main Library, University of California P.O. Box 19557, Irvine, California 92713 USA. (714) 856-4978 E-Mail: DTSANG@ORION.CF.UCELEDU

**Title:** Newsletter - International Association for Social Science Information Service and Technology

**ISSN - United States:** 0739-1137 Copyright 1985 by IASSIST. All rights reserved.
Data Development for International Research (DDIR)

DDIR II: Event Data Research

by Richard L. Merrit and Dina A. Zinnes
University of Illinois at Urbana-Champaign

Numerous scholars of international relations have recently sought to improve the empirical quality of their research. They feel that quantitative approaches, properly designed and applied, can significantly enhance our ability to understand international events and interactions among nation-states. One result has been a plethora of analytic techniques that rely on mathematical bases. Global modeling is an example of this direction. Another result is the generation of new data sources. This article focuses on the latter tack: data development for international research.

Growing emphasis on quantitative data has not been without problems. For one thing, some researchers flat out reject their usefulness or validity. Such intransigence obfuscates a central fact: Our growing comprehension of social scientific knowledge is linked inextricably to the computer-based information revolution. Whether we like it or not, whether we comprehend it or not, we cannot avoid their implications for political analysis. Both developments—new analytic techniques and data sources—demand greater sensitivity. Another problem is weaknesses in early data collections. We cannot deny the fact that some important datasets were flawed, just as we cannot ignore criticisms about some analytic methods researchers have used. Such weaknesses have contributed to misunderstandings, skepticism, and even occasional hostility.

This article describes a particular research project undertaken in the field of international and cross-national relations by a community of U.S.-based social scientists. The Data Development for International Research (DDIR) project seeks to maintain, extend, and develop new data banks for the study and analysis of cross-national and international political phenomena. It was the outgrowth of three years of discussion, correspondence, and seminars involving both data collectors and data users. Funding for 1986-89 by the National Science Foundation enabled the project’s first phase (DDIR I) to focus on four tasks: datasets in the areas of national attributes and interstate disputes, data planning, research organization, and international broadening. New NSF funding for 1991-93 permits a second phase (DDIR II) to concentrate on the area of event data.

The article describes how DDIR began, what it has done, and where it is heading. It seeks neither to assay the often sterile debate on the usefulness of quantitative approaches, nor to offer a definitive answer to the question of what analytic techniques and data sources are appropriate for what purposes. Its concern is rather how the DDIR community envisages the status of quantitative research in international and cross-national relations. It summarizes DDIR’s organizational background, philosophical orientation, and goals.

Origins: Need for Quantitative Data

Four trends in the social sciences are particularly relevant for understanding the need to develop data for international research:

- An explosion in the scientific study of national development and processes of interstate interaction has characterized the last five decades.

Questions concerning the relationship between national attributes and the domestic and foreign policy behavior of nations, the evolving structure of the international system, causes and consequences of international crises and war, and the dynamics of interstate interaction both conflictual and cooperative have come under careful and systematic scrutiny. Many of the cherished maxims of international behavior have been shown to be false; and new insights into causes and consequences of national and international processes have been observed.

- The awareness has grown that datasets are crucial within the context of the entire research process, and integral in the continuing feedback relationship between theory and research.

Contradicting the often trite argument that we allow our data to shape our questions, having large datasets that researchers know exist—and which continue to be maintained—opens up the range of research questions and continues development of theory in international relations and comparative politics.

- Funding for data development has been at best sporadic. This has meant an inability to mount a concentrated and coordinated attack on fundamental problems facing the field.
Currently existing datasets are largely the work of a few dedicated researchers scattered throughout the country, who have been dependent on the vicissitudes of changing national funding strategies. There is no guarantee that these data collections will be continued and certainly no clear opportunity for extending and further developing them in response to the evolving needs of the research community. Furthermore, while data collectors are generally aware of one another, there is no overarching mechanism to integrate and compare their results. This has led to unfortunate duplications of effort, differences in definitions, and differences in usage of sources.

- The data movement of the past several decades has enhanced the methodological expertise for the extraction of data from public sources, development of indicators for basic concepts, and quality control through reliability checks. This, together with the extensive technological advances of recent years in computer technology, makes feasible the future development of considerably more valid and reliable datasets.

These facts—the research record; recognition of the need for systematic datasets; the currently scattered, *ad hoc* nature of data collection activities; and the available methodological/technological expertise—point to the desirability of a large-scale, integrated effort that can contain, extend, and further develop the data resources available to the research community of international relations scholars.

Such perspectives on the state of the art in international and cross-national relations generated an interest in taking action to improve the field's quality. A series of informal meetings, piggybacked on to professional conferences, and workshops at the University of Illinois at Urbana-Champaign and elsewhere led to a remarkable degree of consensus among several dozen researchers. These meetings and workshops eventually honed in on a basic decision. If those interested in using quantitative data did not take action, the participants argued, then opportunities to have such data would atrophy. Accordingly, an effort to organize the relevant community of scholars was warranted and, indeed, long overdue. The researchers then focused on the overall strategy that such an organization, Data Development for International Research, should pursue: Should DDIR serve solely as an interest group, or should it encourage and seek funding for research activities? And, if the latter, which kinds of relevant research should have DDIR's initial attention?

The organizational task was easily resolved provided that some colleagues were willing to devote some of their time and energy. The point of departure for DDIR was in a sense the National Election Study project. As a large-scale, long-term data collection project for the enhancement of social science research, the NES clearly stands as a model. In another sense, however, important differences distinguish, on the one hand, the theoretical framework, goals, and structure of the NES and, on the other, the needs of the research community studying international and cross-national phenomena. The two research communities are diverse in the questions they ask, degree of consensus regarding fundamental methodological issues, and sheer number of researchers currently relying on the data collections.

This diversity suggested the need for more decentralization in the data-collection efforts and communications framework than has been needed in the NES. DDIR thus supports not a single, massive project, but rather individual researchers at different universities carrying out separate—though clearly related—projects. The diversity should be seen as a major strength of DDIR I; and it is this orientation that guides DDIR II since it also points to a multiplicity of research agendas. While the pitfalls of decentralization are apparent, these dangers do not obviate possibilities for successful coordination and integration.

The task of choosing areas for research focus proved to be more difficult simply because the potential areas are many and the competition for needed funding and other scarce resources is even greater. Some of the principle supporters, none of them with any immediate claim for DDIR-related resources, distributed questionnaires and carried out other research to ascertain how members of the potential community evaluated data priorities (McGowan et al., 1988). A solicitation of research ideas, further consultation in meetings and workshops, and much telephoning eventually produced substantial if not complete agreement on a particular strategy.

The informal consensus saw three activities: First of all, DDIR would seek funding to carry out a discrete number of projects in two research areas, national attributes and interstate war, that its growing number of members considered most relevant and likely to be carried out. Second, those administering DDIR, in conjunction with an advisory committee, would assess the prospects for similar research projects in two other areas, event data and international political economy (IPE) data, which DDIR might wish to initiate later. Third, DDIR would also try to improve communications among scientists interested in quantitative research in international and cross-national relations. This meant on the one hand setting up a regular newsletter, *DDIR-Update*, and, on the other, scheduling at professional conferences both research sessions and organizational meetings.

**DDIR I: National Attributes and Interstate War**

DDIR’s first task, aimed at improving the quality of data
on national attributes and interstate war, proceeded from a rich background. Significant international and cross-national data collections were developed well before World War II (Merritt, 1990). Not until the late 1950s and early 1960s, however, did the large-scale data movement begin. As part of the general behavioral movement in political science away from assessments based on intuition or folk wisdom and toward more rigorous, systematic analyses, scholars became sensitive to their need for adequate data bases to study key questions. Is inequality in the world at large becoming more or less intense? Do alliance configurations and power distributions enhance or decrease the probability of war? Do internal domestic problems have specific effects on foreign policy behavior? To move beyond simple, impressionistic answers to such questions, it was necessary to begin collecting data on the attributes of nations and events characterizing their interactive behavior.

It is possible, in retrospect, to trace three broad data collection efforts that sought to provide the evidence necessary to facilitate the scientific study of international processes: data collections that focused on the quantitative and qualitative characteristics of (1) national attributes, (2) major conflicts and wars, and (3) interactive events within and between nations. Intriguingly, these projects all began within a year or two of one another and spread rapidly across the scientific geography of the United States and even abroad.

**DDIR I: National Attributes Dimension**

**Historical background.**

In the late 1950s and early 1960s Karl W. Deutsch at Yale University was arguing for the use of data to confirm or disconfirm hypotheses about international and cross-national politics. He had demonstrated the practicability of the search for such data and their analytic value in his studies of *Nationalism and Social Communication* (Deutsch, 1953) and *Political Community and the North Atlantic Area* (Deutsch et al., 1957), and in a series of articles (most notably Deutsch, 1960, 1961) that showed how important questions were not being addressed because of the absence of valid, comparable indicators based on reliable (that is, replicable), impersonal, and quantitative data.

The year 1963 was a watershed for these innovative ideas. At the Yale Data Conference (Merritt and Rokkan, 1966) held in September international scientific researchers gathered to discuss systematic means to compare nation-states, outline organizational efforts to further such research, and learn at first hand of three major data-collection activities reaching fruition in the United States.

- Russett et al. (1964): Yale Political Data Program.

With financial support from the National Science Foundation, Deutsch and Harold Lasswell created the Yale Political Data Program, which in 1962, under the direction of Bruce M. Russett, had begun the cross-national collection of political, social, and economic data (see Deutsch et al., 1966). Its immediate result was the *World Handbook of Political and Social Indicators*, known as "World Handbook I"; and in later years World Handbooks II and III appeared (Taylor and Hudson, 1972; Taylor and Jodice, 1983).

- Banks and Textor (1963): Cross-Polity Survey. Arthur S. Banks, a political scientist, and Robert B. Textor, an anthropologist, had combined forces to classify 115 polities according to 57 sets of carefully operationalized criteria.

- Rummel (1964): Dimensionality of Nations. At Northwestern University, initially as a component of Harold Guetzkow’s Inter-Nation Simulation (INS), Rudolph J. Rummel had compiled data characterizing nation-states (see Rummel, 1979). (He also—and we shall return to this later—systematically searched the *New York Times Index* and other sources to record domestic-political and foreign-conflict events.)

Years subsequent to this burst of creativity saw three important developments. The first was the growing use of data already collected to examine theoretically interesting propositions. Second, the efforts of the 1960s were continued and expanded during the 1970s. Particularly important here were (1) Taylor et al.’s *World Handbooks II and III*, (2) Gurr’s (1974, 1978) research on polities and segmental groups, and (3) the national characteristics assembled by the Correlates of War project directed by J. David Singer (see Singer and Small, 1972; Small and Singer, 1982). Third, researchers became more sophisticated in both measurement and data-collection techniques. They also met increasingly frequently to discuss their work, as well as to exchange preprints and sometimes tables of data. A “community” of quantitative international relations (QIP) scientists was emerging.

These years of an emerging QIP community were heady ones for scientific advances. Data-generators and users did not simply rest on the intellectual platforms given them in the 1950s by Karl Deutsch, Harold Guetzkow, Harold Lasswell, and others, but used them to push understanding forward. The analyses themselves were not always elegant, but nevertheless established clearly at least two points. First, “the activities of nations,” in Rummel’s (1966:205) words, “are highly patterned behavior,” and, second, it is both possible and intellectually profitable to establish data banks to help ascertain what these behaviors are, how they are structured, and what impact this structured behavior has on such vital
issues as war and peace.

By the mid-1980s the need for aggregate data was growing but so was the reality that past data sources were aging and not being kept alive. It is thus not surprising that research scientists saw in DDIR the opportunity to resuscitate and significantly improve this field. A coordinated effort by both data generators and data users assembled a research design on national attribute data and, through DDIR, submitted it to the National Science Foundation. NSF support enabled DDIR I to carry out the following projects:


DDIR I–2. Correlates of War national capabilities dataset. Principal investigator: Ted Robert Gurr, then University of Colorado, Boulder, and now University of Maryland at College Park. Cooperation with J. David Singer, University of Michigan, to produce an integrated dataset, 1816-1988, on the COW project’s national capability variables (population, urban population, iron/steel production, energy consumption, military expenditures, and military personnel) plus government revenues and expenditures for all states at one time members of the central-state system and, insofar as possible, peripheral systems.

DDIR I–3. Correlates of War dyadic relationships dataset. Principal investigator: J. David Singer, University of Michigan. Collaboration with Michael Wallace, then University of British Columbia and now Simon Fraser University, to update and revalidate the COW dyadic relationship dataset, 1816-1988, on shared membership in foreign alliances, diplomatic representation, and shared membership in international bodies.

DDIR I–4. Political structures dataset. Principal investigator: Ted Robert Gurr, then University of Colorado, Boulder, and now University of Maryland at College Park. For all international-system members, 1816-1985, development of a complete and updated dataset on régime characteristics, collaboration with Mark Lichbach, University of Illinois at Chicago, to transform into time-series data coding on each authority dimension.


These projects are now for the most part complete, reports included in DDIR’s newsletter, DDIR-Update, and the datasets sent to the Inter-University Consortium for Political and Social Research (ICPSR) in Ann Arbor, Michigan, for access to the scientific community.

DDIR I: International Conflict Dimension

Historical background.

Just as the national-development datasets had their predecessors in the many efforts of individual scholars, government agencies, and international bodies, so, too, current efforts to generate data on international conflicts can look back on a tradition of earlier projects. Though flawed, these early studies pioneered the path that modern researchers, with their richer sources and computer-based operations, continue to treat. Among these are four deserving particular attention:

- Woods and Baltzley (1915): Is War Diminishing? Frederick Adams Woods, an eminent biologist, and a young political scientist, Alexander Baltzley, provided a list of wars and their participants for most of the major European states for 1450-1900 (and back to 1100 in the cases of England and France). Chapters on each of eleven countries indicated the years of initiation and termination of national war, and hence the duration of war, and statistical graphs showed national percentages of interstate, imperial, and civil wars.

- Sorokin (1937): Social and Cultural Dynamics. Pitirim A. Sorokin identified "almost all the known wars" for the major European states from antiquity to 1925—including internal disturbances as well as interstate, civil, and imperial wars. He gathered the dates of initiation and termination, the war's duration for each major state, and estimates of the average army size, percentage of casualties, and total number of casualties for each state.

- Wright (1942): A Study of War. For 1480-1936 Quincy Wright listed balance of power, civil, imperial, and defensive wars involving each major or minor state. The dataset includes for each war its initiation and termination dates, identity of participants, their individual day of entry, and number of important battles. Wright also assembled data on the frequency and types of battles, casualties, and internal systemic disturbances.

- Richardson (1960): Statistics of Deadly Quarrels. Lewis Fry Richardson's compilation of conflict data includes all deadly quarrels—imperial wars, civil wars, and other forms of domestic conflict—between 1820 and 1949 that caused death of
humans. The war is the unit of analysis, and wars are organized by magnitude and then chronologically within magnitudes. The data on each war include the magnitude, dates of initiation and termination, participants, identity of the initiator, and ostensible cause or issue at stake.

Each of these data collections, of course, had problems. That by Woods and Baltzly was remarkably incomplete, especially by modern standards. Sorokin gave no explicit operational criteria for interstate, civil, and imperial wars. Wright's use of legalistic criteria for including and excluding wars was questionable. Richardson's research raises serious questions about the reliability of the data and validity of the categorical codings. But each was a significant beginning. Each, in a sense, set the groundwork for the subsequent ones (although the lateness of Richardson's publication makes it easy to ignore the fact that his research efforts were contemporaneous with those of Wright).

More to the point, however, each has been superseded by modern datasets that have not only initiated major data collections but also spawned the use of those datasets to conduct significant, empirical research on questions of war and peace. Some recent international conflict datasets are:

Singer: Correlates of War (COW). J. David Singer's intellectual focus was on the correlates of war, those factors that seemed to covary and thus be associated with the occurrence, duration, and magnitude of wars. His COW project, which began formally in 1963 under the auspices of the National Science Foundation, initially concentrated on obtaining information on the attributes of the international system that theorists had argued were the "cause" of war. COW researchers systematically culled historical texts to obtain as complete a listing as possible of all wars since 1815, together with major identifying characteristics, such as the number of participants, battle deaths, and durations (Singer and Small, 1972; Small and Singer, 1983). Subsequent data collections expanded COW's horizons—to such independent variables as population, iron and steel production, energy consumption, military expenditures, and military personnel, and, with several colleagues, to forms of international behavior, such as crises.

Siverson and Tennefoss (1982, 1984): Interstate Conflict. Unaware of COW's shift in focus and data gathering, Randolph Siverson and Michael Tennefoss independently developed a dataset on major international crises since 1815. They classified into three types the implicit/explicit level of war: threats to use force, unilateral uses of force, and reciprocated military interactions.

Levy (1983): Great Power War. Jack Levy's dataset overlaps that of COW but extends the latter back to 1495 for the great powers. Concentrating on interstate, great-power wars (excluding civil and imperial wars) which had more than 1,000 battle deaths, it provides data on their magnitude, severity, and intensity.

Overlapping but not identical with these efforts were several other projects. Robert Butterworth (1976, 1980), Michael Brecher and Jonathan Wilkenfeld (1982), and Hayward R. Alker, Jr. and Frank L. Sherman (1982; cf. Sherman, in progress) collected data on significant attributes of international crises since World War II. Along a somewhat different but related dimension, Frederic S. Pearson (e.g., 1974) collected data on international interventions in the post-World War II period.

As was the case along the national development dimension, participants in the DDIR enterprise saw a clear need to update and expand data on the international conflict dimension. Major insights had come from the analyses based on the earlier datasets. But, as was true with respect to the national development dimension, inadequate coordination had led to duplication and incompatibility. Members of the emergent DDIR community responded to the need by preparing research proposals that eventually formed a component part of DDIR's main task: research to develop adequate measures of international conflict and systematically to collect relevant data. NSF funding supported the following projects:

DDIR I-6. Great-power war dataset. Principal investigator: Jack S. Levy, then University of Texas and now Rutgers University. Collaboration with T. Clifton Morgan to revalidate and fill in missing data in Levy's dataset on participation, casualties, and initiation/termination dates for all wars among great powers, 1495-1815.


DDIR I-8. Interstate war catalog. Principal investigator: Claudio Cioffi-Revilla, then University of Illinois at Urbana-Champaign and now University of Colorado, Boulder. Completion of a master catalog comparing (with reliability indicators) existing datasets on interstate wars (see Cioffi-Revilla, 1990).

DDIR I-9. Interstate war dataset. Principal
investigator: J. David Singer, University of Michigan. Updating for 1980-88 the COW dataset on the initiation of interstate wars, participation (in nation-months), and casualties; defining and coding additional variables for 1816-1988 on interventions by third parties, war phases, monthly casualty rates, and characteristics of war terminations.

**DDIR I-10. Interventions dataset.** Principal investigator: Frederic S. Pearson, University of Missouri-St. Louis. Filling in the dataset on unilateral, multilateral, and international organization interventions for 1816-1988 on interventions by third parties, war phases, monthly casualty rates, and characteristics of war terminations.

These projects are now for the most part complete, reports on most included in the DDIR’s newsletter, *DDIR-Update* (and one, Cioffi-Revilla’s [1990] interstate war catalog, published), and datasets sent to the ICPSR for access to the scientific community.

**DDIR II: Event Data**
A second, and equally important, DDIR I activity was planning future data-gathering activities on two dimensions: interstate events and international political economy (IPE). For the field of international relations to keep up with and anticipate data needs deriving from new theoretic growth requires imaginative and sustained attention to such matters as conceptualization, indicator validity, and collection procedures. DDIR’s organizational goal was to hold separate sets of conferences on the two dimensions, at which active scholars would discuss needs, priorities, and procedures. The long-term hope was that conferences would produce specific research programs which could be developed for future funding.

Accordingly, with respect to the event-data dimension, planning conferences took place in May 1987 in Columbus, Ohio (Hermann, 1987), November 1987 in Cambridge, Massachusetts (Alker, 1988), and March 1990 in Chicago, Illinois. What emerged was a two-year proposal to the National Science Foundation that included researchers at seven different academic institutions who will carry out distinct but generally integrated research projects. NSF funding, awarded in January 1991, permits the realization of DDIR II. And, as in the past, the Merriam Laboratory for Analytic Political Research, located at the University of Illinois at Urbana-Champaign, serves as DDIR II’s administrative umbrella.

*From Political Arithmetic to Event-Data Research*
Narratively oriented diplomatic historians generally view the course of international relations as a series of events—démarches, protests, treaties, crises, wars, conferences, and the like. An event in this sense is an occurrence that stands out against the gray background of everyday living. In principle an event is a discrete unit of action, with its own beginning and ending points. In practice we often view events as nested sequences of yet smaller events. Thus an historian may view the Franco-Prussian war of 1870-71 in the light of *inter alia* Bismarck’s wars against Denmark and Austria, the Ems dispatch, declaration of war, military hostilities, siege of Paris, conclusion of a peace treaty, and such consequences as indemnification, territorial transfer, and formation of the German empire; and each of these in turn comprises a congeries of lesser events. Is there another, more systematic, way to look at international events?

Analysts have devised various ways to study the events they define as important in our individual and social lives. Indeed, modern statistics finds one of its main roots in the “political arithmetic” used in the 17th century by John Graunt and William Petty to examine mortality tables. Sickness and death are individual events. And yet knowledge of how many of a society’s members suffer from particular illnesses and die of particular causes tells us something about the society itself, and enables us to predict the need for medical services and the proper price for insurance. Similar considerations led Petty and other social philosophers to argue for the collection of criminal statistics (Walker, 1971; Collmann, 1973), and the occasional monarch or cabinet minister undertook a survey from time to time.

Such studies had individuals as their unit of analysis. Not until the late 19th century, with the flowering of labor unions throughout the industrialized West, did government agencies begin to gather data on social events. The target was the strike or lock-out, industrial disputes leading to stoppage of work in some firm or branch of industry. Nor is it surprising, given the general attitude then prevailing toward labor unions as a whole, that data on strikes took on the character of criminal statistics (International Labour Office, 1926). In the United States, the Department of Labor’s Bureau of Labor Statistics combed newspapers and other sources to identify work stoppages, sent questionnaires to key participants to ascertain the dimensions of these events, and reported on the number of strikes, workers involved, duration, days idle, and so forth (see U.S. Department of Labor, 1976: 195-202).

The 1930s saw three major social scientific efforts to collect data on social events. The first, described earlier, focused on aspects of wars. A second was Harold D. Lasswell’s (1936) intentional/instrumental view of politics in terms of “who gets what, when, and how.” The third was Lasswell and Blumenstock’s (1939) study of social unrest and world revolutionary propaganda in Chicago from 1919 to 1934. They recorded the number
and characteristics of communist-sponsored meetings, demonstrations, parades, and other social gatherings; strikes; group and individual complaints about violations of civil rights; and evictions, foreclosures, and arrests of "radicals." Laswell and Blumenstock concluded among other things that communist propaganda was most successful during tough economic times and when it incorporated American symbolism instead of harping on Soviet accomplishments. But at the same time, by giving hard strapped citizens an outlet to vent their frustrations and business a scapegoat to blame for the country's economic woes, communist agitation worked ultimately to deflect any truly revolutionary spirit and hence to strengthen the capitalist system.

Two decades later scholars of international relations renewed their interest in systematically studying events. One starting point was growing concern with processes of political development and the place of violence in them. Cross-national studies using data for a single year (that is, synchronic) aimed at discovering the correlates of unrest and violence; longitudinal (diachronic) studies traced patterns over time among some more limited set of countries. The nation-state was the unit of analysis. Researchers tabulated such events occurring within a state's boundaries as demonstrations, coups d'état, and revolutions.

Another starting point for event analysis centered on foreign-policy decision-making. Scientists conducting simulations of international processes—whether using people only, computers only, or some combination of the two—discovered they needed hard data both to feed into the simulation itself and/or to check the realism of their findings. Eventually the focus shifted from the nation-state as the unit of analysis to interactions between pairs of nation-states: ongoing processes such as trade and diplomatic exchanges as well as more or less distinct occurrences such as a threat or militarized intervention. From there it was a short step to taking seriously the new emphasis on the international system qua system (Kaplan, 1957) and tabulating the attributes of that system as a whole and the events taking place within it.

Still a third and doubtless the most important starting point was a growing concern with international crises and war. In the nuclear age, the possibility of war cannot be taken lightly. If analysts had had the correct tools, scientists asked, could they have recognized the probable outcome of the sequence of events in mid-1914 or in the 1930s early enough to have prevented the outbreak of war? Is there some means to ascertain when international crises are reaching the boiling point? What steps can governments take to de-escalate crises? Answers to such questions seemed to require detailed information on the course of events occurring in the global arena.

Progress in Developing Event Datasets

Initial efforts to assemble data about the events of nation-states electrified the discipline of international politics. They were, broadly speaking, of two types. First, global studies defined events of interest, specified coding rules, and, in such universal sources as the New York Times or Facts On File, coded every single occurrence of such events. (Regional studies pursued the same procedures but focused primarily on regional issues and sources.) Second, event-specific studies proceeded from the opposite direction. That is, they identified critical events of interest, such as the Suez crisis of 1956, and searched a wide variety of newspapers and historical treasuries to describe, in detail, their characteristics and the chronology that preceded the key event.

Not only did these event studies set the standards that subsequent researchers would use and contend with, but they resulted in empirical studies that opened scientists' minds to new modes of research. As the data movement captured the field of international politics a series of datasets were compiled by different researchers. Limitations of one dataset for a new research question being posed led to the development of new datasets. A glance at the history of this evolution suggests at least seven major compilations.

- **Dimensionality of Nations (DON).** Rummel, as we saw earlier, generated one of the original collections of national-attribute data. He also focused his research on interactions within and among states. He used five sources to assemble data for 1955-57 on the domestic-politics and foreign-conflict behavior of 77 nation-states (Rummel, 1964, 1967, 1972). Among other things DON tabulated the presence or absence of guerrilla warfare, number of assassinations, and seven other domestic conflict events. Rummel's thirteen foreign-conflict variables were, besides the presence or absence of military action, the number of anti-foreign demonstrations, negative sanctions, protests, countries with which diplomatic relations were severed, ambassadors expelled or recalled, diplomatic officials of less than ambassador's rank expelled or recalled, threats, wars, troop movements, mobilizations, accusations, and people killed in all forms of foreign conflict behavior.

- **World Event Interaction Survey (WEIS).** At roughly the same time Charles A. McClelland initiated at the University of Southern California an unrelated data enterprise. This collection focused on the events, or interactions, that took place over time between pairs of countries (and in this sense was not dissimilar to the foreign conflict events coded by Rummel). WEIS consisted of a very detailed set of coding categories (63 mutually exclusive and exhaustive categories) designed to capture the type of
hostile or cooperative action that one country directed toward another, but not the intensity of hostile or cooperative behavior. Relying on reports published in the New York Times, McClelland and his colleagues (1971) recorded such acts in terms of initiator, target, type of act, and date of occurrence, covering the period after 1946. The extensive historical chronicle of interstate interactions that resulted made it possible to observe patterns in the activities of states and to determine the degree to which special patterns preceded major crises or wars. That such a monitoring system might facilitate forecasting of the onset of future crises was an integral part of McClelland’s overall research design.

- Conflict and Peace Databank (COPDAB). Edward E. Azar’s particular interest in recurring Middle Eastern conflicts led him to develop a new and somewhat differently focused event dataset (see Azar, 1970, 1980a, 1980b; Azar and Sloan, 1975; Azar and Havener, 1976; Azar and Lerner, 1981). Building on earlier work by Robert C. North, Lincoln E. Moses, and their collaborators (Moses et al., 1967; Choucri and North, 1975), Azar defined events as occurrences between or within nation-states that were sufficiently distinct from the constant flow of “transactions” (such as trade or mail flow) to stand out as reportable or newsworthy against this background. The coding categories were very similar to those of McClelland (see Howell, 1983; McClelland, 1983; Vincent, 1983), but the sources Azar used for coding the events went far beyond the New York Times to include a variety of international as well as local reporting sources.

- Comparative Research on the Events of Nations (CREON). Yet another important event dataset, developed by Charles F. Hermann and his colleagues (1973) primarily at The Ohio State University, sought to examine the correlates of foreign policy behavior. It focused on events that characterized different foreign policy positions of states. The coding categories were therefore somewhat different from those developed for WEIS or COPDAB. Further, since the central question concerned the relationship between certain attributes of states and types of foreign policies, extensive and costly time-series were not necessary. CREON rather provided snapshots at various points in time of the foreign policy behaviors of states.

- World Handbook of Political and Social Indicators (1983). In the late 1970s, Charles Lewis Taylor and David A. Jodice (1983) significantly expanded the data-gathering approaches originally developed, as noted above, at the Yale Political Data Program by Russett et al. (1964) and Taylor and Hudson (1972). World Handbook III provided daily event data for domestic political events only, for 136 nation-states for 1948-77. The event categories include political unrest (e.g., protests, riots), state coercive behavior (e.g., government sanctions, political executions), and governmental change (e.g., elections, executive transfers). The number of deaths from events involving domestic violence is also recorded, and additional codings for event duration, intensity, scale, and impact are included for events from 1968. World Handbook III also separately compiles for each state statistical indicators of political, economic, and social change, thus helping to define the broader context within which coded events occur.

These five event datasets, despite their apparent differences, share two important similarities. First, the definition and coding of an event are in terms of actors (national or subnational) and actions; and events are classified into a set of predetermined categories which provide descriptors of the event. Second, they pursue global coverage, that is, they are concerned with the entire international system.

These were not, of course, the only event datasets to emerge after the 1950s. For the Political Instability Data Bank, Ivo K. and Rosalind L. Feierabend (1966a, 1966b) codified 28 types of events occurring for 1955-61 in 84 countries. In his Comparative Study of Civil Strife, Ted Robert Gurr searched standard sources for the occurrence in 1961-68 of civil violence in 119 polities; this data collection, which he analyzed in various forms and made available to the scholarly community, provided the empirical basis for Gurr’s important, prize-winning theoretic work, Why Men Rebel (Gurr, 1970).

Two other datasets are event specific and thus differ from the others in significant ways. In effect, two levels of “events” characterize these datasets. One is the identification of a key event, for example, an international crisis. The other is a minute examination in considerable detail of all preceding events, where “event” in this second instance is considerably more fine grained.

- Behavioral Correlates of War (BCOW). The BCOW dataset, developed by Russell J. Leng as an offshoot of the Correlates of War project, starts with Leng’s earlier data on militarized interstate disputes (MID)—defined in terms of disputes in which parties on both sides threaten, display, or use military force—but focuses only on a subset of more intense disputes, called militarized crises (Leng and Singer, 1988). It then provides for the time period prior to each militarized crisis a fine-screened description of all events. Unique features of the BCOW coding scheme (beyond the core coding of who does or says what to whom and when) include: location of each event;
duration and variations in intensity of multi-day events; assignment of physical events to one of 103 categories of military, diplomatic, economic, or unofficial behaviors; and detailed analysis of sequential verbal interactions (allowing identification of bargaining strategies). This fine-grained coding of verbal actions allows for a detailed analysis of interstate bargaining and the development of an "hierarchical choice tree."

- SHERFACS. Using criteria to select and merge conflict cases from the FACS dataset (Farris, Alker, Carley, and Sherman, 1980) and nearly 40 other studies, Frank L. Sherman’s SHERFACS produced a combined file of 730 international disputes and 980 domestic quarrels that provide data on, among other things, the identification of conflict phases, means of referrals to management agents, and nature of actions taken by all parties (see Alker and Sherman, 1982; Sherman, 1987a, 1987b, in progress). Sherman then developed a phase structure for domestic quarrels similar to the CASCON structure for international conflicts.

Some related datasets were mentioned earlier: Butterworth (1976), Brecher and Wilkenfeld (1982), and Pearson (1974). Then, too, empirical studies of conflict management, such as SHERFACS, have a rich tradition: Ernst B. Haas’s (1968) disputes referred to the United Nations for management, Joseph S. Nye’s (1971) added conflicts referring similarly to regional international organizations, the joint effort by Haas, Robert Butterworth, and Nye (1972) added to the existing set three new types of conflicts—interstate disputes in which some kind of international organization, e.g., the United Nations Security Council, sought involvement; civil strife in which one side of the dispute enjoyed the support of another government; and “non-managed” interstate conflicts in which fatalities occurred—and the CASCON phase structure developed by Bloomfield and Leiss (1969).

These event-data projects saw enormous use by scholars. This was particularly the case with Azar’s COPDAB, which continued until 1979 to collect data, and, like other event-data collections, was made generally available to users. But these projects—and hence the fundamental idea underlying event datasets—also came under fire by critics, both friendly and hostile. Complaints ranged from the usefulness of particular sources, such as the New York Times, to the modes of categorizing the data. The level of hostility had multiple effects. It diminished funding and shifted intellectual concerns. It discouraged previous and emerging event-data researchers from either undertaking new collections or updating the old ones. The scientific progress of the 1960s soon began to languish. But, at the same time, challenging the past value and uses of event data encouraged researchers to spend time thinking through various dimensions of previous projects, exploring new ideas, and, particularly, adapting their research plans to take advantage of modern computational capabilities.

DDIR II: Developing New Event-Data Research
DDIR’s three event-data conferences sought first of all to assess the state of the art, then to review new data priorities, and finally to develop an effective research strategy. Several considerations shaped a decision to pursue a mixed strategy: the need to (1) generate a rich and general, core dataset; (2) improve the capabilities of key specialized event datasets that already exist; (3) enhance software so as to minimize the time and cost of expanding datasets in the future; and (4) explore the possibilities for new styles of event-data research.

Enhancing existing and generating new event datasets.
If we are to enhance the quality and quantity of some existing datasets, which ones should they be? Our survey of the literature (McGowan et al., 1988) together with a study of each event dataset’s time-span and comprehensibility across a wide range of theoretically interesting issues strongly suggested a central focus on the COPDAB file. Not the least reason for this is the fact that, of the five global event datasets—DON, WEIS, COPDAB, CREON, and World Handbook III—COPDAB best met the combined criteria of past scientific usage, availability over a long time series, and attention to a broad range of new styles of computer-aided, event-data research (Starr, 1987). Other factors included COPDAB’s compatibility with case-oriented datasets (most notably BCOW and SHERFACS), the needs of those initiating regional event datasets, COPDAB’s appropriateness for testing new software, and, by no means least significantly, the fact that the Center for International Development and Conflict Management (CIDCM) at the University of Maryland at College Park was planning to update and expand the COPDAB dataset.

Thus the Global Event-Data System (GEDS) project at Maryland became the natural focal point for organizing DDIR II’s core data-generation part. The CIDCM’s research team will establish GEDS for computer-assisted identification, abstracting, and coding of daily international and domestic events, as reported primarily in comprehensive, on-line news sources such as the Reuters news service. GEDS thus aims at developing a core event-data stream from 1979 forward. It will include:

- the actions vis-à-vis each other of (1) nation-states, (2) major nonstate communities, and (3) international organizations,
- detailed event summaries and coding, including
direct quotations and cross-referencing, and

- information allowing users to access those full-text source articles which are available on-line.

GEDS software will permit partially automated, continuous updating after 1990 of the core event-data stream. In the discussion that follows, the term GEDS refers to the event-data stream generated by using Maryland's computer-assisted coding procedures on on-line news sources. Each of the projects described below produces a specialized dataset based on GEDS.

**DDIR II–1. University of Maryland: Updating and Extending Existing Datasets.** As part of its larger GEDS effort, the Maryland team—John L. Davies, Ted Robert Gurr, and Chad K. McDaniel—will update to 1990+ the existing COPDAB dataset, and incorporate updated WEIS and, as they become available, World Handbook III (and BCOW and SHERFACS) event data. The updated dataset will be compatible with each of these previously-coded datasets, but expanded to include new foci (e.g., inclusion of nonstate actors) and new sources made available through computer-assisted coding.

**DDIR II–2. American University: Foreign Policy Behaviors of Southeast Asian States (SAS).** Llewellyn D. Howell will use the GEDS computer-assisted procedures on regional sources to produce a data bank on 10 Southeast Asia states. The SAS event-data stream, to be added to the Maryland core event-data stream, will thus enrich the latter and provide a check on the comparability of global sources vs. regional sources.

**DDIR II–3. University of Kansas: Kansas Event-Data Sources (KEDS) for Central Europe and the Middle East.** Philip A. Schrot, Ronald A. Francisco, and Deborah J. Gerner have two tasks. First, they will extend their existing software for automated coding. Using the GEDS files as inputs, the current software automatically generates WEIS-coded data. Resources permitting, the software can be expanded to produce COPDAB-coded data. Second, the Kansas team will assemble a high-density, international, event dataset for Central Europe and the Middle East. It uses specialized journals and government publications around the world to increase regional coverage without the time and expense involved in working with regional journalistic sources such as newspapers. Like Howell's SAS project, the use of regional sources will provide the basis for comparing alternative, global vs. regional sources of events.

**DDIR II–4. Middlebury College: Behavioral Correlates of War (BCOW).** For 40-55 militarized crises occurring in 1979-90, and starting with the core data provided by GEDS, Russell J. Leng will apply BCOW data-collection procedures to produce a fine-screened dataset. The BCOW coding manual specifies as many as 103 descriptors of each action (such as alert, mobilization, or evacuation) that could take place during a militarized crisis. Each such event action is categorized according to the date of occurrence, actor, target, location, whether the actor was acting unilaterally or with another state, and "tempo" of the action.

**DDIR II–5. Miami University: Nonstate Actors in Interstate Conflicts (SHERFACS).** Frank L. Sherman at Miami University of Ohio will enhance and bring up to date the SHERFACS dataset, which comprises fine-screened accounts of several kinds of episodic conflict situations. Inclusion is global, but limited to international conflicts and domestic quarrels, especially those involving collective management (e.g., UN mediation) and nonstate actors. The expanded event summaries generated by GEDS will increase the number of international conflicts and domestic quarrels that will be coded using the SHERFACS template. And, like the BCOW dataset, the SHERFACS dataset will augment the analytic capabilities inherent in the expanded COPDAB dataset to be developed by CIDCM at the University of Maryland.

**DDIR II–6. Massachusetts Institute of Technology: Data Development for Interpretive Analysis.** Hayward R. Alker, Jr., at MIT, will develop methods for the interpretive analysis of detailed event summaries by adding narrative depth and variety of interpretive perspectives for specific conflict episodes in the GEDS dataset. The three data components to be studied are (1) explicitly coded WEIS/COPDAB/BCOW/SHERFACS event data, (2) humanly constructed narrative summaries of each event, and (3) quotations attributed to principal actors/interactors of the event being described. In addition, original and secondary source stories will be made conveniently accessible, possibly as part of each record, for the purposes of detailed textual and interpretive analysis of both quantitative and qualitative, political data.

These various data-collecting activities can significantly improve the quality of research in the field of quantitative and textual international politics. First, they will bring up to date and expand the more important event datasets identified by publications and by quantitative and textual scientists. Second, they will provide procedures for routinizing future such event-data collections. This will sharply reduce the need to turn to funding.
agencies every five years or so in the search for new support to update the datasets. Third, they aim at achieving an integrated event dataset. Interaction among the principal investigators through DDIR’s agis can ensure that interchangeable datasets are in the public domain. Fourth, the coordinative thrust nevertheless permits maximum flexibility among these principal investigators to carry out their individual research strategies.

Software Developments to Aid Data Collection and Analysis.
Recognizing the need for a core data-collection effort such as GEDS was only one step. Researchers in recent years also began to appreciate the important role that computerized methods could play. With major international news sources, such as Reuters, Associated Press, and United Press International, as well as local news reports (as translated, for instance, by the FBIS Reports) either now or soon to be accessible on-line, the retrieval of source stories begs for automation. Moreover, the enormously expanded storage capacity, processing speed, and programming flexibility at the microcomputer level now makes it possible to develop an event-coding system which sacrifices neither the comprehensiveness of global coding efforts nor the depth and diversity of coverage of the episodic coding projects.

DDIR II proceeds from the conviction that the development of computerized methods for the collection of data is not only a desirable but a necessary innovation. It includes several projects in this area:

DDIR II–7. University of Maryland: Computer-Assisted and Partially-Automated Coding in GEDS. With a grant from DDIR and backing from their institution, the Maryland team has developed and tested a preliminary version of software for computer-assisted entry, coding, and editing of Reuters on-line source stories to produce GEDS event records. As a significant product of its software development, the team will set in place at CIDCM a process for continuously coding GEDS records.

DDIR II–8. University of Kansas: Partially-Automated Procedures for the KEDS Machine Coding Systems. The KEDS machine-coding systems will be enhanced to permit continued development of event-data generating software, which will use inexpensive, machine-readable data sources and personal computers. The KEDS-X rule-based coding system will (1) add a practical English parser to handle grammatical tasks associated with text analysis, (2) experiment with non-English source text, and (3) implement a parallel processing network for increased coding speed. Schemes for coding time-dependent datasets, such as BCOW, will also be explored. The software developed at Kansas will provide inexpensive, up-to-date, and easily customized datasets on international and domestic conflict and cooperation, and will also aid in developing the partially automated coding software being written by the Maryland team.

In addition, machine-assisted coding procedures will be implemented by two other projects. Howell’s SAS project will make extensive use of the computer-assisted (and ultimately partially automated) methods that the Maryland team will develop. Some of these methods are now in use in the SAS project. In addition, Leng’s BCOW project will use machine-assisted coding software recently developed as a part of that project. This software is specifically designed to use as input for the detailed data records produced by the GEDS project.

The software component of DDIR II also focuses on software for data analysis. Included are four projects at the participating institutions as well as an evaluation to be carried out in Illinois:

DDIR II–9. Massachusetts Institute of Technology: Computerized Textual and Interpretive Analysis of Conflict Episodes. Alker is exploring software development for the interpretive analysis of event histories. This will allow subsequent validity- and reliability-oriented comparisons of original sources, GEDS codings, human narrative summaries, speech fragments, and such computational interpretations as would be produced. Central to redefining available software routines for computational text analysis in the Schank-Abelson tradition are developing and implementing an “event description framework” motivated by Lasswell’s work on interactions, and a translation scheme for “filling in” this framework using, in particular, SHERFACS data. The interpretive routines would then operate on this framework to produce event interpretations computationally.

DDIR II–10. Middlebury College: Extension of Computerized Procedures for the Analysis of BCOW Data. Leng is modifying and enhancing two currently existing software packages developed for analyzing BCOW data. Because of the richness of BCOW coding categories, software is the only efficient way for aggregating the data for subsequent analyses. One program, CRISIS, permits users to select, count, and scale events along various dimensions. Another, INFLUENCE, is designed specifically for analyzing crisis bargaining. Both programs currently exist only in the environment of a (VAX) mini-computer, and the goal is to increase their functionality and availability by converting them to microcomputer environments.
DDIR II-11. Miami University: Computerized Preparation of SHERFACS Data for Interpretive Analysis. Sherman will also explore means to fit the SHERFACS coding schema into Lasswellian frames, which Alker proposes to use for interpretively describing conflict episodes. Computer-assisted or partially automated coding sequences are needed to transform into Lasswellian categories the SHERFACS information (and, by extension, the associated event summaries and event categories of GEDS). The software will be compatible with the GEDS data-collection system.

DDIR II-12. University of Maryland: GEDS User Software. The Maryland team will develop GEDS end-user software for browsing, data selection, temporal and spatial aggregation, graphic display, and to interface with related databases with full-text sources as well as statistical and interpretive software packages.

The Merriam Lab is considering the possibility of enhancing the utility of software developed by the various projects. For example, its numerous computer language compilers (e.g., C, Pascal, Lisp) for several different operating system environments (e.g., IBM, Macintosh, UNIX) are available for coordination tasks; and it can develop simple macros designed to link processing across the different executables so as to reduce the amount of time needed by users to perform multiple research tasks.

Though focusing primarily on data collection, DDIR II can creatively enhance software facilities that expand the usage of such data. To some measure it banks on enhanced hardware and software technologies. An ideal and very "futuristic" automated system for handling unstructured data would provide multiple interpretations of one unstructured data stream—just as ordinary citizens, political activists, and scientists working within different research traditions while looking at the same ordinary language texts might draw different interpretations. While several experimental parsers already exist, more basic research is needed before they can become reliable components of a data development infrastructure. Although a multiple-interpretive parser for ordinary language text will probably not be available for some time, we recognize the need to anticipate future technological advances in the more modest coordination outlined here. Future technological developments undertaken by other researchers will eventually permit some further extensions such as semi-automated technologies for processing unstructured, that is, ordinary-language, text.

DDIR II itself can also contribute to enhancing the hardware and software technologies that are needed. It is also essential, however, to look more closely at the degree to which coding judgments stray from case-study level understandings. The Merriam Lab will thus include some general comparisons across the basic event datasets (COPDAB, WEIS, BCOW, and SHERFACS) to assess their relative validity against original source texts regarding, say, the crisis leading up to the Persian Gulf war. The point is not that these datasets are invalid, but rather that their quality will reflect coders’ perceptions, and that, therefore, independent analysts would have to take this fact into account in using the data for their own research.

Toward the Future: DDIR III on International Political Economy Data

About a dozen years ago, international relations scholars rediscovered the importance of international political economics (IPE). It had of course remained alive and well in some quarters, particularly in Great Britain where the field of political economy was nurtured some two centuries ago. But it tended to interest economists, not political scientists, just as such issues as social change in developing countries tend to interest sociologists. Political scientists, even those concentrating their studies on international relations, by and large treated economic considerations as peripheral to the main struggle for national power and global order. Especially in recent decades the main thrust of their scholarship and instruction had been power politics, with its emphasis on military security, East-West confrontations, and guiding the political development of new nation-states. The long-standing tradition of political economy paled in the perspectives of all but a few of those who were shaping the post-1945 directions of international political research.

The renewed interest in IPE caught empirical researchers in a state of acute embarrassment. As we have seen, QIP scientists had focused on national characteristics, conflict, events, and a wide variety of other topics. By the end of the 1970s, when they looked in the larder of systematically evaluated IPE data, they found the cupboard bare.

A curious sequence of events then took place. The availability of IPE-related data from the United Nations and other agencies posed a delicious dilemma. On the one hand, a wide variety of such data sources existed but, on the other, they were of mixed quality for the type of analysis conducted by QIP scientists. The data were not always compatible, nor did they address some of the key questions relating to the broad domain of IPE research. This led to dismay in some circles. Perhaps scientists had become too accustomed to readily available, reliable, and paradigmatically similar data from such agencies as the ICPSR, the European Consortium for Political Research (ECPR), and the Zentralarchiv für empirische
The assumption sometimes seems to be that desired datasets will drop from the clear blue sky. To the contrary, they must be developed. The question thus focuses on two issues—especially in an international framework. One is, How can we enhance institutional arrangements to facilitate data development? The other is, How can we support or persuade leading IPE researchers to take on leadership roles in these endeavors?

DDIR’s plan to initiate a third research phase on IPE data remains in its pre-planning stage. An earlier effort to organize a team of researchers interested in generating data programs proved to be premature. The reason for this may have been simply that scientists invited to participate were too involved in other projects to undertake new, time-consuming ones. It may also be that the most active IPE researchers view their own roles as chiefs rather than braves, as theoreticians willing to recommend and eventually to use improved data files rather than practitioners willing to dig out the data. But, whatever the cause, the result is that any DDIR effort to encourage IPE data collections will require renewed vigor. In the meantime, word of mouth and conversations at professional meetings have revealed a number of younger and perhaps less well-known scientists with a keen interest in assembling new data collections so that they can use them for their own research. This suggests a revised DDIR strategy. It should doubtless solicit requests for proposals for IPE data programs, first to ascertain the extent to which the community of IPE scientists is interested in undertaking data-gathering activities and, second, if this proves to be the case, to work out joint procedures to coordinate these activities and seek funding.

A key element of a projected DDIR III will be the internationalization of any joint data-gathering activities. DDIR I and II have been directly related to datasets generated and carried out predominantly in the United States. It thus made sense to seek initial funding from the U.S. National Science Foundation. In the future, of course, given the international response to data on national capabilities, interstate conflict, and international events, we may expect more data-gathering activities to emerge in other countries. Accordingly, it will make sense to enhance international collaboration and seek international funding. These conditions already exist in the field of IPE, for both data-producers and data-users; and, indeed, the most significant IPE datasets to be created in recent years came from West Europe (Gron- nink, 1988; Müller, 1988). Going it alone, either for individual researchers or those at a single country’s academic institutions, may continue to be feasible but is not the best research strategy.
Clearly, international collaboration is needed. In April 1989 a study group on QIP data was established within the framework of the International Political Science Association (IPSA). IPSA's 15th World Congress, to be held in Buenos Aires, Argentina, in July 1991, provided an opportunity for the study group to hold sessions on IPE data development and data uses. Letters sent to several dozen U.S. and foreign scientists, however, found virtually no response—and only one expression of interest in participating in such a session (and this a U.S. scientist). Establishing the basis for better international cooperation appears to be something yet in the future.

The scientific field of international political economy is clearly in an exciting state of flux. While it is burgeoning in an intellectual sense, its data needs continue to be substantial. Governmental and nongovernmental agencies create many datasets, of course, but, for theoretic research carried out at academic institutions, these clearly need assessment to ascertain their value and sometimes much reworking to ensure consistency across time and space. An increasing number of scientists working in the field has recognized the need for IPE data to carry out their research activities. Also important is the fact that some of these scientists express interest in improving existing datasets and/or generating new ones. Multi-institutional and multinational organizations can facilitate such research activities. If DDIR's current organizational efforts can be carried out—or modified so that they function more effectively—the prospect is for a new era of data-based research on IPE that can significantly address important human issues.

References


Collmann, Hans-Jürgen (1973) Internationale Kriminalstatistik: Geschichtliche Entwicklung und


Merritt, Richard L. (1990) "Data in international research: Confluence of interest and possibility." *DDIR-Update* 4:3 (April), 1-11.


—— (in progress) *Recognizing and Responding to..."
International Disputes.


Wright, Quincy (1942) *A Study of War.* Chicago, Ill.: University of Chicago Press.

---

1 Richard L. Merritt is Professor of Political Science and Research Professor in Communications, and Dina A. Zinnes is the Charles and Ethel Merriam Professor of Political Science, both at the University of Illinois at Urbana-Champaign. DDIR is housed at: Merriam Laboratory for Analytic Political Research, University of Illinois at Urbana-Champaign, 512 East Chalmers Street, Champaign, Illinois 61820-3696, U.S.A. (Tel: 217-244-0739; FAX: 217-333-4369). Its newsletter, *DDIR-Update*, is currently distributed without cost; as its subscribers expand in number, however, we can expect a nominal fee to be charged. Subscriptions are available from the Merriam Lab.
Unlocking The Census Storehouse For Beginning Undergraduates

by William Bosworth
Political Science Department,
Lehman College - City University of New York

As the industrial revolution gained momentum, the individual artisan gave way to organized, hierarchized factories. Now, academic computer users seem to be evolving backwards: data processing is swiftly moving from an organized mainframe environment to increasingly powerful PC’s in the hands of independent computer users, just like individual artisans. But as we become able to store, manipulate, and display even the most complex forms of data, we can easily become isolated from fellow researchers. Manipulating data on one’s own machine may have its advantages: for example, we can tailor data sets and programs to the specific needs of our students. However, when we deal with data sets as universally necessary as the US Census, there is a danger of “reinventing the wheel.”

This paper discusses specific projects developed for beginning students in an urban setting, in the hope that others can find the work useful for their own academic projects. It is always desirable for the researchers in this field to suggest changes and improvements in their various projects. It really makes no sense for each of us to reinvent the same wheel; we may be independent artisans in our census-oriented research, but with a little cooperation we can perfect different approaches and, by sharing them, change that tiresome wheel into a complete, road-worthy vehicle.

This paper concentrates on data derived from the US decennial censuses, since such data has a number of special advantages for all researchers dealing with social questions in the US. Census items are uniformly labelled, so a program to access and transform them will work everywhere in the US. No data is more reliable. Census data reaches down to the level of city blocks, and, transformed into percentages, can enable us to compare almost any government units with any others, large or small. The Census Bureau itself provides us with hundreds of cross-tabulations so we can characterize in extraordinary detail the qualities of various age, racial, and economic categories of the population. And though most census data is based on geographic units, the PUMS file is based on a representative sample of people for each major county. Finally, census material is available on tape at least from 1960 onwards, so comparisons through time are facilitated. Where tracts and blocks have remained basically the same, such comparisons can be done for very small geographic units. Thus, those of us interested in getting undergraduates started in data analysis have in the census data our richest storehouse.

But a storehouse with a locked door is of no use. Confronted by mountains of census information coming raw from the government, each researcher is tempted to develop his own program to convert the data into usable form. Here we see the sinister danger of simultaneously re-inventing the wheel. At this point, we should inform one another what works in our experience so that others can profit from it.

First, a note about the specific needs and resources that influence our activities here. Lehman College is a public commuter college, 80% of whose students come from one county (The Bronx, a borough of New York City). Thus there is a built-in student interest in studying this area. The Bronx is separated by water and a greenbelt from surrounding counties, so it is easy to identify and analyze through time. It is one of those Northeast urban areas that have changed dramatically over the past thirty years, so demographic analysis through time is particularly rewarding. And when the Bronx is seen in detail (particularly when we examine each of its 4,132 blocks) we see economic and ethnic differences that allow for many other dimensions of analysis.

Resources for data analysis are available from our college and from the City University of New York as a whole (the artisan has his own workshop, but he can also whittle away in the modern factory). We have a powerful university mainframe computer, and individual faculty members can have terminals in their offices. The mainframe provides powerful statistical languages such as SPSS-X and SAS, as well as tape storage and disk space for real-time work. University membership in the Inter-university Consortium for Political and Social Research (ICPSR) enables us to get most census tapes without charge. As a US government depository, our library has most of the technical documentation needed to identify census items on the tapes. New York City’s City Planning Commission has developed a mapping system for computer representation of City features down to the individual block. At the college we have a number of classrooms with networked PC’s and a common elementary statistical language (ABC). There is also a
classroom with Unix-based PC’s connected to an RT file server. Here students can display and manipulate census data directly on maps of the Bronx, down to the level of city blocks. For this work we use the History Machine program developed by Prof. David Miller at Carnegie-Mellon University.

The foregoing inventory of resources shows what we start with. Most colleges probably have most if not all of them. Many undoubtedly have other resources that facilitate the work we shall describe. If other things work better, we would like to hear about it. At this point we present to you, in detail, the story of how we at Lehman College unlocked the Census storehouse for our beginning undergraduates.

I. Development of Data Files
For 1980 census data going down to the tract and block-group levels there are two files: the complete count STF1 and the sample STF3. Only the latter has information on income and poverty, education, and occupation. Using SPSS-X on the CUNY mainframe computer, we selected the STF1 and STF3 files for New York City and suburban counties, identified and labelled each of the original census variables, created new variables for “non-Hispanic White” (an ethnic category which the Census Bureau should have developed itself), created percentages from most of the census variables, changed the order of variables to make the file look more logical (in our judgment), and finally “Matched” the STF1 and STF3 files to create a single master census file for the Bronx and other metropolitan areas that could be analyzed in SPSS-X. The process involved six successive transformations of data. Each of the six programs can be made available to interested researchers.

For the 1980 PUMS file we first rectangularized the file by “nesting,” changed the order of variables to make the file more logical, and recoded ages and ethnic background to simplify for easier analysis. This process involved two successive data transformations, which interested researchers may obtain.

Using the mainframe disks we can quickly generate crosstabulations from the Bronx PUMS files. We have found no PC program that will do so for the largest PUMS, the sample based on 5% of the population (for the Bronx, this sample includes over 58,000 respondents and the datafile on PC would be larger than 5 megs). We have created a PC usable PUMS subset by randomly selecting one-quarter of the PUMS cases. The file is still over one meg in size, and works best from the v disk of one of the more powerful PC’s. For the tract and block group material we have created ABC datafiles on our PC network, so students on their own can do the analyses we shall describe below. This information is also a base for student projects involving maps of The Bronx on our Unix-based PC network. Students can study census items for the 64 Bronx “health areas,” the 356 Bronx census tracts, or the 4,132 Bronx city blocks. We must reiterate that all the data displays are based on the transformations we did for STF1 plus STF3, and for PUMS, and the programs for these transformations can be made available to colleagues.

II. Student Projects
In introductory classes, a neighborhood study is the first project that introduces students to our computerized programs. Students describe their Bronx neighborhoods (those who are not Bronx residents “adopt” a local neighborhood). They draw a map of the neighborhood, indicating its boundaries, and describing why they chose the boundaries they did. The reason for their choice is generally socio-economic rather than spatial, so the students already have generated assumptions about their neighborhood. Next, students are presented with tract maps that approximate as closely as possible the neighborhoods they have described (we must be honest: in this project we try to convince students to tailor their neighborhoods to the boundaries of one or more census tracts). Then each student is given a printout of 63 variables, with figures from New York State as a whole and from the Bronx as a whole. These are selected from items in the computer program for each of 338 Bronx census tracts. They include general population figures as well as items on employment and occupation, education, family structure, income and poverty, and housing. Items for 1970 are included as well as 1980 items. Students thus see on the printout certain “norms.” They then predict what they will find for the census tract or tracts constituting their neighborhood. Then they are shown how to use the simple “list” command in ABC to retrieve the figures for their local tract, so they see how accurate their guesses were. Great differences will stimulate students to make hypotheses about demographic factors they did not consider - or perhaps about change in their neighborhood since 1980. Throughout this first project, students are encouraged to use their own personal experience in the neighborhood to supplement the statistics they find.

The following items are examples of what the students work with (figures for 1980 unless otherwise specified):

Once they have done the first project, students will have mastered the ABC software package and (we hope) will be interested in exploring their local area in other ways. Using the same dataset just described, we next show students how to aggregate items among all Bronx tracts using weighted means. We soon develop rather complex questions. For example, we can consider all the tracts (there are 55 of them) where the 1980 population was less than half the 1970 population. We can then get weighted means for these tracts to see any peculiarities. We find, for example, that in these 55 tracts that lost so
much population, the number over age 65 actually increased by more than a third.

Or we can select areas where few Blacks are below poverty and compare them to areas where many are below poverty, and see the differences in family structure. We can do the same for Hispanics and have a couple of dimensions for interesting speculation (at this point we must remind students that they are working with areas, not with individuals). Illustrative tables are given below.

Those students who are particularly interested in the preceding studies are introduced to a second dataset, based on the Public Use Microdata Sample (PUMS file) from the 1980 census. We use the PUMS file for the Bronx, which, when tailored for the PC, includes over 14,000 individuals, around 1.3% of the Bronx population. Though we cannot look at areas within a county, the PUMS file uses individuals as its cases, so we can define specific characteristics of a population and make comparisons through crosstabulations without fear of confusing people with census tracts.

With PUMS, we can find unexpected differences among populations, which may well call for reconsideration of

### TABLE 1

<table>
<thead>
<tr>
<th>NAME</th>
<th>N.Y. STATE</th>
<th>BRONX</th>
<th>LABEL</th>
<th>GUESS FOR YOUR NBRHD</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLPCT</td>
<td>13.68</td>
<td>31.82</td>
<td>% Blacks in Pop, 1980</td>
<td></td>
</tr>
<tr>
<td>BLPCT70</td>
<td>44.77</td>
<td>24.30</td>
<td>% Blacks in Pop, 1970</td>
<td></td>
</tr>
<tr>
<td>FEMPLOYD</td>
<td>20.40</td>
<td>37.58</td>
<td>% Females, 16+, who are employed</td>
<td>N.Hsp.White 25+: % 4 + Yrs College</td>
</tr>
<tr>
<td>WHCOLL</td>
<td>21.68</td>
<td>44.41</td>
<td>Kids - 18: % in 1-Parent Homes</td>
<td></td>
</tr>
<tr>
<td>KD1PAR</td>
<td>13.09</td>
<td>26.98</td>
<td>% Pop, Income Below Poverty Level</td>
<td></td>
</tr>
<tr>
<td>VACANT</td>
<td>5.35</td>
<td>4.80</td>
<td>% Units that are Vacant</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE 2

| Procedure: | Univariate  |
| Datafile:  | BXCOR78     |
| Partition: | popratio lt 50 |
| Number of cases passing partition: | 55 |
| Number of cases not passing partition: | 283 |
| Variable:  | OLDPCRAT (TotalPop, % 65+: 1980 Compared to 1970) |
| Weight:    | TOTALPOP (Total Population) |
| N total:   | 55 |
| N included:| 55 |
| N weighted:| 110,580 |
| Minimum code: | 0.00 |
| Maximum code: | 274.75 |
| Num. unique codes: | 44 |
| Mean:      | 138.902 |
| Mode:      | 183 |
| Median:    | 140. |
| Sum:       | 15,359,836.00 |
| Standard deviation: | 61.981 |
| Variance:  | 3,841.698 |

### TABLE 3

| Procedure: | Univariate  |
| Datafile:  | BXCOR78     |
| Partition: | biblopop lt 10 |
| Number of cases passing partition: | 120 |
| Number of cases not passing partition: | 218 |
| Variable:  | BLMAKDS (Black: % Fams, No Hsbn, Own Kids) |
| Weight:    | BLACKPOP (Black Population)  |
| N total:   | 111 |
| N included:| 38 |
| N weighted:| 60,138 |
| Minimum code: | 4 |
| Maximum code: | 28 |
| Num. unique codes: | 15 |
| Mean:      | 12.6 |
| Mode:      | 15 |
| Median:    | 13.2 |
| Sum:       | 756,937 |
| Standard deviation: | 3.8 |
| Variance:  | 14.4 |
certain social policies. For example, if we concentrate on the three major ethnic groups in the Bronx (non-Hispanic Whites, Blacks, and Hispanics), we note very significant age differences. We can further divided these groups into those who are native born and those who are not (in the Bronx, the Blacks who are not native born are in their majority Jamaicans. Dominicans are the largest non-native Hispanic group, while almost all native born Hispanics are Puerto Ricans). If we do this, we find that age differences are even more magnified: Kids under 17 are twice as large a constituent of the native born Black and Hispanic groups than of the non-native groups, while those over age 65 actually constitute a majority of the non native White group!

The PUMS tables presented below show only one striking aspect of Bronx population groups. With the PUMS file we can spend hours examining other characteristics and relationships. Does income increase with education in the same way for male Black heads of household as for male White heads of household? How does the specific ancestry of non-native born Whites differ from the ancestry of native born Whites? Does a larger percentage of Bronx residents of Albanian origin have air conditioners in their apartments than Bronx residents of Irish background? If age and marital status are held constant, do Hispanics of Dominican origin still have higher household incomes than Hispanics of Puerto Rican origin? Do Bronx residents who spend over an

---

**TABLE 4**

| Procedure: | Univariate |
| Datafile: | BXCOR78 |
| Partition: | hblhopov ge 40 |

Number of cases passing partition: 98  
Number of cases not passing partition: 240

**Variable:** BLMAKDS (Black: % Fams, No Hsbd, Own Kids)  
**Weight:** BLACKPOP (Black Population)  
**N total:** 98  
**N included:** 94  
**N weighted:** 147,976  
**Minimum code:** 9  
**Maximum code:** 68  
**Num. unique codes:** 34  
**Mean:** 32.2  
**Mode:** 40  
**Median:** 32.0  
**Sum:** 4,758,267  
**Standard deviation:** 7.6  
**Variance:** 57.5

**TABLE 5**

| Procedure: | Univariate |
| Datafile: | BXCOR78 |
| Partition: | hblhopov lt 10 |

Number of cases passing partition: 82  
Number of cases not passing partition: 256

**Variable:** HSMAKDS (Hisp: % Fams, No Hsbd, Own Kids)  
**Weight:** HISPOP (Hispanic Population)  
**N total:** 82  
**N included:** 34  
**N weighted:** 16,143  
**Minimum code:** 4  
**Maximum code:** 46  
**Num. unique codes:** 17  
**Mean:** 10.5  
**Mode:** 6  
**Median:** 8.0  
**Sum:** 169,706  
**Standard deviation:** 5.9  
**Variance:** 34.9

**TABLE 6**

| Procedure: | Univariate |
| Datafile: | BXCOR78 |
| Partition: | hblhopov ge 40 |

Number of cases passing partition: 130  
Number of cases not passing partition: 208

**Variable:** HSMAKDS (Hisp: % Fams, No Hsbd, Own Kids)  
**Weight:** HISPOP (Hispanic Population)  
**N total:** 130  
**N included:** 130  
**N weighted:** 242,455  
**Minimum code:** 5  
**Maximum code:** 51  
**Num. unique codes:** 34  
**Mean:** 33.2  
**Mode:** 31  
**Median:** 32.4  
**Sum:** 8,058,798  
**Standard deviation:** 6.9  
**Variance:** 47.3
hour commuting to work have fewer bedrooms than Bronx residents who walk to work? From the ridiculous to the sublime, one cannot predict which of these questions will stimulate an undergraduate.

The accompanying maps indicate the final stage in our process of introducing undergraduates to census information. Our Unix-based PC network includes Bronx maps showing three sorts of geographic units: the 64 Bronx health areas, the 356 Bronx census tracts, and the 4,132 city blocks into which the Bronx is divided. Available data is displayed for each unit (note that income, education, and occupation figures are available only down to the census tract level; for city blocks our data shows age and ethnic divisions, family structure, and housing). The mapping system is extremely flexible. Students can change the cutpoint values of an item and see the changes instantly on a new map. New variables can be created from two or more existing ones. The screen can be split so that two maps (showing the same or different geographic units) can be displayed simultaneously. For greater detail there is a powerful zoom feature. Most important, all the manipulations are shown instantly and can be printed. The History Machine program, mouse-based and insulating students from the horrors of Unix, is also very easy to learn. We include here three maps to illustrate each geographic unit we are able to present: the health area and census tract maps have a split screen to show changes in variables through time. The block map is just too detailed to reproduce perfectly without zooming in on one part of The Bronx; nonetheless, the complete map reproduced here will give an idea of what we can do with our map program.

We continue to enlarge the kinds of data manipulation as well as the census information available to students. We are just beginning to incorporate the items of the 1980 STF4A file into our systems. We shall soon create new units from the existing city block map (police precincts; community school districts, for example) so that the data associated with these units can be compared visually with our census data. And, of course, we are preparing to plug in the 1990 census data as soon as it becomes available. Whatever the research potential for all this material, we shall not forget that in the first instance it was designed for use in undergraduate teaching. We stand ready to share the materials we have developed with others who think like us.

---

1 Presented at the IASSIST 90 Conference held in Poughkeepsie, N.Y. May 30 - June 2, 1990.

---

### TABLE 6

| Procedure: | Xtables |
| Datafile: | BXPUMS |
| Partition: | citizen eq 0 |
| Number of cases passing partition: | 11820 |
| Number of cases not passing partition: | 2736 |

| Row: | AGEGROUP (Age in 1980, Categorized) |
| Column: | SIMPLRAC (Racial/Ethnic Categories) |
| N total: | 11820 |
| N included: | 11696 |

<table>
<thead>
<tr>
<th>NH</th>
<th>His</th>
</tr>
</thead>
<tbody>
<tr>
<td>Col %</td>
<td>White</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>4 AndUnder:</td>
<td>4.5</td>
</tr>
<tr>
<td>5-12:</td>
<td>8.0</td>
</tr>
<tr>
<td>13-17:</td>
<td>6.9</td>
</tr>
<tr>
<td>18-24:</td>
<td>12.8</td>
</tr>
<tr>
<td>25-29:</td>
<td>7.3</td>
</tr>
<tr>
<td>30-39:</td>
<td>10.4</td>
</tr>
<tr>
<td>40-49:</td>
<td>8.4</td>
</tr>
<tr>
<td>50-64:</td>
<td>22.5</td>
</tr>
<tr>
<td>65 And Up:</td>
<td>19.3</td>
</tr>
<tr>
<td>Total</td>
<td>100.0</td>
</tr>
<tr>
<td>N's</td>
<td>3682</td>
</tr>
</tbody>
</table>

### TABLE 7

| Procedure: | Xtables |
| Datafile: | BXPUMS |
| Partition: | citizen eq 0 |
| Number of cases passing partition: | 2736 |
| Number of cases not passing partition: | 11280 |

| Row: | AGEGROUP (Age in 1980, Categorized) |
| Column: | SIMPLRAC (Racial/Ethnic Categories) |
| N total: | 2736 |
| N included: | 2522 |

<table>
<thead>
<tr>
<th>NH</th>
<th>His</th>
</tr>
</thead>
<tbody>
<tr>
<td>Col %</td>
<td>White</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>4 AndUnder:</td>
<td>0.3</td>
</tr>
<tr>
<td>5-12:</td>
<td>2.1</td>
</tr>
<tr>
<td>13-17:</td>
<td>1.8</td>
</tr>
<tr>
<td>18-24:</td>
<td>3.4</td>
</tr>
<tr>
<td>25-29:</td>
<td>2.5</td>
</tr>
<tr>
<td>30-39:</td>
<td>8.3</td>
</tr>
<tr>
<td>40-49:</td>
<td>8.4</td>
</tr>
<tr>
<td>50-64:</td>
<td>19.9</td>
</tr>
<tr>
<td>65 And Up:</td>
<td>19.3</td>
</tr>
<tr>
<td>Total</td>
<td>100.0</td>
</tr>
<tr>
<td>N's</td>
<td>1195</td>
</tr>
</tbody>
</table>
MAP OF THE 4,132 BLOCKS IN THE BRONX, SHOWING PERCENT HISPANICS: lightest to darkest cross-hatches:

Population 0 - 2% 3 - 9% 10 - 39% 40-74% 75 to 94% 95% and above

(Blank areas contain no population)

Spring 1991
MAP OF CENSUS TRACTS IN THE BRONX
Electronic Records: The Challenge to Archives

Introduction
The information society has had a major impact on the activities of traditional archives. The prediction of the "paperless office" in the eighties has not materialized as yet and in fact the amount of paper has increased substantially over the past decade. Rather than reduce paper, the introduction of computer technology has increased the number of products and copies of those products. The work of the archivist in the identification of the archivally valuable records has increased due to the paper burden. One must sift through far more records to identify those of historical value.

Added to the increase in the number of records created is the pressure of the research community to retain more rather than fewer records. Prior to the mid-seventies, the major factor in the appraisal of records was that of evidential value - the evidence the records contain of the organization and functions of agencies. Archives (and I restrict my comments mostly to North American and particularly Canadian Archives) have acquired many more records based on their informational and research value over the past fifteen years. In Canada, this coincides with the growth of social and economic programs of the federal government.

The computerization of many government programs began in the early sixties and has increased ever since. The centralization of edp expertise was very evident during this time and continued until the arrival of the micro computer. The large database systems were, in most cases, built by the EDP experts and used to service the program managers' needs.

Archival Programs for Machine Readable Records
The use and importance of computers was recognized by the large National Archival repositories in many countries in the establishment of machine readable record programs. Over the years standards were developed for the appraisal, acquisition, processing, conservation and servicing of machine readable records. Due to the small number of archivists involved in these programs, a great deal of co-operation and sharing of information lead to the development of procedures to handle these new records.

Only a minimum amount of success was achieved in the identification and subsequent transfer of computer records of archival value from government agencies. The control of these records was in the hands of the edp area and outside the normal channels of the control of paper records (Records Managers).

Efforts to mimic the systems in place for the control of paper records met with limited success, mostly due to the lack of familiarity with the archives by those in charge of the development of the systems. Machine readable programs, in traditional archival settings, although recognized as important, lacked the focus and strength required to affect the overall organization of records.

Trends
In recent years, this trend is changing due to a variety of reasons. In the discussion which follows, I will outline some of the changes and trends which will have a profound impact on archival repositories. These changes range from new understandings of the importance of information; technological change; and major changes in the ways data are created, stored and used. The next decade will require archives to focus on electronic records or risk losing the electronic cultural heritage.

Technological Change
It is not my intention to provide an overview or history of the changes we have experienced in technology in the past decades. It is, however, important to review some of these changes in light of how records are created, why, and how archives must adapt to these changes. The major trend which has affected the way records are created results from the rapid penetration of microcomputers into the market in the last five years and, in particular, into government departments and agencies. The centralization of edp services is disappearing with the use of microcomputers in the office environment. Managers, officers and clerks have now as much computing power on their desks as the mainframes of the seventies provided. The ability to create, manipulate, access and disseminate data has been decentralized. The trend to purchase "off the shelf" software has taken away from the centralized edp shops in the creation of in-house software and database systems. Linked to the penetration of microcomputers is the development of local area networks. The linking of staff provides for the creation
and revision of documents which can be done on-line with only the final version being available in either hard copy or electronic format. The ability to provide for the development of documents relating the evolution of the policy, to changes in administration, or to data is now in the hands of the creators of those documents. In most cases, LAN's do not provide for the traditional records management approach to control of records.

Individual workers make decisions regarding the disposition of the records. Such a system existed for the development of large database systems under the control of edp professionals. The difficulties experienced in gaining control over what information is being created and destroyed can be magnified as all employees become responsible for their records. The program of economic restraint experienced in most western countries is also leading to the use of microcomputers, as it is seen as a way of increasing productivity and decreasing personnel costs.

A major effort can be seen in the development and interest in communication standards. The lack of compatibility between hardware has been, and continues to be, a major problem to the increased usage of data. The efforts now seen in the development of International Standards is encouraging. The trend to the Open Systems Interconnection standard protocols provide for the possibility of connecting systems with different hardware. Other standards will have an impact of the increased sharing of data. Map and Chart Data Interchange Format, or MACDIF, data is an attempt to provide a standard format for the transfer of chart and graph data from and to a variety of systems. Office Document Architecture/Office Document Interchange Format (ODA/ODIF) provides for similar transferability of text. The work towards developing and implementing such standards must be followed closely by archivists, as it is through such efforts that some of the technical issues such as making valuable data accessible in the future may be resolved.

New techniques for software development such as fourth generation languages and expert systems techniques are becoming important tools providing faster and more flexible software development and more user interfaces. No longer must the design and development of databases be the sole responsibility of the edp professional. Databases can be created and used by those who have access to D-Base or other such software. Expert systems potentially pose major problems for the archivist. In the past the acquisition of data tried to steer away from software dependent systems. Expert systems which can be defined as “an intelligent computer program that uses knowledge and inference procedures to solve problems that are difficult enough to require significant human expertise for their solution” are only in the early stages of practical application. Their potential to assist managers with complex planning and scheduling tasks, diagnose diseases, etc. is great. The impact of such systems on the documentation of the decision making process is evident. How archivists will respond to such systems is a major challenge.

Types of Data
Part of the technological changes, but one which should be highlighted, is the trend towards integrated systems and applications. Two specific types will be discussed:

The Geographic Information System (GIS)
Geographic Information Systems are beginning to play a major role in the information society. Today's systems only superficially resemble the automated mapping systems of the sixties. GIS are increasingly being used to conserve and manage a wide variety of data from natural resources to environmental pollution as well as in the planning and management of cities - such land information systems cross organizational and sectorial boundaries and represent an opportunity to develop new information based products and services.

Compound Documents
The move to more integrated systems is seen in the “compound document”. The integration of voice, data documents and graphics oversteps the traditional media boundaries. All are reduced to the common language of binary code. Not only is it feasible to create the compound document, but it may also have been created from information which was only accessible on the screen for a brief period. The source of that information cannot be traced. The accessibility of data from other systems through local area networks and the merging of data from a variety of databases will create documentation problems for the archivist. The ease with which such information becomes available and usable will be reflected in the move to adopt communication standards, more user oriented software, and more computing power.

Information As A Resource
The information society has created a new awareness of information as resource. The major expenditures on hardware and software development of the seventies has created a new awareness of the value of the information which these systems manipulate and store. In Canada, Access to Information and Privacy legislation led to the acceptance of a computer based record as a record. In the definition of a record for the purpose of the legislation, machine readable is included.

The requirement to account for information regardless of the machine on which it was stored was an important step in the recognition of computer records. The new National Archives Act passed in 1987 also uses the same definition of record. The Act stipulates that no records of
the Government of Canada can be destroyed without the consent of the National Archivist. It further stipulates that those records deemed to have archival value must be transferred to the National Archives. The definition of record to include machine readable records ensures that electronic records are part of the National Archives' responsibility. More recently two new policies have added to the importance given to information: the Information Holdups Policy (just recently approved) which will direct government departments to manage their information in a holistic manner and account for that information through the development of directories to it; and the Information Management Plan which outlines to departments the importance of the planning of information management rather than the justification of new equipment purchases.

All of these policies, acts, and planning strategies are the result of an awakening to the importance of information as a resource.

Another interesting trend in the field of information is the move by the public and private sectors to develop co-operative databases. The Geographic Information Systems are an example of this type of co-operative effort. Similar to the compound document such co-operative databases will have an impact on the archival organization of records.

Much more could be said on the trends and changes which technology will initiate. It is important to look briefly at the impact such changes will have on the traditional functions of archives.

**The Challenge to Archives**

"The digitization of information through the common language of the binary code is bringing about the convergence of voice, usage and data - and of the telecommunications, electronics and computing industries based upon them".

Over the last decade, Archives have tended to expand according to media-based responsibilities: textual records, cartographic, film and television, photographic and machine readable. Practices and procedures were developed to acquire, process, store and service the different forms of information as each had its special requirements. The fundamentals of archival theory were common to all media. Appraisal criteria, were based on the principles of evidential, informational and legal value. It was in practices for arrangement and description where the differences became more evident. Procedures and practices for the long term preservation of machine readable records were developed in the seventies. The procedures were based on large mainframe systems and proved successful for the conservation of data in systems.

Technology is now the driving force behind the integration of the different types of records. Just as media divided Archives into specific units, it now will play a large role in integrating these units. With the use of electronic technology in the creation of all types of records, the media on which the information resides becomes the common element.

Information is created and transmitted in so many forms that archivists must now look at program activities as a whole and identify those records which have archival value as well as the most appropriate form in which they should be stored. Electronic records provide many research possibilities. As more types of records are created in electronic form - more such records are likely to be of archival value. The major obstacle is, of course, the long term accessibility of the information in electronic form. For other records-paper, photos, maps established techniques have been developed which preserve the records for future use. Technology did not have a major effect on long term preservation except for improving the techniques. With electronic records, the media, the software and hardware are constantly changing - Evidence of this can be seen in the experience gained to date. Electronic records created in the seventies are different from what is now being created. Procedures, valid for data in systems, must be modified and reevaluated to cope with compound documents and GIS. The technological requirements put pressure on established procedures. The focus of any archival program for electronic records must focus its resources on resolving the technical issues of how best to transfer records to the archives, how to process these records to ensure their accessibility. Archivists will be required to support efforts to ensure standards; to become involved with systems as they are being created; and to keep abreast and knowledgeable about the changing technology and how it affects the creation and use of records.

These are major changes for institutions which have traditionally dealt with the past.

Finally, new methods in records creation may have fundamental effects on traditional archival theory and principles. Archivists must become active participants in the creation of information, in many instances identifying elements of archival value before they are created, in order to ensure the preservation of the historical record. Archives have, to date, been concerned with documenting the activities of an organization or business. How does this new role affect the documentation of activities when the archivist has participated in the creation stage?

As more organizations undertake co-operative efforts in information creation, how do we determine which records originate with which organization? Who has the ultimate responsibility or control of the records? Such
systems break down the barriers between public and private sectors; federal, provincial and municipal governments. The clear lines of origin become blurred.

Conclusion
Today’s presentation can only briefly mention the issues and resulting challenges to traditional archives. Issues such as these are of utmost importance to the archival community. The “paperless office” is not yet a reality but signs of its existence are much more evident today than they were two to five years ago. Efforts to resolve the problems are imperative if records documenting the nineties are to be available for future generations.

Presented at the IFDO/IASSIST 89 Conference held in Jerusalem, Israel, May 15-18, 1989

Complex Data, Simple Tools: An Introduction to Text Retrieval Packages

by Andrew Marchant-Shapiro
Departments of Political Science and Sociology
Union College, Schenectady, N.Y.

Introduction
Before the advent of personal computers, qualitative researchers could often be found "waste" deep in typed interview transcripts. Dealing with these transcripts often meant hours of searching for a particular passage or pattern. For many, little has changed — except that the transcripts are now word-processed instead of typewritten. But precisely because they are word-processed, these transcripts open up new possibilities for computer-aided retrieval and analysis.

This article provides an overview of one class of software programs, text retrieval packages (TRPs), that can provide significant assistance to qualitative sociologists with minimal investments of both time and money. Using a hypothetical text retrieval package, I suggest some techniques that sociologists can use to maximize the utility of TRPs. I outline the basic characteristics of TRPs, and describe a few commonly available software packages that present variations on the TRP theme. The techniques introduced here are not specific to any one system, and may be used to advantage with a wide variety of text retrieval packages.

I should make clear at the outset that this article is about improving access to textual data, with specific application to qualitative data such as transcripts from unstructured ("conversational") interviews. None of the TRPs described in the following pages provides for analysis of qualitative data. While such programs are readily available, and are used by a growing number of qualitative analysts, they are not my concern in this article. Rather, the programs discussed below are simple tools that provide qualitative researchers with greatly improved access to their complex data.

Thinking of the Interview as a Database
When a sociologist hears the word "database," he or she is likely to think of a collection of coded data arrayed in rows (cases or 'records') and columns (variables). Normally, such databases are manipulated with software programs known as database management systems (DBMS). The DBMS makes it possible for the researcher to gain rapid access to specific sections of his or her data. For example, a researcher using DBASE IV, a popular DBMS program, might want to see the ages of all those individuals in her database who lived in Illinois in 1970. Depending on how the data are structured, she might give the following command:

```
list age for "IL"$STATE_1970
```

The DBMS program would first locate those rows of data for which the variable STATE_1970 had the value "IL," and then isolate the variable AGE in each such record and print its value on the screen. A typical output might look like this:

<table>
<thead>
<tr>
<th>STATE_1970</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
</tr>
<tr>
<td>28</td>
</tr>
<tr>
<td>19</td>
</tr>
<tr>
<td>41</td>
</tr>
<tr>
<td>30</td>
</tr>
</tbody>
</table>

Note that by using the DBMS program, the quantitative researcher has gained great power in interrogating her data. She no longer needs to manually search for each case in which a subject was living in Illinois in 1970. This ability to isolate particular records for inspection is one of the reasons that DBMS programs have gained popularity with quantitative researchers. On large surveys, such programs are often used to ease cleaning of data and allow for the isolation and closer inspection of outlying cases.

The usefulness of similar strategies should not be lost on the qualitative researcher. There are many instances in which it is desirable to move rapidly to a section of an unstructured interview that is marked by the occurrence of one or more key words or phrases. These range from the early days of a research project, when one is exploring the transcripts of recent interviews, to the final stages, when a researcher may need to find one or more quotations to reinforce her point. One may even want to test the notion that two words or phrases verbalizing particular concepts occur only (or most frequently) in conjunction with one another. In a set of interview transcripts that can run to hundreds of pages and millions of words, how can you find the particular passage, or passages?

One approach, to be recommended for its economy and simplicity, is to use the search function of your word processing program to look for the text in question. But
with a few exceptions (noted below), this limits you to searching for a single word or phrase in a single text file at a time. Moreover, more complicated searches (such as searching for all paragraphs of text that do not contain a particular word or phrase or combination of words and phrases) are beyond the capabilities of word processing programs.

This is where text retrieval programs come into play. The grandparent of modern TRP programs is a widely available program called GREP. It originated on UNIX mainframes, and is today available on all computers that run the UNIX operating system. Moreover, various public domain versions of the GREP program are available for most microcomputers, and a limited version of GREP, called FIND, is distributed by Microsoft with every copy of MS-DOS.

GREP is a simple program, but extremely powerful. In essence, you give GREP a word or phrase to search for, and it compares each line in a data file with that specific word or phrase. Lines that match can be counted, printed to the screen, or saved into a new file for further manipulation (alternatively, you can do the same thing for lines that don’t match). A researcher might be interested in seeing how often the word ‘credibility’ appears in a particular interview transcript. The transcript is stored in the file TRANS017.TXT, so our researcher invokes GREP this way:

```bash
grep -n credibility trans017.txt
```

GREP scans each line of TRANS017.TXT for the pattern of letters forming the word *credibility*. The ‘-n’ in the command causes GREP to number the lines in the file as it scans them. When it finds a line containing the pattern *credibility*, it prints that line to the screen, forming an output like this:

```
0023 and that was a serious problem for our credibility
0040 was it a credibility problem? No, credibility was
0215 Credibility. Plain and simple. If it hadn’t of
```

The researcher now knows how often the word appears in the transcript, as well as where it appears. Since GREP works very quickly, a few such searches can give the researcher significant insight into his data in a very short time.

Note, however, that there are some important drawbacks to the way that GREP handles the data. First of all, what we have retrieved are lines of the file, not sentences. From a computer’s point of view, lines are a sensible units to use because it is easy to tell where one line ends and the next begins. The computer treats each line as an independent record. But from a human perspective, lines are not very useful as records: they may contain anything from a few words to a few short sentences, and they do little or nothing to establish the context within which any particular datum is found. GREP can rapidly locate all lines of text containing matching patterns; we need programs that can retrieve entire chunks of text (sentences or paragraphs, for example), context and all. At a minimum, we need to locate not only the statement containing the word or phrase for which we are searching, but also the stimulus that evoked that statement.

A second problem is that GREP is limited to searching for a single word or phrase at a time. While a skilled user can compensate somewhat for this limitation through the use of complex ‘regular expressions’ or root searches (see below), GREP is incapable of searching for phrases that break over lines and cannot examine text chunks for the presence and/or absence of multiple words and phrases. GREP is limited to searching for a single word or phrase (a sequence of words); we need programs capable of looking for combinations of words and phrases that may or may not be sequential.

Modern TRPs answer both of these needs and more. Rather than operate on single lines of text, they can operate on paragraph-sized chunks and can make use of Boolean operators (see below) to allow for a variety of ways to combine search texts. Moreover, unlike word processing programs, TRPs can search many files with a single command, greatly speeding up the retrieval process.

In themselves, these improvements over word processors and GREP make TRPs powerful if simple-minded tools. To get optimum performance from a TRP, however, requires more than just aiming the program at a file and telling it to go to work. By adding a modicum of structure to the transcript of an unstructured (or structured) interview, we can realize significant benefits. Structuring a transcript actually involves three different aspects: structuring, in the sense of organizing a conversation into meaningful ‘chunks’; identifying concepts, adding keywords to the record that either amplify the content of the conversation or actually represent analytic categories; and problem prevention, a process analogous to the ‘cleaning’ of quantitative data.

**Structuring the Transcript**

If the chunks of data that we wish to retrieve are larger than a single line, we need to structure the text so that the TRP we are using understands where a particular chunk begins and ends. How we structure chunks (or records) depends on the particular data we are looking at and how we plan to use it.

Consider the unstructured interview. Such an interview is a conversation, typically made up of paragraphs; the
first party speaks, then the second, then the first, and so on. Typically, the interviewer asks a question, then the informant replies, as in Figure 1.5

Figure 1

Q: Were you particularly worried about extremists coming into the movement at that time (1978)?

A: Not especially, no. Not until we got word that the news media had mixed up some of our group in the west with the Posse Comitatus. That cut into our credibility something fierce, and made it very difficult for us to get sympathetic press coverage.

The question and answer — sometimes with followup questions and answers or other interactions — provide the context within which to understand a particular statement. If we can treat each question-and-answer set as a record, that is, as an independent datum, then we are well on the way to transforming what may be a long (and sometimes rambling) interview into a useful database. Within each record, we should have not only the full text of the answer, but the stimulus that evoked that answer. Bear in mind, however, that the real challenge is not understanding for ourselves what constitutes a record, but organizing our data in some way so that the computer’s notion of what constitutes a record is identical to our own. Once we have arrived at a definition of a record that is adequate for our own use, we have to think about the structure of the records as the computer sees them.

From a computer’s perspective, the most useful form of raw data is a file that consists of text (letters, numbers, punctuation and spaces) and a few special characters, such as carriage returns and line feeds. Such a file is known as an ASCII (American Standard Code for Information Interchange) text file, and uses characters in a standardized fashion.5

To divide a file into records that both the researcher and the computer/TRP will understand, use single spacing within paragraphs, and double spacing between paragraphs, as in the following example:

```
XXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXX <-record 1

XXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXX <-record 2

XXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXXXXXXXXXXX <-record 3

XXXXXXXXXXXXXXXXXXXXXXXXX X
```

In this way, each paragraph of text becomes a distinct record, and the TRP can easily distinguish where one ends and the next begins.

Records should include, at a minimum, a question and the response it invokes. These should be divided in some way, however, so that we can tell at a glance at which part of the interaction we are looking. One way to divide between the two is to insert a line of hyphens (-) between question and answer. The one way not to divide the question and answer is with a double carriage return; this will make the question and answer appear to the TRP as two independent records.

Identifying Concepts

Not infrequently a conversation has more meanings than would be apparent from the text of the interaction itself. In such instances, we may wish to add still another section to the record — again, divided by a string of hyphens or other special characters — that consists only of keywords or comments pertaining to the interaction. Such keywords may simply clarify the meaning of the text of the conversation, or they may be analytical categories you have assigned to the particular record.

If you do use keywords, it is a good idea to use some special character to mark them so that the TRP can differentiate keywords from the rest of the text. For example, all keywords might be preceded and followed by the ’*’ character — *aggression*, *money*, and so forth. This helps to avoid confusion between words that are part of the text per se and others that are introduced by the researcher once the interview has been completed.

Preventing (and Resolving) Common Problems

While we are busily creating the perfect data record, however, we need to be aware of complications that we can introduce that may make searching difficult or impossible. The major problem is one that afflicts all text processing: misspelling and inconsistent spelling. This is a particularly nasty problem if someone other than the interviewer transcribes the interview. Computers are powerful but inflexible creatures; if you search for the name Kamin, you may find that the name doesn’t come up in the database — because it has been entered variously as Kemin, Kamin, Camin and/or Camyn. There are flexible TRPs (discussed below) that may find one or more of these misspellings through the use of ‘fuzzy’ search criteria,8 but it is probably best not to rely on technology to fix this problem after the fact.

The most straightforward answer to this problem is the spelling checker. These programs, often integrated with word processing programs, scan the text, either while it is being entered or once the file is complete, and locate words that do not match a dictionary file. Most spelling checkers have provision for an auxiliary dictionary,
which contains words not in the main dictionary but which are used by the writer. This is the place to establish a list of names of persons and organizations, so that misspellings will be detected at once and corrected. You should also supply yourself, or the person transcribing the interview, with a list of names and special terms that appear in the interview(s) with which you are working. As you proofread the transcripts, you can add to this list (and the spelling checker list) as you go along.

Sometimes, however, a new name or term comes up, or a name is garbled. You should have provisions for alternative spellings, such as the following:

after that [Mankoff/Makov (0213)] told me that

An early step in going through the interviews could be to search for the ‘I’ character, so as to locate and resolve problems. Alternatively, you can leave the variant spellings in place, in case context later allows you to interpret the garbled passage. If you are not transcribing the interview yourself, have the person who is insert the digit counter value for the point on the tape where the garble occurs. In this way, even imperfect transcriptions (and there are few perfect ones) will be usable by computer search programs.

It may sound like a lot of work to structure text in this way, but it is not really that difficult if the structuring is done when the data are first entered into a word processing program. If the researcher himself or herself is entering the interview, then keywords can even be added at the same time. The additional labor imposed by a simple data structure is a small price to pay for the ease of access that will result. In the next section, we turn to the issue of access in order to get a sense of what can be accomplished. Whether coded or not, once the data have been structured, the hard part is over.

Searching for Data

If we think of our interview data as now consisting of paragraph-sized records, each record consisting of a question and its associated answer and constituting a context for the statements therein, we are in a position to consider how we would like to specify which records to retrieve. Searches may be simple (for one word or phrase) or complex (for various combinations of words and/or phrases). If we have added keywords to the interview data, we may search for these as well.

Simple Searches

Recall the example of the quantitative researcher. Her search began by specifying a subset of possible records — those records that contained the value ‘IL’ in the variable STATE. 70. The qualitative researcher does not have variables to work with in the same sense; instead, he has a chain of verbalized (and perhaps coded) concepts. While these are not consistent from record to record — only a few members of the set of all possible concepts are present in any given record — we can test for the presence or absence of particular words. Consider Figure 2. This is the same paragraph shown in Figure 1, but now structured as a record and stored, with other records, in the ASCII file INTRVW.001:

**Figure 2**

<table>
<thead>
<tr>
<th>Q: Were you particularly worried about extremists coming into the movement at that time (1978)?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: Not especially, no. Not until we got word that the news media had mixed up some of our group in the west with the Posse Committed. That cut into our credibility something fierce, and made it very difficult for us to get sympathetic press coverage.</td>
</tr>
</tbody>
</table>

This record contains a stimulus, a response, and a set of keywords that both overlaps (e.g., *media*) and categorizes (e.g., *perception*) the information contained in the interaction. The record shows the presence of such concepts as extremist, movement, media, credibility, 1978, and so on. On the other hand, it does not contain terms indicating such concepts as electoral politics, formal organization, or legitimacy (to name just a few possibilities). So, if we wanted to retrieve only those records that included a verbalized or keyworded conception of credibility, we might give a hypothetical TRP a command like this:

**list ‘credibility’ in INTRVW.001**

The result would be a listing of all of the records in INTRVW.001 that contain the term credibility, including, of course, the record shown above. If we wanted to search for all records that contained the concept electoral politics, the TRP would not retrieve this record. Conversely, if we searched for all records that did not refer to electoral politics, this record would be among those retrieved.

But suppose that we wanted to search more broadly — for variations on credibility. Suppose that our informant didn't actually use the word credibility, but said something like 'it was hard for us to be credible.' Since credible is not the same pattern of letters as credibility, the computer would not have found that record. But we can modify the search in one of two ways so that we are more likely to find appropriate records. We can either search using roots, or we can search using multiple terms.
If we are searching for variants on a single term, searching for a root can do the job. For example, we might search for the pattern common to both words, i.e., *credibile*. To do a root search, consider all of the similar terms you want to retrieve and search for the common portion of those words. *Legitimacy, legitimate*, and *legitimation*, as well as variations such as *illegitimate*, can all be retrieved through a common root. If you use this approach, though, be careful not to shorten the root too much; if you do, the search results may be useless because large numbers of records containing ‘noise’ words satisfy the search.

Some TRPs allow for a variation on the root search method using *wildcards*. These are special characters that can be inserted into a search that will match any other character or combination of characters. If ‘*’ matches any single character and ‘*’ matches any group of characters, then ‘gr*’ matches words such as grew and grew, and ‘il_’ would match anything from stimulent to impossible. Obviously, wildcard searches are also subject to ‘noise’ problems, and should be undertaken with care.

**Complex Searches**

While searching for a single word or for variations on a single word can be helpful in plowing through long transcripts, it is often more useful and more interesting to be able to choose records based on the presence of two terms, or on the presence of one and the absence of another. *Boolean operators* are ways of specifying logical connections between words and/or phrases. Online information services such as Lockheed’s DIALOG service make use of these operators, as do more common, PC-based systems such as WilsonDisc, and virtually all DBMS programs. The basic Boolean operators, AND, OR, and NOT, can be used singly or in combinations to set exacting criteria that records must meet before the TRP will retrieve them.

**Widening the Search: Logical OR**

A root search works by using a single, less rigorous criterion for matches. In contrast, a multiple term search expands the search pattern by allowing a record to be retrieved if it satisfies one or more elements of a set of criteria. To construct such a set, we use the Boolean logical operator OR. For example, if we give a command to our TRP to:

```
list for 'credibility' OR 'credible' in INTRVW.001
```

A record that contains either word will be retrieved. Obviously this technique can be expanded so that concepts that may be expressed in a variety of ways can be searched. We might want to search for terms like ‘credibility’ OR ‘legitimacy,’ for example. Using the logical operator OR always widens the search, since a record that satisfies any part of the expression that has been ORed together is retrieved. Sometimes, however, ORing things together gets us more than we want. It is then that we can use another logical operator to tighten our search criteria.

**Narrowing the Search: Logical AND and Logical NOT**

When we AND things together, we are telling the computer to retrieve only records that meet multiple criteria. For example, we could exclude the example record shown in Figure 2 from a search by asking our TRP for the following:

```
list for 'credibility' AND 'organization' in INTRVW.001
```

The record satisfies one criterion but not the other, so it is not retrieved. Only those records will be found that contain both words. Using AND takes care, because it is possible to quickly reduce the number of records that match the search to zero.

The utility of AND and OR is increased by adding the third logical operator, NOT. NOT allows the TRP to retrieve a record only if a particular term is not present in the record. NOT is seldom useful alone, but in combination with AND and OR, it allows for very precise specification of searches. If we wish to find only those records that refer to ‘this’, but not those that also refer to ‘that’, then we can search for ‘this’ AND NOT ‘that’.

**Grouping Logical Operators with Parentheses**

While many searches are easy to specify with one or two logical operators, searches can become quite complex, and it is important to specify the priority in which logical operators act. Fortunately, most TRPs allow the use of parentheses, which allow the researcher to specify the order in which the TRP evaluates logical relationships. We can develop searches such as (‘credibility’ OR ‘legitimacy’) AND NOT ‘organization’. This particular search would first retrieve the subset of all records in which either ‘credibility’ or ‘legitimacy’ were present, and then reject the sub-subset of records which also contained the term ‘organization’. If we had instead defined the search ‘credibility’ OR (‘legitimacy’ AND NOT ‘organization’), the TRP would first find all records containing ‘legitimacy’ but not ‘organization’, and then retrieve as well all records containing ‘credibility’ regardless of whether or not they included ‘organization’.

An example of the logical operators’ power to differentiate among records may be in order here. Consider the following one-line records:

1. Then Bob told Carol and Ted.
2. But of course Alice and Carol told Bob and Ted.
3. Alice and Ted were outraged at that.
4. Finally, Alice left with Ferdinand.

Below are some search criteria and the numbers of the records that each search would retrieve. These should demonstrate clearly the different behaviors of the various operators.

'Bob' OR 'Carol' OR 'Ted' OR 'Alice' 
(1,2,3,4)

'Bob' AND 'Carol' AND 'Ted' AND 'Alice'
(2)

'Alice' AND NOT ('Bob' OR 'Carol' OR 'Ted')
(4)

'Alice' AND ('Bob' OR 'Ferdinand')
(2,4)

**Searching Using Keywords**

With the use of Boolean operators, keywords take on a special significance. They are more than merely additional tags that we can use when our informants use varying terms to discuss a single concept. Through the use of AND, OR, and NOT, we can examine the relationships that exist between keywords that indicate coded concepts and the content of the conversation itself.

Recall that keywords are marked with special characters (*). These markers affect searching in particular ways. For example, a search on the term 'legitimacy' will be satisfied whether the term occurs in the text or in the keyword section. But '*'legitimacy*' will only be satisfied by the term in the keyword section. By combining keywords and logical operators, we can do searches like this:

list for '*media*' AND 'credibility' in INTRVW.001

This search would find only those records noted and marked by the researcher as having some bearing on media issues, and then only the subset of these records that had verbal and/or keyword relations to credibility. By adding keywords to our records, we begin to approach the same kind of specificity and power in searching that DBMS programs afford quantitative researchers.

**Making Use of the Output**

The goal of all these manipulations is, of course, to find specific records within a large body of information. What you do with that information once you find it is up to you, but you should be aware that not all TRP programs allow you to save the data that you find. Some merely allow you to view the records that the program has retrieved. Most have provisions for saving some or all of the retrieved records to an ASCII file. Some, such as Golden Retriever (reviewed below), take you to the point in your transcript where the match occurred and allow you to save as much or as little of the surrounding material as you desire.

Since the usefulness of TRP programs lies in their ability to winnow data, as it were, you should probably avoid programs that do not allow you to save output to a new file. For example, SeekEasy, one of the programs reviewed below, has no provision for placing the retrieved text into a new file. This limits its usefulness in anything other than exploratory research, since the only way to record the results of your search is with pencil and paper (or the print screen key).

Once you have an output file, you can do several things with it. You can simply include the file, or an edited version, in a paper or article you are working on. Or, if the number of records retrieved is large, you may be able to treat the new file as a second-order database — searching more specifically within the file.

In any event, you should always take a look at the output file before including it in other documents or doing further searches. Computers are wonderful servants, but they take everything — including our mistakes — literally. If the results look strange to you, review your search commands carefully. The difference between

'(Bob' AND 'Carol') OR ('Ted' AND NOT 'Alice')

and

'Bob' AND ('Carol' OR 'Ted') AND NOT 'Alice'

may turn out to be significant. A good way to check the search results is to make certain that a randomly-chosen record within the output actually does satisfy your search request.

**From Ideal to Real: Some Inexpensive TRP Programs**

Up to this point, we have been dealing with a hypothetical TRP. None of the programs that I discuss below does exactly what our hypothetical model does. Rather, each emphasizes one or more features described above. None of these programs costs more than $50, and most cost significantly less; some are available for the asking.

For each program, I give a brief summary and then a
description and evaluation of how the program works. These descriptions are summarized in Table 1. I also include information on how to obtain each program.

Originally, I intended to begin this section with a speed comparison across the programs, and to this end I tested each program using the ASCII transcript from a three hour unstructured interview — approximately 22,500 words. The slowest program I tested was a version of GREP, which took 70 seconds to go through the file; the other programs all had times of 30 seconds or less. Consequently, I have not included a speed comparison. The differences here are negligible.

Rather than focus on speed in deciding which program might meet your needs, I suggest that you consider the features that particular programs emphasize that might make them most useful in your particular work. One important factor to note is that, while the hypothetical TRP described above is controlled through command lines, many TRPs are menu-driven: You select the actions you want from a list, and the computer does the rest. These may be simpler to use for those unfamiliar with computers, or in classroom situations.

**Golden Retriever**

Golden Retriever, version 4.0, shareware — $39.95. Golden Retriever is a powerful TRP: it can be menu or command driven, and it has the capability to search for multiple-word phrases. It even has an adjustable fuzziness level. That is, you can make close guesses at the spelling of terms you don’t quite recall, and Golden Retriever will often find them. The degree of “fuzziness” Golden Retriever will allow in a search comes preset to a reasonable level, but you can make the search more or less rigorous through a menu choice. Golden Retriever can make use of logical operators, as described above, but only in a very limited fashion. If you AND words together, for example, they will only match exactly the same pattern in the file — ‘Bob’ AND ‘Carol’ will only match Bob Carol; it will not match Carol Bob or Bob Alice Carol. The words in the file must not only appear in the same order as in the search criterion, but they must also be adjacent.

Golden Retriever’s menus are clear and easy to understand. When Golden Retriever finds a word in a file, it takes you to the appropriate record and highlights the word on the screen; you may then use the cursor keys to choose how much of the surrounding material, if any, to save into an output file. One unusual feature allows Golden Retriever to run in the background, while you work in your word processor or other text entry program. Pressing a special key shifts you into and out of the Golden Retriever program, allowing you to search for data while you are working on a report, for example. There is a preview version of Golden Retriever available, the Golden Retriever Pup. Golden Retriever Pup works exactly the same way as does Golden Retriever except that it will not read data files on a hard disk, which limits its usefulness considerably.

The Pup version is available via modem from computer bulletin boards, or for $10 from the National Collegiate Software Clearinghouse (NCSC), Duke University Press, 6697 College Station, Durham, NC, 27708. The full version can be ordered for $39.95 from Wesware, 42 Epping Street, Lowell, MA, 01852.

**GREP**

There are dozens of versions of GREP available, most of them in the public domain, posted on computerized systems across the country. If you have access to a modem, this is one way to locate a GREP program. If you don’t, find a colleague or computer center person who can help you. Most microcomputer GREPs will explain themselves to you if you enter GREP or GREP ?. GREP is a good place to start looking at TRP systems because it is simple and cheap; you should be able to obtain a copy for free. Most (but not all) versions of GREP can save output to a file by appending the command ‘>’, followed by a file name, to the end of the search request. Hence,

```
GREP 'bob' intrvw.txt >save.bob
```

saves the results of the search to the ASCII file SAVE.BOB.

**ResNoter**

ResNoter, version 1.0, (c) NCSC — $35. ResNoter is one of the most technically sophisticated programs I evaluated. It is the only one of the TRP programs reviewed here that uses indexing. This means that using ResNoter is keyword-intensive: if you want to use this TRP, you must insert extensive keywording in your data; ResNoter will not search raw text. From the keywords that you supply, ResNoter constructs a list of code words and their locations in the database. If you ask for ‘bob’ AND ‘carol’, ResNoter need only look at the locations in the ‘bob’ list and compare them to those in the ‘carol’ list. It can then jump directly to the records that satisfy the request.

Because of this indexing, all logical operators are available and ResNoter is very fast. However, you must remember that if you add a new keyword you will not be able to use it until you have re-indexed the database. Indexing does not take long, but it is a step you must not forget in working with ResNoter. Another drawback is that ResNoter is loaded with menus. Menus should make life easier for the user, but ResNoter’s menus are posi-
tively frightening because their operation is highly inconsistent. Still, if you need rapid access to large amounts of data, and if you are willing to insert keywords, ResNoter is worth learning. You can save output to a file, and you can choose what portion of the record to save (keywords, raw data, or both). ResNoter is one of a series of text retrieval and analysis programs published through and available from the NCSC.

Search

Search, version 1.3, public domain — free/$25. Search is one of the more complete TRP programs reviewed. It uses all three logical operators and, unlike Golden Retriever, is not limited to 'exact' logical matches. That is, Search scans the whole record to see if the logical requirements are matched, not just adjacent sets of words. 'Bob' AND 'Carol' will match not only Bob Carol but also Carol Bob and Bob Alice Carol. Search supports multiple levels of parentheses and can search for any combination of up to fourteen words and/or phrases.

One particularly nice feature, useful with logical OR searches, allows Search to note, either on the screen or in the output file, which of the logical search terms it matched in a given record. You can also have Search ask you whether or not to save a given retrieved record to its output file. An option allows Search to work like GREP, if you want to look only at line-sized records.

Search has two relatively minor drawbacks. First, it is mainly command-driven. To use it, you must learn to type in a sequence of commands. For example, if you gave this command:

```
SEARCH INTRVW.001 B =bob&carol > OUTPUT.TXT
```

Search would look for paragraph records containing both 'bob' AND 'carol' and saves the results to OUTPUT.TXT. These commands are not hard to learn, but may intimidate a first time user. Search provides a second, more limited search mode for beginners, which allows for only AND and OR operators. In this secondary mode, the TRP asks the user for search terms and filenames. Still, it is not as friendly as a menu-driven system like Golden Retriever.

A second drawback is that matched words are not highlighted in retrieved records when they are printed to the screen — if you are dealing with large records, this can make it difficult to find the exact point at which the match occurred.

Search is available free on computer bulletin boards or directly from its author for $25, which includes a subscription to future versions. Note however that, if you obtain SEARCH from a bulletin board, no donation is expected. For further information, contact Eric Bohlman, 1921 Highland Avenue, Wilmette, IL, 60091.

<table>
<thead>
<tr>
<th>PROGRAM NAME</th>
<th>VERSION</th>
<th>BOOLEAN LOGIC</th>
<th>FUZZY SEARCH</th>
<th>SAVE OUTPUT</th>
<th>MAX SIZE PER RECORD</th>
<th>PRICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Golden Retriever</td>
<td>4.0</td>
<td>yes(1)</td>
<td>yes(2)</td>
<td>yes</td>
<td>no limit</td>
<td>$39.95 (3)</td>
</tr>
<tr>
<td>GREP</td>
<td>various</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>1 line</td>
<td>free</td>
</tr>
<tr>
<td>ResNoter</td>
<td>1.0</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no limit</td>
<td>$35.00</td>
</tr>
<tr>
<td>Search</td>
<td>1.3</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no limit</td>
<td>free/$25.00</td>
</tr>
<tr>
<td>SeekEasy</td>
<td>5.0</td>
<td>no</td>
<td>yes (4)</td>
<td>no</td>
<td>2 lines</td>
<td>$30.00</td>
</tr>
</tbody>
</table>

(1) Golden Retriever uses Boolean logic to match only adjacent words.
(2) Golden Retriever allows the user to adjust the level of 'fuzziness.'
(3) A "sample" version is available through computer BBSs.
(4) SeekEasy's fuzzy search is not adjustable.
SeekEasy

SeekEasy, version 5.0, shareware — $30.00. SeekEasy is a slightly speedier but much less successful implementation of “fuzzy” searching than Golden Retriever, with considerably less flexibility. You cannot adjust the “fuzziness” level, and the program is limited to two lines of context around each word or phrase it finds. There is no way to save the results of the search to a file. In its favor, SeekEasy is extremely easy to use; type what you are looking for and, if it is in the file, SeekEasy will find it. Unfortunately, since it will retrieve the 100 closest matches (in no particular order), it will find a great deal of material you don’t want, and you may have to search through all of that to locate what you asked the program to find for you in the first place. This program would be more useful for keeping an address list than for data searching.

SeekEasy is available from computer bulletin boards or for $10 from the National Collegiate Software Clearinghouse. The author requests a $30 contribution if you make use of the software, and that also entitles you to updated editions, when they are released. For more information, contact Correlation Systems, 81 Rockinghorse Road, Rancho Palos Verdes, CA, 90274.

Other Programs

In the course of this section I have limited myself to a discussion of public domain and shareware programs, with the exception of ResNoter. All of these programs are available for less than $50, and some can be had for free.

Potential users should be aware, however, that a large number of commercial programs exists designed for similar purposes. These include Ask Sam, Gofer, ZyIndex, Notebook II+, FYI3000, and the word processing package Nota Bene, which includes an interface to the FYI3000 text database system. Some conventional DBMS packages, such as DBASE IV, have added features that allow them to cope with large bodies of textual data as well. Potential users should also be aware, however, that the price of these programs can range from the moderate to the stratospheric. While I would not discourage anyone from investigating some of these programs, I have not found that the increased costs purchase significant increases in power or sophistication in TRPs.11 What the increased costs do buy is support. If you are uncomfortable with, or inexperienced in the use of computers, it may be worth spending some extra money to gain access to software support personnel. For those who have moderate computer experience, however, public domain software and shareware come very close to being the proverbial free lunch, and I would encourage you to investigate those sources first.

Summing Up

Quantitative researchers, with their relatively simple data, have been the first to benefit from the computer revolution. But the increasing speed and power available through microcomputers makes even the complex textual data of qualitative researchers more accessible. This article has described the ways in which common, inexpensive, TRP systems may be useful in dealing with large quantities of interview data. The approaches described here can be applied as well to other textual data — field notes, for example, or archival research entered through text scanners. Any textual data can be made more useful through the application of simple computerized tools.

The availability of these tools does not, however, absolve the analyst of his or her responsibility. TRPs can only retrieve and display data — they cannot understand what those data mean, and they will willingly supply answers to queries whether those queries are motivated by theoretical understanding or conceptual blindness. Computers are always increasing in power, but never in intelligence, and it is worth remembering the first principle of data processing — GIGO 12 — whenever one sits down at a keyboard. Always think of the computer as an exacting but unimaginative research assistant, and you will not go far wrong.

Finally, you should be aware that TRP programs are rapidly increasing in power and flexibility and that, by the time you read this, there will probably be new versions available of most of the programs discussed here and a host of new TRPs as yet undreamed of. To find out about the latest programs, contact your computer center or local users’ groups. A little time spent looking at the available TRP programs will be rewarded with a simple but powerful data retrieval tool.

* For helpful comments on earlier drafts of this paper, I would like to thank Theresa Marchant-Shapiro, Charles Tidmarsh, Martha Huggins, Renata Tesch, and several referees, who shall remain nameless.

1 Presented at the IASSIST 90 Conference held in Poughkeepsie, N.Y. May 30 - June 2, 1990.

2 GREP stands for general regular expression print. Regular expressions are ways of expressing complex patterns of letters and numbers. GREP was originally designed to search through lists using these expressions and print the results on a teletype terminal.

3 Public domain software is a body of programs placed by their authors into free public circulation: the programs can be freely copied, used and given away but cannot be sold for profit. Computer hobbyists often trade these programs and they are also available through electronic bulletin boards and services such as Compus-
serve. Finally, there are companies that sell public domain software through catalogs for a 'copying fee,' which is usually no more than a few dollars per disk. Public domain software should be differentiated from shareware, where the author of the software freely distributes his or her programs, but asks for a contribution from those who use them. Both public domain and shareware are excellent sources for useful and unusual programs. Note however that, for your own peace of mind, you should carefully test such programs. Never test new programs on the machine you use for storing interview transcripts and book chapters: Programmers sometimes, though rarely, accidentally release programs with bugs in them, and it's best to find out without destroying irreplaceable materials.

4 GREP is an extremely flexible tool, capable of rapidly seeking out particular patterns in your text database. Rather than go into details here, however, I refer you to the support personnel at your institution. If you have access to a UNIX-based computer, however, you should be able to get a comprehensive overview of GREP with the following command:

`man grep`

This will display the pages of the UNIX manual dealing with GREP on your terminal. These will give you some sense of the power of the program. If you are using an MS-DOS based computer, on the other hand, your MS-DOS manual will give you an outline of how to make use of the FIND program.

5 For computational purposes, and for the purposes of this paper, a paragraph includes all single-spaced text that occurs between sets of double carriage returns.

6 All of the quotations in this article are constructed based on a series of unstructured interviews I conducted in 1988-89.

7 Typically, you have an ASCII file if, when you use the MS-DOS "type" command to show your file on the screen, lines end without wrapping around from the right to the left, and you see only alphabetic, numeric, and punctuation characters on the screen.

Unfortunately, most of the more powerful word processors do not create pure ASCII files. WordStar and WordPerfect, to name two popular word processing programs, include special codes in their files to make printing easier. Such codes must be stripped out if the file is to be searched by most TRPs. Some word processing programs solve the problem of these codes with a built-in option to save files in ASCII format. In WordPerfect, for example, you should save your transcription into a "DOS TEXT" file. This will be an ASCII version of the file, with all special characters removed. For many other word processors, you will need a special conversion program. For the most part, such programs are available free or at a nominal charge. If your word processing program is incapable of writing an ASCII file, go to your college or university microcomputer lab or computer center, and explain what you need to do. They should be able to help you find a suitable conversion program.

8 Fuzzy searching is a term that covers a great deal of ground. In general it means one of two things. If they do not find an exact match, some programs will look for words or phrases that contain many of the same characters in the same order as the search phrase. Others will seek words or phrases that are phonetically similar to the search phrase.

9 We might think of each 'record' as being made up of a chain of dummy variables (words). Each word in the record indicates the presence of a characteristic, and the absence of a word indicates the absence of that characteristic. In any given search, therefore, we are trying to discover whether particular dummy variables are present or absent. At the same time, we will be ignoring most of the variables in a particular record — all of the words that do not appear in a search command.

10 Given that you can only search for words that you have explicitly coded, you may want to think carefully about the amount of work involved in such coding before choosing this TRP. Its power comes in large part from a great deal of time and preparation on the part of the user. Coding 80 pages of interview (the outcome of the three hour interview I used to test programs) is, to say the least, a nontrivial investment of time.

11 While some commercial programs may have slight speed advantages over their public domain competitors, the major constraint on search speed is likely to be the access speed of the hard disk in your computer. Since this affects all programs equally, and is the major constraint on data retrieval, retrieval speed should not be given undue weight in deciding between two TRPs.

12 "Garbage in, Garbage out."
Report On Possibilities For A Photograph Database

by Adam Engst
Consultant
901 Dryden Road #88
Ithaca, NY 14850

Introduction
As Gould Colman explained it to me, the transfer of the bibliographic information on the archival photograph collection to a computer database is currently open to a large number of possibilities. While there are advantages and disadvantages to programs on a number of machines, there are no external forces currently requiring a certain solution. I was retained to research the possibilities and present my findings, either recommending hardware and software combinations to test or recommending that the Archives wait several years before repeating this process.

I have gone through several steps to come up with this report. First, I tried to determine precisely the needs and desires of the Archives. Second, I researched the software possibilities on three different hardware platforms which are available to the Archives, the Macintosh series, the IBM PC line, and a general category of mainframe. Third and finally, I weighed the advantages and disadvantages of various combinations, adding in my knowledge about the Cornell community, the state of the software industry, and the history of some companies in particular. Most of the information below comes from my notes on telephone conversations held with representatives of the various companies.

Questions
Since the Archives is not locked into using any specific program or computer, I was left to figure out what sort of a system would best fit the needs of the department. As I see it, there are some requirements placed on any system by the size of the data, the nature of the data, the use to which the data is put, and the cost of the hardware, software, and programming time.

Size and speed
Gould told me that the Archives currently has between 50,000 and 100,000 photographs. Assuming one record in the database for each photograph, the system must be able to handle 100,000 records with decent searching speed. This was the first question I asked of the various database companies. Their replies must be taken at face value though, since the only way to really test the each system is to put 100,000 representative records into each and do some searches.

Keywords
Since a small number of photographs are the end result of any search through the database, the system must be able to handle a relatively large number of keywords, or else researchers will have difficulty narrowing down their searches. A number of databases require programming convolutions to be able to deal with a field containing an unknown, but potentially large, number of keywords. Such a limitation does not rule out a database, it simply downgrades it in terms of ease of setup and programming. In addition, selecting the keywords is an extremely important task which must be thought out carefully.

Graphical information
Bibliographic information is useful for providing a brief description of each photograph and locating it within a collection, but for a researcher who is trying to find a certain photograph, possibly out of hundreds of similar ones, bibliographic information will not allow that researcher to select a certain photograph with surety. A graphical method of describing the photographs would decrease the amount of time it would take a researcher to find the right photograph for the use he or she has in mind. The main possibility is displaying images on a videodisc. Few of the database systems can directly control a videodisc player. There are serious cost drawbacks to displaying visual information though, so inability to control a videodisc does not disqualify a database.

Initially, scanning the images and storing them on a CD-ROM would seem to be feasible because each CD-ROM can hold 600 megabytes of information. Unfortunately, CD-ROM is not feasible because a scanned photograph has an average file size of 300K, which, when multiplied by 100,000 photographs, would force you to use close to 50 CD-ROMs holding 600 megabytes each. In comparison, a single videodisc can hold 108,000 images.

Costs
The cost of the software are minimal in comparison to the costs of transferring the images to videodisc, although the purchase of expensive mastering equipment can reduce the overall costs. Another cost which cannot
be ignored is the cost of programming and setup with whatever software is decided on. As a result, ease of programming does play a financial role in the final decision as well.

So the questions that I asked each database company were as follows:

• Can your program handle 100,000 records with a fast search speed for a single record, say under 10 seconds as a worst case scenario?

• Can your program handle unlimited length text fields in an index (to retain searching speed) or is there a simple way around the program’s inability to do so?

• Is there any way for your program to access images stored on a standard videodisc player?

• How hard would it be to set up your program with a simple interface for researchers who may be inexperienced with computers?

I also tried to get a feel for each company—how easy they would be to work with, how much help they would be if we needed any technical support, and whether or not they would still be in business in several years. These are intangibles, but potentially useful pieces of information.

A note before I get into the details. I’ve tried to write this so no technical knowledge is required to understand it. I’m sure that in some places I have failed because there is simply no other way to talk about certain features and actions of computers. In those places, I’ve included a footnote or tried to explain the term I use within the text. If at any time, you are confused reading this, please call me, and I will attempt to clear up the source of the confusion.

Hardware

The companies with whom I spoke have database programs that run on the Macintosh series of microcomputers, the IBM PC line of microcomputers, and (in the cases of Oracle and NOTIS) almost all minicomputers and mainframes. The Archives currently has several IBM PCs and clones and will be getting several Macintosh SE/30s shortly. In addition, I gather that the department has access to the mainframe resources of the library and the university. So existing hardware does not bias the decision.

With a few exceptions, all of the software packages I researched can handle the large size of the database without a loss in searching speed. Obviously, the minimum (and preferred) hardware configuration in each case does vary slightly, although some packages run fine on less powerful machines, which is a bonus since it will reduce the costs.

In general, and like all generalizations this one is not to be trusted completely, the Macintosh will be the easiest to set up and for both researchers and staff members to use. IBM PC clones have the advantage of being in the majority, although powerful systems are not really much cheaper than Macintosh systems. Microcomputers have the advantage (and disadvantage) of local control—if something goes wrong with the computer you can have it fixed quickly if necessary, whereas you must wait for another department to respond to your problem with a mainframe. On the other hand, if something with a microcomputer fails, you must deal with it, unlike with a mainframe, which will have a staff to deal with problems. Mainframes often suffer from poor interfaces as well, although there are ways of avoiding the poor interfaces.

Using a Mac and HyperCard with a videodisc is simple, while using a PC with a videodisc requires a special device driver, which is a small program which allows the computer to control the videodisc. Such programs are available, often from the videodisc maker, and there are also programmers who could write a custom device driver if necessary.

General hardware conclusions

Based on my experiences with the various types of computers and my knowledge of the Cornell user community, I recommend using a Macintosh. Macs are predominantly easier to work with in the setup phase, and they are far easier for inexperienced users to work with. Since the entire point of this project is to provide easy access to information, I think that the interface is one of the most important parts of the system, and better interfaces can be created on the Macintosh. In any case, my research covers all three platforms, and I hope that my recommendation of a hardware platform is born out by the software possibilities on the Macintosh. In addition, the Macintosh database companies were far more knowledgeable about controlling videodiscs, which is why I often have more information on the Macintosh databases.

Macintosh Software
Company: 1stDesk Systems
Program: 1stTeam
Hardware: Mac
Price: $795

Their powerful relational database, called 1stTeam, is compatible with HyperCard and can store up to 255 characters in each field. Offhand, 255 characters doesn’t sound like it would necessarily be enough for our keywords, but
perhaps their literature will shed more light on the subject. Otherwise, 1stTeam is certainly a possibility because it should be fast enough and can control a videodisc through HyperCard, although it is much less well-known than either 4th Dimension or Omnis 5.

Company: ACIUS  
Program: 4th Dimension  
Hardware: Mac  
Price: $695  

4D can handle 100,000 records with no problems, but it would have trouble with indexing. Without indexing, the search speed slows tremendously, but 4D cannot index its unlimited length text fields, which we would use for holding keywords. So setting up the keywords would be a little tricky in 4D. There are a number of different ways around this problem, but they would require a bit more work programming. In the first French version, there was some kind of external command which could control a videodisc, although they may not still exist in the current version. There is a demo database, called Minifans, which we could look at if 4D turned out to be a likely candidate. Against 4D, I’ve heard that it is one of the slower databases for the Mac, which is a problem for this project. A test of its speed would definitely be needed before I could recommend it any farther.

ACIUS is one of the major database companies for the Mac, but I have been unable to get through to them at all, which may indicate mediocre customer support. While this is not a complete argument against using 4D, it doesn’t bode well for future support needs. I can’t really recommend them unless I can get through to talk to them.

Company: Blyth Software  
Program: Omnis 5  
Hardware: Mac and PC  
Price: $695  

Omnis 5 from Blyth Software certainly has the power to deal with 100,000 records, and it can be extended to do even more such as control a videodisc, although the representative didn’t think such an external command had been written so far. Alternately, either Blyth could do it for us for free if it was small and fairly easy or an independent programmer might be willing to write such a thing for a fee. Omnis can index variable length text fields, so it would have no problem with a field containing a variable number of keywords.

If the software to control a videodisc was difficult to write or acquire in other ways for Omnis, it can work with HyperCard so that HyperCard uses the Omnis database while acting as a front-end. However, Omnis can also create simple interfaces easily, so it should not be necessary to link the two together on that account. Omnis’s language is supposedly English-like and easier than most database programming languages. An advantage of Omnis over any of the HyperCard extensions is that Omnis is a full-fledged database, and as such, can generate reports and display multiple windows, which would be good for displaying a number of records which met the search criteria.

Omnis needs a minimum of 1 megabyte of memory and is happier with a fast machine and more memory. Overall, I was quite impressed with the possibilities of using Omnis 5, since it seems to meet all the requirements and be fairly easy to work with in addition. The representatives have been extremely knowledgeable and responsive, unlike some of the other companies, such as ACIUS.

Company: Fox Software  
Program: FoxBase Plus and FoxBase/Mac  
Hardware: PC/Mac  
Price: $395/$495  

FoxBase can have up to 254 characters in text fields, and can search on unlimited length text fields, but they aren’t indexed which slows the search. However, Fox claims that FoxBase can handle up to 1 billion records, and that it is the fastest of all the Mac databases by a great deal (some 30 times faster than 4D), although some of the PC databases come close in speed. Reportedly, a new version of FoxBase/Mac can use HyperCard’s external commands (such as the ones to control a videodisc player) directly, which would be a major point in its favor.

FoxBase is generally accepted to be better than DBase III+ on the PC and the Mac version is correspondingly good, if not better since the Mac version can handle unlimited length text fields. FoxBase cannot control a videodisc, although it could work with a CD-ROM. Despite FoxBase’s speed and file compatibility between machines, I think it is somewhat too limited in this situation because of its inability to index unlimited length text fields and its inability to control a videodisc. In addition, if it crashes for any reason, it will often corrupt the entire database rather than just losing the last record entered. This is a serious problem because you can never predict crashes.

Company: Odesta Corp.  
Program: Double Helix II  
Hardware: Mac
Price: $395

Double Helix cannot link to HyperCard and cannot control a videodisc, although there is a version that runs on Vax mainframes (for about $5000) which would help the speed and storage problems. Despite the fact that Double Helix can handle 100,000 records and has a simple method of programming, I doubt that this program is a real answer. Double Helix simply doesn't have enough to recommend it over any of the other major databases except its idiosyncratic programming environment, which may be easier than most.

HyperCard extensions

All of the following products require HyperCard, or one of two HyperCard clones, SuperCard or Plus, which provide the same basic features as HyperCard but with significant extensions. If a HyperCard system is decided on, it would be well worth the time to investigate creating the database in SuperCard or Plus rather than in HyperCard itself. The various extensions listed below may or may not work with SuperCard or Plus, although there is a good chance that they will. The areas in which SuperCard and Plus go beyond the capabilities of HyperCard include reporting, graphics, multiple windows, and color. Whether or not these features are worth moving away from HyperCard is another question entirely, and one that need only be asked if the Archives decides to go with HyperCard rather than one of the full-fledged databases.

Company: Answer Software
Program: HyBase (under HyperCard)
Hardware: Mac
Price: $150

Size and speed are not problems, since HyBase can handle up to 2 billion records and can usually find a single one in about 5 seconds. All fields are unlimited in size, or at least very large. The company claimed that it is not difficult but that some programming experience is helpful. Answer Software could set up the database for us if necessary. However, Gregory Crane at Harvard said that he used HyBase on Project Perseus for a while and found it very difficult to work with.

Dealing with the people at Answer Software was rather difficult and based on Gregory Crane's advice, I don't think that HyBase is a good possibility. It suffers from difficult set up, which is unnecessary for this project.

Company: Discovery Systems
Program: HyperSearch (under HyperCard)
Hardware: Mac
Price: $99

I have not yet received any information from Discovery Systems regarding their HyperSearch package, so I cannot make any specific statements for or against it. However, Library of Congress is using it in their American Memory project, and they seemed pleased with its speed and ease of use.

Company: KnowledgeSet Corp.
Program: HyperKRS (under HyperCard)
Hardware: Mac
Price: $195

HyperKRS works completely within HyperCard so it would be simple to design the database. Nothing else need be done in terms of setup except for generating the index, which is fairly slow, but only needs to be done once. A Mac Plus is all that is required for searching. HyperKRS was designed for CD-ROM, which accounts for its speed. I tested the demo software they sent me and I wasn't remarkably impressed. I had trouble finding anything, mostly because I was unfamiliar with the information for which I was searching. The speed was good but not great, but my Macintosh is not that fast, which is certainly an issue with this program.

On the whole, HyperKRS sounds like it may be the simplest of all the HyperCard extensions to set up initially. After that, I have no real numbers to compare its speed with HyperHIT or Xearch.

Company: NovaSoft Engineering Group
Program: GridFile (under HyperCard)
Hardware: Mac
Price: $195

NovaSoft has a sample application called ClipFile for GridFile which is being used right now to access pictures on clip art CD-ROMs. We would have to do the indexing and setup ourselves, which would be difficult without the aid of a relational database expert. GridFile is extremely fast, though, and is able to search any database for a unique record in 3 disk reads (certainly under 1 second).

The cons of GridFile include the fact that it requires 2 megabytes of memory and a fast hard disk; it does not provide as good data packing as some other databases, which makes the file larger; it slows down on smaller databases in comparison to the others; it doesn't support split files over two or more hard disks; and it would be hard to set up.

The pros of GridFile are that it is bindingly fast (faster even than some mainframe databases) and that it uses HyperCard as a front-end, which can then control a videodisc.
On the whole, I think GridFile is very powerful, but possibly too difficult to work with. There are other programs which provide similar speeds, but are easier to work with and require less hardware.

Company: SoftStream International
Program: HyperHIT (under HyperCard)
Hardware: Mac
Price: $195

Steve Hannaford, the technical support representative for HyperHIT said that HyperHIT has extremely fast searching speed (~1 sec) and can handle unlimited length fields with no problems. The information does not need to be textual—it could be pictures or sounds. The setup is not trivial but not that hard, and the HyperHIT system is entirely contained in external commands that work within HyperCard. Steve didn’t think it would be hard for someone without database training to use.

The data file is external to the HyperCard stack which would control the videodisc and thus requires only a small amount of space. Another advantage to the external data file is that there could be a number of different interfaces since the HyperCard stack does not have the data embedded in it. Search time is usually under 1 second with a Mac Plus, and it would drop with a faster machine and hard disk. There is little speed degradation when the file size increases (3 hundredths of a second when going from 1,000 records to 10,000 records). In fact, the videodisc might be the bottleneck, depending on how fast it can find each frame.

Steve Hannaford was very helpful and said that he wasn’t getting many calls as the technical support person for HyperHIT, which could mean that people aren’t having any problems worth calling about. The advantages of HyperHIT are that it is extremely fast despite what sort of machine it runs on, it supposedly isn’t difficult to set up (although Gregory Crane will be testing it for Project Perseus soon and will have an opinion on its ease of use), and it will allow simple interfaces and videodisc access through HyperCard. Overall, HyperHIT sounds like a good possibility.

Company: The Voyager Company
Program: VideoStacks
Hardware: Mac
Price: $99.95

VideoStacks is a set of external commands to control a videodisc for a number of different videodisc players. There is a possibility that some of the external commands would be available from Apple free of charge or they might be distributed with the videodisc itself. Some sort of videodisc drivers will be necessary.

Company: Xiphias
Program: Xearch
Hardware: Mac
Price: ????

Xearch is an external command for searching in HyperCard which Xiphias uses in their CD-ROM-based product, Time Line of History. It sounds like it would be fast enough and they do have a licensing agreement, although I don’t yet have the details.

Initially Xearch sounds like it could be quite useful, although I don’t have a sense of how easy or fast it is in comparison to HyperHIT or HyperKRS.

General HyperCard software conclusions

I think that all of the various packages mentioned above will probably provide HyperCard with the searching speed necessary to use the database. The main distinction then, lies in the ease with which each is set up. HyperKRS and Xearch are probably the easiest, with HyperHIT, HyBase, and GridFile lining up in increasing order of difficulty. More specific research and testing would need to be done to determine speed and ease of use in order to choose between the various extensions. HyperHIT may be the best compromise between speed and difficulty.

General Macintosh software conclusions

I am of two minds in this category. I think that HyperCard is a wonderful program (not to mention the fact that it is free with all Macs), and it will become integrated into the Macintosh hardware and system software in the next few years, making it even stronger and faster. On the other hand, it really is not a database and does not provide the features that a full-fledged database provides, such as reporting and fast searching. It may be necessary to use HyperCard in some fashion to facilitate access to a videodisc, which lends strength the cases of those database products that can link to HyperCard, such as 4D, Omnis 5, and 1stTeam. On the other hand, the structure of the proposed database is very simple and does not really require the full power of a relational database. In the final consideration, I think I would currently recommend Omnis 5 because of its power, flexibility, and ability to link to HyperCard, not to mention the quality of the customer support, with which I was very pleased.

PC Software

Company: Ashton-Tate
Program: DBase IV/Dbase Mac
Hardware: PC/Mac
Price: $795/$495

Neither DBase IV nor Dbase Mac have any internal way of controlling a videodisc, and the representative didn’t know of any external ways either, although he thought one might be possible. In addition, neither can index on variable length text fields, which would slow them down a great deal for this purpose. Add these problems to the fact that Ashton-Tate is undergoing major problems as a company and has publicly announced that they will not be upgrading Dbase Mac at all, and you get a company to stay away from.

Company: Borland International
Program: Paradox/Reflex Plus
Hardware: IBM PC/Mac
Price: $725/$279

The Borland representative didn’t think that either Paradox, the more powerful PC program, or Reflex Plus, a decent Macintosh database, could control a videodisc. It took several phone calls and some time on hold to get that much information, so I didn’t pursue it farther. However, Tim at Turquoise Film/Video Productions (one of the mastering services) said that he was thinking about re-writing his custom database in Paradox because it was fast and fairly easy to work with. He also said that Paradox runs on a number of machines and is probably file compatible with Reflex. As a result, Paradox sounds like the best of the PC databases that I’ve looked into. Using Paradox would require some additional device driver to control the videodisc, but such a program might be available from a number of sources, including Turquoise Productions. Paradox won most of the speed tests I saw in the course of my research, so I would recommend it over the other PC databases based on what I currently know.

Company: DataEase International
Program: DataEase
Hardware: IBM PC
Price: $700

DataEase cannot control a videodisc, although it supposedly can interface with three scanners for using pictures. However the storage of those scanned images would be ridiculous and dealing with graphics on the PC is more difficult than on the Mac. DataEase does have long text fields which can be searched, although the representative wasn’t sure about whether or not they were indexed, which is a major concern. I have a demo disk from them which may answer the indexing question, although I see nothing special about DataEase otherwise.

Company: Image Concepts
Program: C-Quest
Hardware: PC or Unix mainframe
Price: $6000 or $25000

C-Quest is a proprietary system for storing photographic information and controlling a videodisc. It has been around for several years, but doesn’t seem to have a devoted following.

Clif Nickerson of Image Concepts was somewhat helpful, although his system is designed more for a stock photograph collection than a historical research collection. The main evidence of this is the way it uses synonyms of keywords, a method which allows the user to search on “Stream” and get “Brook” and “River” and “Run” and “Creek”. Unfortunately this is not nearly as useful with proper names of people and places, since they tend to be specific. The only use I can think of it is use modifiers, so you could have Frank Rhodes walking, talking, shaking hands, or making a speech, and search on the action involved. I don’t know if that is too much trouble to set up and key in or not.

C-Quest runs under Unix mainframes as well as PC clones. Under the Unix system, C-Quest can display 18 pictures at once; on the PC it can only display one at a time. Its speed is dependent on the number of subjects used in the search, but Clif said something about speeds of under 1 second, which he said was faster than the mainframe database Ingres (and he thought than Oracle). The C-Quest interface is menu-driven and not particularly good. It does not have a simple interface for researchers to use, although one is being proposed. Image Concepts will change, add, or remove fields from the menus for a nominal fee, which is not as good as setting it up oneself. C-Quest is not cheap, by any means, at $6000 for the PC version of the software and $25000 for the Unix version.

Clif said that the easiest way of getting images on disc were to buy a video camera and a writable videodisc, at which point you could do it all in-house. I suspect that quality wouldn’t be as good, although there is no way to know without trying. He recommended making a 35mm film image in case the resolution of the monitors increases enough to make it worthwhile to re-master a videodisc.

C-Quest has an impressive list of clients, although I suspect that is from being the only game in town for 4 years, since no one else does this on the PC at all. Evidently, the Library of Congress system is slower than C-Quest, although that doesn’t really

Spring 1991 51
mean much without more details. C-Quest can control videodiscs from a number of companies, such as Sony, Pioneer, Philips, and Panasonic.

Overall, I find their system to be somewhat clumsy, expensive, and not really suited to the needs of the Archives. The Archives photographs have specific subjects without synonyms and need a very simple interface for researchers. While $6000 is not truly expensive in relation to the cost of mastering the videodisc, I think it is quite a bit more than you would pay for any other system. It might require less setup initially, although it is still a generic program that would require some customization. I cannot recommend it, especially since I heard from another consultant that the version of C-Quest at the United Nations was actually quite slow.

Company: Microrim
Program: Rbase for DOS
Hardware: IBM PC
Price: $725

Rbase will handle an unlimited number of records but the representative didn’t know if it could handle an unlimited length text field. I didn’t want to hold any longer to find out if it might be able to control a videodisc since the representative didn’t think so. I see no reason to specifically recommend Rbase.

Company: Symantec
Program: Q&A
Hardware: IBM PC
Price: $???

Q&A does not have variable length text fields, but it can have large ones which are indexed so that the search speed doesn’t suffer. The speed isn’t good, though, at 15-20 seconds average, partly because Q&A is not a high-powered relational database. There is no videodisc access, although the representative thought that an external program might work. Considering the speed problem, I can’t recommend looking any further at Q&A.

General PC software conclusions

I think most of the major PC database programs will handle the textual part of the database without trouble. However, it seems as though it will be more difficult to link the textual information in a PC database to the frame numbers of a videodisc. These device drivers do not seem to be readily available or supported by the database companies. For instance, the representative of Ashton-Tate knew nothing about linking to a videodisc, yet supposedly the Library of Medicine is using DBase III+. However, a device driver might come with the videodisc player. I also feel that it will be more difficult within these PC programs to create a foolproof interface for researchers who are inexperienced with computers. I missed at least two major databases for the PC, Revelation and Nutshell, because I was unable to find phone numbers for them. However, I am not particularly worried that they are the perfect database because none of the other PC database companies had much of an idea what a videodisc even was, much less if their program could control it. The PC database companies were also much harder to reach on the telephone and much less willing to talk. If someone else turns out to be using Revelation or Nutshell, it would be worth checking them out. Otherwise, I think Paradox will be the best on the PC side.

Mainframe Software

Company: Oracle
Program: Oracle (runs under a HyperCard front end on the Mac)
Hardware: Mac/PC/minicomputers/mainframes
Price: variable depending on version—from $299 to $1299

Oracle for the Macintosh is a port of the most popular database program in the world. It retains complete compatibility with all other Oracle databases on all other machines, which is a plus if this data will be shared with other people. In addition, a Macintosh running the HyperCard front-end to Oracle can use any Oracle database on any machine. Because HyperCard is the front end to the actual database, Oracle can control a videodisc through HyperCard. Should the Archives wish, they could probably find a mainframe or minicomputer on which they could use Oracle.

Oracle’s advantages are speed, portability, and ease of use with HyperCard, although it might be a bit more expensive than the Archives would want initially. There is a Developer’s Version for the Mac for $299, which would allow us to test its capabilities (with the only limitation being that this version cannot link to other Oracle databases on other machines). The main disadvantages to Oracle are that it is potentially more expensive (although the Macintosh version is quite cheap) than other databases, and that it may simply be too complicated for the relatively simple database information we have. I gather that setting up a database in Oracle is not all that easy. In addition, I’ve heard that Oracle for the Macintosh is not as fast and occasionally does strange things to data files.

Oracle as a company is excellent, with toll free support and guaranteed stability. They are the
largest database company in the world and the third largest software company in the world.

Company: Northwestern University
Program: NOTIS
Hardware: IBM mainframe
Price: free

NOTIS has a number of advantages, although it also sports major several disadvantages. NOTIS is currently installed and running in the library, so there are no added software or hardware costs to the system other than a Macintosh and videodisc from which people can search in the Archives. It is relatively fast and can certainly handle another 100,000 records in its database. NOTIS has the advantage of being accessible from anywhere on campus, but researchers may not use it unless they can also see the videodisc images because it is difficult to search for photographs based solely on bibliographic information. It has been in use at Cornell for some time now, so many people are familiar with its interface, although its interface is also one of its main disadvantages.

Searching and moving between the various results of a search in NOTIS is difficult and completely not intuitive. Its other main disadvantages include the fact that it would be very difficult to link it to a videodisc, if it is possible at all, and the problem of portability of data since NOTIS does not have the ability to export its information to another program, something which all of the microcomputer databases can do and which is very important for future expansions or modifications. It might be possible to sidestep NOTIS’s poor interface with a HyperCard interface currently being worked on at Mann Library. In addition, there is a commercial product that will be available soon from Texas A&M and Apple, called MacNOTIS, which also provides a better interface to NOTIS. Because both of these products use HyperCard, it is theoretically possible to have the HyperCard interface control a videodisc while using the information from the NOTIS database. Howard Curtis of Mann Library thought that this was possible, although extremely clumsy and prone to break whenever either NOTIS or HyperCard changed much. Other people thought that it would be an unworkable situation even if it was theoretically possible. Howard also said that it might be possible, though difficult, to program HyperCard to download records from NOTIS to the Mac, which would allow the records to be used by microcomputer databases.

NOTIS is an easy solution because it requires no new hardware or software, but putting the records into NOTIS removes them from a certain level of accessibility. It would be difficult and clumsy to attach a videodisc to a Macintosh running one of the HyperCard interfaces, if it is indeed possible at all. More research would need to be done to determine the reality of such a setup. Even worse, it would be hard to transfer those files to any microcomputer system. However, there are some ways of moving from microcomputer databases to a format which NOTIS can read, which points towards putting the records into a microcomputer database first, and then, if there is interest, transferring a copy to NOTIS. As much as NOTIS seems like the simplest solution, I don’t feel comfortable recommending it given the possibility for videodisc access and the inaccessibility of the data once it is in NOTIS. I realize that NOTIS data can be shared by other mainframe cataloguing databases, but they don’t (on the whole) provide the kind of features that microcomputer databases do. Being able to move data between systems is important, and customized mainframe databases are a blockade to such a move.

Videodisc Mastering Services

Company: Image Premastering Services
Program: Videodisc services
Hardware: NA
Price: variable

Image Premastering Services claims they are known for having the highest image quality for still frame transfers. Some of their main clients have been the United Nations, the Library of Congress, the Mayo Clinic, and the American College of Radiology. They can handle absolutely any original—for the Library of Congress they laid down 30,000 glass plate negatives without cracking any. Of course, slides are the cheapest method, and run anywhere from 55c to $1.35 per slide. Other original media are correspondingly more expensive, although presumably the cost goes down with quantity. In addition, there are some basic initial costs which cannot be avoided. These costs total $3150, although that is minor compared to the cost of transferring 100,000 photos to the disc.

$2000 for the master disc
$500 for a check disc
$150 for the videotape
$150 for a duplicate/backup, kept at their site
$350 as a basic setup fee

They claim that Stokes is mainly a slide copy service and makes a 35 mm film negative, which is a second-generation picture of the original. If the
image is from a print, then the videodisc image is third-generation picture and suffers correspondingly in quality. Stokes does color-correction, so the colors may be bright, but they are likely to be inaccurate. Stokes is also generally cheaper because everything is automated in their process.

Image, on the other hand, is specifically dedicated to mastering videodiscs and they have patented technology for the process. They use a 2 foot lens over a 12 foot optical bench, which gives them two advantages.

1) The light comes in at a perfect 90° angle, which gives much better edge definition to the image.

2) They use an aerial image transfer, which somehow projects the image so that there is no film grain in the resulting videodisc image. It also allows them to easily perform custom sizing.

The Image representative recommended the Mac and said that a military project used HyperCard with Oracle. He had heard something about 4D, but didn’t know of anyone who was using it. He didn’t recommend using the PC at all because the hardware is more expensive and is harder to set up the software to interface easily with the videodisc.

Company: Stokes Mastering Services
Program: Videodisc services
Hardware: NA
Price: variable

I spoke with John Stokes and Jim Couch of Stokes Mastering Service. In regard to costs, Stokes estimated that a basic image transfer of positive images would be somewhere between $2 and $3 per image, although his estimate for complete costs (ie. in-house handling and database work) was closer to $4.50 per image. The work is done on-site and includes a person to come and do it with Stokes’s somewhat specialized equipment. There is not much difference between Stokes’s doing the work and it being done in-house except for the fact that he claimed they had higher quality control, which is fairly likely. I suspect this is somewhat cheaper than Image Premastering’s prices, although not by as much as I had originally thought.

They have a number of projects going on, the most notable of which is for the Library of Medicine, whose cost was about $2.40 per image. That price is slightly inaccurate because it was a test run in some ways and the library got two sets of negatives and slides for each of 70,000 images. The person to talk to at the Library of Medicine is Lucy Kiester, phone number 301-496-5962. The Library of Medicine is using a PC with DBase III+ for their database. Stokes claimed that videodisc access was incredibly simple with any database and that you didn’t need a custom driver, but he did admit that you had to write some software. The Library of Congress is also working with Stokes, which is curious since the representative at Image Premastering said that the Library of Congress was working with them. Perhaps there are two different departments? In any case, Stokes claimed that the Library of Congress is using some in-house computer system rather than an off-the-shelf software package. Stokes said something about how that was their policy. This is not necessarily true since the American Memory project is using a Macintosh and HyperCard.

One advantage of Stokes’s method is that you can get negatives of each image as well, which allows you to reduce handling of the original images by making additional negatives. The Library of Medicine uses these negatives for public access to avoid giving out their originals.

Interestingly enough, Stokes said that more people are doing the imaging first, then the database work, partly because Stokes can transfer the images to videodisc faster than the database can be set up. I was unsure about the real reasons for this, but they could be determine by talking to some of the people Stokes referred me to.

As far as hardware goes, Stokes sounded like he doesn’t really know very much about the Mac. He claimed that the Mac had no advantage over the PC in ease of use if the software was designed well, although I disagree with that rather strongly. Based on a number of years of working with novices on both systems, the PC is less intuitive and clumsier than the Mac when it comes to user interfaces. In any case, Stokes has developed a database package under Informix (which is Oracle compatible, or so he said) which runs on a number of different machines. If we decide to use Oracle, we could buy the Oracle package and then Stokes would provide us with his custom database for only the cost of support. This is curious because he could very easily create a stand-alone Oracle database and then just sell that (or give it away if he wanted) without the customer having to buy their own copy.

Stokes wouldn’t really comment on Image Premastering except to give me the name and
number of Bill Perry (202-857-7537) at National Geographic, which did independent tests of both Stokes and Image Premastering (and one other, actually whose name Stokes did not mention). In addition, Stokes claimed that their quality has improved since then. Evidently, the American College of Radiology had 11”x14” X-rays which Stokes claimed were optimized for the Image Premastering system and those came out better.

As far as quality goes, any transfer from a positive image will lose quality in the transfer process, much as copying a tape or videotape loses quality. The contrast of a videodisc is usually around 20 to 25 with a maximum of 45, whereas a transparency is about 1000 and a slide about 250. Thus a great deal of contrast is lost when going to videodisc in any case. The transfer to a negative reduces this contrast lost by spreading out the contrast rather than clipping it, although it also puts it through several generations of imaging. Stokes is aiming at a contrast factor of two to three times better than high definition video, which is as good as a monitor will get in the near future. He said that it is very difficult to match the original exactly, and that matching the original better is their main task right now.

Company: Turquoise Film/Video Productions
Program: Videodisc services
Hardware: NA
Price: variable

Turquoise said that they can provide anything up to a turn-key system. Their background is in motion picture processing, and they moved from that to providing software and hardware as well. Their database is an in-house one currently but it can import and export to a number of other formats. They are thinking about re-writing in Paradox, which can also run on a number of different machines. They charge an average of $2 per image, including hardware, and they can shoot either in St. Louis or on-site. They use a special motion picture film to get better quality images, but I have no sense how the quality of their images compares to the quality of either Stokes’s or Image Premastering’s images. It doesn’t seem that there is anything remarkable about Turquoise in relation to the other two mastering services, but should the Archives decide to have a videodisc mastered externally to Cornell, it would be a good idea to talk more specifically to all three companies.

Other Cornell Projects
I spoke with Anne Carnell about the project in the University Photography Department, and she said that they are having someone in Publications develop an in-house program. This program will catalog and store all of the information on their photographs, but it will also provide billing, usage, and reporting capabilities. They didn’t think one of the commercial programs could provide all that, something which I doubt, given the power of some of these databases. They are looking at videodisc in the near future, for much the same reason as the Archives, perhaps in the next year or so. She didn’t seem to have a wonderful grasp on what is entailed with the entire technology, since she didn’t know about the problem with file sizes for scanned images, and she didn’t know how many images could be stored on a videodisc.

I also spoke with Dave Watkins, who is the head of Media Services and in doing so found my way back to the original Stokes project mentioned in the memos to and from Chris Pelkie. They are currently selecting slides, negatives, and prints for a free sample videodisc to be supplied by Stokes. Partly to test the quality of Stokes’s service, they are trying to assemble a number of different types of images for inclusion. If the Archives wishes participate in this project to see how it works out, they should contact Dave Watkins. He is looking for up to three hundred of the most difficult type of images. The deadline for submission to this project is January 1st, 1990, which is fast approaching.

Other departments that may be interested in this project include the Department of Entomology, which has 30,000 slides that they wish to use for diagnostic work, obviating the need to go to the slide collection itself. These slides must be of the highest quality because of their use and the fact that the disc will be sold to other universities. Plant Pathology and Veterinary Medicine may wish to do similar things. Evidently the Hotel School has a videodisc of wine labels and the Vet School and the Law School are still looking into the possibilities of some sort of image database on a videodisc.

If the Archives wishes to look more closely at the specifics of a videodisc system in future, I strongly recommend that the department provide a number of photographs to Dave Watkins for this test project. The project will provide a videodisc which can be shown to potential donors and with which we can test the pros and cons of various software packages. Such an opportunity should not be passed up lightly!

I met with Margaret Webster, who runs the Architecture School’s Slide Library. She has for some time been planning a videodisc project to keep track of the 350,000 slides in the library. They hired an outside consultant, Nancy Humphries of ETECH, to research the possibilities and provide a system. Margaret has been putting records into the database and will be setting up the pilot project after the Slide Library moves in January.
of 1990. They will be using a PC database called bricive/xtrieve (with which I'm not familiar because it was never compared in the literature with the more well-known databases) along with a specialized graphics board in the PC that will allow them to manipulate the images electronically. Manipulation of images is something which I did not explore particularly because of the expense involved, but is certainly a possibility for the Archives. The question that must be answered to justify the cost of such a board is the use to which these photographs are being put. If the photographs are ending up in publications designed and executed on a personal computer, or the images must be frequently manipulated, then a graphics board makes sense. However, if the publications in which these photographs appear use traditional methods of production, then additional 35mm negatives would be more useful. Margaret Webster knows more than most people on campus about videodisc systems because their system has been in the research phase for close to two years now.

The final person from Cornell with whom I spoke was Mike Oltz from the Interactive Multimedia Group. He gave me some bits of information that may be useful. He thought that the Architecture School and the History of Art department were looking into something similar and were probably working along different lines. As it turns out, History of Art has dropped their project entirely, whereas the Architecture project is the closest to reality of any of the ones I've heard of. The exception to this is the Medical School, which has a system for pathology training using a Macintosh pseudo-database called Guide. They started out with no funding at all and now have close to five million dollars of computer equipment, which does lend hope to the Archives getting funding for a videodisc.

The Interactive Multimedia Group has a sample videodisc from Image Premastering Services, which is currently lost, but Mike will try to find it and let me know via email. He mentioned that Revlon, the makeup people, are also doing something like this. If we wanted to learn more we should talk to the advertising photography department.

There is no reason to assume that any of the other systems are better than a system the Archives could come up with, although the ability to perform information transfer might be useful in the future to avoid re-keying records. Similarly, standard information in each record would help in translating the records from another system when other departments wished to archive various photographs. Database information can be shared between PCs and Macs without too much trouble, so the specific machines used by different departments should not really matter, although the ability to transfer between the various databases matters.

**Cornell Coordination**

One problem that has come up time and time again in my research is that there are a number of Cornell departments working completely independently on similar videodisc projects. While such a lack of communication is not unusual at Cornell, it is regrettable, particularly in a field such as this where the information really is fairly finite. It would be extremely useful if there could be a single person who would, if nothing else, have copies of all the various pieces of information collected by the different departments. That way, whenever anyone was thinking about starting such a project, the information would be more or less at hand and would include names of the people at Cornell who are good resources. This person would merely disseminate information and would refrain from making any recommendations as far as hardware or software go in order to avoid the politics.

The Interactive Multimedia Group would seem to be a logical group to coordinate or various videodisc information, but because they exist completely on soft money for specific projects, they are not set up to handle any sort of coordination. Geri Gay said that Media Services was one place coordination could come from, and some part of CIT Services would be another. People to talk to in CIT include Larry Fresinski, Donna Tatro, and if all else fails, Stuart Lynn.

Another area in which the various departments could pool resources would be in setting up facilities at Cornell for transferring images to videodisc. I gather that there are close to a million images at Cornell that could be put on videodisc if the process was cheaper and easier. Dave Watkins in Media Services is the person to talk to about such a project. Margaret Webster in the Architecture Slide Library would also be very interested.

**Other Non-Cornell Projects**

I've found names of people at other institutions who have done something along these lines or are thinking about it. Talking to them might help the final decision because you can get an opinion from someone in a similar position. I did not get more detailed information from these people since it is often easier in this situation to use academic channels for sharing information, and the Archives already has contacts in some of these institutions, whereas I would be going in cold. So it doesn't make sense for me to talk to everyone immediately unless it seems that they have something important to offer to the decision-making process right now. That step can come if and when the Archives decides on a specific system or type of system. If I know of a way of contacting the people below, I've mentioned it. Most of this information comes via electronic mail, so I can ask for additional contact information if desired. My apologies for the lack of organization, but no method proved
Elizabeth Wood mentions joint project between the Emergency Medicine and Radiology Department of Los Angeles County and the University of Southern California Medical Center that had their mastering done by Image Premastering.

Elizabeth H. Wood
Computer Services Librarian
Norris Medical Library
University of Southern California
ewood%phad.hsc.usc.edu@usc.edu

The AV Department of Hornbake Library at the University of Maryland is developing a videodisc in conjunction with the National Agricultural Library. I wonder if this is the Forestry Service collection mentioned above.

David Austin mentions two other projects. First, Andrew Eskind at the Eastman House is working on something to do with a videodisc. Second, Jim Sheldon at the MIT Media Lab is working on a videodisc of Edward (sic) Muybridge motion pictures in conjunction with the Addison Gallery of American Art, Phillips Academy, Andover, MA, 01810. Finally, David says “Also, make sure you check the SN/G: Report on data processing projects in art (1988). It is not yet on-line but available in hard copy, maybe even at Cornell. It is a list of projects registered with the Scuola Normale Superiore, Pisa, Italy and the Getty Art History Information Program, Los Angeles.”

David Austin
U29716@UICVM

Jim Sheldon
jls@media-lab.media.mit.edu

The AVIADOR (Avery Videodisc Index of Architectural Drawings on RLIN) project at the Avery Architectural and Fine Arts Library at Columbia sounds very similar to what the Archives might want to do. In addition, RLG is working on a way of linking a videodisc to an RLIN terminal, which would be very interesting. Janet Parks sent me a copy of their literature on the videodisc system.

Janet Parks
Curator of Drawings
Avery Architectural and Fine Arts Library
Columbia University
New York, NY 10027
212-854-6738

Jane Kleiner mentions several videodisc projects, one of which is the Emperor I collection done by Ching Chi Chen at Simmons. It is quite sophisticated and includes sound as well. She thinks MIT has an architectural collection on videodisc and adds that the National Agricultural Library has a collection of historical photographs from the Forestry Service on videodisc.

Jane Kleiner
notjpk@lsuvvm

Lennie Stovel mentions that there would be more information in the Library of Congress’ literature on their Prints and Photographs Division’s videodiscs, although he does not give a specific contact.

Lennie Stovel
Library Systems Analyst
Research Libraries Group
bl.mds@rlg.bitnet

David Finkelstein at Stanford University Academic Information Resources says that they are currently digitizing a large slide collection, which will eventually reside on videodisc. They are currently using HyperCard because of its ease of use but are looking into more powerful database programs such as Ingress, which in use but is not necessarily well-liked at MIT’s Athena Project.

David Finkelstein
Academic Information Resources
Stanford University
davef@jessica.stanford.edu

Steve Cisler from Apple Computer has available a technical report on basic videodisc production. It is called “Multimedia Production: A Set of Three Reports” and includes: “Casual Multimedia Production”, “Videodisc Basics”, and “Videodisc Production of the Visual Almanac”. It was done by Apple’s Multi-media Lab for the production of the forthcoming Visual Almanac. It is written for the non-technical person and includes mastering costs, sources of replicators, techniques. Steve mentioned that some people at the Visual Resources Association felt that the image quality was not high enough for scholars. He also said that there is a new method of distributing videodisc images with a certain type of network called Broadtalk. Steve can be contacted for more information on Broadtalk, and he will send a copy of the videodisc report to anyone who sends him a request on university letterhead along with a self-addressed mailing label. I have the report and recommend it highly for anyone who is actually starting on the specifics of producing videodisc.

Steve Cisler
Apple Library
In response to another question, Steve Cisler gave the address of several replicators for videodiscs which I have yet to contact. These are as follows.

Crawford Communications
506 Plasters Ave
Atlanta, GA 30324
404-876-8722

Pioneer Communications
1058 E 230th St.
Carson, CA 90745

3M Optical Recording
223-5S 3M Center
St. Paul, MN 55144
612-733-2142

Cynthia Read-Miller at the Ford Museum is using a videodisc and microcomputer catalog set up by a company called Argus.

Bernard Littau at UC Davis is putting together a radiology learning system for the veterinary school. He warns about several problems with videodisc production. First, it is extremely expensive to master the first disk from videotape. It requires a great deal of staff and equipment time just to make the videotape, and then the frame numbers on the resulting videodisc must be matched with the photograph database. Second, he feels that videodiscs are more suited to video sequences since that was what they were originally designed for, and, videodiscs have limited resolution for displaying still images in comparison to a digital image stored on CD-ROM and displayed on a computer monitor. Unfortunately, as subsequent conversations with Bernard proved, 100,000 photographs is simply too many to put in CD-ROM format because it would require 30 or more CD-ROM discs. Third, he said that when they did the videodisc, they were forced to use three frames for each image by the mechanics of the process of transferring the images to videotape. As such, they ended up using three frames of the videodisc for each image, reducing the storage capacity by three. Using three frames per image had the advantage of safety if one or two of the images were bad for some reason or other. I wonder if this problem appears if a commercial mastering service does the work since Bernard’s project was done in-house, I believe.

Bernard Littau
VM Radiological Sciences
School of Veterinary Medicine, University of California
Davis, CA 95616
916-752-0184
Internet: vmrad@ucdavis.edu
BITNET: vmrad@ucdavis

There is an integrated image database package with its own Programmers Application Language developed by PCM, Inc. The package is called PC ALBUM and runs on an IBM-PC.

PCM, Inc.
8330 Boone Blvd. Suite 430
Vienna, VA 22180
703-356-4600 or 800-654-5845

Ernst Robl recommends an expensive system called INMAGIC. It is sold by a company of the same name in Cambridge, MA. The Los Angeles Public Library uses it to catalog their extensive photographic collection and speaks very highly of it. The version which runs on IBM-PC type microcomputers is $1000, and there is a version which runs on VAX mainframes as well. Ernst says that INMAGIC allows a considerable amount of individual configurations and handles variable length data well. It can accept data from other sources, which is good for compatibility reasons. In fact, the LA Public Library has staff members do the cataloguing on laptop computers in the stacks rather than bring the collection out to a terminal.

Ernst has served a couple of terms as chair of the Picture Division of the Special Libraries Association and has authored an introductory book on picture librarianship, Organizing Your Photographs [Amphoto,1986]. In connection with the above, he has visited a large variety of institutional and commercial picture collections. (The Picture Division no longer exists as an individual entity with SLA, but its interests have been taken over by several other divisions.) His book points out some general issues to consider in the cataloging of photos, although the sections on computers are fairly basic because of its audience.

Ernest H. Robl
Systems Specialist (Tandem System Manager),
Library Systems
027 Perkins Library, Duke University
Durham, NC 27706
(919) 684-6269 w; (919) 286-3845
ehr@ecsvax

Russell Grau mentions a project he worked on with a company called Laser Recording Systems. The project consisted of taking images, scanning them onto a WORM drive, and then accessing the images via bibliographic information stored in a database. The whole
thing ran on IBM-PC type microcomputers. The person Russell worked with was named Tom Corsten, but he may not be there any more.

Laser Recording Systems, Inc.
270 Sparta Ave.
Sparta, New Jersey 07871
201-729-3055

Russell Grau
916-920-9092

Gordon Fair mentions that Oracle for Macintosh can work with SuperCard as well as HyperCard. Unfortunately, SuperCard is much slower than HyperCard and a project like this does not require SuperCard’s color and animation abilities.

Gordon Fair
gf07+@andrew.cmu.EDU

There is a package called Videodisc ShowMaker that will allow you create a database of entries with keywords and then search over the fields in the database. It is intended for a substitution for a slide projector in classes that require many images (it was originally designed for graphic arts education). It is a collection of stacks for the novice HyperCard/Macintosh user; it interacts with the videodisc players using videodisc drivers from Apple; and it can handle large databases of images (around 2000). The current version of ShowMaker uses the HyperCard Find function and is not very fast, but it does what it is supposed to. It is going to be released in December through a company called Ztek, which supplies interactive-video software. If interested, get in touch with them or with the professor in charge, Mark Sanders. My only problem with VideoDisc ShowMaker is that it is definitely not fast enough for 100,000 images. Some sort of HyperCard extension software would be required to increase the search speed.

Mark Sanders
msanders@vtvm1.cc.vt.edu
msanders@vtvm1
(703) 231-6480

Bob Samson at the University of Texas at Arlington might be setting up a system using the Series 2000 Laser-Optic Filing System from TAB Products. I don’t know much about this project, but I gather that the system is a digital system, so I don’t know how they are getting enough storage space for 350,000 photographs even though it comes with either a 5.25 or 12 inch optical disk. The system also includes a computer (no indication of what kind), a scanner, a high-resolution monitor for viewing the images, and possibly a laser printer for creating hard copy. He would be using this to store 350,000 images from the photographic archives of a local newspaper. Since many of the photographs are quite old, he wishes to avoid physical contact when possible.

Bob Samson
University of Texas at Arlington
B366RCS@UTARLVM1
817-273-3000

Lucy Kiester (phone: 301-496-5962) at the National Library of Medicine is finishing up a videodisc project and used Stokes Mastering Service to transfer her images to disc.

Bill Perry (phone: 202-857-7537) at the National Geographic Society has also done some work with Stokes Mastering Service.

Mike Segel recommends using the Informix database (which runs on many different microcomputer and mainframe systems) because it allows you to store BLOBs (Binary Large Objects) in the data base. I don’t know if storing images as BLOBs would take up less space, but if it didn’t the space requirements would be prohibitive.

Mike Segel
sege@quanta.eng.ohio-state.EDU

Ed Heath is an intern at the Library of Congress and is working on the American Memory project. He sent me quite a bit of information on American Memory. The project uses a Pioneer Laservision Player and the machine is controlled by a Macintosh IIX using HyperCard and Discovery Systems’ HyperSearch. The photographs were mastered onto the videodisc before the project started for another reason so Ed didn’t know too much about the specifics. American Memory deals with keywords by using free text searches with a “Visual Materials” thesaurus developed by the Library of Congress. Otherwise it is The American Memory setup also includes a CD-ROM.

Ed Heath
Special Projects
University Computing
George Mason University
Fairfax, VA 22030
(703)232-2941 EHEATH@GMUVAX

Lloyd Davidson tells of a article in BYTE magazine (January 1988— (“A Better Way to Compress Images”, BYTE 13/1, 215-218, 220-223)) in which a method using fractal geometry achieves graphic compression at ratios of over 10,000 to 1. Such compression ratios would easily allow a CD-ROM to store a great many images and would make them far more feasible for
extremely large image collections. He also mentions a second article about the same researchers in the November 4, 1989 issue of the New Scientist, p.40.

Lloyd A. Davidson
Seeley G. Mudd Library for Science and Engineering
Northwestern University
Evanston, IL 60208
L_Davidson@nuacc.acns.nwu.edu

Overall Conclusions and Comments

Taking everything I currently know into consideration, I would recommend using a Macintosh SE/30 with 2 megabytes of RAM and at least an 80 megabyte hard disk. That will satisfy any of the programs and leave plenty of room for expansion. As far as the programs go, I currently recommend Omnis 5 with HyperCard to provide videodisc access if no external routine for this are easily available. Of course, testing would be necessary before a final decision. No matter what software is used, there will be a fair amount of programming time necessary to set it up and get it running. In addition, the time it will take to enter 100,000 records into the database will be considerable. I cannot make any recommendations as to the mastering services because I do not have enough hard evidence to work with. Ideally, Cornell would set up its own facility for transferring images to videodisc.

Since transferring images to a videodisc is so expensive, I can only recommend that the Archives search for donors. In the meantime, deciding on a database and starting to enter the data would be useful whether or not a videodisc is ever monetarily feasible. Once the data is entered into a microcomputer database, it would not be too difficult to move it to another system, should a standard appear or merely a better method of working with a videodisc. I see no reason to wait on starting the database for this reason, and the cost of transferring images to videodisc will not drop much in the future, if at all. If Cornell set up a facility to transfer images to videodisc, it might be cheaper, although one never knows.

An important procedure for the moment is to think about the format of the database. The fields of bibliographic information are set, but some thought must be given to the keywords. The problem is with keywords because there are simply too many different possibilities, since everyone thinks different keywords are important. The Architecture School has a thesaurus, which helps, but they will still need to add some keywords and ignore others. Three to four levels of hierarchical keywords (i.e. Post 1905 - People - Professors - Professor Kaplan) are probably as detailed as you want to go at first, since it is too easy to come up with keywords which only make sense to the cataloguer after four levels in the hierarchy.

A good way to figure out a system is to find a picture and then work backwards so you see what steps you went through to find it. Hopefully this will be solved easily by simply using the categories of information that are already set up for the current system. Looking at the Architecture School's system might be helpful in determining the number and type of keywords.

Some more questions to keep in mind while designing such a system include who will be using the system, how will they be using it, and what is the end result going to be? Answering these questions before the database is designed will help in the design stage to make sure that the database is really set up correctly for the purposes at hand.

As a caveat, let me merely mention the problem of copyright. I assume that Cornell owns the copyright on the photographs in the Archives, but if not, it is technically a breach of copyright to transfer these images to a videodisc.

Further questions

I have left a number of questions unasked in my inquiries because of the preliminary nature of the investigation. Most of these deal with the specifics of the videodisc access, since that is the great unknown in the whole project. The questions into which I have yet to delve are as follows (along with my current opinions).

• Is the videodisc access feasible soon or farther in the future?

My feeling is that the videodisc access will be very nice once it is set up, but it will be expensive to master the disc. Assuming the prices quoted by the mastering services, a videodisc of 100,000 images could easily run over $300,000, if not more. Unless a munificent source of funds appears, I suspect that the videodisc will simply be too expensive for the moment. I don't think prices will drop much in the future, because the imaging and material handling work involved will remain more or less the same. In the event that several hundred thousand dollars should become available, the database should be able to support a videodisc. Otherwise, the files would have to be exported and imported into another program, a procedure which can be difficult and time consuming.

• What would be the best videodisc players for the Archives's purposes?

I didn't check into this at all, although I know there are a number of models that would work with either a Macintosh or a PC. All the prices that I've seen are in the $2500 range. The writable videodisc which you can
record to directly is quite a bit more expensive, at $13000 to $15000.

Would the Archives wish to use an outside mastering service or set up an in-house mastering service?

The outside services are likely to be more expensive, although they would also probably give higher quality results. If there is enough interest, Media Services might set up a videodisc mastering system at some point, which would be ideal. Their system might be somewhat cheaper and would certainly be closer. In any event, price and image quality seem to be directly related, so the nicer your pictures look, the more it will cost. If in-house mastering is deemed unfeasible, then some testing between the three mastering services would be in order.

In addition to these specific questions, I’m afraid that this report brings forth more questions yet that can only be answered by careful thought on the part of the Archives. I’ve made recommendations and hopefully discovered sources of information that will help answer these questions, but much more work will need to be done before a project like this becomes reality. For instance, it took Margaret Webster almost two years to start her pilot project. If I can be of assistance at any later point in the project, please feel free to call me.

Appendix
Addresses and phone numbers

1stDesk   Systems
7 Industrial Park Rd.
Medway, MA 02053
508-533-2203
800-522-2286
Makers of 1stFile and related programs.

ACIUS
20300 Stevens Creek Blvd
Cupertino, CA 95014
408-252-4444
Makers of 4th Dimension. They are very hard to reach.

Answer   Software
20045 Stevens Creek Blvd.
Suite 1E
Cupertino, CA 95014
408-253-7515
Makers of HyBase, a HyperCard database extension.

Ashton-Tate
20101 Hamilton Ave.

Terrance, CA 95052
213-329-9989
213-329-8000
Makers of DBase IV and DBase Mac

Blyth Software
3655 Campus Dr.
San Mateo, CA 94403
415-571-0222
Makers of Omnis 5. I spoke with Jennifer Blome.

Borland International
1800 Green Hills Road
Scotts Valley, CA 95066-0001
408-438-8400
Makers of Paradox and Reflex Plus

Anne Carnell
1159 Comstock Hall
Cornell University
Ithaca, NY 14853
255-7675
Anne works in the University Photography Department.

Howard Curtis
Mann Library
Information Technology Section
Cornell University
Ithaca, NY 14853
255-9570

DataEase International
7 Cambridge Ave.
Trumbull, CT 06611-9983
203-374-8000
Makers of DataEase

Discovery Systems
7001 Discovery Blvd.
Dublin, OH 43017
614-761-4197
Makers of HyperSearch, a HyperCard database extension.

DucSoft
238 Columbus Ave
Sandusky, OH 44870
419-626-6797
Makers of Applications and Routines for 4th Dimension, but nothing for videodiscs.

Fox Software
27493 Holiday Lane
Perrysburg, OH 43551
419-874-0162
Makers of FoxBase+ and FoxBase/Mac

Spring 1991
Clif Nickerson, Marketing Manager. Note that the phone number is different from the old literature. Image Concepts make C-Quest, a videodisc program for the PC and Unix boxes.

Image Premastering Services
1781 Prior Avenue North
St. Paul, MN 55113
612-644-7802
A videodisc mastering service.

Interactive Media Center
Geri Gay or Mike Oltz
Cornell University
Ithaca, NY 14853
255-5530

KnowledgeSet Corp.
888 Villa St, Suite 500
Mountain View, CA 94041
415-968-9888
Makers of HyperKRS and HyperIndexer, HyperCard extensions.

Microrim
3925 159th Ave NE
Redmond, WA 98073-9722
206-885-2000
Makers of Rbase.

NovaSoft Engineering Group
2343 Ridgewood Ave.
Edgewater, FL 32032
904-423-5189
Makers of GridFile, a HyperCard database extension.

Odesta Corp.
4084 Commercial Ave.
Northbrook, IL 60062
312-498-8852
312-498-5615
Makers of Double Helix II

Oracle Corp.
20 Davis Drive
Belmont, CA 94002
800-345-3267
Makers of Oracle database software. Spoke with a Robert Silverberg, ext. 2019

SoftStream International
19 White Chapel Drive

Mount Laurel, NJ 08054
800-262-6610
609-866-1187
Marketing company for HyperHIT, a HyperCard database extension

SofStream—Steve Hannaford
19 White Chapel Drive.
Mount Laurel, NJ 08054
215-543-5194
Technical support representative for HyperHIT.

Stokes Mastering Service
Austin, TX
512-458-2201
A videodisc mastering service.

Symantec Corp.
10201 Torre Ave.
Cupertino, CA 95014
408-253-9600
Makers of Q&A

Turquoise Film/Video Productions
St. Louis, Missouri 63088
314-843-1998
A videodisc mastering service.

Voyager Company
239 Manning Ave.
Los Angeles, CA 90025
800-446-2001
Makers of VideoStacks, a set of videodisc drivers and other software.

Dave Watkins
Media Services
B-27 MVR
Cornell University
Ithaca, NY 14853
255-5431
Dave is the head of Media Services and is working with the videodisc project.

Margaret Webster
Architecture Slide Librarian
B-30 Sibley Dome
Cornell University
Ithaca, NY 14853
255-3300

Xiphias
12464 Washington Blvd.
Marina Del Rey, CA 90292
213-841-2790
Makers of Xearch, a HyperCard searching extension
A videodisc is an optical disc which is read by a laser. The images are analog, which means essentially that they are stored as a snapshot consisting of shades of gray or color, rather than being divided into individual dots which can be either on or off, which is how an image would be stored in digital format.

CD-ROM stands for Compact Disk - Read Only Memory. It is a digital format, which means that any pictures are made up of individual dots which can be either on or off, black or white. As a result, pictures take up a great deal of space on a CD-ROM, so much space that a project this size would not be feasible. Mainframe storage systems would be required to store so many photographs.

There are two types of databases, relational databases and flat-file databases. Flat-file databases work just like a file cabinet in that each record is stored separately. Relational databases can share information between files, so you would not need to duplicate information if you had a database of addresses and a database of phone numbers because the two files could share the person’s name. In addition, relational databases tend to be faster and more powerful.

HyperCard is a program described as a "software erector set" by its author. It is free with every Macintosh and allows non-programmers to create sophisticated programs, called stacks. HyperCard works on the metaphor of a stack of note cards, although it has a great deal of easily-accessed power which seems unrelated to a stack of cards. HyperCard is not a database, but it is an information manager and manipulator.

An external command is a small program that can be inserted into another program to give the second program additional functionality. They are extremely common with HyperCard and provide numerous ways of enhancing HyperCard.

A front-end is what you see and work with, whereas the back-end is the part which actually does the work. For instance, the front-end of a washing machine is the control panel where you set the type of wash and the amount of time. The back-end is the drum and vibration mechanism which actually washes the clothes. You have to be able to use the front-end, but you don’t have to know how the back-end works to get your clothes clean.

I don’t quite understand their technology and am merely trying to repeat it verbatim in hope that someone more well versed in the photographic arts will understand.

I don’t know what the units in question are since Stokes didn’t mention them.

---

POSITION AVAILABLE

PUBLIC SERVICES LIBRARIAN.
Albert R. Mann Library. Cornell U.

Responsibilities include providing reference, instruction, and computerized search services as part of a nine-member public services professional staff reporting to the Head of Public Services. Each of this staff has additional responsibilities. This librarian will work with the coordinator of numeric files in expanding users' access to statistical information in computerized databases. These responsibilities include: evaluating data for readability and validity; comparing and selecting storage and access options such as magnetic tape, floppy disk, optical disk, and online; consulting with users and developing strategies for data extraction and presentation; and providing training in using numeric files systems and in using statistical software. Interested in candidates who are not afraid of computers, able to learn multiple retrieval languages and at least one database management program, interested in statistics, and who know how data are used in research. Will provide training to applicant interested in developing expertise in numeric files, an important growth area for Mann Library's collections and services.

This Librarian will also participate in research and development projects involving accessing, retrieving, and managing electronic information. Mann's working environment is characterized by cooperation and teamwork among staff members. All library staff are involved in implementing an electronic library.

Mann Library holds the nation's second largest collection of agricultural and life sciences information resources in print and electronic form. This is supplemented by a substantial number of related social sciences publications. The library serves students, faculty, researchers, extension personnel, and staff of Cornell's College of Agriculture and Life Sciences, the College of Human Ecology, and the Division of Biological Sciences. Mann has a staff of 65 FTE assisted by over one hundred student employees. Operations and projects are supported by Mann's systems staff of six systems analysts, programmers, and technicians.

Knowledge/Experience: Master's degree in library or information science required. Excellent communication skills and interpersonal abilities required. Interest in statistics or management of research data required. Experience in working with the public highly desirable. Two years' library work experience highly desirable. Desirable experience: use of SAS, SPSS, or a database management program for microcomputer or mainframe; use of BRS, DIALOG or SilverPlatter; classroom teaching. Academic background in life sciences, social sciences, or business desirable.

Send cover letter, résumé, and the names, addresses, and phone numbers of three references by May 10, 1991 to

Ann Dyckman,
Director of Personnel,
201 Olin Library,
Cornell University, Ithaca, NY 14851.

Applications accepted until position is filled.

Cornell University is an equal opportunity, affirmative action employer.
Free Workshops on Using Computers for History

by David L. Clark
History Computerization Project
24851 Piuma Road
Malibu, CA 90265

Free, one-day workshops featuring hands-on training in the use of computer database management for historical cataloging and research are now being offered through the History Computerization Project of the Regional History Center of the University of Southern California and the Los Angeles City Historical Society. No prior computer experience is required. The workshops are held on the last Saturday of each month, except for holidays. Workshops have been scheduled in 1991 for June 29, July 27, August 31, September 28, October 26, and December 7. The project is building a Regional History Information Network of historical organizations who share a common subject interest. The project employs the History Database program and the Pick database system running on IBM PCs. The program is intended for use by historical organizations and researchers at the lowest levels of financial resources and computer experience. The multi-user capability of Pick makes it suitable also for larger repositories. The system at USC consists of 10 PCs connected to a shared database. The course textbook, Database Design: Applications of Library Cataloging Techniques, written by David L. Clark, is published by McGraw-Hill. For information or to sign up for a specific date contact: David L. Clark, History Computerization project, 24851 Piuma Road, Malibu, California 90265. Telephone: (818)888-9371.

Book on Computer Database Management for Research and Cataloging

Database Design: Applications of Library Cataloging Techniques by David L. Clark is due for publication in July 1991 by McGraw-Hill. The book shows how to organize information by combining computer database management with library, archival, and museum cataloging methods. Database Design is a how-to guide for people who have information to manage, but who have no previous experience with database management or cataloging. It includes step-by-step instructions for managing collections of photographs, manuscripts, museum objects, or print materials, for keeping a membership roster, and for organizing research notes intended for publication. The book teaches the use of standardized cataloging methods as a natural complement to database management. The History Computerization Project of the Regional History Center of the University of Southern California and the Los Angeles City Historical Society uses Database Design as the course textbook for free workshops on the computer-cataloging of historical materials. For information on the courses or book contact: David L. Clark, History Computerization Project, 24851 Piuma Road, Malibu, California 90265. Telephone: (818)888-9371. To order the book include a check to David Clark for $34.95 + $5.00 for shipping and handling ($10 outside of the U.S.). (California residents please add $2.62 for sales tax.)
ONLINE COMMUNICATION WITH THE SSDA CATALOGUE

The SSDA online catalogue, a part of Israel inter-university ALEPH network, is now accessible to foreign archives/users via INTERNET (TELNET, TCP/IP).

The following instructions should enable the user to establish a connection and run a searching session. 2 HELP screens provide more information about the catalogue structure and its searching modes. However, if you face either communication or searching problems, you are most welcome to contact Miko Levy at the SSDA, E-mail: MAGAR1@HUJIVMS, Tel. 972-2-883181.

As of now, connection is not available from IBM mainframes, due to terminal emulation differences with ALEPH computers (VAX/VMS).

CONNECTING INSTRUCTIONS FOR INTERNET / TCP/IP

INTERNET letter address is: HAR1.HUJ1.AC.IL and number address is: 132.064.176.002. Although both are usable, it is advised to use letter address, as number address may change.

1. Logon your own computer.

2. Enter TELNET [ALEPH address].

3. After the VAX/VMS - ALEPH logo is displayed, you will be asked to type in a USERNAME. Enter SSDA.

4. Select terminal no. 2 at the Terminal Selection Menu.

5. Select function 2 (ALEPH FUNCTION+IUTIL) at the Function Selection Menu.

6. The Catalogue main screen is now displayed. Start your session with one of the search codes or type M1 for help.

GOOD LUCK!
The International Association for Social Science Information Services and Technology (IASSIST) is an international association of individuals who are engaged in the acquisition, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, database librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompasses hard copy as well as machine readable data.

Paid-up members enjoy voting rights and receive the IASSIST QUARTERLY. They also benefit from reduced fees for attendance at regional and international conferences sponsored by IASSIST.

Membership fees are:
- Regular Membership: $20.00 per calendar year.
- Student Membership: $10.00 per calendar year.

Institutional subscriptions to the quarterly are available, but do not confer voting rights or other membership benefits.

Institutional Subscription: $35.00 per calendar year (includes one volume of the Quarterly)

I would like to become a member of IASSIST. Please see my choice below:
- $20 Regular Membership
- $10 Student Membership
- $35 Institutional Membership

My primary interests are:
- Archive Services/Administration
- Data Processing
- Data Management
- Research Applications
- Other (specify)

Please make checks payable to IASSIST and mail to:
Ms Kay Worrell
Treasurer, IASSIST
% The Conference Board
245 Third Ave
New York, NY 10022

Name/title
Institutional Affiliation
Mailing Address
City
Country/zip/postal code/phone