
INTERACTIVE, DIRECT-ENTRY APPROACHES TO EVENT FILES

R.A. Schweitzer

Steven C. Simmons

University of Michigan

September 1981

CRSO Working Paper No. 245
GBS Briefing Paper No. 15

Copies available through:
Center for Research on Social Organization
The University of Michigan
330 Packard Street
Ann Arbor, MI 48109

GREAT BRITAIN STUDY BRIEFING PAPERS

1. "Great Britain, 1828-1834: Historiography and Selected Bibliography," by Michael Pearlman, June 1977: issued as CRSO Working Paper #159.

2. "Some Political Issues in Nineteenth-Century Britain. Part One: The Government, Catholic Emancipation," by Michael Pearlman, July 1977: issued as CRSO Working Paper #160.

3. "Some Political Issues in Nineteenth-Century Britain. Part Two: The Rights of Collective Association and Assembly; Parliamentary Reform; Industrial Conflict," by Michael Pearlman; issued as CRSO Working Paper #163.

4. "Contentious Gatherings in Great Britain, 1828-1834: Provisional Plans for Enumeration and Coding," by Charles Tilly and R.A. Schweitzer, revised version, September 1977: issued as CRSO Working Paper #165, November 1977.

5. "British Contentious Gatherings of 1828," by John Boyd, R.A. Schweitzer, and Charles Tilly, March 1978: issued as CRSO Working Paper #171.

6. "Interactive, Direct-Entry Approaches to Contentious Gathering Event Files," by R.A. Schweitzer and Steven C. Simmons, October 1978: issued as CRSO Working Paper #183.

7. "Source Reading for Contentious Gatherings in Nineteenth-Century British Newspapers," by R.A. Schweitzer December 1978; issued as CRSO Working Paper #186.

8. "A Study of Contentious Gatherings in Early Nineteenth-Century Great Britain," by R.A. Schweitzer, January 1980: issued as CRSO Working Paper #209.

9. "Enumerating and Coding Contentious Gatherings in Nineteenth-Century Great Britain," by Charles Tilly and R.A. Schweitzer, February 1980: issued as CRSO Working Paper "210.

10. "The Texture of Contention in Britain, 1828-1829," by R.A. Schweitzer, Charles Tilly, and John Boyd, April 1980; issued as CRSO Working Paper #211.

11. "How (And to Some Extent, Why) To Study British Contention," by Charles Tilly, February 1980: issued as CRSO Working Paper #212.

12. "British Catholic Emancipation Mobilization, Prototype of Reform?" by R.A. Schweitzer, December 1980: issued as CRSO Working Paper #220.

13. "Britain Creates the Social Movement," by Charles Tilly, March 1981: issued as CRSO Working Paper #232.

14. "Nineteenth-Century Origins of Our Twentieth-Century Collective-Action Repertoire," by Charles Tilly, September 1981: issued as CRSO Working Paper #244.

15. "Interactive, Direct-Entry Approaches to Event Files," by R.A. Schweitzer and Steven C. Simmons, June 1981: issued as CRSO Working Paper #245.

INTERACTIVE, DIRECT-ENTRY APPROACHES TO EVENT FILES*

Great Britain Study
Briefing Paper No. 15

R.A. Schweitzer
University of Michigan
September 1981

*This working paper is reprinted from Social Science History, Vol. 5, No. 3, Summer 1981.

An earlier more extensive version of this paper was published as CRSO Working Paper #183, October 1978. It is now out-of-print.

Interactive, Direct-Entry Approaches to Event Files

British Contentious Gatherings

R. A. SCHWEITZER
STEVEN C. SIMMONS
University of Michigan

As recently as a decade ago, an authoritative introduction to computing for historians recommended an approach which essentially employed the computer as a rigid, if very large, tabulator. Edward Shorter's *The Historian and the Computer* (1971) described how to reduce complex information to simple fixed-choice codes, transfer the coded data to punched cards, read the cards into fixed-format package programs, and prepare large tabulations or statistical analyses from the data. Shorter's advice made sense: it encouraged historians who knew little about computers or quantification to move ahead, and enabled them to produce useful results without becoming programmers. During the 1970s, however, three important changes in computing made the sturdy old procedures obsolete. The first change was the increasing availability of flexible, inexpensive microprocessors—small machines with memories as big as many large computers of the 1960s, which would operate by themselves or in conjunction with powerful central computers, which came with a great variety of prepared programs, and which would serve for the entry,

Authors' Note: *The authors are grateful to Tim Beckett, John Boyd, Laurie Burns, Keith Clarke, Phylis Floyd, Harry Grzelewski, Dave Hetrick (ILIR: MICRO), Chris Lord, Debbie McKesson, and the other GBS staff members who worked on this material. We also wish to thank Dr. Charles Tilly for his encouragement and assistance. We would like to especially thank Bill Golson, without whose work on design and writing of the data entry program, this article would not have been possible. The National Science Foundation supports the research herein described.*

SOCIAL SCIENCE HISTORY, Vol. 5 No. 3, Summer 1981 317-342

transmission, storage, editing, manipulation, analysis, and presentation of many different sorts of information, including ordinary words. The second change was the improvement of interactive computing, in which a relatively inexperienced analyst could carry on a prompted "conversation" with a sophisticated machine while searching or analyzing a complex machine-readable file. The third was the development of data base management systems which, from the user's point of view, greatly simplified the storage and manipulation of large bodies of machine-readable evidence.

Taken together, the three changes enormously increased the ease, flexibility, and power of many sorts of computing. Those sorts included the standard problems posed by historical research: the need to search and reorder large sets of rich but irregular nonquantitative observations.

Up until recently, most social science computing was carried out by means of rigid codes, 80-column code sheets, punch cards, and excessive data cleaning. This was a long, costly process that involved many steps which could easily result in raising the number of errors in one's data. Hopefully this system has gone the way of the hand pump and the ice box.

The historians of the 1980s have a system that allows them virtually to eliminate the 80-column code sheet, card punching, card verifying, and that trip in the snow to the computing center to batch in the job. They can sit in a comfortable office and in one step enter data, either numerical or textual, directly into the computer files. The system can have built-in checks to eliminate some and limit almost all data cleaning. This system works on a TV screen and is available to call up packaged programs to analyze the data without ever having to seek out "THE PROGRAMMER."

This type of system is in operation today for the Great Britain Study at the Center for Research on Social Organization at the University of Michigan.

This new modern system eliminates many of the "Old Problems" but creates problems of its own: file storage costs, equipment, connect time, staff training, and a host of other related bugaboos. Nevertheless, on the whole the system to be described here appears to have promise for a wide variety of social science computing.

BRIEF DESCRIPTION OF THE STUDY

Over the past few years, our group has been studying patterns of conflict in Western European countries. We have been trying to learn how large-scale change in industrialization and state-making affects the capacity and propensity for collective action of different segments of the population. For example, we have analyzed strikes and collective violence in Italy, France, and Germany in the period 1830 through 1968 (Tilly et al., 1975). What we have done is collect a uniform, comprehensive enumeration of events meeting preset criteria in a specified area (sometimes a region but usually an entire country).

Our latest effort involves the study of collective action in Great Britain from 1828 through 1834. In this undertaking we hope to expand our knowledge of nonviolent gatherings as compared to the more normal violent events that we studied in other European countries (Tilly and Schweitzer, 1980).

DATA COLLECTING

We are collecting data on what we call a "CONTENTIOUS GATHERING": an occasion on which ten or more persons outside the government assemble in the same place to make a visible claim, which, if realized, would affect the interests of some other specific person(s) or group(s) outside their own number. This definition captures just about any event that contemporaries

or historians would call a "riot" or a "protest," and also a variety of meetings, rallies, demonstrations, celebrations, and so forth in which people clearly make claims of some kind.

There are a number of items that we are not interested in having included. These are such things as casual gatherings, festivals, crowds at accidents or fires, or strictly social or entertainment functions such as balloon ascents. Some small-scale violent actions we do not include, e.g., the common crimes of housebreaking, stagecoach robbing, and pickpocketing. We do include the action of ten or more actors who articulate any sentiments into their actions, such as stating that they are starving or that they dislike the rich.

We began our search for these contentious gatherings in seven especially selected sources. They are two London-based but nationally read newspapers, the *Morning Chronicle* and the *Times* of London; two periodicals, the *Annual Register*, published yearly as a summary of the most interesting events of the past twelve months, and the *Gentleman's Magazine*, a monthly tabloid; and three serials dealing with the workings of Parliament, *Hansard's Parliamentary Debates*, the *Mirror of Parliament*, and the *Votes and Proceedings* of the Houses of Parliament.

From the diligent reading of the sources for the period 1828 through 1834, we have collected approximately 150,000 articles or what we call our coversheets (forms that denote an article that may pertain to a contentious gathering). From these coversheets we expect to assemble about 6,000 to 10,000 qualifying contentious gatherings. Our next process is to collate these scattered materials into a set of dossiers or event files that detail all the information we can collect about any particular event in our sample.

DATA COLLATION

The idea behind collation is to organize the vast amounts of materials on one event into an effective unit with which to understand the flow of the gatherings. What we do first is to

remove any materials that are clearly not qualifying, such as described plans for meetings which never occur. From there we try to decide, using a predetermined set of rules, if the remainder of the articles have all the criteria that make a qualifying contentious gathering. There are such criteria as occurring in Great Britain, and containing people who are not government officials and who have a visible claim affecting the interests of some specific person(s). The idea of claims is the most difficult area in which to be specific, so we have some specific rules of thumb.

(1) In the absence of contradictory information, collective violence is prima facie evidence of a claim.
(2) Purely organizational activities do not qualify.

The source articles that are of the types we wish to study, and that pertain to the same event, are logged into a book and given a special coversheet to denote that they are available to code. Once we have a number of the events logged, we can start matching the leftover articles to those main events. In some cases, articles about a particular event occur in our sources months after the actual occurrence of the contentious gathering. Before beginning any coding we try to have sorted most of a particular year so as to have all the matching done before beginning the actual coding work.

There is an intermediate step prior to beginning the coding process. We enumerate the event to show the groups (Formations) and the claims (Action Phases) that are taking place within the confines of the event. The Formations are enumerated on the basis of being either the makers of claims or the objects of those claims. Action Phases are a scenario-type rendering of the claims and related actions that the Formations are making during any event. Each set of enumerations is done once and is rechecked by at least two other persons who are either enumerators themselves or are familiar with the criteria.

To review, we begin with seven sources and have researchers read through them in a systematic way, looking for articles that pertain to contentious gatherings. We gather the articles together

and match the ones that pertain to the same event and place all of them in a dossier event file. This material is then given an identification number that reflects the year, month, and day the event occurred. It is then logged into a book for a further check on reliability of the matching. Next, the event is enumerated to denote the Formations and Action Phases that are involved in the series of claims made by the actors in the event. Now the event is ready to be coded.

CODING

When we have a sufficient number of events enumerated (usually 200 or one quarter of a year), the next step is to determine in what form we will code them. We have developed three separate systems for coding events. The first is the LONG form, which includes all the questions on printed forms that will appear on the computer-simulated entry session and will be included in the record. A record consists of information about: (1) the EVENT as a whole, (2) each separately enumerated FORMATION, (3) each ACTION PHASE, (4) all known SOURCES that describe the event, and (5) a space for COMMENTS. Figures 1 and 2 are examples of the EVENT section and one page of a three-page FORMATION section. We wish to know certain common characteristics of each event, such as the date it began, how many people were involved, what type of event it was, if anyone was arrested, wounded, or killed, and many hundreds of similar questions. We chose not to use the standard numeric coding but rather to develop an alphanumeric coding system.

Our forms for coding look like a questionnaire. We have a system that will take coded answers that can be understood by relatively untrained coders because they are basically written in English. Most of the questions are answered in English that is later entered into our data files with a computer program which transfers them into a numeric form to store. The main advantage

GREAT BRITAIN STUDY E
EVENT SECTION
Bobbi 6-77, Rev. 6-78

GBS

CG ID# (9 digits) Year | Month | Day | No.

Day of Event 2: /1 S | /2 M | /3 T | /4 W | /5 Th | /6 F | /7 S | /0 NA

Accuracy of starting date: 3 Default= Exact /1 = 1Wk /2 = 2Wks /3 = 1Mo /4 = 2Mos /5 = 3Mos

Date event ends: 4 Default= Same /1 = 1Wk /2 = 2Wks /3 = 1Mo /4 = 2Mos /5 = 3Mos Date:

Duration: Days 5 /1 = Guess Duration: Hours 7 /1 = Guess Default = ITJ

General event type 9 /1=Violence /2=Meeting /3=Gathering /4=Delegation /99=Other

Type of Event 10

Major issue, or claim: 11

Total Locations:	Location information	Grid numbers:
	1	/
		/
	2	/
		/
	3	/
		/
	4	/
		/

Sources: 22 MC LT GM AR HPD MOP Other

Total participants: 29 Low High Best Guess

How determined: /0=ITJ /1=#s in report /2=words in report /3=QWO /4=Other:

# of person-days: Estimate 34 Margin of error ± default = ITJ

# of person-hours: Estimate 36 Margin of error ± default = ITJ

Arrests during event: 38 Margin of error ± default = ITJ

Arrests after event: 40 Margin of error ± default = ITJ

Wounded: Total 42 Margin of error ± default = ITJ

Killed: Total 44 Margin of error ± default = ITJ

Assembler 46 Date / / Coder 48 Date / /

Check coder 50 Date / / Enterer Date / /

Total # form. enumerated 52 # of form. participating directly in the event

Section completed E F S C I (Circle for coded, / for entered)

Figure 1 Event Section questionnaire coding form

Formation Section: Fill our one 3 page section per formation. Page F-1 F

Coder 2 CG ID# Year | Month | Day | No.

Total # of formations

Number of this formation Summary name of formation 4

Does this formation overlap with any other formation(s) in the same event? 5

No Yes: Which ones? Give formations #s

What is the relationship between this formation and the Contentious Gathering? 6

/1 = Participants, making a claim
/2 = Participants, object of a claim
/3 = Participants, both making and receiving claims
/4 = Spectator, bystander
/5 = Object of a claim, not present
/6 = Object of a claim, some participating, some not present
/7 = Involved in action before or after CG only, no comment
/8 = Involved in action before or after CG only, comment:
/99 = Other, comment: 7

Names given to this formation in account(s): 8 None = Default

1 2 3 4 5 6 7 8

If the account(s) list any individual names of formation members, list them: last name first, @ sign, first, @ sign, then all other information, title, etc. 9

None = Default

1 2 3 4 5 6 7 8 9 10 11 12 13 14

List continued on next page

Figure 2 Formation Section questionnaire coding form

of this is having the coders answer such questions as "What is the location of this gathering?" by "Middlesex, London, St. Luke's parish, London Tavern assembly rooms" rather than having to look up on lists of special codes for each county, town, parish, or specific place in Great Britain. Before we began the procedures to develop the machine part of the entry system, we deliberately undertook a hand simulation of the coding procedures in order to separate the logical and technical problems of coding in general from the problems of building a computer-based system for editing and analysis.

A SHORT form, a scaled-down version of the long form, is for those events that are not as complicated or do not have large amounts of Sources, Formations, or Action Phases. This form has only information contained in two of the long form sections: (a) the event as a whole and (b) the formation section. Figure 3 is an example of page 2 of the short form, which is a compacted version of the three-page (long form) Formation section.

Last, there is a DIRECT form, which requires no precoding of the material but directs the coder to enter the data belonging to this event directly into the computer via our CRT terminal.

INTO THE DEPTHS OF THE MACHINE

Through the foregoing processes, there has been a gradual change in the structure of the event. It initially began as a typical nineteenth-century British newspaper or periodical story, full of editorial asides, snide remarks, and occasional flights of fancy, and has now reached the point where every event has been fit into a very rigidly defined format, where the same questions were asked of each event, all answers are of the same type from event to event, and the human individuality is gone from the reporting and description of the event. This reduction of an emotion-filled human occurrence to a mechanical description is needed, for in

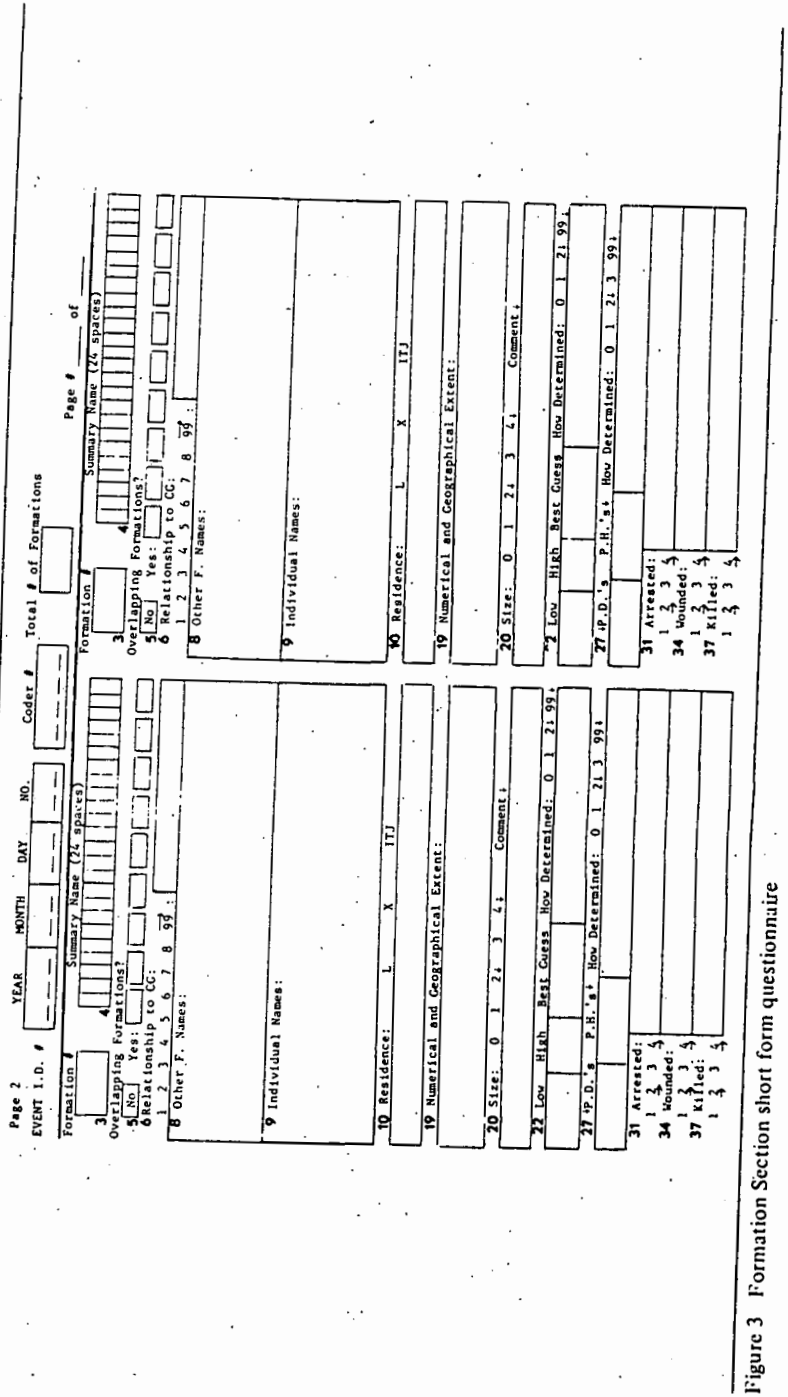

Figure 3 Formation Section short form questionnaire

the next step the event will be handled by a device that has been removed to electronic form but remains mechanical, the computer.

You will notice that there is a distinct difference between what the standard computer input is and what we have reduced the coded event to. These differences illustrate the innovations that we have made in our computer entry system.

There are no cases where we have reduced data to a simple number when those data are not intrinsically a number. Categories have been left with their names intact; exact points along ranges have been left; direct quotes and lists of names have been left in their literal forms; in short, a person reading the coded record still has an excellent idea of what went on in the event.

This differs radically from the more typical "card image" input to a computer. A person reading the data that are normally keypunched would have no idea what is going on and might not even know what field the information was covering. Instead of having the preparation done from our existing format to a standard card image by hand, we use the computer to take the data in the coded form and make a special sort of image from them. What we create here is a much more flexible and efficient version of a card image. Since this image is created from entry program input rather than punched card input, we call it an entry image.

There are three major functions that occur in the program. The first is the input of data that can be broken down into categories and represented by a number. This is done by the program in such cases as the "General Event Type," and whether an "action" occurs before, during, or after some other action. In all cases, the data are taken by the program in the form that is on the coding sheet and are converted into the appropriate number for storage in the entry image.

Second, there are data which cannot be categorized but can be broken down to a form that will occupy much less space in the machine record. An example of this is the contentious gathering identification number (CGID). A CGID is a nine-digit number

we use to identify events uniquely. In literal form on a card image, it would occupy one space per number in the CGID (i.e., nine columns or nine bytes). If it is converted into a different form by changing a numeric string to a binary form, it can be stored in a mere four spaces (four bytes). This is a savings of over 50% on storage and handling costs for the CGID.

Last, we have the input of literal data. This is handled in one of two different ways by the program. In the cases where we have foreknowledge that there will be a very limited length of description (say, 40 characters or less), we will input and store the description directly. It will take up 40 spaces in the input image, but will be directly accessible. We can take advantage of what is called an external field to handle the cases where there is great variance in the length of the data. With an external field the data is written into a file consisting of a series of numbered lines. Then the file name is associated with that field of the input image and the line number on which our particular data are entered is placed in the image. With proper use of the system editing facilities, we can have in excess of 32,000 characters in a single line (which of course can consist of a number of grammatically separate lines), allowing us to put in extensive quotes, lists of names, places, comments, and so forth. Conversely, if there are no data associated with a particular instance, nothing is stored in the external file, and the line number of a blank line is assigned to the input image. Therefore, only as much storage as is needed is used for any particular item.

The card image prepared from the entry program is very complex, probably far more complex than any that would be hand produced. However, the preparation of the card image is only part of the function of the program. With it we also do error checking, error correction, and provide stimulus to the enterer to insert the proper data at the proper time.

THE ENTERING SYSTEM

A researcher will sit down at a terminal with a coded event in hand and will call up a series of specially designed programs that

will assist the person in entering the event. There is no need for the researcher to have any programming experience. The system has been designed in such a way that the person need know only three statements outside of the answers to the questions the program will ask: (1) the signon account name ($SIG SJOA), (2) the password (a system security device), and (3) the signoff ($SIG). With the exception of the password, all of this information is kept posted on the wall above the terminal.

Figures 4 and 5 depict sections of an interactive entering session (Action Phase). Please note that all the items right of the colon were entered by the enterer, and the items to the left are the system or program responses. A few deliberate errors have been included so that the errorchecking facilities and limitations of the machine will be illustrated.

After the researchers have signed on (the program is called up automatically), they need no longer give any active instructions until they wish to make corrections on items that have been entered in this particular session. The only datum kept unchanged during the entry session is the enterer's name. It is the first question asked by the program and is automatically placed with every section entered. In this way we can track the quantity of work done by a given person and determine the quality of that work.

Next, the program asks for which type of entry mode is desired, E or S: Event or Section. An E mode is used when an entire event with all its subsidiary information is being entered. In this case, the program will go from section to section (Event, Formation, Action Phases, Sources, and so on) automatically until the entire record has been entered. If the S mode is requested, the program will respond with a request for the particular section, and the coder can enter either a single Source, an Action Phase, or any other section that was previously omitted.

When an entire event has been entered, the program returns to the "WHICH MODE" question so that additional events may be entered. The coder can continue entering until finishing all events, answering the last "WHICH MODE" question with /Q for quit. The coder then signs off.


```
      ENTER YOUR CODER ID :  RAS1

      WHICH MODE?    (E/S):  S
      WHICH SECTION?       :  A


      CGID                 :  830112303
      CODER                :  MEM1
      ACTION PHASE NUMBER  :  0101

      SUBJECT FORM NUMBERS:
         FORMATION :  01
         FORMATION :

      ACTION VERB          :  ASSEMBLE

      OBJECT FORM NUMBERS :
         FORMATION :

      ENTER TEXT           :  "ON TUESDAY LAST, A LARGE BODY OF FARMERS'
                               LABOURERS ASSEMBLED"
      WHEN          (B/D/A):  D

      ACTION PHASE NUMBER  :  0201

      SUBJECT FORM NUMBERS:
         FORMATION :  01
         FORMATION :

      ACTION VERB          :  BREEK

***THE VERB, BREEK     , IS NOT CURRENTLY DEFINED.
***DO YOU WISH TO HAVE IT DEFINED (Y/N)?  N


***FIELD INPUT ERROR***

      ACTION VERB          :  BREAK

      OBJECT FORM NUMBERS :
         FORMATION :  02
         FORMATION :

      ENTER TEXT           :  "...THEY PROCEEDED...BROKE MACHINES."

      WHEN          (B/D/A):  C

***FIELD INPUT ERROR***

      WHEN          (B/D/A):  D
```

Figure 4 Sample of interactive data entry session for CGID 830112303, Action Phases

We have also developed a system by which we can enter data when the main computer is not operating. We have written a program in a language called BASIC that is very similar to our regular data entry program. The primary difference is that the data are stored on a cassette tape and the program is run on an in-

```
   FIELD CORRECTIONS    : 7
***FIELD INPUT ERROR***

   FIELD CORRECTIONS    : /7
   ENTER TEXT           : "...THEY PROCEEDED... AND IN FORCE...BROKE MACHINES.

   FIELD CORRECTIONS    :
VERIFY CGID:

   CGID                 : 8302304
***FIELD MUST BE SPECIFIED***

VERIFY CGID:

   CGID                 : 830112304
***VERIFICATION FAILED.  REENTER CGID***

   CGID                 : 830112303
VERIFY CGID:

   CGID                 : 830112303
(RECORD ADDED)
```

Figure 5 Sample of further interactive data entry session for CGID 830112303, Action Phases

house minicomputer which is not connected to the main computer. The one big disadvantage is that the BASIC program simply stores data without any of the larger program's error-checking (described below). We use this system to enter simple events quickly and without any connect time, thereby lowering costs. When computing rates are cheaper, we send the data via a high speed-line to our regular disk files. Through this system we have cut the costs of entering a single record by as much as 75%.

BENEATH THE SURFACE: ERRORCHECKING

Obviously if the only purpose of the program were to take straight data and put them into the computer in a literal format, there would not be much advantage in using it as shown.

Fortunately, there is far more going on in the program than appears on the surface. The most important of these items is errorchecking.

At this point we should draw a distinction between legal and correct answers to questions. A legal answer is one that fits the format that the computer expects (certain answers should be numbers; certain answers should be letters; sometimes a specific line length or numeric range is needed) and will therefore be accepted by the computer. A correct answer is one that is within the specified range of legal answers such that it also happens to be the true one. A close examination of the first two questions and answers in the entering session will help clarify this.

The coder ID is a specifically prepared set of initials and numbers that will identify the coder of the program. There is a certain fixed amount of space set aside for this information, and therefore the code must have a certain size. Examples of valid coder IDs are RAS1 for Robert Andrue Schweitzer or SCS1 for Steven Charles Simmons. If by chance a Ruth Ann Sloan came along, her ID would be RAS2.

There is a list of permissible coder IDs stored in the computer and numbered with values from one to the maximum number of coders. When the enterers give a coder ID, the program searches through its memory to check the given code against its list of valid ones. If the given code is invalid, the machine will print an error message and will re-ask the question. It will continue asking the question until a valid answer is received. Once a valid answer has been received, the program will take the numerical value of the code and will enter that into memory. This value is smaller than the code (i.e., it occupies less storage space) and therefore its use saves on machine storage space and charges.

Once a valid code has been processed and stored (a matter of microseconds and imperceptible to the user), the second question, "CGID," arises. A brief explanation of the CGID is needed here and will further illustrate the varieties and limitations of error-checking. A CGID is the identification number of the contentious gathering. It consists of the last three numbers of the year of the event's occurrence, the number of the month, the number of the

day, and the event number on that day. Thus 829112405 is November 24, 1829, event number 05. A CGID is always a nine-digit number, so the first check is to insure both that the year is a valid one and the number is of the correct length. This is done by a simple arithmetic comparison, which in English would be expressed, "Is the CGID between 828000000 and 834123200?" If this answer is yes, the next comparison is to see if there is a valid month. The fourth and fifth numerals are examined to be sure that they are between 01 and 12 inclusive. Then valid dates are checked by examining the sixth and seventh numbers to be sure they are between 00 and 31 inclusive. It should be noted that when the day is unknown, it is assigned 00 as a value. From all of this we can insure that the CGID expresses a date that could possibly exist in the period of time that we are studying, and that the proper number of digits are present. This, unfortunately, does not guarantee that the correct CGID has been entered. The machine will happily accept 829040499 in place of 828112405, and bad data will have gotten through, but nowhere near as much as we would have had without the errorchecking facility.

With these two items—the coder ID and the CGID—we have also illustrated the difference between categorical and analytical, or literal, data. The former assumes a certain finite number of correct answers to a question and thus each fall into a certain preset category. The second assumes so many (potentially infinite) possible correct answers that the data cannot be assigned categories without simplifying it to death. A CGID, even as limited as it is, has 226,400 possible correct answers, so assigning categories would be a thankless task.

Since a CGID is the identifying item for the entire event and all its subsections, it was felt that additional errorchecks were needed. The method chosen was to require the CGID to be entered twice, once at the beginning of the event and once at the end. The entries are then compared and if they do not match, the person entering will be repeatedly asked for the CGID until two consecutive identical entries are made (see Figure 5). By having the CGID entered redundantly we hope to prevent mistakes.

Naturally, some mistakes will still slip through, but this keeps them to a near-irreducible minimum.

The three methods described above are the major ones used in the study's error prevention procedures. In one form or another, they can be applied to dates, numerical sizes, and any categorical entries. In addition, there are a few specialized errorchecks that are installed but that are so peculiar to the data that their description would be excessively tedious.

INTERMEDIATE DATA HANDLING

After the entry program has coded the data into a basic acceptable form, they are written into a standard line file, just as a punched card would be. The advantage of using the program is that the data are all in the proper columns for their interpretation, and many of the standard types of errors have already been prevented. However, there are often new variables that we wish added or additional information that can be obtained by sources not as prone to human errors as are humans. Among these are the data of the entry session, complex numbers that must be looked up in tables, and arithmetic operations. All of these are done by the GBS programs interacting with the main computer and are added to the individual entry images. In this way, new variables are created and old ones modified or expanded. Once this has been done, a second set of programs called "packers" is run. The packer converts all of the data from the entry image file into binary data in a sequential file. This is a concession to the MICRO system, which has its most efficient data input from a sequential binary file. Finally, these binary data are entered into the MICRO files. At this point the next stage of cleaning can begin.

CLEANING AND REFURBISHING IN MICRO

The computing work for this project is being done on the Michigan Terminal System (MTS) at the University of Michigan.

We make especially heavy use of two programs: a data storage system called MICRO and a data analysis system called MIDAS.[1] For a clear understanding of what is done in MICRO, it will be helpful to have a brief overview of the MICRO system. Any organized collection of data, together with a dictionary to interpret it, is called a data set in MICRO. The Great Britain Study uses a number of data sets at one time, as it is much simpler to keep relatively associated information together (for example, the event entry session goes in the event data set and the formations go in the formation data set) in small, easy-to-manipulate sets.

Each data set is divided up into fields, very similar to the way that the coded event was divided up into a rigid format of questions. Due to the intermediate processing that has already been done, there are now more fields with more detail than were originally input by the enterer. These fields contain all of the information that was originally on the entry image, but it is contained in a more compact and organized form that is much more suitable for performing analysis. Fields come in four major types, each with subtypes that are trivially different: analytical, categorical, literal, and external. They serve four significantly different functions.

Analytical fields contain numbers such as would be found in a temperature scale or a head count; nearly infinite gradations are possible. For this reason, analytical numbers are stored simply as numbers, with no categories possible. For an example of an analytical field, see "CGID" in both the entry session (Figure 4) and dictionary (Figure 6).

Categorical fields are those that can be divided into a sufficiently small set of categories, with each category assigned a numerical value. For an example of a categorical field, see the "when" question in the dictionary and entry session.

Literal fields are similar to analytical in that they contain the information exactly as entered (literally), but they generally


```
ACTION PHASE SECTION DICTIONARY DOCUMENTATION

FOR: ACT
DATA SET LOCATION: SH9H:ACT
DICTIONARY LOCATION: SH9H:ACT$
DATE: APR 28, 1980

THIS DATASET CANNOT BE DESTROYED.
THIS DATASET CAN BE REPLACED.
USE COUNT:          1

DATA SET DESCRIPTION:
DESCRIPTOR LINE ACT IN FILE


F(#)   FIELD NAME        ABBR       VALUE   DESCRIPTION
-----------------------------------------------------------------------

F(1)   CGID              CGID     REQUIRED  CGID

F(2)   CODER             CODR     REQUIRED  CODER

F(3)   PHASENO           PHAS     REQUIRED  ACTION PHASE NUMBER

F(4)   SUBJECT           SUBJ     REQUIRED  SUBJECT FORMATION NUMBERS

F(5)   VERB              VERB     REQUIRED  VERB
         CATEGORIES
           ABATE         957          957   ABATE
           ABUSE         807          807   ABUSE
           ACCEDE        226          226   ACCEDE
           ACCEPT        610          610   ACCEPT
           ACCLAIM       413          413   ACCLAIM
           #WARN         881          881   #WARN
           #WATCH        833          833   #WATCH
           #WOUND        771          771   #WOUND


F(6)   OBJECTO           OBJ                OBJECT FORMATION NUMBERS
         CATEGORIES
           NONE          NONE           0   NONE

F(7)   EXTERNAL          EXT      REQUIRED  ACTION PHASE DETAIL
         FILE NAME
          SH9H:ACT.EXT

F(8)   WHEN              WHEN     REQUIRED  WHEN
         CATEGORIES
           BEFORE        BEFO           1   BEFORE
           DURING        DURI           2   DURING
           AFTER         AFTE           3   AFTER

F(9)   ENTERER           ENTR     REQUIRED  INITIALS OF PERSON ENTERING

F(10)  ENTRY-DATE        ENTD     REQUIRED  DATE ON WHICH THE ENTRY IS MADE
                                             INTO THE PRELIMINARY DATA SET
```

Figure 6 Action Phase section dictionary documentation with samples of a few verbs

contain character information, not numerical information. For an example, see the "Major Issue" field, Figure 1 and "Formation Summary Name," Figure 2.

External fields are a special provision of MICRO for cases where comments, descriptions, or any other written information

can be appended to a data set without taking up huge sections of the set for large literal fields. An external field is simply a line number for another file elsewhere in the system where text for that particular field are stored. The data stored in external files may be of any length including zero, which frees us from many of the rigid format problems otherwise associated with computerized data sets. For an example of an external field, see the "Detail" section of the Action Phase session, Figure 4.

With each set of fields is a dictionary (see Figure 6). The dictionary contains all of the fields' names, their acceptable (to MICRO) abbreviations, whether that information must be put into the fields (if it is not, a default value is assigned), a description of the field and its categories, which type of field it is, and how many "columns" it takes up on the punched card. Scale and factor are not used in this study.

When operating, MICRO uses the dictionary to put all of the data back into a format readable in English. In addition, MICRO is designed to proceed as much as possible like English so that it may be used by nonprogrammers. It has been our experience that nonprogrammers can quickly become efficient users of MICRO for most purposes, provided that they are sufficiently clear on how each data set is put together and have some concept of what MICRO is doing with the data. It is by no means necessary to have a full knowledge of the programs behind MICRO, but merely to know what each of the commands will do to the original data set and the subsequent copies that they make of it.

PRELIMINARY ANALYSIS

When a data set has been cleaned it is ready for analyses to be run. We have included two samples, Figures 7 and 8, run on the MICRO and MIDAS facilities. Figure 7, which is produced from MICRO, is basically a descriptive chart. It sets out eight geographical areas in Great Britain and lists types of events from simple brawls to complex preplanned meetings in those areas for the year of 1829. The areas of highest contention are "Other

Type of Event	Middlesex	Dorset	Hampshire	Kent	Lancashire	Other England	Wales	Scotland	Total
Poachers vs. Gameskeepers	0	0	0	0	1	13	0	0	14
Smugglers vs. Customs	1	0	0	0	0	4	0	0	5
Brawls in Drinking Places	3	0	0	0	2	1	0	0	6
Other Violent Gatherings	33	1	0	2	10	47	1	9	103
Attacks on Blacklegs and Other Unplanned Gatherings	0	0	0	0	1	0	0	1	2
Market Conflicts	0	0	0	0	0	0	0	0	0
Other Unplanned Gatherings	18	0	0	2	6	14	0	1	41
Authorized Celebrations	2	0	0	1	0	4	0	0	7
Delegations	1	0	0	0	1	0	0	0	2
Parades, Demonstrations, Rallies	9	0	0	0	1	10	0	0	20
Strikes, Turnouts	1	0	0	0	1	1	0	0	3
Pre-Planned Meetings of Named Associations	49	0	3	9	16	49	5	18	149
Pre-Planned Meetings of Public Assemblies	62	0	0	7	5	42	4	9	129
Other Pre-Planned Meetings	35	2	2	8	8	85	6	14	160
Total	214	3	5	29	52	270	16	52	641
Percentage of Total	33.4	0.5	0.8	4.5	8.1	42.1	2.5	8.1	100

Figure 7 Sample MICRO data analysis of Event Types in eight selected geographical areas in Great Britain in 1829

England" (not surprising) and "Middlesex county" (also not surprising, as the capital city of London is located in the county). The types of activity that are the most common are the meetings categories. These descriptive results are mildly interesting, but if more complex analyses are desired, then the data must be converted from MICRO through an interface into MIDAS data sets for manipulation. Figure 8 has been produced on the MIDAS facilities. The top section of the figure shows participant, arrest,

Regions	Total Events	% Reporting Participants	Mean Participants	Mean Arrests	Mean Wounded	Mean Killed
Middlesex	215	24.2	1583 (52)	0.19(192)	0.04(188)	.00(214)
Dorset	3	0.0	-- (0)	0.50 (2)	0 (2)	0 (3)
Hampshire	5	20.0	120 (1)	0 (5)	0 (5)	0 (5)
Kent	28	28.6	143 (8)	0 (26)	0 (26)	0 (26)
Lancashire	52	23.1	544 (12)	2.26 (43)	0.82 (40)	0.10 (50)
Other England	270	30.4	1069 (82)	0.46(226)	0.12(226)	0.01(259)
Wales	16	12.5	125 (2)	0 (15)	0 (15)	0 (15)
Scotland	52	17.3	249 (9)	0.02 (41)	0 (39)	0.05 (44)
TOTAL	641	25.9	1086(166)	0.43(550)	0.13(541)	0.02(616)
Poach	14	57.1	17 (8)	2 (6)	0.67 (6)	0 (13)
Smuggle	5	40.0	67 (2)	0.25 (4)	1.50 (4)	0.2 (5)
Brawl	6	33.3	87.5 (2)	4.33 (3)	0 (1)	0 (6)
Violence	104	36.0	586 (38)	4.478 (46)	0.79 (38)	0.043 (92)
Attack	2	0.0	-- (0)	0 (1)	-- (0)	0 (2)
Unplan.	41	26.8	3532 (11)	0.171 (35)	0.11 (36)	0.025 (40)
Celeb.	7	14.2	500 (1)	0 (7)	0 (7)	0 (7)
Deleg.	2	100.0	386 (2)	0 (1)	0 (1)	0 (1)
Parade	20	30.0	742 (6)	0 (19)	0 (19)	0 (19)
Strike	3	66.7	65 (2)	0 (2)	0 (3)	0 (3)
Assoc.	149	10.0	100 (15)	0(146)	0(147)	0(147)
Public	128	18.8	976 (24)	0(126)	0(126)	0(125)
Meets	160	34.4	1599 (55)	0.06 (154)	0.16(153)	0.03 (156)
TOTAL	641	25.9	1086 (166)	0.43 (550)	0.13(541)	0.02 (616)

Figure 8 Sample MIDAS analysis of 1829 events

wounded, and killed levels in the eight areas of analysis used in Figure 7. The bottom half gives information on those categories by event type. There are few surprising results; however, a couple of points stand out. First is the much higher arrest rate in Lancashire county as compared to the other seven areas surveyed. Second is the remarkably large number of mean average participants in unplanned gatherings. These two points lead us into further study of the whys and hows of Lancashire and unplanned gatherings.

These two charts were run on full year data sets. Because our final sets will be huge (with some 12,000-15,000 events plus background information), we have begun to create a series of relatively small, unbiased subsets of the master file for the purpose of preliminary analyses of the data. This will also allow all interested parties (graduate students and other researchers) to sample test their ideas and theories on small data files and thereby save time and computer funds by not having to access the larger data files (for further analysis see Schweitzer et al., 1980).

INTO THE FUTURE

The use of an interactive direct data entry system for social science has already spread beyond the Great Britain Study. Additional inquiries are already under way at the University of Michigan and Carnegie-Mellon University which are making use of the techniques developed here plus additional components. One such project is a collective biography of Scandinavian architects and their works, while another deals with collective action in Ireland.

Upon completion of the Great Britain Study, the data gathered will be made publicly available in computer-readable form through the Inter-University Consortium for Political and Social Research. By using the MICRO system, other researchers may select from our data only the information that interests them, add their own new information, and illuminate still more portions of British history. We have left clear tracks behind us so that the researchers of the future can easily find our sources and see our criteria for selection of information. Finally, we hope that this methodology as outlined will be useful to any researcher in preparing a computer-assisted study.

NOTES

1. The Michigan Terminal System is one of the oldest timesharing systems in the world. Its ease of use and relative low cost has promoted the development of a number of special packages by groups not directly connected with the University Computing Center. MICRO AND MIDAS are both examples of this.

MICRO is a "data base management system" (DBMS), an organized system for handling and manipulating large data bases so they can be stored, corrected, modified and expanded as easily as possible. MICRO does have some statistical functions, but they are primarily simple descriptive ones. MICRO was developed by the Institute of Labor and Industrial Relations of the University of Michigan. A copy of the "MICRO Reference Manual" may be obtained through the Institute at P.O. Box B-1, Ann Arbor, MI 48109.

For complex statistical analysis we turn to MIDAS, the Michigan Interactive Data Analysis System. MIDAS has an exhaustively complete set of statistical functions available by use of relatively simple commands. It is not as efficient or flexible as MICRO in its data base capabilities, and thus we are taking advantage of an interface between the two that will convert a MICRO data set into a MIDAS data set.

MIDAS was written at the Statistical Research Laboratory, University of Michigan, Ann Arbor, MI 48109. The Laboratory provides a number of manuals on the use of MIDAS, including some for the user who has a good knowledge of statistics but little or none of computing.

REFERENCES

SCHWEITZER, R. A., C. TILLY, and J. BOYD (1980) "The texture of contention in Britain, 1828-1829." Ann Arbor: Center for Research on Social Organization. Working paper #211, April.

SHORTER, E. (1971) The Historian and the Computer. Englewood Cliffs, NJ: Prentice-Hall.

TILLY, C. and R. A. SCHWEITZER (1980) "Enumerating and coding contentious gatherings in nineteenth-century Britain." Ann Arbor: Center for Research on Social Organization. Working paper #210, February.

TILLY, C., L. TILLY and R. TILLY (1975) The Rebellious Century, 1830-1930. Cambridge: Harvard Univ. Press.

R. A. Schweitzer is a Research Associate II at the Center for Research on Social Organization of the University of Michigan. He is currently involved in the study of collective action and contention in nineteenth-century Britain and eighteenth- and nineteenth-century London. Among his publications are "Studying the Contentious Gatherings in Early Nineteenth-Century Great Britain" and "The Changing Geography of Contention in London, 1755-1835," with Charles Tilly. He is currently working on a manuscript on early London business directories.

Steven C. Simmons is a Systems Analyst supporting interactive graphic systems at Automatic Data Processing in Ann Arbor, Michigan.

WORKING PAPERS OF THE CENTER FOR RESEARCH ON SOCIAL ORGANIZATION

The Center for Research on Social Organization is a facility to the Department of Sociology, University of Michigan. Its primary mission is to support the research of faculty and students in the department's Social Organization graduate program. CRSO Working Papers report current research and reflection by affiliates of the Center; many of them are published later elsewhere after revision. Working Papers which are still in print are available from the Center for a fee of 50¢ plus the number of pages in the paper (88¢ for a 38-page paper, etc.). The Center will photocopy out-of-print Working Papers at cost (approximately 5¢ per page). Recent Working Papers include:

235 "Protoindustrialization, Deindustrialization, and Just Plain Industrialization in European Capitalism," by Charles Tilly, March 1981, 14 pages.

236 "Reagan's Social Services Block Grant: What It Is and What You Can do About It," by Deborah K. Zinn, May 1981, 25 pages.

237 "Flows of Capital and Forms of Industry in Europe, 1500-1900," by Charles Tilly, June 1981, 19 pages.

238 "Social Movement Sectors and Systemic Constraint: Toward a Structural Analysis of Social Movements," by Roberta Garner and Mayer N. Zald, July 1981, 27 pages.

239 "Center for Research on Social Organization - Annual Report, 1980-81," by CRSO personnel, July 1981, 13 pages.

240 "Selected Readings on Political Change," by Charles Tilly, August 1981, 26 pages.

241 "How One Kind of Struggle - War - Reshaped All Other Kinds of Struggle in Seventeenth-Century France," by Charles Tilly, August 1981, 37 pages.

242 "The Political Culture of Social Welfare Policy," by William A. Gamson and Kathryn Eilene Lasch, Revised August, 1981 from Wo. P. #221, 23 pages.

243 "Role of Informal Networks and Medical Care Organizations in Helping Families Cope with Childhood Cancer; Prospects for Collaboration," by Mark A. Chesler, Oscar Barbarin, Joan E. Chesler, Diane Hughes, Judy Lebo, August 1981, 135 pages.

244 "Nineteenth-Century Origins of Our Twentieth-Century Collective-Action Repertoire," by Charles Tilly, September 1981, 12 pages.

Request copies of these papers, the complete lists of Center Working Papers and further information about Center activities from:

Center for Research on Social Organization
University of Michigan
330 Packard
Ann Arbor, Michigan 48109