4. Planned several projects: (a) Development of a list of recommendations addressed to potential researchers on study design as it relates to data management (the "do's and don'ts" list). Group members are to send in their suggestions to the DOM AG coordinators. A consolidated list will be produced in Toronto for publication in the Newsletter and elsewhere. (b) Collation of information on relevant monographs, technical reports, program writeups, etc., which the AG members would like to share. The intent of the collection is to publish this information in a "What's New In Data Management" section of the Newsletter.

5. Prepared an agenda for the Toronto meetings; items for that agenda have been referred to above.

The DOM AG coordinators would like to suggest the following revised mandate for future discussion at international IASSIST meetings:

> "This group addresses the problems of data organization and management confronting those archiving or using social science data. The AG will investigate and evaluate software and procedures for data and documentation preparation and management; recommend guidelines for preparation procedures and software development; and, sponsor workshops and seminars for professional training and the exchange of information in these areas."

[The original mandate includes "hardware" as an area to be addressed by this AG; see Newsletter Volume 1, Number 1 for the full text of the mandate.]

AN OVERVIEW OF PROBLEMS ASSOCIATED WITH PROCESS-PRODUCED DATA / paul müller

For the August 1976 IASSIST meetings, Paul Müller prepared a report which provided a broad overview of problems associated with process-produced data. What follows is an edited version of this report. [Future issues of the Newsletter will contain additional Action Group reports. The membership is encouraged to begin a dialogue on this and other issues of concern to the data archive community.]

"Administrative Bookkeeping as a Social Science Data Base"

by
Paul Müller
Institute for Applied Social Research
University of Cologne

1.0 [...] I will try to give a rather broad overview of problems associated with the "production, acquisition, preservation, processing, distribution, and utilization of machine-readable" process-produced data. This report is necessarily biased by my own viewpoints and experiences; other experiences may well be fundamentally different. But, the function of this paper is to initiate discussion and later, joint actions.

1.1 Process-produced data

1.1.1 Definitions

Within this action group we should be concerned with "process-produced data" as defined by Rokkan, as well as with "official bookkeeping data" (e.g., administrative registers). A common problem with these data is that they are/were not originally collected for scientific purposes and/or within explicit statistical routines, thus creating the research situation that the data are to a great extent already "given". The generic term, "process-produced data," would encompass all data that are/were not collected for statistical or scientific research (e.g., censuses or surveys), but are instead by-products or traces of the daily routines of private or public organizations or persons.

1.1.2 Priorities

As a first step, we should concentrate on already-created machine-readable data within public administration. This would imply structured mass-data.

1.2 Problem areas

1.2.1 Inventory of machine-readable administrative data bases

The first task is to obtain an overview of the existing data bases within the restricted domain as defined in 1.1.2. In West Germany, QUANTUM plans a pilot study for North-Rhine-Westfalia, "A continuous inventory of administrative machine-readable data bases." Prior documentation projects have identified some 185 machine-readable data bases at the state level (North-Rhine-Westfalia), as well as some 655 files at the local government level (City of Cologne) [...] as by-products of EDP use in public administration. These inventories should be [...] descriptions of the data bases: e.g., coverage, period of correction/update, variables, etc. The experiences with the project in Germany show that updating would not occur without the active participation of the administration. The first step towards this end should be an overview of existing EDP routines within public administration. This can be rather easily achieved by using the information of the existing coordinating committees/institutions within public administration and by jointly establishing a check list.

1.2.2 Documentation

Existing documentation of machine-readable process-produced data e.g., National Archives and Records Service, Catalog of Machine-Readable Records in the National Archives of the United States, (Washington, D.C., 1975) or Directory of Computerized Data Files & Related Software, (National Technical Information Service, 1974) are models far less ambitious than the envisaged codebook-like documentation. Good examples for this would be: Department of Health, Education, and Welfare, 1973 Current Population Survey - Summary Earnings Record Exact Match File Codebook, Part I - Basic Information Studies from interagency data linkages, by F. Scheuren, D. Vaughan, and W. Alvey. Report No. 5 (1975) and De-

partment of Health, Education, and Welfare, 1973 Current Population Survey - Summary Earnings Record Exact Match File Codebook, Part II - Supplemental Information, Studies from interagency data linkages, by F. Scheuren, B. Kilss, and C. Cobleigh. Report No. 6 (1975).

1.2.3 Preservation

It is time to follow up the initiatives that were made by the Ruggles-, Kaysen- and Dunn Report in the United States. Specifically, we have to define research needs vis-à-vis national, state, and local archiving institutions. In so far as some 95% of data generated within public administrations are physically destroyed, joint actions must be launched to define worth-while material to be preserved. In Germany, we have an ongoing discussion concerning an archive law which should take into consideration the specific interests of social science research (criteria for preservation, the archiving of temporal and/or cross section samples of linked files). The German data law will have an effect on the quality of the archived data, in so far as it allows for selective destroying of individual records within a register. It is yet very unclear whether this contamination effect can be avoided in some way.

I propose beginning with an overview of existing archiving criteria (Kassation, physically destroying of data) within governmental archives and investigating whether these criteria are compatible with the kinds of uses that are not case-studies or oriented towards identified persons (e.g., historical).

1.2.4 Data Laws

The data laws will have an impact on accessibility to administrative data for research purposes, not only with regard to access to identifiable individual data (very often necessary in the data collection and management phases). As these data laws do not provide for exceptions, serious research projects will be effectively hampered when these projects are not in the interests of a data providing agency. Data laws (or their drafts) are increasingly used for the secretion of public administration data (even in the cases when anonymous information are required/sought)....

1.2.5 Uses already made of process-produced data

To ensure response to research needs for process-produced data, we should survey the uses made of administrative registers (i.e., those uses that were beyond the simple drawing of samples).

In Germany, we will take another look at around 6000 identified research projects within the INFORMATIONSZENTRUM social science research project. Similar endeavors should be made in other countries as far as there exists information on research projects.

1.2.6 Quality and characteristics of administrative bookkeeping

There has been very little attention paid to problems with the quality of process-produced data (e.g., how are these data created, in what ways are the collection or validation processes biased) or to the characteris-

tics of official bookkeeping. There is a plethora of "validity studies," which compare sample survey data with official statistics, but as a specific kind of study are not very cumulative.

The quality of process-produced data and characteristics of administrative registers are the objects of an ongoing research project at the Institute for Applied Social Research in Cologne (Wolfgang Bick and Paul J. Müller) in which we analyse the laterality of representation of the individual's (client's) social context and the temporal changes in registration by public administration of everyday activities.

1.2.7 Record linkage

Because these data are very often "meager" due to the limited administrative purposes for which they are collected, linked registers or data sets should be of the greatest interest. To the extent that administrative registers are organized within different life sectors, the potential of record linkage, (or family reconstitution as it is called within historical demography), for supplementing existing large scale, but limited registers is promising. A codebook-like documentation of linked files should by envisaged (cf. 1.2.2)

Research on record-linkage techniques increasingly concentrates on the problems associated with statistical versus exact links (cf. Department of Health, Education, and Welfare, Some Observations on Linkage of Survey and Administrative Record Data, Studies from interagency data linkages, by J. Steinberg. (1973). The 60-odd exact-matching studies done in the last decades showed that without a unique standard indentifier (SSN, PK, person identification number) exact record-linkage would continue to be a very expensive task achieving only an average of 85-90% matches. We need a methodological breakthrough for synthesizing different data bases according to configurations of socio-economic characteristics.

1.2.8 Instruments

We will repeat the mistakes made in earlier phases of the archive "movement" if we neglect the need for specific computer programs to handle process-produced, often "ragged" data; therefore, we should consider the problems of analyzing strings of data (e.g. within CROSSTABS or TROLL, perhaps) hierarchies of relations (e.g. area-block-house-family-respondent), or "life histories".

It is good to hear that the SPSS-survey has already brought these issues to the attention of the SPSS-people. [...]

1.2.9 Interdisciplinary communication

There is a real opportunity for intensified communication between those researchers working in quantitative history (the analysis of tax registers, birth certificates, marriage licences, etc.) and social scientists who have already worked with or are interested in using process-produced data for their research purposes. This communication, which is already institutionally organized within QUANTUM, should enable us to document,

store, and distribute process-produced data in such a manner that the kinds of uses have not to be invented after completing all these tasks. In particular, the development of a "source criticism" for mass data, analogous to the development of the methodology of survey data, can only be achieved in an interdisciplinary way. Such efforts are necessary for the envisaged descriptors for machine-readable process-produced data. Similarly, cooperation with those people working on record-linkage problems (e.g., Oxford Record-Linkage Study on medical record linkage) or with large-scale process-produced data (e.g. criminal statistics utilizing court-records) should be initiated....

1.3 Quantities

It is not possible to deal with these problems solely within the existing social science archiving movement which has so heavily concentrated on survey data. Other institutes must be brought into a network of archives and information centers in which coordination and a division of labor must be planned. (Examples of these include the "Sozialdatenbank" [Social-Information-System of the Department of Labour and Social Affairs, West Germany], the proposed "Zentrum für Aggregatdaten" [Centre for aggregate data of the German National Science Foundation], and the National Archives.") In Germany, the prospects for coordination are a little bit better because of the pioneering work being done within the Information- and Documentation Program of the Federal Government.

BOOK NOTICES / kathleen m. heim

Introduction

This column is a preliminary step in defining the literature of data archiving. Those of us who have tried to assess the state of the art in order to formulate annual reports, write articles, or keep professionally informed have been frustrated by the lack of bibliographic control over our area of concern. Indexing and abstracting services such as Social Science Citation Index, Information Science Abstracts, Library Literature, Social Science Index and Resources in Education are unsystematic in their assignation of subject headings to pieces of literature related to data archives. The problem is further confounded by the fact that seminal information concerning the establishment of data archiving has often been distributed informally at conferences or in unpublished papers. When our numbers were small we could depend upon an invisible college network to disseminate important information. However, as our numbers grow and as new archivists enter the field without access to the established network, it becomes mandatory that we define and organize the literature of our profession.