DOCUMENT RESUME

ED 055 319 EA 003 651

AUTHOR	Boruch, Robert F.
TITLE	A Class of Administrative Models for Maintaining Anonymity During Merge of Data Files. A Draft.
PUB DATE	Feb 71
NOTE	47p.; A portion of this paper presented at American Educational Research Association Annual Meeting (55th, New York, New York, February 4-7, 1971
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Administrative Principles; *Confidentiality; *Confidential Records; *Data Bases; Data Processing; Information Science; Information Storage; *Models; Research Needs; Research Problems

ABSTRACT

This report examines a series of general models that represent the process of merging records from separate files when it becomes essential to inhibit identifiability of records in at least one of the files. Models are illustrated symbolically by flow diagrams, and examples of each variation are taken from the social sciences. These variations cover simple situations such as that of soliciting anonymous data from previously identified respondents as well as more complex merge operations, e.g., merging files from mutally insulated data banks and merging data under code linkage systems. Characteristics of the models are discussed with emphasis on their benefits and disadvantages. (Author)

A CLASS OF ADMINISTRATIVE MODELS FOR MAINTAINING ANONYMITY DURING MERGE OF DATA FILES

Robert F. Boruch

American Council on Education

DRAFT: February, 1971

Note on "A Class of Administrative Models for Maintaining Anonymity During Merge of Data Files."

Robert F. Boruch

A portion of this paper has been presented at the annual meeting of the American Educational Research Association Meeting, New York, February 4, 1971.

A Catalog of Administrative Models for Maintaining Anonymity During Merge of Data Files

Robert F. Boruch
American Council on Education

Merging different kinds of information, which concern exactly the same individuals, is frequently a crucial element in social science research. Some kinds of inquiry require most research, e.g., the union between previous identification of a respondent and the record of his response to a later inquiry. Such efforts are growing rapidly in the magnitude and variety of information which is consolidated. For example, unions of data from different archives or merges of archival data with data solicited directly and ad hoc from subjects of the records, are being emphasized heavily in education and psychology (e.g., Schoenfeldt, 1970), law and sociology (Schwartz and Orleans, 1968), economics and public administration (Dunn, 1967), statistics and epidemiology (DuBois, 1969) and elsewhere.

Two social factors have strong implications for this expanded use of archival data: existing constraints on accessibility of identifiable records and the increasing public and professional interest in maintaining confidentiality of records. Both factors are evident in early governmental restrictions on access to identifiable archival data, exemplified by the U.S. Census Bureau (1968) and Social Security Administration (e.g., Steinberg and Cooper, 1967). More recently, new guidelines for enhancing privacy have been developed in the private sector (e.g., Russell Sage Guidelines, 1970) as well as the public domain (e.g., new H.E.W. screening procedures for experiments involving human subjects). Relevant professional efforts to appraise the legal and administrative accessibility of many kinds of institutional records have been conducted by Professor Alan F. Westin (1967) and are currently being conducted (independently) by Professors Westin and D. T. Campbell.

Objectives and Basic Definitions

Private and governmental restrictions on access to identifiable archival records often prevent the outside social scientist from merging his data with archival records. Given this constraint plus the desirability of merging records for research purposes, what strategies can the researcher legitimately use to conduct social research on merged files?

The primary objectives of this paper are to examine, appraise, and improve some techniques which serve to answer this question. A secondary objective is to furnish a series of specially constructed administrative models which accurately represent the simple data merge process and the techniques for merging data under legal or social restrictions.

Figure 1 illustrates the prototypical model for match-merge operations with data which is identifiable to the researcher. Model 1.0 should be interpreted as an abstraction, i.e., the data sets A1 and B1 may represent physical records or recollection, the information flow may be concrete or intangible, the identifying information may be alphanumeric or visual. The first element in the nomenclature A, indicates that the file contains a particular kind of data (A and the second element (1) indicates that identification is part of each record in the file.

Insert Figure 1 About Here

From the diagram, we infer that Agency A (e.g., a researcher) merges its own data file (A1) with identifiable data held by Agency B (e.g., an institutional data archive, B1). This merge operation depends totally on Agency A's ability to acquire the <u>identifiable</u> records from file B1, and any such acquisition is presumed here to infringe on the privacy of persons whose records are maintained in the file.

The alternative models considered below permit Agency A to merge the data while preserving anonymity (with respect to himself) of file B and anonymity with respect to Agency B of file A). By using a class of general models, we can easily systematize existing strategies for achieving the objectives, develop new strategies by manipulating the model and, finally, appraise the shortcomings of the strategies. The effectiveness of all the models depends, in part, on cryptographic encoding of statistical data, or of identifying information, or of both. Encode-decode systems, or specialized forms of character substitution, appear to be most pertinent to the administrative models considered here. Complete individual records, or discrete elements within a record are replaced by arbitrarily assigned characters (termed code groups). With very long lists of identifiers or of numerical records, infinite key transforms (see, Carroll and McClelland, 1969) present a convenient mechanism for generating nonrepetitive code numbers.

Three major subclasses of models are developed here. The first involves a single encode-decode operation on statistical data, conducted by one agency controlling a data file. Although the second subclass also relies on an encode-decode strategy, a broker is introduced to perform the merge operation; a broker is employed in the third subclass, primarily to maintain code linkages which are necessary for the merge. The subclasses cover simple situations, such as soliciting data from subjects identified by alias alone, as well as more complex operations, e.g., mutually insulated data banks and link file systems. Starting with the most elemental model -- the mutually insulated data bank concept -- we examine the utility and corruptibility of each class of models in succeeding portions of this report.

Model 2.0 The Insulated Data Bank Concept

Model 2.0 (Figure 2) is a formalization of a technique developed by Professors D.T. Campbell and R.D. Schwartz for merging two separately maintained data files, without permitting either agency to interrogate the other agency's identifiable records. Figure 2 represents the following activities.

Insert Figure 2 about here

Agency A, controlling data file A1, initiates action to merge its data with Agency B, controlling file B1. In order to conduct the merge operation, the first agency creates file A'1, containing complete identification (1) on each individual and a cryptographically encoded record (A') of each individual's attributes (e.g., academic performance, personality characteristics, etc.). Encoding is based on a computing algorithm which must be unavailable to the Agency B.

Data File A'1 is then transmitted to Agency B and is merged with file B1 by this agency. The merging is based on the common identification included in files A'1 and B1. As a record from one file is matched and merged with a corresponding record from the second file, Agency B deletes the identifier in both records.

As a result of the merge and delete operations, the file labeled A'B is produced. Each encoded statistical record from file A' is associated with the proper statistical record from file B, and the records are virtually anonymous.

File A'B is returned to Agency A, which then decodes the records previously encoded. The decoded statistical file AB is then ready for editing and analysis by Agency A.

Under optimal use of Model 2.0, Agency A has the data it requires in statistical form (i.e., devoid of identifications), and has never had access to identifiable records controlled by Agency B. On the other hand, Agency B has had no knowledge of the statistical attributes maintained in Data File A'1 or A1. Before considering the possibilities of corrupting Model 2.0, suppose we examine the flexibility of the model; inhibiting abuse of the model is conditional on this flexibility.

Variations on Model 2.0

One can devise at least four useful variations on Model 2.0 by manipulating the identity of Agency B and imposing minor constraints on the flow of information implied by the model. Assuming that Agency A represents a single researcher who initiates action, we can consider Agency B as (1) a single institution; (2) several independent institutions; (3) a specific researcher or research group; and (4) the respondent. Each identity of Agency B suggests different administrative regulations and different reference groups to which anonymity is most pertinent; these characteristics are discussed in the following section.

Single Institution: In many instances, the social scientist may wish to merge his own data with information controlled by public or private institutions. Municipal, state and Federal agencies may, for example, maintain demographic and medical data on individuals from whom the researcher has already acquired data. Private agencies, including schools, medical institutions, market research and polling organizations, may also have obtained data of interest to the researcher. Insofar as these institutions have formal regulations for preventing third party access to identifiable records, the usual merge operation implied by Model 1.0 is not acceptable. In this situation, Model 2.0 becomes a convenient device for merging the researcher's data with institutional records without violating any institutional regulations. That the model is feasible is evident from the social experiments conducted by Schwartz and Orleans (1967). These researchers employed the model in merging experimental data with IRS records on the same individuals without compromising the anonymity of the individual with respect to his own tax record.

More subtle uses of Model 2.0 concern those institutional records which fall into the category of public information or into the more ambiguous area which Lister (1969) designates "pseudo-public records." In either case, the bona fide legal difficulties which the researcher confronts in accessing these records may be exacerbated by ambiguous institutional regulations, by vaguely defined statutes and laws, or by idiosyncratic enforcement of regulations, e.g., by institutional administrators.

If the researcher can anticipate such difficulties in research which is endorsed but is also impeded by (ostensible or real) concern about confidentiality of records, then Model 2.0 can be used to resolve the issue and achieve research objectives.

Multiple Institution Case: A schematic diagram, representing the multiple-institution variant of Model 2.0, is presented in Figure 3. It should be evident that logistical problems become much more complex when more than one separate institution is involved with Agency A in merge operations -- at least twice as many encode-decode operations are implied if the pattern of Model 2.1 is used with Agency A and Agency B. Specifically, Agency A must encode its data AI; Agency B must encode its data file A'BI and transmit the encoded file A'B'I to Agency C.

Insert Figure 3 about here

Encoding of files AI and A'BI are necessary in order to prevent personnel at Agency C from interrogating identifiable records. When File CI is merged with the encoded data, any identifiers are removed and the resulting file, A'B'C is returned to Agency B for partial decoding. Having decoded data pertaining to File B, the File A'BC is returned to Agency A for further decoding, editing and analysis. Given ample time, funds and accurate processing of data files, all these tasks can be performed easily. However, I know of no good example to illustrate an actual application of the model.

It is interesting to note that this variant of Model 2.0 provides a kind of primitive resolution to the problems and issues implied in the abortive proposal for National Data Center (Dunn, 1968). Rather than accessing all data under the auspices of one governmental agency, i.e., the National Data Center, the independent researcher could, for example, solicit and merge identifiable information from both the U.S. Census Bureau and the Inte . Revenue Service without violating rules for confidentiality, by using the model. A similar variation might involve separate social research agencies or social scientists all participating in a cooperative program which depends on a common pool of subjects. Each agency, for example, could maintain a unique set of data on the same subjects, or, each data file might represent one descriptive time frame or cross-section for static, descriptive research enterprises; the total merged data file constitutes an empirical basis for longitudinal research. The implicit assumption here is that all agencies would cooperate in providing the resources to implement the model or to permit outside manpower to actually merge files under cooperative surveillance.

<u>Independent Researcher</u>: When an independent researcher or research agency constitutes the auspices under which Data File BI is maintained, several kinds of constraints on Agency B's operations can make Model 2.0 a useful one.

In many circumstances, the social scientist promises the respondent that his response will be used only for research purposes and that summary data will be presented only in statistical form. The implication, for many respondents at least, is likely to be that the data will be kept under the auspices of the researcher and that identified records will not be handled in a way which permits a possibility of disclosure of data to any other parties.

A researcher may, however, choose to furnish identified data to a professional colleague for research use, usually with a verbal agreement that the colleague must not abuse or disclose identified records. This kind of exchange is, of course, a cause for ethical and legal concern if full confidentiality was promised initially. Should the respondent or his representatives view this practice as a violation of confidentiality, based on <u>their</u> interpretations of the original promise, then the use of Model 2.0 may help in ameliorating ethical problems. In essence, only statistical information is exchanged under the model, while identifiability of records is preserved in accordance with the original promise of confidentiality. Note that identification of <u>membership</u> in a sample (on which Files A or B are based) is presumed here not to be a violation of the promise except under extraordinary circumstances.

<u>Respondent</u>: Now consider the situation in which Data File B is managed under the auspices of the respondent himself. That is, the respondent is presumed to have some information about himself which is of interest to the researcher. Moreover, this information must be linked with data previously obtained from the respondent in order to maximize its utility. In this situation, the researcher constitutes Agency A (with previously obtained Data File AI) and the respondent constitutes Agency B with information BI.

Under ordinary circumstances, direct inquiry to B from A is a convenient mechanism for soliciting survey information and, when these conditions prevail, Model 2.0 is fatuous. However, there are several situations in which Model 2.0 may become essential. Consider any inquiry to which a response furnishes very unique and socially undesirable facts about an individual. In addition to the response's potential unreliability (or complete absence), the question itself may become illegal, in the extreme according to some experts (e.g., Goldstein, 1969). There are several alternative strategies, based on Model 2.0, which the researcher can employ in circumventing these problems.

For example, the researcher can punch information from each of his records into a single perforated EAM card for each individual. Some of the card columns are left blank for the data to be solicited. The researcher may then furnish each member of his (potential) respondent group with a card, with instructions on its function and use, and with the questions of interest to him. By punching out the appropriate columns and by punching out <u>all</u> perforations in the identifier columns, the respondent, in effect, merges his own data file with the researcher's while maintaining his anonymity. The cycle implied in Model 2.0 is completed when each member of the respondent group returns his card to the researcher.[3]

The researcher's original data set may or may not be encoded. Decoded information would be warranted if there was no reason to expect the information to influence the individual's decision to respond or the substance of his response. The decoded information, together with an explanation of its meaning may be essential if there is some distrust of the purposes or methods of the researcher. On the other hand, the data should be encoded if there is some risk of disclosure to third parties during the process of punching cards and transmitting them to the respondent.

Note that, if the encode-decode operation is eliminated from this paradigm, the model is analogous to the classical mailout-mailback questionnaire scheme when the questionnaires are mailed back anonymously.

Utility and Corruptibility of Model 2.0

The Campbell-Schwartz model, when employed correctly, is attractive in several respects. Its logical basis and composition and the necessary flows of information are all quite simple. Yet, as we have seen, the general concept is quite flexible in that it can be generalized to a variety of organizational situations. Furthermore, the objectives and the steps for implementing the model are clear enough to facilitate communication with researchers, administrators of data files, and with the intelligent layman who expresses a reasonable apprehension about the union of data files. These properties suggest that the model can be a reasonable for merging data when record identifiability in any one file must be eliminated relative to the agency have no control of the file.

There are, however, two major potential weaknesses in the model, which can undermine and perhaps destroy any utility it may have. The first disadvantage is a logistical one: few agencies or individuals who are placed in the role of Agency B may be capable of accurate match-merge operations even when the volume of data is small. Merging large data files can be very expensive, particularly when search and match strategies, whether computerized or manual, are inefficient (see DuBois, 1969, for discussion of this point). When the respondent plays the role of Agency B, implementing the model may be very difficult because of his resistance or indifference to the research, communications problems between researcher and respondent, etc.

A second disadvantage involves possible corruption of the model by Agency A or Agency B. When the encoding transform is a good one, it is impossible for Agency B to corrupt the system unless it had access to the decipher key or to the actual file AI. I will assume that any such access can be prevented by the usual physical safeguards and personnel checks, otherwise there is no real justification for encoding (Peterson and Turn, 1969 describe and evaluate these safeguards).

Agency A, on the other hand, may corrupt the model in at least three ways: encoding duplicate identifiers, using dummy records, and merging data sequentially. Using the first method, Agency A duplicates identifiers in each record, producing a file AII; then, data set A and one set of identifiers are encoded, producing Data File A'I'I. The deletion of I after match-merging by Agency B is fatuous, since Agency A can decipher Data File A'I'B and acquire identifiable merged records.

The second mechanism for corruption involves the use of attribute data as partial identifiers. If each individual's record is completely unique, the statistical record itself constitutes an identifier. Again, the deletion of formal identifiers after the match-merge process by Agency B is fatuous; Agency A's duplicate file of AI can be used with the unique statistical records to disclose the association between the formal identifiers (I) and elements from Data File B. A variant on this method of corruption is also possible through sequential match-merge operations. That is, one can solicit sequential merges of data, using different elements in the B file to construct a dossier on specific individuals in the AI file. Although time-consuming, the strategy is feasible and well-documented by some researchers, notably Hoffman and Miller (1969).

Considering these potential weaknesses of Model 2.0, how can the social scientist (or other agencies) eliminate or minimize risks attached to them. Of the two categories of weakness, perhaps corruptibility is most serious and least amenable to easy solutions. Under this assumption, we attempt to develop mechanisms for inhibiting and eliminating threats of corruption in the next section. An obvious device for ameliorating the logistical problems -- shifting merge responsibility to an independent brokerage agency -- is discussed in the succeeding section of this paper.

More Secure Versions of Model 2.0 and Model 2.1

For inhibiting the possibility that Agency A will subvert the purpose of Model 2.0, three kinds of counter-measures appear to be reasonable -- trusting and/or licensing the initiating agency, monitoring the merge process, and extending the responsibilities of Agency B to limit the access which Agency A has to raw data files. Of these three activities, only the last two can seriously be considered as counter-measures to corruption and only the last activity (resulting in Model 2.1) reduces physical threats with economy.

Trust in the social researcher has been a classical basis for his activities. This trust is often an essential element in soliciting, maintaining and merging data on individuals overtime. It appears to be particularly necessary to the conduct and evaluation of ameliorative programs, be the program directed toward unified groups of individuals or toward a single person. The sociolegal formalization of this trust, or licensing, has also been commonly employed as a mechanism determining the trustworthiness of a particular researcher or research agency. Insofar as trust in Agency A or formal licensing of the agency are justified, and criteria for justification can be agreed upon by involved agencies and the subjects of records, then Model 2.0 is likely to be a useful one. The potential for this trust may be further undermined if punitive measures can be applied to an agency which attempts to interrogate identifiable records by corrupting the model (provided, of course, that corruption of the model is detected). On the other hand, the continuous availability of the merge privilege may furnish research findings which are much more valuable to the agency than benefits of corruption are, making punitive measures less relevant.

Although we admit that the integrity of Agency A is one kind of safeguard against corruption, we must also observe that accurate appraisals of integrity and rigorous licensing requirements are often difficult and time consuming. If we can discover other safeguards we may be able to eliminate entirely the need to rely solely on the apparent integrity of Agency A.

One possible strategy for detecting and preventing corruption of the kinds described relies on the use of monitors during the merge process. That is, Agency B might continuously observe the conduct of the merge and examine the physical contents of data files supplied by Agency A for the merge. The examination of contents must be focused on detecting uniqueness of each and every statistical record and to prevent match-merges of de facto identifiable statistical records. Also, sequential merges can be monitored so as to inhibit attempts to employ the 20 questions strategem in building dossiers on identifiable subjects of the merged files. Monitoring, however, may be too expensive, time consuming, or weak to detect and prevent all but obvious attempts to corrupt Model 2.0. In fact, it would be difficult if not impossible for a monitor to detect the presence of encoded identifiers (i.e., Data File A'I'I supplied by Agency A) if sophisticated enciphering techniques are used. Given these limitations of a monitor process, a more effective counter-measure may be constructed by simply extending Model 1.0.

This extension, Model 2.0, is illustrated in Figure 4.

Insert Figure 4 about here

Model 2.0 implies that Agency B not only match-merge the data, but also furnish a statistical <u>summary</u> rather than raw records to Agency A. Agency A must specify, prior to merging data, all the summaries necessary for its own analysis of the merged data. Provision of statistical summaries reduces markedly the potential for corruption (by encoding identifiers or by de facto statistical identifiers) when certain conditions are met. The conditions are simple and depend on the kind of data condensation which is prescribed and developed. Cross-tabulations and associated relational statistics (e.g., Chi-square and phi coefficients) can be produced under the constraint that the observed frequencies within all cells be above a certain number. Similarly, Agency B might require that all parametric statistics be based on at least 30 observations within a given group. Note that monitoring is still necessary to detect and prevent the sequential method of corruption when frequencies counts of cross-tabs are solicited periodically by Agency A.[4]

When statistical summaries, rather than raw data, are furnished, cryptographic encoding takes on a bit different cast. Some of the classical encoding mechanisms -- infinite key transforms, for example -- change the character and mathematical properties of the record completely. Statistics based on such transforms are meaningless. In lieu of such key transforms, the researcher can exploit one obvious option which depends on the kind of scales (nominal, ordinal, interval) characterizing the data and on the type of summary prescribed. For nominal scales, a short key transform which isomorphically relates actual categories and encoded observations is useful. For ordinal or interval scales, common statistical transforms (e.g., Thoni, 1966) of data or simple linear functions of the actual data can be employed. The statistical transformations may, in any event, be essential for intelligent analysis of the data. Both statistical and linear transforms <u>disguise</u> rather than cryptographically encode data. However, this strategy ought to be sufficient to inhibit overt interrogation of data files by the broker, Agency B, and Agency A <u>when each of these groups monitors the merge process</u>. In addition to on-site monitoring, the usual precautions against interrogation of files stored (temporarily) in a computer or EAM equipment, can be used to prevent duplication of files for later interrogation, etc. (see Peterson and Turn, 1968 for a complete list of precautionary measures in a computer environment).

One additional safeguard can be employed by Agency B to minimize the utility of the potentially identifiable records in File BI, in Model 2.0, or Model 2.1. Agency B can simply innoculate errors into the records on that copy of file BI which is involved in the merge. It is possible for Agency B to control the statistical properties of the random error which is introduced and, although the integrity of any particular record is undermined, the statistical condensations of the merged (imperfect) data file can be corrected for errors using common mathematical techniques (see, for example, Cochran [1968]). Corrections may be made by Agency A as part of data summarization in Model 2.0, and by Agencies A or B in Model 2.1 when distributional properties of the error are known by both agencies. For a description of the limitations of this technique, see Boruch (1970).

Using a Broker: Models 3.0 and 3.1

In this section, we incorporate an intermediary agency into the formal structure of the earlier models; two principal functions of this broker include match-merging data (Model 3.0) or maintaining code linkages (discussed below as Model 4.0).

Model 3.0 is illustrated in Figure 5; the figure represents a direct extension of Model 2.0, containing most of the same elements and flows of information.

Insert Figure 5 about here

In this model, Data File AI is generated by Agency A, and the statistical portions of each record are encoded (i.e., AI becomes A'I). Similarly, Agency B generates encoded Data File B'I, using a different enciphering algorithm. The two resultant files, A'I and B'I, are match-merged by the broker, based on the unique identification portion of each record. Encoding, of course, protects the files against interrogation by the broker during the merge process. Following the match-merge operation, all identifiers are deleted and Data File A'B' is returned to Agency B for decoding. This partially decoded file, A'B, is then sent to Agency A for decoding, editing, and analysis.

By moving responsibility for match-merging from Agency B to the broker, we have reduced some of the technical expertise and manpower required of Agency B, thereby ameliorating a disadvantage of Model 2.0. A decoding operation has been added but this is likely to be no more of a problem for Agency B than the original encoding. If Agency B considers this operation to be an unwarranted imposition, the agency can simply provide Agency A with the decipher code and let Agency A decode the B' sections of the merged file.

Although an economic problem is resolved by Model 3.0 and the encoded data are secure against disclosure to Agency B and the broker, the potential for corruption of the system by Agency A still has not been reduced. Model 3.1, an obvious extension of Model 2.1, presents one resolution of this problem. The broker, in this case, is assigned responsibility

---

Insert Figure 6 about here

---

for summarizing the data (where the summary is specified a priori by Agency A) as well as merging the files. As in Model 2.1, monitoring is necessary to prevent use of the 20 questions strategem in corrupting the system. Also, a transformation of the data and secrecy of file contents are essential for eliminating the possibility of the broker corrupting the system. Also, innoculation of random error with known parameters will help to minimize the utility of identifiable records to the broker and to each agency.

Perhaps the best method of further inhibiting the broker's ability to interrogate identifiable records is to cryptographically encode the identifiers in each file, using an encoding scheme developed jointly by Agency A and Agency B. So long as the same encode system is used in each matching identifier, the merge can be conducted yet the possibility of interrogation is virtually eliminated.

Variations on Models 3.0 and 3.1 and Their Corruptibility

Models 3.0 and 3.1 can be manipulated in the way prescribed earlier in order to demonstrate the variety of situations to which the models are applicable. Instead of varying the identifiers of Agencies A and B, however, we may change the identity of the broker more conveniently. Three such variants are considered here: a "neutral agency," respondent, or researcher, each considered as the broker in the system.

<u>Neutral Agency</u>: In some instances, it may be possible to engage an agency which is relatively independent of the other agencies involved in Model 3.0 and 3.1 and of any third parties which might attempt to interrogate merged files. For example, a governmental agency such as the Census Bureau can play the brokerage role when the effectiveness of the intermediary is dependent on constitutional protection of potentially identifiable merged files. A need for such protection is evident if the union of files jeopardizes respondents more than separated files do, or if the data for each separate file had been gathered initially under statutory or constitutional protection. The use of the Census Bureau in a more generalized brokerage role, and the use of a specially created government agency to fulfill a similar role for social scientists has been discussed by Dunn (see Westin, 1965) and recommended in some published legal opinions, e.g., in the <u>Valparaiso Law Review</u>, (1969).

One of the problems here is that Federal agencies are not likely, at least in the near future to regard themselves as brokers for social scientists who wish to merge data. Unless legislation or regulations are created to specify that this must be one of their missions, the agencies will probably not have the manpower, computer facilities or other logistical support to implement Models 3.0 or 3.1.

Under these conditions, commercial service organizations might fulfill the role of broker with dispatch and with a good deal of security for the data. Highly confidential and secret records are processed routinely by computer service groups, for industry and for municipal, state, and Federal governments.[5] When the identifiers and statistical data are encoded by Agencies A and B and when there is strict monitoring of the merge process (with safeguards against secret reproduction of files, merged or otherwise), there appears to be no critical problem in using such an agency. The agency, of course, cannot furnish statutory protection for the files it processes, as the Census Bureau or similar variants might be able to do.

Respondent: Suppose that Agencies A and B, be they independent researchers or institutions, cannot agree on a choice of institutional broker. Their unwillingness to do so may be caused by general distrust of the candidates for brokerage or their suspicion of the model, by the expense and logistical problems involved in implementing the model, or by the difficulties in monitoring the merge (and perhaps statistical summarization) process.

Under these circumstances, the individual on whom records are maintained (i.e., the respondent), can substitute as a reasonable broker. That is, the respondent can merge data through mailout-mailback methods or through more controllable techniques within institutional environments, when his record from each file is presented to him in appropriate physical form. This strategy is analogous to the one presented earlier -- match-merging data when the respondent is identified as Agency B in Model 2.0. As in Model 2.0, encoding-decoding operations are optional, depending on the potential for unwarranted disclosure of information during the record's processing and transmission.

Using the respondent as broker is inconvenient and inferior to other strategies insofar as nonresponse rates are likely to be high and logistical problems are serious. Moreover, any of the corruption strategies mentioned in connection with Model 3.0 are applicable in this case. The respondent-broker substitution is not a good one unless there are some other guarantees that Agency A is not interested in obtaining identifiable records. If these guarantees are absent, a fourth agency might be introduced to the system; the agency must be dedicated entirely to computing summaries of the data, destroying merged records, and furnishing the summaries to Agency A under the safeguard conditions prescribed earlier.

Researcher: Using the researcher as broker in Models 3.0 or 3.1 requires a slightly different interpretation of the information exchanges described earlier. Specifically we can impose the constraint that Agency A and Agency B are actually the respondent at two different points in time. Rather than encoding statistical portions of the record each individual encodes his identification uniquely and in accordance with his own enciphering technique. The consistent use of this alias at points A and B in time, in conjunction with the researcher to act as broker permits match-merging and summarizing the data. Aliases can be constructed systematically using a variety of instructions (see Boruch [1970]) and so long as the researcher lacks the ability to link aliases with true identification, the anonymity of the respondent is protected.[6] (Note that flow lines in Models 3.0 and 3.1 must be adjusted so that merge, summarization and analysis of results are conducted under the auspices of the researcher.)

Code Linkage Systems: Model 4.0

In some research programs, code linkages between different data files may be maintained indefinitely for possible use in merging the files. The justification for the linkage and the physical generation of the linkage seem to depend in a large measure on the kind of research which is being conducted. Therefore, employment of code linkages is discussed primarily in terms of published examples of the systems. The basic composition of the code linkage model is given in Figure 6.

Insert Figure 7 about here

The model is characterized by three basic elements; the two agencies which maintain the data files and a broker to facilitate match-merging. If we delete the broker from the system, this model becomes closer to Model 2.0 in conceptualization; the benefit of having the broker depends on the broker's ability to implement those processes which Agency B cannot. The model works in the following way.

Each element of statistical data in each record of Data File A is encoded by Agency A; identification has been previously encoded under a different encrypting technique. Similarly, Agency B encodes its own statistical data using a unique encoding technique; identifiers in this data file, as in Data File A, have been encoded previously (I") using an encoding scheme which differs from all others used in the process. The two resulting data files, A'I' and B'I", are transmitted to the broker, which then merges the data based on its knowledge linkage between coded identifiers (i.e., I'I"). The resultant Data File A'B', is returned first to Agency B for decoding and then to Agency A for further decoding and analysis.

This model exhibits several potential benefits over Models 3.0 and 3.1. Protecting the records against corruption by Agency A is unnecessary under optimal operation of the model, since the model specifies that Agency A maintains only encoded identifiers in its own statistical record.

The likelihood that the broker can decipher both encoded records and encoded identifiers is low, suggesting that the broker can be monitored under less stringent conditions (i.e., requiring less manpower) than in previously suggested models. The opportunity for third parties to penetrate any files during the processes implied by the model is also minimal. Finally, if the code linkage is maintained under very secure auspices (free from third party interrogation, legal or otherwise), the routine maintenance of data as well as merge process is virtually free from the possibility of any disclosure of information.

So far, I have not mentioned the actual mechanism for generating the encoded identifiers and code linkages. This mechanism is crucial to the integrity of Model 4.0 and to its distinctiveness relative to other models. How might such a code linkage be generated and maintained?

Two published descriptions of code link use are examined below, with special regard for the method of generating code linkages and the corruptibility of models implied by each description. The Manniche-Hayes system is an early variation, developed well before the interaction among social research, computerized records, and the privacy issue became important. A second model, exemplified by the ACE LINK FILE System, was created in direct response to public and professional apprehension about maintaining identifiable records in a longitudinal research program.

Manniche-Hayes System

Figure 8 illustrates a system developed by Manniche and Hayes which permits a researcher to solicit and merge information on a pool of individuals, using two sources of data. The two sources include a broker who obtains information from identifiable archival records, and the respondent himself. The broker's function is to control solicitation of data and to construct the code linkage system which is used by the researcher to merge data furnished by the respondent and by the broker.

Insert Figure 8 About Here

Figure 8 is interpreted as follows. The broker compiles Data File AI from existing identifiable records. Then, identifiers in the file are encoded and the resulting file AI' is supplied to the researcher. The broker simultaneously creates a file linking true identifiers with the encoded identifiers; this dictionary file is designated II' in the figure. Each respondent also creates two kinds of records, where identifiers in records are encoded arbitrarily by the respondent himself. Data File BI is then transformed to BI" and supplied to the researcher. Each element in a second dictionary file II" is supplied to the broker by each respondent.

The broker, having both dictionaries, II' and II", match-merges these on the basis of common true identifiers (I) and supplies the resulting code linkage file to the researcher. Given Data Files AI' and BI" and the code linkage between the files, II", the researcher can merge the files easily.

Utility and Corruptibility of the Manniche-Hayes Model

Assuming that the broker is not corruptible, and it would be difficult if not impossible for the researcher to obtain any identifiable records on the respondents. The usual physical safeguards and monitoring devices can be used to inhibit overt attempts by the researcher to corrupt the system; the absence of access to any identifiable records makes corruption via encoded identifiers almost impossible. The 20 questions strategem by the researcher can probably be detected by the broker if the broker monitors the data which it supplies to the researcher and the data supplied by the respondent.

The most obvious weakness in the system is the broker, simply because this agency does have access to fully identifiable records in one file and to the complete code linkage system. If the broker is officially responsible for maintaining file AI, then there is no particular threat unless the broker has a definite interest in expanding its information system to include File BI; if there is little physical security for the BI" file, the broker may gair access to it and conduct it's own data merge using it's dictionary files.

The potential for collusion between researcher and broker is also evident. If, as Manniche and Hayes suggest, the broker is a professional colleague of the researcher, the likelihood of collusion is bound to be perceived as high, regardless of it's true likelihood. In order to lower the probability of collusion, we might employ some of the strategies described earlier. The brokerage role can be limited to, say, neutral agencies which can gain nothing by collusion and may suffer punitive action as a result of collusion. For example, a school registrar might be required by administrative regulations and/or municipal law to insure that his records are never identifiable to third parties. Punitive action can be taken against the broker if its attempts at corruption of the system are detected. In this case, the Manniche- Hayes model is not substantially different, in advantages and limitation, from the Campbell-Schwartz model.

ACE Link File System

One of the most interesting variations on Model 4.0 has been developed recently at the Office of Research of the American Council on Education (Astin and Boruch, 1970; Boruch, 1969). Illustrated in Figure 9, this experimental system employs a foreign intermediary (i.e., a broker) to maintain the code linkage file (I'I"). Data File BI represents information gathered by the research agency at time T_1, while information for Data Files AI is collected and cosolidated at some later point in time, T_2.

Insert Figure 9 about here

The link file itself is created at T_1, in conjunction with the transformation of identifiers in file BI. Also at time T_1, a dictionary is created (in effect) with three kinds of data: true identifiers (I) and two sets of encoded identifiers I' and I"). The two encoded sets differ from one another in physical contents and in the manner in which codes are generated. File BI" is constructed by replacing true identifiers (I) with one set of encoded identifiers, resulting in the file maintained on-site by the research agency (BI"). After this operation, the dictionary is used to construct the link file I'I" and a second dictionary II'. The link file I'I" is then sent to the broker; the first dictionary, II'I", is destroyed as are any researcher's copies of Data Files BI or I'I".

Later merge operations are conducted in two stages. The broker merges Data File AI' with the link file, I'I", and deletes the set of identifiers I'. When this file AI" is returned to the research agency, it is merged with file BI" by the researchers on the basis of common identifiers, I".

Utility and Corruptibility of the Link File System

When the model is adhered to rigorously, the Link File System demonstrates some important ways for preventing interrogation of identifiable records during the merge process. Data File BI" is virtually free from penetration even by Office of Research staff, since identifiers are encoded and the decipher key (II'I") has been destroyed. Similarly, the process of merging AI' with BI" is free from threat of the broker's penetration since only encoded identifiers in File AI' are supplied to the broker. The physical merge process appears to be quite safe because the researchers themselves cannot decipher the encoded identifiers in AI" and BI". A final benefit is that data file BI" (and all succeeding data files) can be maintained without risk of extra-legal or legal interrogation of files. True identifiers are legally inaccessible by subpoena, if the broker is a foreign agency and if the agreement between broker and researchers specifies that the linkage be kept secret and secure, even from the researchers themselves.

These and other advantages described by Astin and Boruch (1970) are impressive. However, this model is vulnerable to some of the same corruption strategies mentioned in the context of Model 2.0. The problems described below are based on a few of the writer's own perceptions, and on two very professional critiques supplied by Dr. Rein Turn of Rand Corporation and Dr. Lance Hoffman of University of California at Berkeley (both personal communications).

Suppose we consider possible corruption of the system by members of the research agency. First, there is no real guarantee that the agency actually destroys copies of files BI or the code linkage I'I"; given the files AI and II', of course, completely identifiable records (of the form ABI) can be constructed, subverting the purpose of the system.

Actually, covert duplication and maintenance of files BI and I'I" by a member of the research agency or the failure to destroy original files at the appropriate time is not really necessary to permit later interrogation of identifiable records. One need only construct a dummy variable in each and every record of File BI"; the dummy must contain covertly encoded true identifiers and link file characters. This strategy is a simple extension of one mentioned in connection with Model 2.0 -- encoding identifiers.

The brokerage agency constitutes a second potentially weak element in the system. The 20 questions strategem can be employed here to corrupt the system. In this case, the broker's objective may be to construct identifiable records corresponding to Data File AI. If the broker has access to the list of individuals whose records are maintained it may then construct its own file of commonly available data about those individuals. Given these data, its copy of Data File AI', and the documentation for the file, the broker may be able to interrogate the file and build its own dossiers, using the 20 questions strategem. This would be particularly easy to do with relatively small numbers of individuals and a large number of elements in each record. One convenient way to ameliorate this difficulty has been suggested by Lance Hoffman: The researchers must encode the statistical position of each record (that is A is transformed to A") using a unique encoding scheme which is unavailable to the broker.

Dr. Turn has emphasized the weaknesses of foreign brokerage as opposed to domestic maintenance of link files. He contends that one objective of the system -- keeping link files secure from legal penetration -- would not be met if certain plausible events occurred. In such occurrence, foreign courts may submit quite readily to our government's requesting the linkages. Normal international regulations may be quite unnecessary, if informal disclosure of files is perceived as being a friendly understanding between governments or as an amicable political gesture.

If the foreign agency chooses not to abide by its contract to maintain the link file (or if it decides to sell the information), then the system's functional utility is destroyed. Moreover, successful prosecution of the broker may be so difficult and time consuming that the system's utility would be imparied considerably, if not destroyed entirely.

These kinds of weaknesses in brokerage are improbable (although still possible), if the broker is selected carefully, and if there are some external guarantees of adherence to the model.

In the ACE system, one such guarantee is the ACE agreement to provide exactly the same link file services to researchers at the foreign agency. If the foreign broker ignores its own responsibility toward ACE, then presumably, ACE can make similar reprisals. This kind of countermeasure is not particularly appealing (if only because it is so destructive) but it may be a useful mechanism for deterring violation of formal contract or informal agreements.

<u>Variations on the Manniche-Hayes and ACE Systems: Relation to Earlier Models</u>

Both Manniche-Hayes and ACE Link File Systems were developed with a specific purpose different from the function of models considered earlier; the reader will recall that Models 2.0-3.0 were dedicated to preventing disclosure of <u>one</u> file used in a merge operation. On the other hand, the Manniche-Hayes paradigm eliminates the need for the researcher to maintain any identifiable record for any length of time. The ACE System limits the maintenance of identifiable records to short periods of time (i.e., during the period a link file is created).

Both models can, with minor adjustments, be treated as variations of the early models and contrariwise, building on the earlier models results in systems that provide many of the same services that the Manniche-Hayes and

Link File Systems provide. In the first place, one could adjust the Manniche-Hayes approach to permit the researcher full access to one set of identifiable records; this eliminates the need for the link file and makes this situation identical to Model 3.0 in function, benefits, and shortcomings. Building from the earlier models, specifically the multiple institution (or time) variant of Model 3.0, we have a situation which is identical to the longitudinally operated Link File, except for the maintenance of a code linkage (and associated benefits and shortcoming of the linkage strategy).

Both the Manniche-Hayes and Link File Systems can be manipulated in much the same manner as earlier models. Research agencies, formal institutions or the respondent himself, can be used to complement the researcher and broker in each model. In each model, the broker may be manipulated; when the respondent himself is used as a broker, the Link File System is quite similar in form and function to the situation in which the respondent plays the same role in Model 3.0.

Difficulty in Applying the Models and Consequences of Their Use

Three kinds of problems -- technical, contextual and logistical -- are inherent in any implementation of the models described here. At the core of technical problems is the need for encoding alphanumeric information in each one of the models. Techniques for cryptographic encoding are likely to be unfamiliar to most social scientists, computer scientists or managers of data files. Moreover, there appear to be no standardized criteria for appraising the adequacy, efficiency and costs of the techniques currently employed by commercial and military organizations. Although informal guidelines are currently available, it is likely that the nature of privacy transformations and their effectiveness will change considerably in the near future as the algorithms used in code generation are more closely linked with computer control systems developments and microminiature circuitry advances (Taylor and Feingold, 1970). Under these circumstances, the social scientist who wishes to employ one of the models must learn to develop his own encode-decode systems based on existing information. A brief description of encode techniques and a selected bibliography is provided in Appendix I.

The second problem, a contextual one, involves appraisal of the need for a model and of potential corruptibility. Need is obviously a function of the nature of the data being merged and the interest of a participating agency or some third party in gaining access to identifiable records. These factors are not easy to evaluate themselves, much less with respect to the costs of employing one of the models and ancillary safeguards. One examination of this issue is given by Boruch (1970), but much more systematic and empirical exploration is needed. The comments made earlier on shortcomings of the models represent only <u>one</u> kind of appraisal technique, based essentially on examination of important elements in the models information flow. Even in this context certain kinds of corruptibility have been ignored, e.g., collusion among agency personnel. Other methods for appraisal developed and these may be much more effective insofar as they permit detection of attempts to corrupt the systems, and insofar as they furnish use with meaningful quantitative indices of the risk of corruption. Taylor and Feingold (1970) present an approach to quantifying the feasibility and utility of certain safeguards which function as counter measures to corruption of computerized record systems. Still another approach involves the creation of prototype systems, coupled with a devil's advocate group whose function it is to penetrate the systems. At MIT, for example, students play this role in effect, when they succeed in entering a "secure" resource-sharing system, without authority (knowledge of pass words, etc.) in order to get their homework done. The appointment of a devil's advocate as a formal position as a secure data environment has been suggested by a number of computer experts and social scientists.

A third major problem concerns the accuracy and the ability to manipulate data files. The researcher who merely assumes that the institutional files (or his own files) being used in the merge is likely to be disappointed. We know that administrative records are subjected to distortion in a variety of ways and that documentation on accuracy is, as a rule, absent (see, for example, Campbell, 1969).

If the data are known to be accurate, however, a second problem arises -- overload in demands on institution data files. Since the number of data banks is small relative to the number of available respondents, at least, and relative to the number of social scientists, the risk of swamping institutions with requests to match-merge data is high. Without a formal (expensive) mechanism to meet a high demand, few projects are likely to be completed. Unless researchers are willing to pay for personnel and machine time used on the project, as well as overhead and service charges, official cooperation by institutions cannot reasonably be expected.

Assuming that these problems can be solved at least partially, we can anticipate certain benefits from wide-spread use of models by the social research community. Acting on the recommendations made by Miller (1970), I will try to list the important implications of the methodology presented here and to evaluate them relative to a more general reference system.

The most obvious useful result is the enhancement of the social researcher's ability to obtain and analyze data without infringing on the privacy of the individual. Expansion of the pool of data -- in kind, magnitude, and quality --

is perhaps one of the more useful benefits to the social science enterprise. The conduct of research will, in some instances, be rendered much more economical and efficient: There are fewer political and administrative problems in collecting the data and the cost of merging the data is negligible by comparison to the cost of actually soliciting and obtaining it through a formal survey.

The availability of these models may stimulate more secondary analyses of the data -- another economic benefit for the researcher, funding organizations and, hopefully, society. In addition, the data may be of sizeable volume and stable enough to permit cheap replication, an opportunity which cannot be considered trivial in the social sciences.

A more generalized benefit concerns the need for explaining science to the public, where "public" means institutional administrations. The cooperation between administrators and researchers, their information exchanges, and the benefits which both groups derive from this cooperation may contribute substantially to the integrity and to the development of social science.

References

Astin, A. W. and Boruch, R. F. A "link file" system for assuring confidentiality of research data in longitudinal files. American Educational Research Association, 1970, 7, 615-622.

Boruch, R. F. Maintaining confidentiality in educational research: A systemic approach. American Psychologist (in press).

Campbell, D. T. Reforms as experiments. American Psychologist, 1969, 24 409-429 (a).

Campbell, D. T. Measuring the "opportunity effects" of university teaching by means of eleven-plus decisions. Mimeo Research Proposal-Report. Psychology Department, Northwestern University, 1969 (b).

Carroll, J. M. and McLelland, P. M. Fast "infinite-key" transformations for resource sharing systems. Paper #191. London, Ontario: University of Western Ontario, 1970.

Cochran, W. G. Errors of measurement in statistics. Technometrics, 1968, 10, 637-665.

DuBois, N. S. A solution to the problem of linking multivariate documents. Journal of the American Statistical Association, 1969, 64, 163-174.

Dunn, E. S. The idea of a national data center and the issue of personal privacy. The American Statistician, 1967, 21 (1), 21-27.

Feige, E. L. and Watts, H. W. Protection of privacy through microaggregation. In R. L. Bisco (Ed). Data Bases, Computers, and the Social Sciences. New York: Wiley Interscience, 1970.

Goldstein, A. S. Legal control of the dossier. In Wheeler, S. On Record: Files and Dossiers in American Life. New York: Russell Sage Foundation, 1969.

Hoffman, L. J. and Miller, W. F. How to obtain a personal dossier from a statistical data bank. Datamation, May, 1970.

Lister, C. The confidentiality of pupil's school records: A background paper for the Working Conference to Consider Certain Legal Aspects of the School Counselor Role. Unpublished Manuscript. Yale Law School, May, 1969.

Manniche, E. and Hayes, D. P. Respondent anonymity and data matching. Public Opinion Quarterly, 1957, 21 (3), 384-388.

Miller, G. A. Assessment of psychotechnology. American Psychologist, 1970, 25, 991-1001.

Peterson, H. E. and Turn, R. System implications of information privacy. Paper presented at Spring Joint Computer Conference, Atlantic City, April 17-19, 1967, (Mimeo report reproduced by Rand Corporation, Santa Monica, California).

Schoenfeldt, L. F. Data archives as resources for research, instruction, and policy planning. American Psychologist, 1970, 25, 609-616.

Schwartz, R. D. and Orleans, S. On legal sanctions. University of Chicago Law Review, 1967, 34, 274-300.

Steinberg, J. and Cooper, H. C. Social security statistical data, social science research and confidentiality. Social Security Bulletin, 1967 (October), 3-15.

Russell Sage Foundation. Guidelines for the Collection, Maintenance and Dissemination of Pupil Records. Russell Sage Foundation Conference Report, 1970.

Taylor, R. L. and Feingold, R. S. Computer data protection. *Industrial Security*, August 1970, 20-29.

Thoni, H. Transformations of variables used in the analysis of experimental and observational data: A review. *Statistical Laboratory, Technical Report No. 7*, Ames (Iowa): Iowa State University, 1967.

U.S. Bureau of the Census. Policy governing access to Census Bureau's unpublished data and special services. *Data Access Description, Policy and Administrative* Series, PA-1, Washington, D. C.: U.S. Census Bureau, 1968.

Valparaiso Law Review. Legal note: Social research and privileged data. *Valparaiso University Law Review*, 1970, *4* (2), 368-399.

Westin, A. F. *Privacy and Freedom.* New York: Antheum, 1967.

Wheeler, S. (Ed.) *On Record: Files and Dossiers In American Life*. New York: Russell Sage Foundation, 1969.

Footnotes

[1]Supported by NIMH Grant 1 R12 MH17, 084-03. I should like to thank both Eli Rubenstein, D. T. Campbell, A. W. Astin and A. E. Bayer for providing advice or cirticism on earlier drafts of the paper. However, views expressed in this paper do not necessarily reflect this advice nor should the views of the sponsoring agency.

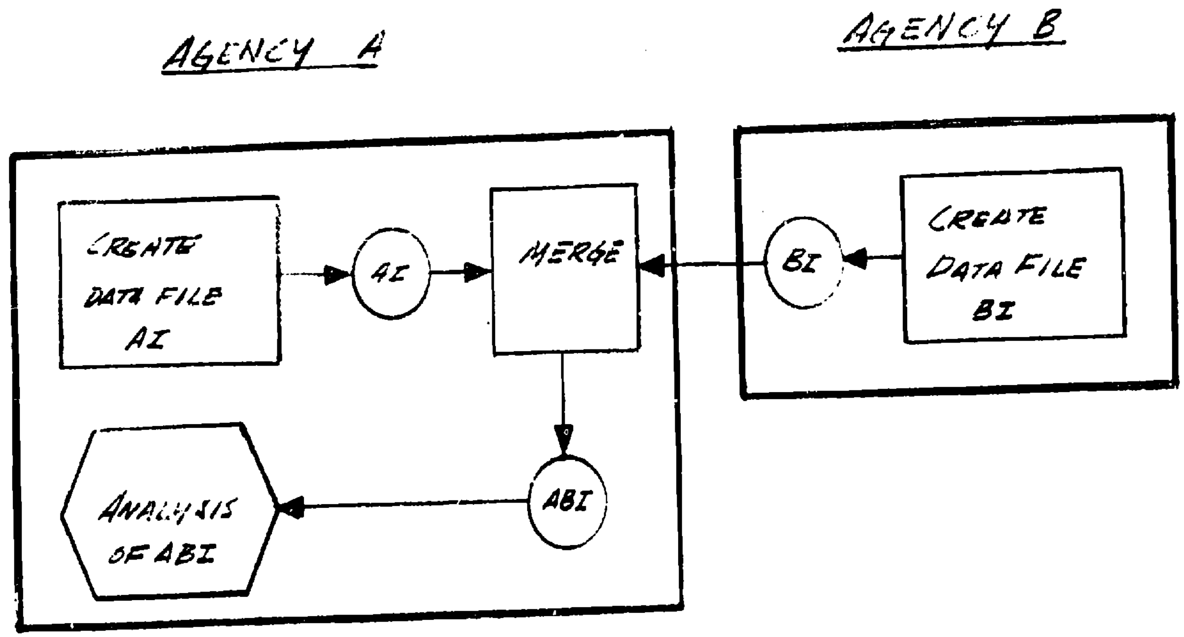
[2]For excellent discussions of the current legal and professional restrictions on accessibility of a variety of organizational files, see Wheeler (1969).

[3]In order to appraise validity of the sample in each case where individual subjects volunteer to respond, a post-card for each subject can be constructed containing only statistical information. Return of the post-card by the subject indicates that the subject responded to the inquiry and returned this response under separate cover.

[4]In some cases, the agency with responsibility for summarizing the data may have the computing facilities necessary for sophisticated ad hoc data condensations, e.g., covariance-correlation matrices, nth order statistics, etc. More typically, however, this capability is likely to be absent. One potentially useful strategy, suited for these conditions, involves micro-agregation of data, where the kind and degree of aggregation is fixed by policy and limits of computing facilities. Sample statistics (e.g., means) are then supplied for groups, rather than individual subjects, and the size and kind of group must be specified a priori for maximum efficiency. Although micro-agregation techniques are still at a primitive stage of development and generally lead to inefficient estimates of parameters, the techniques do appear to be generating interest and research simply because they are a convenient device for preserving anonymity of records (see Feige and Watts, 1970).

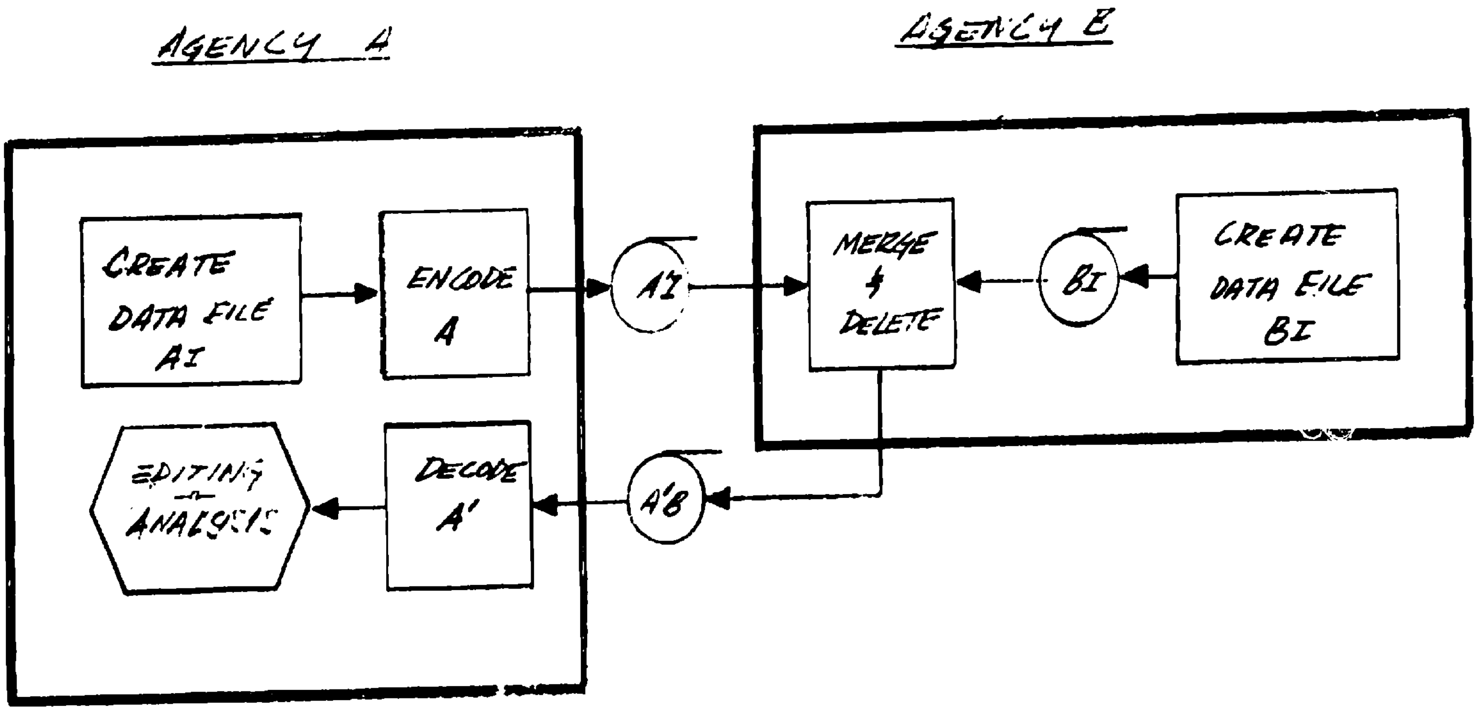
[5]Price-Waterhouse (New York) fulfills such a brokerage role for the Board of Medical Examiners; Agency A corresponds to the Board and Agency B corresponds to a Medical School aspirant who participates in an experimental testing program.

[6]Numeric aliases, created by the subject on the basis of prescribed formula, have been used by Professors Peter Rossi and Eugene Groves in mailout-mailback surveys of college students. Problems in minimizing replication of numbers in such a group suggest that simple alphabetic aliases may function at least as well; with Dr. John Creager, this writer has successfully used subject-created alias names in studies of the same kind of population.



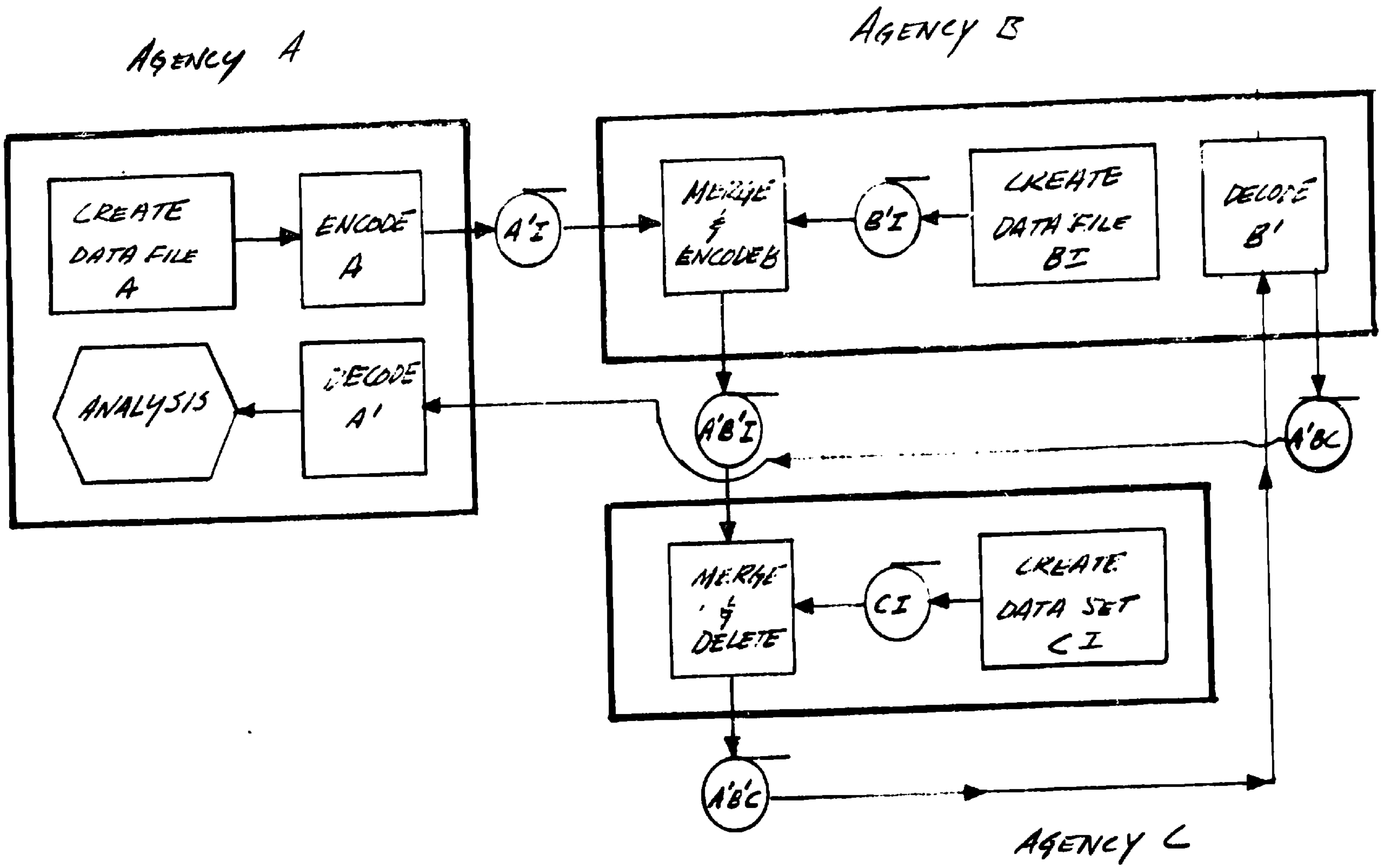
MODEL 1.0

FIGURE 1



MODEL 2.0

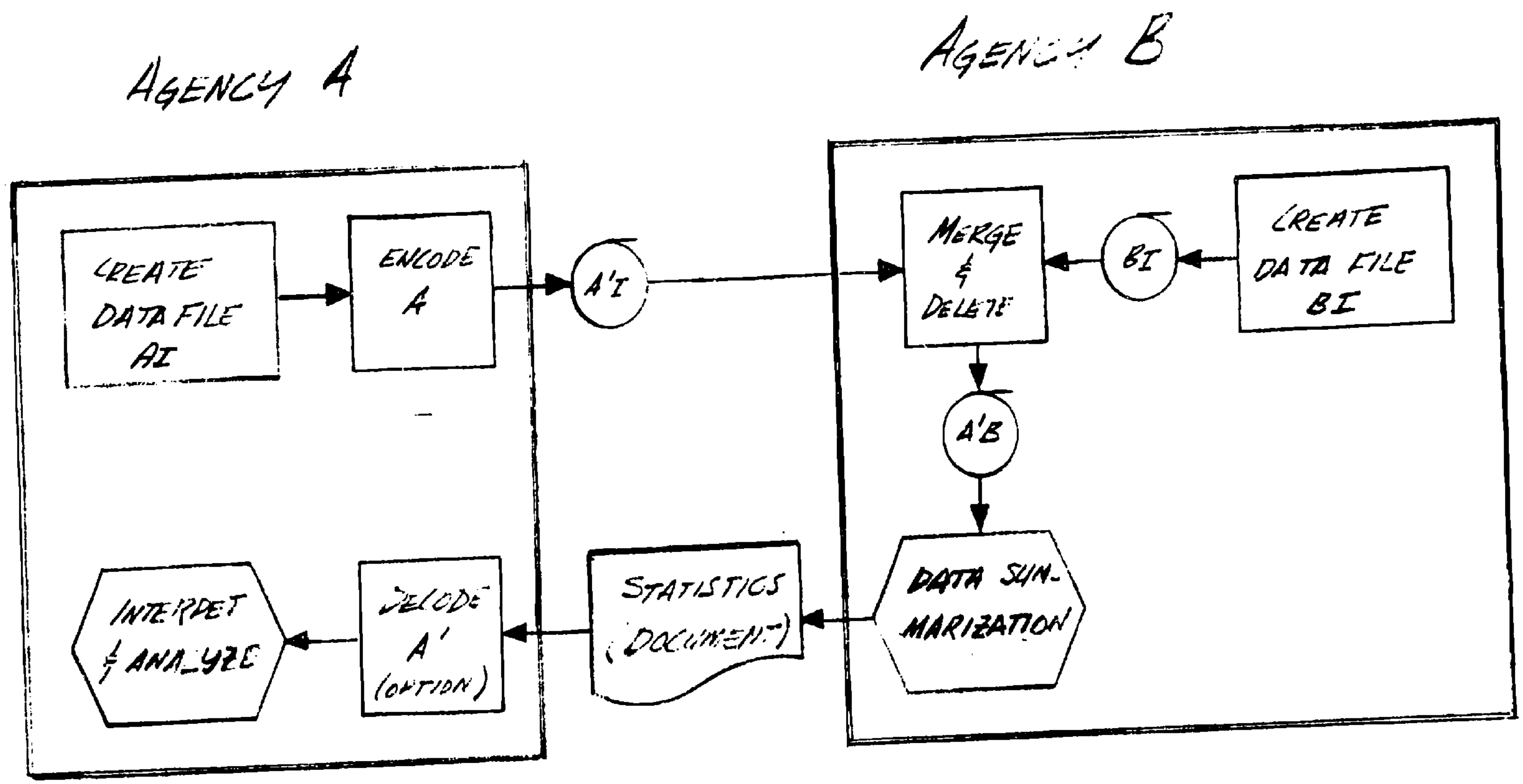
FIGURE 2



MULTIINSTITUTIONAL VARIANT OF MODEL 2.0

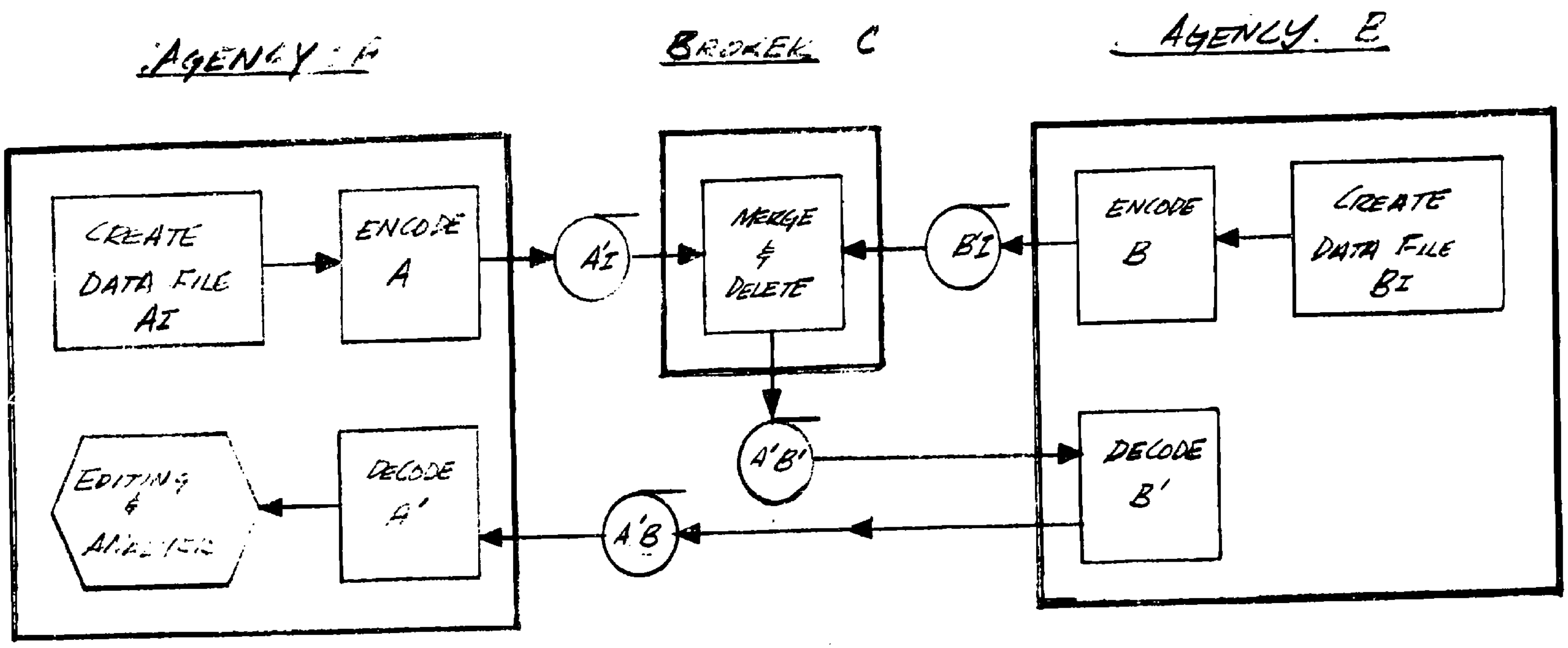
FIGURE 3

BORUCH
1/10/7



MODEL 2.1

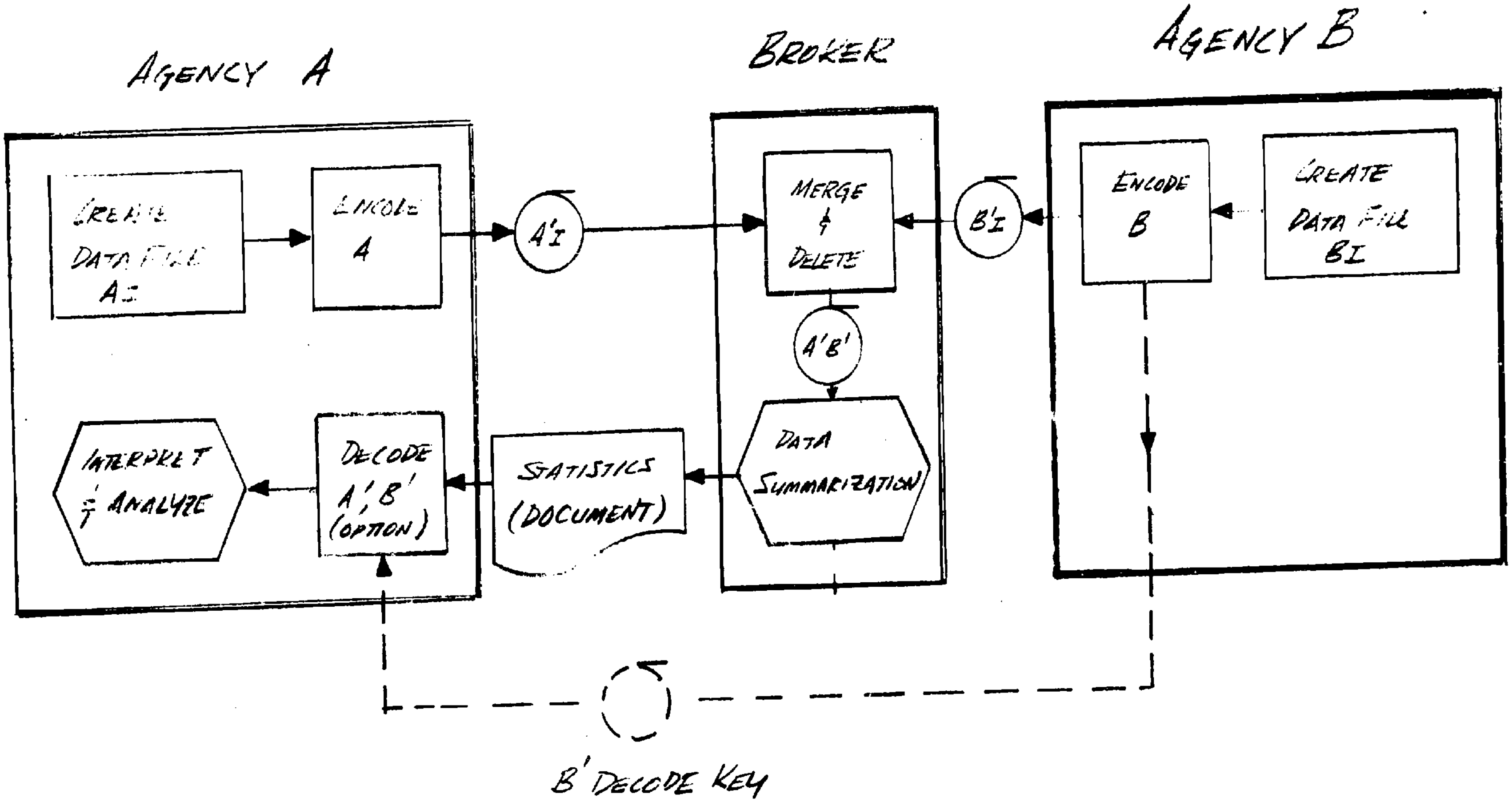
FIGURE 4



MODEL 3.0

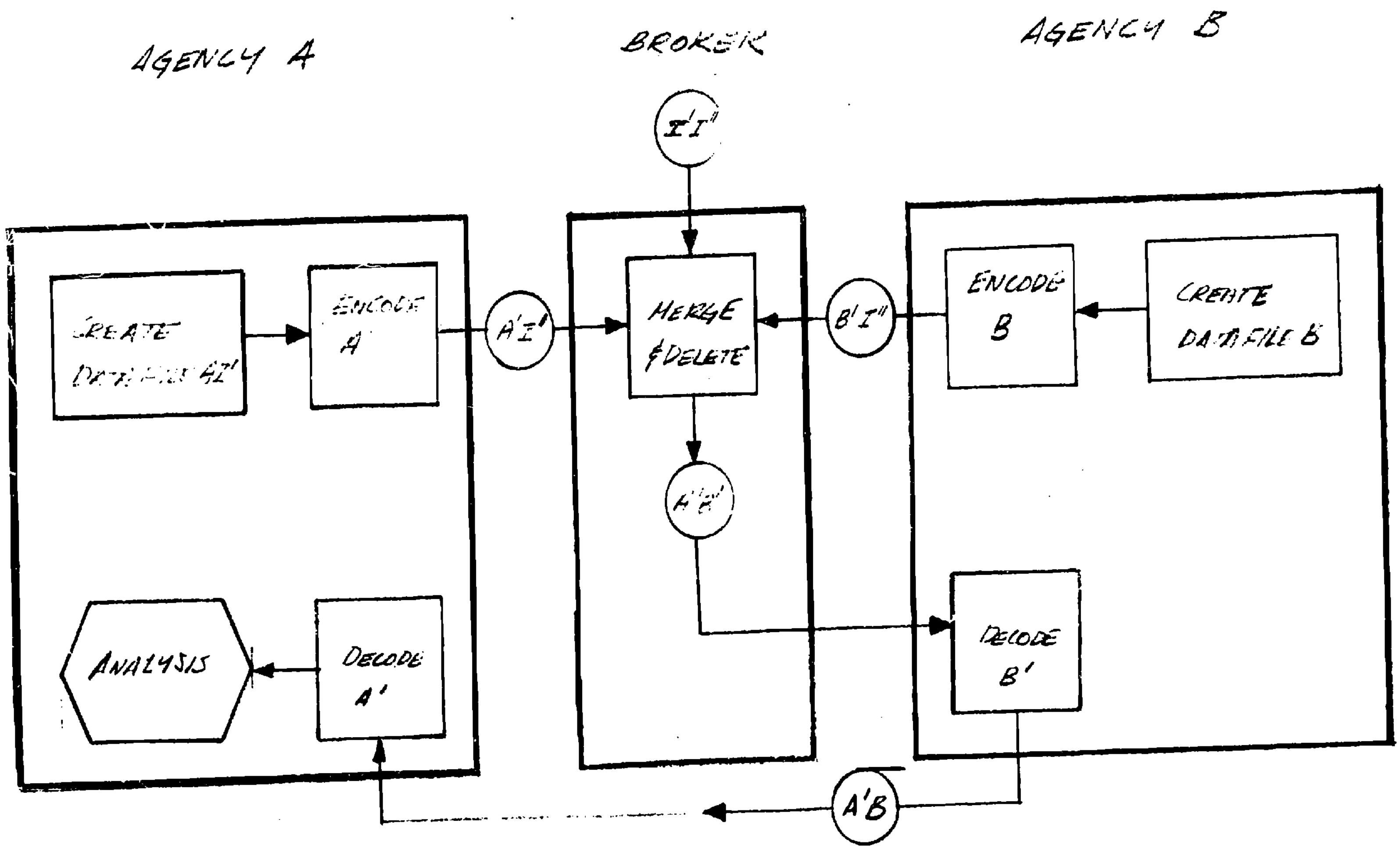
FIGURE 5

11/10/70



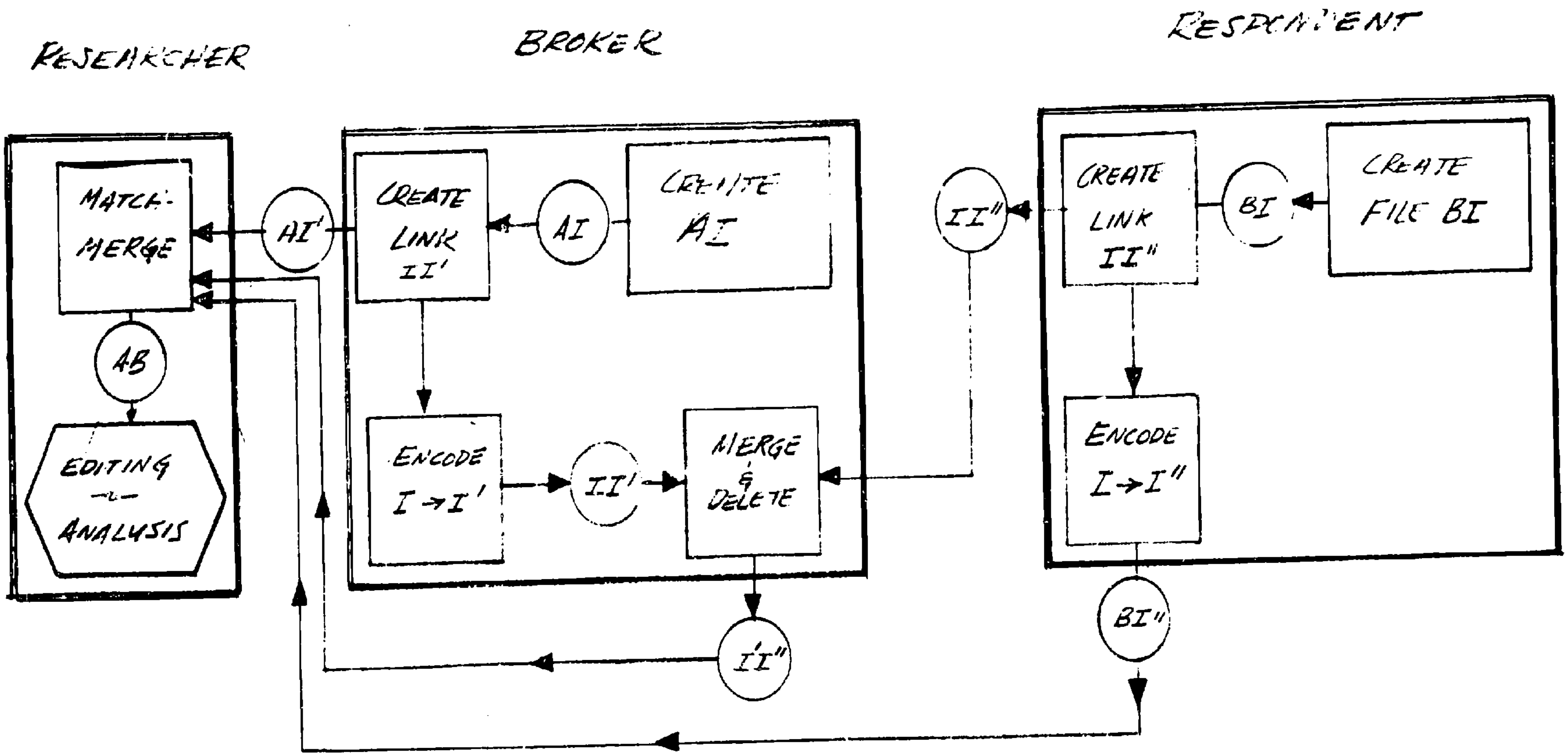
MODEL 3.1

FIGURE 6



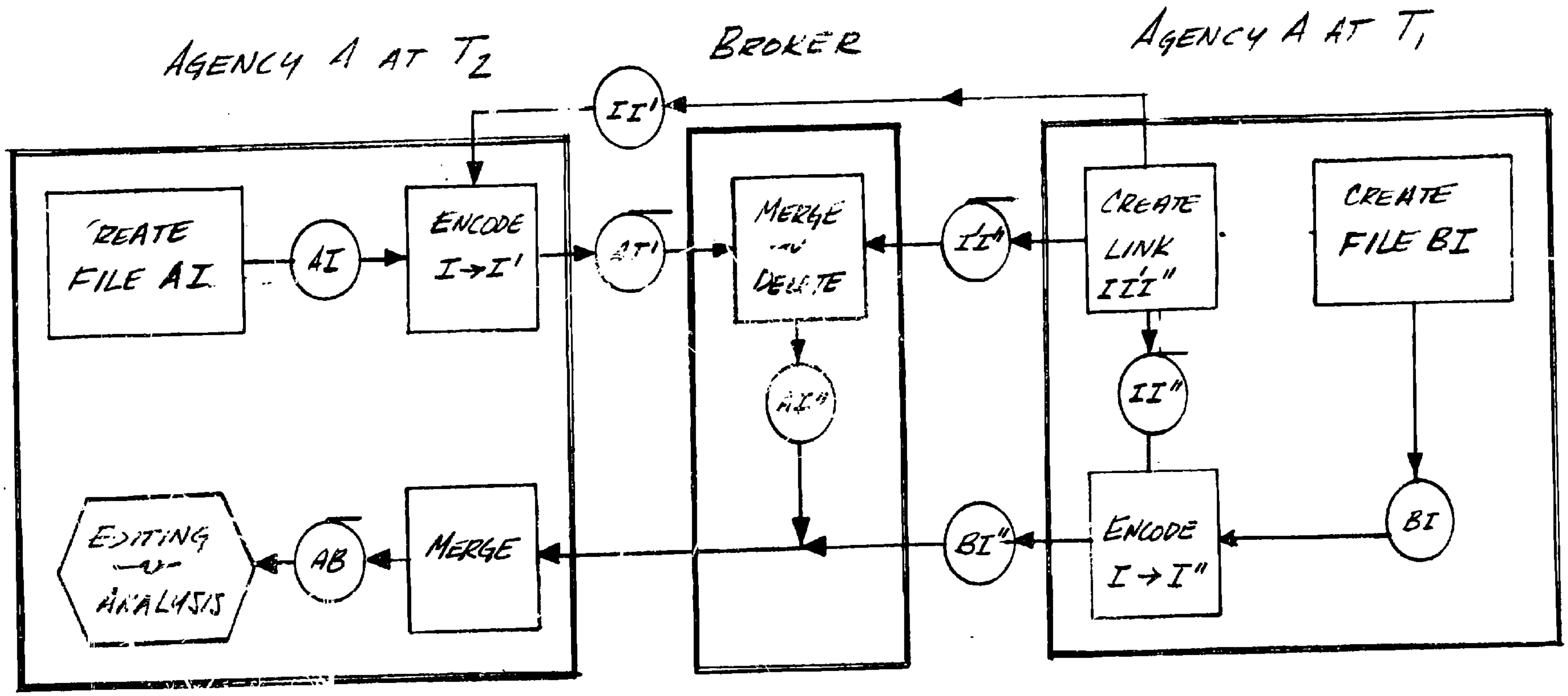
MODEL 4.0

FIGURE 7



MANNICHE & HAYES MODEL

FIGURE 8



ACE LINK FILE SYSTEM

FIGURE 9