

We have already seen how constructing legitimate disclosure by-passes forces one back to a manipulation of highly disaggregated components or even respondent units as building block elements. The collection and tabulation procedures of the agencies generate constraints on data use that lead in a similar direction. This arises out of some fundamental problems in dealing with the coding and classification of original source data.

As has been noted, a common form of information usage in analysis requires the matching of an attribute for two or more statistical sources or the association of two different attributes. This may require matching between different historical sets or between the files of different programs or agencies. This association of records is rendered difficult or impossible by at least three classes of record incompatibility that stem from current production practice. One difficulty in associating records from the different sources stems from the noncompatibility of classification of the data by several collection agencies and information systems. In the process of condensing and summarizing source records from initial respondents, information is tabulated on the basis of classification schemes that group items into classes. Often these classifications are inconsistent. (This is sometimes the result of the failure to develop general purpose standard classification codes applicable to all programs for these intermediate aggregates. It is often a result of the fact that standard codes are applied differently by different agencies so that there is no assurance that each agency (or program) will assign the same respondent to the same cell. Either of these cases often makes a comparison of the cells meaningless or difficult for purpose of analysis.

A further difficulty grows out of the fact that the basis for classification applied by the collection agency in defining the cells may be inconsistent with the analytical or descriptive requirements of the user.

When either type of problem occurs, one solution is to return to the initial respondent unit or some other disaggregate building block and reconstruct consistent boxes of data. This yields the same class of by-pass procedure identified with the disclosure problem. At this point one may encounter another common problem in the form of the noncompatibility of the definitions of the respondent unit. This is a class of noncompatibility that not only renders questionable the comparison of seemingly similar cells for different systems (as in the other two classes of incompatibility) but may render difficult or even impossible the reconstruction of compatible cells.

The anomalies that grow out of these compatibility problems can be tackled in two ways. One often hears it proposed that general purpose standards for the classification of intermediate aggregates be considerably extended and aggressively applied to all agencies. This may not be the most fruitful line of approach. Existing standards may possibly be improved and made more general purpose by a more intensive analysis of user requirements and a concern with the issue is not unimportant. However, an attempt to force all uses into a common standards mold for intermediate aggregates has attendant disadvantages from the point of view of the user as well as the producer and the agency vested with the responsibility of formulating standards. General purpose classifications for intermediate aggregates always require some compromises in taxonomy that reduce the utility of the data for special purposes. Furthermore, a great deal of the data generated by the Government comes from programs that have a special purpose mission and restrictive legislative authorities and requirements that go with it. Forcing on these agencies a rigid application of general purpose codes for intermediate aggregates may be impossible and even undesirable because they conflict with special purpose missions. It does not appear to be a helpful possibility that all data sets can be arrayed in compatible boxes that will anticipate all uses. The attempt to deal with standards in this context will place impossible strains and burdens upon the machinery for making and enforcing policy with respect to standards.

A more fundamental way to handle this problem may be a progressive move in the direction of compatible building blocks that can be reassembled to provide compatible and relevant aggregate sets for special uses and can be used as a bypass for disclosure problems and other procedural obstacles. This suggests that the problem of standards of greatest importance in the emergent situation is the need for uniform identification, definition and coding of the respondent unit as a basic building-block unit. The absence of a uniform system of coding and classification for geographical areas is also a serious deficiency and is an important part of this same problem. This also suggests the importance of procedures for assuring that every agency puts each respondent in the same cell