Some reflections on coding

Jarausch, Konrad H.

KONRAD H. JARAUSCH

Some Reflections on Coding

Coding seems to be one of the necessary evils of historical social research. While there has been much progress with textual data bases or with formalized data entry in non-numerical fashion [1], most quantitative historical research does transform sources into numerical expressions. The absence of universal standards and the decentralization of historical research have led to a widespread "coding chaos" and to the continual reinvention of the wheel by individual scholars. The waste of time and effort involved has produced periodic appeals for standardization of codes [2]. But in my own area of research (modern German social history) they have not yet had much success.

While some of the coding frustration must stem from the contrariness of human nature, lack of agreement on general codes has deeper reasons. One set of causes involves the divergent institutional foundations of historical research. While on the continent *Grossprojekte* tend do dominate the scene, in Anglo-American countries individual cottage industry seems to be the prevailing mode. The former tends to foster some degree of local uniformity while the later leads to greater idiosyncracy. More important for the coding problem is the basic ambivalence of the process itself, which will not simply yield to appeals for good will: On the one hand coding is generalizable and repetitive, dealing with uniform topics across locale and time (such as the distinction between male and female). On the other hand creating a codebook is heavily dependent upon the specification of the research hypothesis and therefore peculiar to each project (such as an investigation of the mobility vs. the working class). This inherent intellectual difficulty will continue to defeat calls for greater uniformity.

Perhaps it would be more practicable to distinguish three different aspects of the coding process, each amenable to a different degree of systematization.

1) First there are the technical rules (*Kunstregeln*) which can be found in any suitable textbook[3]. They concern such principles as *Eindeutigkeit*,

---

[1] See the papers by *G. Jaritz / A. Müller* in this volume on pp. 93ff and *E. Mergenthaler*: Text Base Managment Systeme - Werkzeuge zur Archivierung und Analyse sprachlicher Daten, in: Angewandte Informatik 6 (1983) 262-267.

[2] Cf. the papers by *H. Reinke* and *H. Schultz* in this volume on pp. 125ff and 179ff.

[3] *M. Thaller:* Numerische Datenverarbeitung für Historiker, Wien, 1982, pp.

Systematik, completeness, efficiency and *Rückverfolgbarkeit* on which there is bound to be widespread agreement.

2) Second, there are many areas in which it is preferable to employ codes developed by historical contemporaries – especially when case studies are later to be compared with aggregate figures available in published sources. The geographical/administrative classifications of the Prussian or German statistical office are a good example. Would it not be more efficient simply to use the numbers for the Prussian provinces and administrative districts (starting with 01 for East Prussia) than to make up a set of one's own, especially if one wanted to compare these findings with data from the relevant *Volkszählung*?

3) Third, there are other sizeable fields in which a uniform set of codes would be counterproductive. For instance, in social stratification research the purpose of the investigation determines the structure of the coding scheme. It would be useless to distinguish between different levels of nobility in a staudy of the Berlin working class of the 19th century. Conversely, it would be nonsense to draw fine lines between manual and non-manual labor in an investigation of the social origin of German students during the last two centuries – since less than 2% stemmed from these strata, whereas it might be interesting to differentiate the dominant *Bildungsbürgertum* into educated officials and free professionals [4].

Because such problems are inherent in the coding process, three different steps are necessary to reduce the chaos. First, there needs to be more explicit discussion of practical techniques. For instance, the basic principle that coded figures ought to be as diverse and close to the source as possible (whether it be numerical or nominal) is worshipped more often in the breach than in the observance. There is no need to mention individual names (and to embarrass colleagues) lest they turn around and accuse the accusers, but examples of thoughtless and premature aggregation could easily be multiplied. This is where the pilot study can be helpful and where non-numerical data entry is often essential, so that early and sometimes mistaken coding decisions can later be reversed. Second, it seems essential to develop, in those areas where there will be little ideological or methodological disagreement, not only com-

---

126 sqq. and *K.H. Jarausch, G. Arminger and M. Thaller:* Quantitative Methoden in der Geschichtswissenschaft: Eine Einführung in die Forschung, Datenverarbeitung und Statistik, Darmstadt, 1985. See also *Floud, Dollar-Jensen or Shorter* for earlier examples.

[4] *K.H. Jarausch:* Students, Society and Politics in Imperial Germany: The Rise of Academic Illiberalism, Princeton, 1982, 114ff.

mon codes but also common data sets on tape or diskette. Why does someone using urbanization or place of birth as variables have to look up the size and location of each German village, town or city in some volume of the *Deutsche Statistik*? It seems not only possible but imperative to transform that information into machine-readable form and make it accessible so that the computer can do the matching during data entry (when the name is typed in) - and not the tired researcher or the assistant! Third, coding more controversial and complex variables could be simplified greatly if scholars understood the beauty of multiple codes more widely. Instead of having to argue about the one and only stratification system, why not classify the same datum, namely occupation, according to different principles? For instance one consideration could be Max Weber's conception of power over others (which might divide professions into elite, intermediate and dependent groups). Another might be the economic functional classifications of the German statistical office such as agriculture, commerce, industry, servants, government or professions as well as no-profession. A final consideration might involve social strata - as perceived by contemporaries themselves. This could lead to a stratification scheme starting with the *Bildungs-* and *Besitzbürgertum*, including the *Alte* as well as *Neue Mittelstand* and the working class ... There is no need to argue the specifics of this case developed for research into 19th century German higher education [5]. It is principle that matters. The advantage of a multidimensional approach is not only that it reproduces the complexity of social space better than a one-dimensional scheme, allowing the researcher to try different conceptual dividing lines on the data empirically and letting him thereby determine where the significant distinctions in the data lie. It also makes it possible to use competing sets of classifications which make results more comparable between researchers (if someone else's scheme is also employed).

While the above recommendations themselves do not magically solve all coding problems, they suggest a direction for debate which ought to be more fruitful than endless argument about the various merits and flaws of a single scheme. Clearly, coding is a crucial step in quantitative historical research, since categories tend to predetermine results. As it begins the interpretation process, it will be left to uninformed student assistants only at considerable intellectual peril. If data-sets, whether they be institutional or individual, are not to be buried with each project but be used in secondary analysis,

[5] *W. Hubbard and K. H. Jarausch:* Occupation and Social Structure in Central Europe: Some Notes on Coding Professions, in: Historical Social Research, 11 (1979), 10-19.

they must be clearly documented as well as cleanly coded. In some cases exchanging the original documents might be necessary in order to introduce more complex analytical distinctions [6]. But in most instances careful coding techniques, generally agreed upon contemporary schemes and multidimensional classification ought to suffice for secondary analysis and a comparison of results. There is no need for coding over-kill by spending two thirds of the research effort to elaborate a "perfect" code-book. But if quantitative historians are to reap the full benefits of their computer usage, they need to begin cleaning up their coding act!

[6] See the secondary analysis of the Kater NSDAP sample for a subsample of professionals, semi-professionals and proto-professionals by this author forthcoming in the *German Studies Review*.