
Problems in the Use of Business Data on Tape

by E. Kay Worrell,
Director, Survey Research Center
The Conference Board, New York

Introduction

This will be a description of some types of business data available on magnetic tape, the uses made of these data by companies and business researchers, and some problems that may be encountered. Examples will be taken from experience with Conference Board applications and from discussions with business users of such data among Conference Board associate companies.

The Conference Board publishes research reports, economic forecasts and newsletters, and organizes conferences for the business community. It is supported by subscription income from associate companies, conference fees, and publication sales.

The Survey Research Center provides assistance to research staff in the collection and processing of survey data, including maintenance of research-related mailing lists. We also provide statistical support, training in microcomputer software, and assistance in database design and interface between the mainframe and microcomputers.

Types of business data

There are three broad types of business data available on magnetic tape.

- 1 Directory information, including names, titles and addresses of executives
- 2 Financial and other descriptive information about individual companies
- 3 Aggregate statistical information on the economy, the workforce and industrial production and services.

I will focus on the first two types: directory information on executives, and descriptive information about individual companies.

Two of the most commonly used sources of data on tape are Dun & Bradstreet Corporation and Standard & Poor's Corporation. Both companies were formerly best known for their printed directories, and tape products were derived from computerization of these data. Dun & Bradstreet also offers extensive business data processing and mailing services. Both companies are producers of proprietary on-line databases.

Standard & Poor's and Dun & Bradstreet rely to some extent on self-reporting by companies, in addition to publicly available financial reports and other published sources. The titles of individual executives are coded in some detail. Standard & Poor's codes up to four separate titles for each individual, using a two-part code consisting of 35 discrete designations of rank (e.g. asst. V.P.) and 25 functions (e.g. finance). A maximum of 20 unique standard industrial classification (SIC) codes describing products or services produced can be provided, though some tape products may not contain all.

An administrative database containing information on all publicly-held companies is now available in machine-readable form from Disclosure Inc. All companies whose stock is publicly traded in the United States must file annual and quarterly reports with the Securities and Exchange Commission (SEC).

The annual forms — the 10-K for U.S. based companies, 20-K for foreign companies — and the 8-K quarterly forms have been available on microfilm from Disclosure Inc. for more than ten years. Data are extracted from these, as well as from annual reports and proceedings of annual meetings, and are now available on-line from Disclosure, as well as on magnetic tape and microdiskette.

Information on the chief executive officer and directors of a corporation is also included. The cash compensation awarded to the top five officers is provided, along with their names and titles. Little other information on individuals is available, however.

A recent entry into the tape product market is the National Register Publishing Company (NRPC), publisher of the Directory of Corporate Affiliations, the Directory of Advertisers, and the Corporate Bluebook of Financial Executives. NRPC is an especially good source of detailed information on subsidiary units of corporations.

Tape files from NRPC are now available as parallel products to each of the printed directories currently produced. Each directory product is tailored to one specific business market. Like Standard & Poor's and Dun & Bradstreet, public data are supplemented by data solicited directly from companies, and all data are submitted annually to these companies for verification.

Uses of Business Data on Tape

Three types of information may be of interest to the purchaser of business databases on tape, and the type that is of primary importance may determine the choice of vendor. The types of information are:

1. ■ classificatory information, such as SIC code, or even sales or assets of corporations,
2. ■ names, titles and addresses of individuals, most often required coded by title — both level and function,
3. ■ detailed financial information on corporations.

Companies purchase data on tape for a variety of reasons, but primarily for purposes related to marketing. In addition, they may use these data for analysis of financial and other information for planning and comparisons relating to investment, mergers, acquisitions and divestitures.

Marketing activities involving use of such data fall into three categories:.

- 1 production of labels for direct mail,
- 2 update and improvement of in-house lists,
- 3 analysis of information for marketing comparisons.

There is some overlap among these activities. Information may be used for mailing purposes directly from the purchased tape, or may be run against in-house computer files — a "merge/purge" run — to produce a non-redundant mailing. This merge/purge run may be done by the company's in-house EDP department or may be subcontracted to a list broker or service bureau. Dun & Bradstreet

offers this service to purchasers.

Data from purchased tape may be merged with other machine-readable data, such as the corporation's own list of clients, for update purposes or for the addition of analytical variables which the user company may not collect and maintain, such as sales, assets, or SIC classification.

An example of the latter was a recent request from one of our associate companies seeking advice on matching SIC codes from a purchased tape file with their in-house client list. They wanted to be able to analyze their client list stratified by industry groups. Two problems were encountered.

An alphabetic match on company name resulted in only a 60% match with their own list. Use of a unique identification number might have facilitated a more complete match.

Also, the company was specifically interested in SIC codes in the retail trade series. The dataset they had purchased included 6 SIC codes per corporation. Only the 6 SIC codes that described their primary activities, reflected in the proportion of the annual revenue generated by those activities, were listed. There was a distinct possibility that the codes in which they were interested were not always included for companies involved in retail trade activities.

Our own most recent use of a purchased tape was in conjunction with a survey on corporate benefits. The tape was acquired from Standard & Poor's, and the SAS System was used to produce mail labels (in a time-sharing system), selecting subjects by title within certain industry group and size parameters. We wanted to mail a questionnaire to senior human resource officers. Although a great variety of functions were coded, that particular release of tape product had a disappointingly low number of high-ranking officers coded as this function. Only about 65% of the largest 1000 corporations

had an identifiable senior human resources officer coded.

A final example is an investigation we concluded recently on the feasibility of using another external tape product. The Conference Board has recently begun to identify a second level of major corporate leadership, the chief executive officers of large subsidiary companies. In our exploratory work we used the printed version of the Directory of Corporate Affiliations. A complimentary product is now available on tape. This is the most comprehensive source we have found of information on levels of corporate ownership.

The vendor relies heavily on descriptions of corporate levels provided by individual companies, which are polled annually. Because of the variety of reporting procedures used by the companies, it is difficult to extract reliable lists of second-level executives which are consistent from company to company. We have resorted to checking the annual reports and organizational charts that the Conference Board collects annually from cooperating companies. Many of the companies listed in the National Register directory appear to be "paper" entities rather than corporate profit centers. Several may be headed by the same individual.

Preliminary work suggests that these tapes will not be useful for our purposes.

Summary of Problems and Solutions

Types of problems one encounters in using commercially available business databases on tape are:

- a. the quality of the data
- b. the appropriateness of the content of the

data for the intended purposes

- c. logistic problems related to implementation of the intended use, e.g. matching external data with internally produced data
- d. format and documentation of the tape.

The industrial classification codes assigned vary from tape source to tape source, as do printed sources. Often, these codes are assigned by clerical staff on the basis of reported descriptions of business activities provided by the responding company; these may be incomplete or erroneous. They may also be inconsistently or incorrectly interpreted by the coders. The composition of the corporation may change through merger, acquisition, divestiture, or change in the revenues generated in specific sub-units, so that the original "primary SIC-code" no longer applies.

To facilitate merging information from an externally purchased list with one's own in-house company list, it may be advisable to add a unique identification number to the companies in one's own list. The CUSIP number, used by the SEC, the D-U-N-S number assigned by Dun & Bradstreet, and the FORTUNE number used in the Fortune Data Bank, have all been incorporated into at least one other database besides that of the generator of the number. One or more ticker symbols (stock market codes) may also be included, but these are not consistent from exchange to exchange. Standard & Poor's includes the CUSIP number; Disclosure includes both the Fortune and D-U-N-S numbers, as well as the CUSIP.

There are also list service houses which will merge the data and do some specified checking, based on text-matching of company names. This is far less successful than matching on a unique identification number, which allows the end user substantial control over

quality-checking.

Computer-produced, customized print lists and labels are also available from these same suppliers, and are frequently the most cost-efficient way to purchase the information. The list brokers use a variety of trade journal subscription surveys and other resources.

In order to update certain specialized lists, we have purchased sets of labels or print-outs from specialized vendors. Periodically, in order to update our research mailing list of banks, we have purchased a printed listing from R. L. Polk & Co., whose primary research focus is on banks, and compared it to our own in-house list. We find the printed copy easier to work with, as we have limited database management capability in-house. Our own database, maintained on a Burroughs system, can not readily accommodate external files.

We request from Polk's print-out of the 3,000 largest banks, in alphabetic order by state and city, showing the total deposit income of each. We then select for our list those banks with total deposit income of e.g. over \$500 million (or \$100 million or \$1 billion). These we check against our own in-house list, especially noting possible bank mergers or name-changes. These are easier to identify in listings ordered geographically.

The most difficult problems facing those who would merge external information with internal information are the problems of subsidiary units, and, more recently, the even more complicated problem of joint ventures.□