INTRODUCTION TO THE DATABASE OF IRISH HISTORICAL STATISTICS 1911-1971

ABOUT THE PROJECT

The project to construct a Database of Irish Historical Statistics 1911-1971 was funded by the Economic and Social Research Council. The aim of the project was to extend and expand a database of Irish historical statistics covering the years 1821-1911 presently under completion in the Department of Economic and Social History at Queen's University of Belfast. The twentieth-century extension of the nineteenth-century database therefore follows closely the form and content of the latter. The combined database 1821-1911 includes a wide range of social statistics from Irish censuses of population, agricultural statistics, vital statistics, crime statistics, censuses of industrial production, and trade statistics.

COMPONENTS OF THE DATABASE

In addition to this introduction (INTRO.DOC), the database comprises the following components:

(1) A CODES diskette containing the Geographical Unit Code Book, which gives the coding scheme for the spatial units in the database. This is available in both ascii format suitable for importation into the Ingres DBMS (CODEGEOG.DOC) and in WordPerfect 6.0 for DOS format ready for printing (CODECD.DOC). This disk also contains files for other codes used in tables of the database (CODEBPN.DOC, CODEOCC.DOC, CODETR1.DOC, CODETR2.DOC).

(2) A DOCUMENTATION diskette containing five documentation volumes, which describe the series of tables in the database, the sources used in each table, and the notes and alterations associated with those sources. These are available as document files in WordPerfect 6.0 for DOS format (REPORTCP.DOC, REPORTVT.DOC, REPORTIN.DOC, AND REPORTTR.DOC). This disk also contains the file for this introduction (INTRO.DOC), and a file containing an inventory of the tab-delimited ascii files for each table in the database (INVENT.DOC).

(3) Several diskettes containing the source-specific tab-delimited ascii data files referred to in the documentation volumes.

(4) Several diskettes containing graphic image files of prefaces and introductions in the sources which give important background information and in some cases analysis of the data contained in the database. These are available in *.gif format.

The twentieth-century database contains over 100 tables of data. These are listed below. A general description of the documentation for each table in the database is provided here. A more detailed description of the contents of each table will be found in the five documentation volumes. The methods used for capturing, processing, and verifying the accuracy of the electronic data are then described, and an account of the coding schemes used is also given.

THE DATABASE AT A GLANCE

The data is divided into the following five categories: Census of Population data, Registrar

General data, Agricultural Statistics data, Census of Industrial Production data, and Trade statistics. The database is composed of tables holding data from each of the five sources. These tables are briefly described as follows:

I. CENSUS MATERIAL (county and county district spatial units at decade intervals)

POPULATION_TOTALS: Population by sex of each county district.

HOUSING_TOTALS: Housing by status and value of buildings in each county district.

RELIGION, RELIGION_AGE: Religion by age and sex of the population of each county district.

BIRTHPLACE_CD, BIRTHPLACE_COU, BIRTHPLACE_NI: The county of birth and county of residence of the population.

IRISH_LANGUAGE: Speakers of the Irish language in each county district

AGE: The age structure of the population of each county district.

CONJUGAL: The conjugal status of the population of each county district.

FAMILYAGE1_IR, FAMILYAGE2_IR, FAMILYCOU_IR, FAMILYREL1_IR, FAMILYRELR_IR: The classification of families by duration of marriage, age and religion of spouses, and children born in each county.

DEPENDENCY_IR: The number of dependent children in families in each county.

OCCUPATIONCOU_IR, OCCUPATIONCOU_NI: Occupational classification of the population of each county.

INDUSTRYOCCCOU_IR, INDUSTRYOCCCOU_NI: Industrial classification of the occupations of the population of each county.

II. REGISTRAR GENERAL MATERIALS (county data at yearly intervals)

VITALS: The births, deaths, and marriages by sex occurring in each county.

DEATH_AGE: The age of death by sex in each county.

DEATH_CAUSE: The cause of death by sex in each county.

INFANTMORT_1, INFANTMORT_2: Infant mortality in each county.

MARRIAGE_AGE: The age of marriage of brides and grooms in each county.

III. AGRICULTURAL STATISTICS (county and county district data for discontinuous sets of yearly intervals)

CROP_CD, CROP_COU, CROP_COU, CROPSIZE_1, CROPSIZE_2: The various types of crops on various farm sizes in each county district or county.

STOCK_CD1, STOCK_CD2, STOCK_CD3, STOCK_COU1, STOCK_COU2, STOCK_COU3, STOCK_COU4, STOCK_COU5, STOCK_SIZE1, STOCK_SIZE2, STOCK_SIZE3, STOCK_SIZE4: The various types of stock on various farm sizes in each county district or county.

FARMLABOUR_1, FARMLABOUR_2A, FARMLABOUR_2B, FARMLABOUR_3, FARMLABOUR_4A, FARMLABOUR_4B, FARMLABOUR_4C: The various types of farm labour (family, non-family, permanent, temporary) by age and sex in each Eire county and Northern Ireland.

FARMMACHINES_1, FARMMACHINES_2A, FARMMACHINES_2B, FARMMACHINES_2C, FARMMACHINES_2D, FARMMACHINES_2E, FARMMACHINES_3A, FARMMACHINES_3B, FARMMACHINES_3C, FARMMACHINES_3D: The numbers of various types of farm machinery employed on farms in each EIRE county and Northern Ireland.

IV. CENSUS INDUSTRIAL PRODUCTION MATERIAL (national data at yearly intervals)

OUTPUT_IR, OUTPUT1_NI, OUTPUT2_NI, OUTPUT3_NI: General tables showing gross and net output, costs of inputs including wages and salaries by industry in Eire and Northern Ireland.

CAPITAL1_IR, CAPITAL2_IR, CAPITAL3_IR, CAPITAL1_NI, CAPITAL2_NI: Working fixed capital in each industry in Eire at year end. Fixed capital increases and decreases in each industry.

FIRMSIZE_IR, FIRMSIZE_NI: Size of firms in each industry.

FIRMSCOU1_NI, FIRMSCOU2_NI, FIRMSCOU_IR: Number of firms in each county in each industry.

WAGES1_IR, WAGES2_IR, WAGES3_IR, WAGES4_IR, WAGES5_IR, WAGES6_IR: Persons employed in each industry in Eire distinguished by wages paid. Average earnings hours worked per week in each industry in Eire.

V. TRADE STATISTICS (national data at yearly intervals)

TRADECOM_1, TRADECOM_2, TRADECOM_3, TRADECOM_4, TRADECOM_5, TRADECOM_6, TRADECOM_7, TRADECOM_8: The value of imports, exports, and

reexports of various of each commodity to and from Eire and Northern Ireland.

TRADECOU_1, TRADECOU_2, TRADECOU_3, TRADECOU_4, TRADECOU_5, TRADECOU_6, TRADECOU_7: The total value of imports, exports, and reexports between Eire and Northern Ireland and various countries.

DESIGN OF DATABASE

The remit of the project was to construct a relational database housed in a Database Management System (DBMS) called Ingres. The database will be held in this form at Queen's University of Belfast and at the University of Essex Data Archive. The Ingres software holds the data in a series of tables that are related to each other by a common field of geographical units. The DBMS allows data to be retrieved either through a "Query by Forms" command menu or via SQL query language. While the query language allows for sophisticated manipulation and retrieval of data, user's interested in simple retrieval of data for use in a spreadsheet or other analytical software should have little difficulty mastering the necessary vocabulary.

However, it is not necessary to be familiar with the DBMS system to make use of the twentieth-century database. The tables in the database have been constructed from bibliographically tagged tab-delimited ascii files. Users of the data may consult the documentation and acquire only those data from the bibliographic source tables that interest them. The database exists in two forms: as a series of Ingres tables and as a collection of source-specific ASCII files. Both are clearly presented in the documentation volume.

The tables in the database have been structured according to three constraints:

(1) The database must be geographically relational, so that all tables must contain a column with spatial unit codes. In some cases this required significant manipulation of the original source.

(2) The database must dovetail in content and form with the nineteenth-century database of Irish historical statistics currently being completed in the Department of Economic and Social History at Queen's University of Belfast.

(3) The database must be consistent in structure through time for given types of data, so that meaningful temporal comparisons are possible for the user.

(4) The database must be as comprehensive as possible, anticipating the needs of as wide a research community as possible given time and resource constraints.

The requirement to create a relational database housed in the Ingres DBMS has both advantages and disadvantages. The relational database software and SQL query language are powerful tools for the selection of data from particular geographic units, particular years, a specific range of values, etc. However, the database is not structured for immediate statistical analysis in spreadsheet or other statistical software, and some manipulation of the data will be required of the statistically-minded user once the data is downloaded from Ingres.

Tension between the latter two constraints is also evident in the contents of some tables. While in a well-designed database, each table contains information that is related to a single item of interest, the source material itself is often more densely organized. In addition, variations in source material over time means that some database tables are not as comprehensive as the original source in certain years.

METHODOLOGY OF DATA CAPTURE, VERIFICATION, TRANSFORMATION

The transformation of printed source material into electronic form proceeds in the following steps:

(1) The tables selected are scanned. The project's heavy investment in scanning hardware and optical character recognition software introduced a bias to the selection of data. Data that was easily scanned and transformed into ascii format was given a higher priority than data that is not systematically presented in printed sources or for which the physical reproduction of the original source was not clear enough for the software. Some small groups of data that were not amenable to scanning were manually typed.

(2) The image files created by the scanner are then processed with optical character recognition software. Once the software has adapted to the particular font of the original source, this process proceeds more quickly. The process is slower if the original source is not clear and in good condition. The end product of this process is an ascii text file composed of the numbers and text scanned in the original source.

(3) The numbers in these files are then verified to ensure against errors in the scanning and recognition process. The ascii files are loaded into spreadsheet software and column and/or row totals in the files are compared with those given in the original source. The error rate is quite low. Occasionally this process uncovers errors implicit in the original source. In twentieth-century printed material computational errors are extremely rare.

(4) The text in these files are then edited and checked for accuracy or replaced by pre-established codes. All geographically oriented text are replaced by code. In addition, codes for the various industrial groups in the Census of Industrial Production data and the occupational groups in Census of Population data are inserted into the files.

(5) After similar data from various sources (e.g., population tables from all the census years) have been processed and coded, their structures are homogenized so that they reside in one or a small number of database tables. This may involve the transposition of columns and rows, or the summation of rows where the description of variables is more specific in some sources (for example, if age or wage bands become wider, or the different types of cattle are more generally specified). In some cases, slightly differently defined data are placed in the same column and the difference is specified in the notes to the table.

(6) The homogenized ascii files are then given a final check for accuracy and inclusiveness, stripped of word processing code and prepared for copying into the Ingres database table. Once the data is held in database format, a number of simple verification procedures using

SQL query language (counts, lists of distinct variables, sums, etc.) are employed as a final check.

SPATIAL UNITS IN THE DATABASE

Where possible data has been collected for spatial units smaller than the 32 counties. The data in the nineteenth-century database applies either to Counties, Baronies or to Poor Law Unions. In the twentieth-century, the data applies either to the County or the County District, both of which are historically and geographically related to the spatial units of the nineteenth century database. In addition, the twentieth-century database includes data which has been recorded or published only at the national level. These include the trade statistics of Northern Ireland and Eire, and the respective Censuses of Industrial Production. While these data cannot be spatially related to data in other tables in the database, they will hopefully be valuable for many academic users.

The dominant spatial unit in the database is the County District (CD). The county districts were small enough to give detailed geographical coverage for a significant range of data. The database also includes a wide range of data which is available only at the county or national level. The CD is an amalgamation of District Electoral Divisions (DEDs). CDs also have the advantage of being directly related to the boundaries of Poor Law Unions and Towns, the dominant spatial units in the nineteenth-century database. CDs are located within counties but not within the seven County Boroughs in Ireland.

County Districts were established under the Local Government Act (I) Act of 1898. The spatial units to which they refer were originally Sanitary Districts established by the Public Health (I) Act of 1878. The Public Health act divided Ireland into Urban Sanitary Districts and Rural Sanitary Districts. Urban sanitary districts were created for the City of Dublin, all Corporate Towns, and all towns with a population of 6,000 or more. The area of every Poor Law Union, excepting those portions included in urban sanitary districts, formed rural sanitary districts. Under the Local Government Act sanitary districts were redesignated county districts, with the additional requirement that CDs to be situated entirely in one County.

CDs were therefore created out of DEDs all of which are in the same county and the same PLU. In all cases where Poor Law Unions span counties or county boroughs, the county boundaries within the Poor Law Unions were used to create the CDs. CDs are divided into rural and urban districts. While the former were created from the DEDs, the latter were created from the wards of the towns, the urban equivalent of the DED. Urbanization and rural depopulation have resulted in the creation of many new urban CDs and the amalgamation of adjacent rural CDs in the twentieth century. By virtue of the municipal corporations act of 1840, a number of urban CDs in Northern Ireland have been redesignated Municipal Boroughs.

CODING OF SPATIAL UNITS

In a relational database, it is necessary to make the selection of variables through a query language such as SQL as simple and efficient as possible. To this end the counties, county boroughs, and county districts have been coded. The coding scheme follows closely that

devised for the Poor Law Unions in the nineteenth-century database, so that users interested in long-term change occurring in particular areas may easily bridge the century. The database includes the table CODEGEOG which identifies the codes for the county districts, counties, and provinces, respectively. The ASCII file for this table is CODEGEOG.DOC. This information is also available in WordPerfect 6.0 for DOS format in the file CODEBOOK.DOC. The table also identifies boundary changes, the creation of new county districts, and the amalgamation of county districts. The table ACRES gives account of the change in size of county districts resulting from these boundary and amalgamating alterations. The ASCII file for this table is ACRES.DOC. The structure of these two tables is given below.

STRUCTURE OF TABLE CODEGEOG:

COLUMNS	CONTENT
CODE	codes of places
PLACE	provinces, counties, county boroughs, and county districts
TYPE	P = province, C = county, R = rural cd, U = urban cd, CB = county borough
CHANGE	O = status in 1911 (for those cd's for which 1911 values are changes from 1901, a further CHANGE entry is included in the table) A = amalgamated E = enlarged to contain an amalgamated district B = other change of boundary N = newly created R = renamed T = change of TYPE
YEAR	census year by which CHANGE applies
NOTES	textual description of CHANGE

STRUCTURE OF TABLE ACRES_CD

COLUMN	CONTENTS
CD	Code of each county district
ACRES1901	Acreage of each county district in 1901
ACRES1911	Acreage of each county district in 1911
ACRES1926	Acreage of each county district in 1926
ACRES1936/7	Acreage of each county district in 1936 or 1937
ACRES1946	Acreage of each county district in 1946
ACRES1951	Acreage of each county district in 1951
ACRES1956	Acreage of each county district in 1956
ACRES1961	Acreage of each county district in 1961
ACRES1966	Acreage of each county district in 1966
ACRES1971	Acreage of each county district in 1971

Note: the figures for Northern Ireland in 1971 are given in hectares.

OTHER CODE TABLES

Aside from the geographical units, a number of other variables have been coded to allow for more efficient retrieval of information. These are:

CODE_BPN: coding scheme used for international birthplaces in the table BIRTHPLACE_NI

CODEOCC: Coding scheme for general occupational categories used in the tables

OCCUPATIONCOU_IR, OCCUPATIONCOU_NI, INDUSTRYOCCCOU_IR, and INDUSTRYOCCCOU_NI

CODETR1 and CODETR2: Coding scheme used for the trades and industries listed in the tables of the Censuses of Industrial Production.

Further detail on these code tables is given in the respective documentation volume.

DOCUMENTATION

The five volume of documentation are available as the following WordPerfect 6.0 for DOS files:

(1) REPORTCP.DOC: Census of population tables
(2) REPORTVT.DOC: Registrar General's reports of vital statistics tables
(3) REPORTAG.DOC: Agricultural statistics tables
(4) REPORTIN.DOC: Census of industrial production tables
(5) REPORTTR.DOC: Trade statistics tables

Each table in the Ingres DBMS is described in one of the above files. The following information is given:

(1) A description of the columns of data giving their titles and a brief description of the data residing in those columns. The column titles are those used in the Ingres installations of the database in the Department of Economic and Social History and at the ESRC Data Archive at the University of Essex. The titles are highly abbreviated and a short description of each heading is given.

(2) A bibliography of the source tables included in the Ingres table and the name of the ASCII text file which holds the data from each source table. These ASCII files allow users interested in particular source material direct access to that material without the use of the relational database and its query language. The ASCII files are also the building blocks of the relational database.

(3) A description or quotation of footnotes, headnotes, or other material in the original source pertaining to the Ingres table.

(4) An account of the alterations to the original source table aligning them with the structure devised for the Ingres table. The general rule for the construction of the database has been to restrict data selection to certain geographical units (see below) and to exclude subtotals, totals, and percentages which can easily be recreated within spreadsheet software. Columns or rows in the source table which have not been included in the Ingres table are noted. In addition, the data in many source tables have been manipulated to conform with similar tables in earlier or later years so that similar data could be housed in one table to allow for comparison and analysis through time. These manipulations include the summation of rows or columns in the original source, or the transposition or other rearrangement of the source table data.

GRAPHICS IMAGE FILES

Each of the documentation volume also includes a list of graphics image files of the tables of contents, prefaces, and other notes relevant to the source tables in the original printed material. With both the individual documentation files and the graphic images, all relevant background material given in the original sources are available to the user. The graphic image files are scanned pages of the original source which cannot be accessed through word processing software. It is necessary to view these files with graphics image software such as LView Pro.