Public Data In Use: A Case Study of Ireland

by John Blackwell[1]
Resource and Environmental Policy centre,
University College, Dublin

Introduction

This paper is concerned with the policy issues which arise in enabling the public data which are most useful for research purposes to be disseminated in a timely fashion. Case study material from Ireland is drawn on, outlining problems which have arisen in identifying and in meeting the needs of users, and the extent to which it is possible to meet the requirements at a time when there is no prospect of an increase in real Government expenditure on the provision of statistics. Particular examples in the areas of earnings, employment and social welfare statistics will be given. The extent to which administrative records can be a fruitful source of data, and can substitute for purpose–built surveys and censuses, especially in view of the cost constraints, is assessed. At the end of the paper, some policy issues are posed.

[1]Presented at IASSIST/IFDO International Conference May 1985, Amsterdam

The background – data provision in Ireland

In Ireland there is a centralized Central Statistics Office (CSO), whose output ranges from population and vital statistics to labour force, employment and unemployment statistics, industrial production, prices, earnings and hours worked in industry, trade, transport and distribution. Many of these statistics are produced by means of censuses and surveys. The methods of dissemination used by the CSO vary from published volumes to regular mimeographed series (issued usually monthly or quarterly); data on Ireland which appear only in publications of international organizations such as Eurostat and OECD; small area data such as from the Census of Population and from the Census of Agriculture which are disseminated directly to the users; and special analyses which the CSO does in answer to specific requests e.g., from the labour Force Survey and the Household Budget Survey, which are usually not charged for.

However, in recent years there has been an increase in the amount of data available from public sector bodies outside the CSO, which in the main complement the existing range of data from the CSO. The monetary and public expenditure data, produced by the Central Bank and the Department of Finance respectively, can be put with the national accounts as the basic raw material for macroeconomic analysis. Much of the information on education and manpower flows and science and technology, on activities such as transport, energy, and tourism, on

education, health, housing: and social services, and on taxation are produced by a variety of bodies in the public sector outside the CSO. In a number of cases, these data are produced by government departments (e.g. Education and Social Welfare); in other cases the data comes from government agencies or state bodies (e.g., in the case of science and technology statistics).

Problems of resource allocation

There has been no increase in the volume of resources provided to CSO in recent years, rather, a decline. While there is no coherent account of the amounts of resources which are put into provision of statistics outside CSO, these are unlikely to have increased. The budgetary problems in Ireland rule out any prospects of an increase in resources for statistics. Indeed, one of the elements which worked to the advantage of data users in Ireland in recent years was Ireland's entry into the European Economic Community. This meant that EEC regulations and directives on statistics became binding. This provided, for example, a much needed set of statistics on the distribution of earnings through the 1979 Structure of Earnings Survey. As a reflection of EEC budgetary pressures, a proposed 1986 survey will not now take place, and there is no domestic survey which can fill the gap.

Especially at a time of pressure on the government budget some means of ranking the statistical output by reference to the intensity of demand, is required. This is very difficult to achieve, for the following reasons:

- There is a heterogeneous body of users, ranging from government to research institutes, individual researchers and private sector firms. The implied value of attributes such as timeliness, frequency, accuracy of estimate, level of disaggregation, will vary among these groups of users. Many private firms, and those in the financial sector, put a premium on timeliness. Hence, as a result of timeliness problems in the case of the Irish Census of Industrial Productions (problems which are currently being righted, slowly), there has been minimal use of industrial statistics by business firms. There are other reasons for this neglect of industrial statistics on the part of the private sector. Firms will typically want highly disaggregated data in order to answer questions such as market share or productivity change for a particular product. In a country such as Ireland, where one or two firms may dominate some sub-sectors, there often there can be no possibility that CSO would be able to provide the data for reasons of confidentiality. Hence, the popularity among business users of the foreign trade statistics, which are quite disaggregated and are more timely than production data.

- Related to the preceding point, it is difficult to establish the nature of the "trade-offs" which exist between improvements along these different dimensions of statistics. For example, timeliness could be improved by sacrificing some accuracy of estimate. For some firms, such as those in the financial sector, timeliness is crucial. For them, it is often more important to have an early estimate which is subject to later revision, than to have a purer number which arrives after a long interval. By contrast, in the case of social policy, the time-scale

is different. Here, the analysis of large data sets, gathered at longer intervals, concerned more with details at household or family level, is often needed. With regard to frequency, there is a danger that in a perverse way, an over-concentration on frequency could slow up decision-making, especially if estimates are subsequently revised to a marked extent. In the case of national accounts data – where there is already a time-lag due in part to slowness in provision of local authorities' data – there has in Ireland been a good deal of data revision long after the event.

- The price mechanism is not used to ration demand so that those uses of highest value (or which use up least scarce resources) are effected. That is partly because information has many of the attributes of a "pure public good". This is, information which is made available to one person (through, for example published sources, in the widest sense) can be made available to the community as a whole, and one person's consumption of information does not "take away" from the consumption of another person. (Of course, this must be qualified: information can be withheld from certain consumers and can be bought and sold on the market.) Moreover, a number of the statistics are provided at close to zero prices, which gives rise to difficulties in assessing the intensity of preferences among actual and potential users.

In summary, therefore, a key issue is how many more, or how many fewer, resources should be devoted, at the margin, to producing particular data series. This is, in a sense, a problem of providing a surrogate for the market system in guiding resource allocation to the provision of public statistics. Difficulties abound in assessing the intensity of user preferences, partly because of the various preferences of the different users. Furthermore, it is difficult to get people to reveal their preferences, a reflection of the 'public good' aspect of statistics. Users will be tempted to overstate their preferences, in the knowledge that if information is provided by somebody else, they in turn can benefit. In other words, people can obtain information at no cost to themselves, even though, if they were confronted with the stark alternative of "pay up or do without the information" they would be prepared to pay something.

The "classic" perception of the way in which information is gathered and organized has often centred on the following sequence by which decisions are made: identification of a problem; collection of data to illuminate that problem; identification of feasible solutions; choice of the best solution from the alternatives. However, in practice, this logical sequence is not always followed. Information may be gathered only after a proposal has been decided on, either to monitor the implementation of policy, or to co-ordinate activities across Departments, or to serve as a back-up for particular cases which a public body may wish to make.

Why does the classic model not apply? In part it reflects the fluctuating pressures for decisions, often made under time pressure. There is inevitably a time lag before information can be gathered, by which time decisions may have already been made. In part it reflects the lack of a clear set of signals from the users of data to the producers. In Ireland, the CSO has found it difficult to get users to articulate their requirements an unambiguous way. While there has been a generalized feeling of dissatisfaction among users, there has been no formalized method whereby the reactions of users can be channelled to the producers. At mid-1984, for the first time, a one-day statistical users' seminar was held in order to help communications between users and producers. Even here, there was a disappointing turnout

from private sector users. CSO does consult with major users, but this consultation gives little weight to non-institutional users.

It is almost trite to observe that the demand for statistics is a derived demand. Yet it should be noted that these derived demands are governed by the ecomic structure and by the way in which policy analysis and research occurs. Thus, for example, most social science research in Ireland occurs in large research institutes. In the absence of a funding body for social science research, and given a weak tradition in team research in the social sciences, the countervailing power of the lone researcher, not to speak of ability to use large data sets, may be weak. Indeed, for some uses such as macro-economic models, the data supply/demand position is more or less one of one monopoly producer (CSO) and one or two large users such as the Economic and Social Research Institute.

Another example is the use, in other countries with more decentralized of government, of sub-national statistics to govern the spatial allocation of public resources: with the degree of centralization in Ireland, this use is virtually absent. A final example: the more rapid the pace of social change (such as women's participation in the labour force), the greater the likelihood that the full censuses will (again, given current methods) lag behind what would be needed to monitor change. This has implications for the potential use of sample surveys or else of simpler censuses.

There will tend to be a chronic excess of user demands over the supply of statistics (given existing resources and methods). The question arises whether it would be possible to decide on allocating resources to statistics by using formal criteria such as cost-benefit analysis. The answer is "no". While it is possible to estimate costs, the diffuse nature of the uses of statistics would make the estimation of benefits extremely difficult. Except in a few particular cases, the monetary value of statistics will be almost impossible to compute. Moreover, given the links between information, policy and "background" knowledge of the world, the potential returns from statistics may be greater than the actual returns if there is potential to improve the linkage between public policy-making and the provision of statistics. This can be illustrated by reference to the Irish Census of Population. The cancellation of the 1976 Census had some direct effects on other data provision – it led to increasingly inaccurate estimates of population and of the labour force and consequently of employment and of national output. At the same time, there seems to be almost no accounting for the individual decisions, including those in physical planning, which have been affected. But the measure of these effects would be different with a different type of public administration, one that was more responsive to changes in economic and social information. In this sense there are interaction effects which complicate any attempt to ascertain a pure return to statistics.

It is difficult, in fact, to point to particular policies in Ireland which were critically influenced by the provision or absence of certain data. There are a number of reasons for this. First, the provision of data *per se* is best seen as contributing to the understanding and explanation of economic change and social processes. Second, even if a starkly empirical view were (mistakenly) taken – that the matter begins and ends with gathering "the facts" – the dissemination of data would work its effects in a diffused way, with time lags. Third, there are usually time pressures on policy-makers – research findings and associated data must be speedily available, and this is not always possible. Fourth, it is difficult to predict data needs ahead of time – the focus of policy attention tends to change over time, while the provision of statistics lags behind. Finally, there are insufficient contacts between researchers and policy-makers (including lack of effective dissemination of findings) and there is certainly too little interaction between the three groups of

policy-makers, researchers and producers of data.

However, it is possible to get a feel for the value of statistics by asking in any particular case:

- what are these statistics for?
- who is using them?

In anwering these questions, due account should be taken of the nature of the benefits, many of which are indirect, rather than relating to specific uses by those who decide on public policy. And a number of benefits relate to users outside Government.

In particular, there is need to concentrate on decisions at the margin - what additional information should be provided, and what should be dispensed with? This would enable the identification of those data which are most crucial to policy decisions. It would also enable the recognition of those pieces of information which are redundant, or where a lowering in frequency or a degeneration in timeliness could be tolerated, especially if this meant that resources would be freed to devote to the provision of information for which there would be a high return. In Ireland it has, however, been extremely difficult to identify "redundant" data series. This may in part reflect a system which is producing the "minimum critical mass" of statistics. It may also reflect a "ratchet effect" whereby, once a series is put in place, any attempts to discard it are met with opposition.

Scope for Co-ordination and Rationalizaton

While the collection of statistics by a wide variety of public bodies outside the CSO has filled a number of gaps in knowledge, it has given rise to a number of problems.

- No explicit co-ordination of this output of statistics occurs, nor is there any indication of the amount of resources which go into this activity. There is no "statistics budget" for the public sector as a whole, with an associated set of outputs. CSO itself has no "leverage" over the gathering of statistics by other public sector bodies. The activities of CSO could be greatly influenced by the data gathering of other bodies, either because they complemented the work of the CSO, or because the CSO could use the data for purposes such as national accounts compilation, or because data could be used either as a means of verification, a sampling frame, or a register for coding purposes.
- There is no quality control with regard to non-CSO statistical output. This has in part reflected the way in which data collection began - sometimes as a by-product of filling internal planning needs and the need for co-ordination, sometimes as a result of implementing administrative schemes. Unpublished data are available from an employment survey conducted annual by the Industrial Development Authority, from a quarterly consumer survey undertaken by the Agricultural Institute for the EEC, and in an information system on public sector employment in the Department of the Public Service. The first and last cases exemplify data sets which have been compiled primarily for internal

management purposes.

- There are no principles of access to non-CSO data. While there has been some dissemination of the Industrial Development Authority and Department of the Public Service data, this has been as a by-product of secondary analysis by "outside" workers. The Industrial Development Authority data have been used as a source for information which, in the absence of CSO industrial employment data on a "component of change" (job gains, job losses) basis, cannot be obtained from other sources. They could, in part, be substituted by CSO data if the basic CSO data were analysed in a form which facilitated the "component of change" work. Government departments hold a great deal of information which is of interest to policy analysts, some of it in the form of raw data. A certain furtiveness in even disclosing the presence of data seems to be endemic. The ultimate barrier is when a Department or agency does not release a body of data. For example, twice since 1973, the results of housing conditions surveys, undertaken on behalf of the Department of the Environment have been kept within the Department. This is a notable case, because the information would complement the census data on housing and, in the areas of unfitness and obsolescence, provide information for policy which is more relevant than could ever be obtained from the census.

- While there are many *ad hoc* surveys conducted by Departments and other bodies, surveys conducted by the Economic and Social Research Institute for the Institute work, and surveys conducted by other research institutes and by individual researchers, none of this body of material is pulled together in one place. Nor is it even catalogued. There is no central user's guide to the body of data.

- There is danger that the current round of computerization which is occurring in public sector bodies will result in further fragmentation. There is, for example, a danger that computerization among local authorities will miss out on the possibility of providing the means of a data bank. To take another example: the Department of the Public Service is to expand its personnel information system. Ideally, in this process, the CSO should be consulted and compatible codes for occupations and other entities should be used.

This leads to the issue of whether there would be potential for rationalizing the data-gathering on Irish by State-sponsored bodies. A number of these bodies - national training agency, the Irish Export Board, the Industrial Development Authority, the Institute for Industrial Research and Standards, and the National Board for Science and Technology - compile data on Irish firms, as does the CSO itself. On the surface, rationalization could contribute to the problems of non-response and of slowness in response, which in part account for the time lags in the issuing of CSO industrial statistics. This may be the case up to a point, but it has to be qualified in view of the client relationship which exists between many of these firms and the State body in question, which is likely to facilitate response. A more compelling reason for seeking some rationalization is the richer potential for cross-tabulation which would ensue, as at the moment this is limited to the variables on each individual inquiry. Moreover, there would arise a common set of standards with regard to methods of compilation and of analysis, which would ideally be in accord with CSO categories as regards coding. There would, of course, remain a sub-set of data which would be confidential to individual agencies.

Some co-ordinating role is needed in relation to the output of statistics from sources other than the Central Statistics Office. There are a number of ways in which the desired co-ordination of non-CSO sources could occur. One would be through expanding the responsibilities of CSO to include the co-ordination of all public statistics. Another would be a co-ordinator of non-CSO. The co-ordinator could be in a Government department, or outside the departmental system, or could work through an inter-departmental committee.

The main objectives of the co-ordinating function would be as follows:

- To encourage public sector bodies to consider the statistical possibilities of administrative records and to consult on this with bodies such as the CSO. This is especially important at a time when new computer facilities are being installed.

- To ensure that statistics which are of importance to policy-makers but which are of little or no importance to the collecting agency, are produced. This may at times require that such a subordinate objective be given explicitly to the agency.

- To ensure quality control, i.e., some minimum standards and compatibility between individual data sets. The more standard classifications and definitions are used, the lesser the likely burden on respondents as their records would be in conformity with all requirements.

- To examine the potential for rationalization in the data-gathering activities of State bodies. This could lead to richer possibilities of data analysis, as it might be possible across the headings which are contained in the existing surveys. There are two ways by which this could occur: through an "umbrella" survey, or through the linking of records. The latter would be much more difficult to achieve. Some move towards an "umbrella" survey would be the most fruitful route.

- To consider the use of these statistics as a means of verfication of sampling frames or registers for coding purposes.

- To take part in a revised and explicit system of resource allocation for statistics, influencing the allocation of resources and pointing up the areas where attention needs to be focused.

- To fulfill some gatekeeping functions with regard to the increasing volume of surveys of Irish firms by those in the public sector.

Among the cases which are candidates for action are manpower and energy. Another instance is the road vehicle file of the Department of the Environment which is computerized but which is not being used to produce data on the vehicle fleet by age, make, type of fuel used. Perhaps the prime candidates are the Department of the Public Service data on public sector employment (mentioned above) and Revenue Commissioner's data derived as a by-product of tax collection. This leads to the next section, on the potential for using administrative records.

Using Administrative Records

In part because of the budgetary problems which face us, the use of administrative records as sources of statistics is an attractive option. Limited resources can be devoted to the

gathering of statistics by surveys; statistics based on administrative records can be collected at relatively low cost, in contrast with surveys. In addition, there has been increasing resistance on the part of respondents to provide information to the statistical office. (Indeed, this was the main reason the pre–1968 survey on the distribution of industrial earnings by the CSO was dropped). In addition, income data derived from administrative records have potential for being more accurate than income data derived from surveys, which are subject both to sampling error and to understatement of income. This potential is particularly evident in the case of the service sector where relatively little data are available on earnings and employment – even for the public sector, where the only readily available information is on civil service employment.

The potential for making greater use of administrative records as a data source, or as a basis for sampling, will increase as existing records become more computerized, thereby enhancing the potential for cross–tabulations. A number of departments and agencies are now apprised of the potential for decentralized collection and dissemination of data via micro–computers. The idealized picture, thus, is one where the contraints on resources which currently bind us are loosened through a combination of the use of administrative records allied with computer collection and analysis. There are a number of potential snags which could disturb this pleasant sequence of events.

First, the definitions and categories used in administrative records reflect the administrative process (e.g., entitlement of benefits, being in the tax net). They may not be ideal for statistical analysis, and many change over time. Ideally, statisticians would be involved at an early stage, either when new systems are being put in place or when manual records are computerized. One feels apprehensive that the current wave of computerization could lead to a rash of incompatible variables and codes.

Experience to date with some of non–CSO sources–with their lack of compatibility with CSO data and changes over time in categories which impede time series analysis – does not inspire confidence that this can be avoided without some policy initiatives. The Department of Social Welfare data, which have a good deal of potential for providing data on social welfare recipients, are tied to benefit recipiency. Hence, it is difficult to explore problems of lack of take–up of benefit. And in the case of unemployment data, in Ireland (as in other EEC countries) a substantial minority of the unemployed (from the labour Force Survey) receive neither unemployment benefit nor unemployment assistance.

Second, much of policy analysis requires linkages – either to CSO data or to departmental data. An example would be the relationship between social welfare payments and patterns of work. Such linkages are either quite difficult to effect or can raise problems of confidentiality.

Third, the only aspects of administrative records which are reliable are those which are regularly used by the authority in question and are, therefore, kept accurate and up to date. Considerable clerical resources may be required to maintain administrative records for statistical analysis purposes (i.e., keeping designatory details up to date, elimination of duplication, etc.) and authorities have other priorities in mind. If administrative records are to be fruitful, they need constant attention. Important administrative details required for analytic purposes (e.g., designatory details such as business descriptions, occupation, age) are not usually complete nor up–to–date because they are not frequently used by the authority. This is the biggest handicap to the use of these records, and arises in particular in the case of the Irish Revenue Commissioners data which have the potential to provide earnings and employment data and to fill notable gaps in coverage, especially on the service sector.

Fourth, there is need for close coordination between statisticians and administrators in order that the needs of both administration and provision of statistics are met. Decisions have to be made about classifying data in cases where there may well be conflicts between the desires of different users. At times when schemes are designed, changed or computerized, there is need to be sensitive to the possibilities of using these records to provide statistical tabulations. This would mean consulting with statisticians in the design stage. If consultation were to occur too late, a costly redesign of systems could be required. In the past in the case of the Irish health sector, where there is a certain centralizaton of the functions of Health Boards, computer systems which were incompatible across Health Boards were acquired. While there has since been a rethink of information requirements within the Department of Health, this has involved too little contact with those with a planning function in government.

Fifth, a major constraint to using the Revenue Commissioners' data relates to objectives. This is linked both to the issue of resources and the need to use clerical resources. Understandably, the Revenue Commissioners regard their main purpose as getting in money from the tax system and would regard information provision as, at best, a subsidiary objective, or at worst having no justification at all. In the case of a tax with high collection costs per yield, such as the residential property tax, the minimum amount of information would be provided. From the point of view of the public weal, and from the point of view of cost-minimization for the State as a whole, there is need to use the Revenue Commissioners data. But there is no incentive for the Revenue Commissioners themselves to take account of this. Rather the opposite: if people are switched from revenue-raising in order to provide more information, the Revenue Commissioners will be in danger of receiving criticism for failure to maximize tax revenue.

If the basic data were "clean", it would take little resources to produce the required data, by using the automatic data processing facilities which Revenue has. The biggest obstacle to producing tables which would help policy-making outside the specific remit of Revenue is the need to deploy clerical workers to "clean" data with regard to designatory codes. This would especially be the case as Revenue is prepared to consider the provision of a subset of basic data on magnetic tape to outside agencies - with the individual identifying characteristics removed. Users could then analyse these data using either a software package or a custom-built program. This would be a particularly valuable facility.

There will of course, always remain areas of inquiry where there is no substitute for survey or census data. An example arises in the case of labour force statistics.

It is manifest that existing data are often inadequate to answer the analytical and policy questions which arise. This is partly because most of the labour force data are given in terms of stocks. Many of the questions raised need an analysis of flows between states. Another reason for difficulty is that existing data come, in part, from administrative records which are tied to benefit criteria, which differ over time, and in part from self-description of states which are inherently ambiguous, such as "unable to work due to disability". An instance of this from the Irish Labour Force Survey is that in 1977 a category "unable to work due to permanent sickness or disability" was introduced, but led some people who described themselves as "unemployed" in 1975 to opt for this new category. There is an interplay between disability, unemployment and disability, and between retirement and disability. If one is to capture the impact of social security payments on the labour market, there is need to complement the existing stock data with flow data.

Ideally, flow data (shown in Table 1) by sex would be required. In this table, the intersection of a row and a column gives the flow between states. From this table, a table showing transition rates (for, say, one or two years) between labour force states could be constructed. This would show entry rates into a particular state from all types of states, together with exit rates, and continuation rates (the proportions of people who remain in a particular state between one period and another).

In Ireland there is already one survey which in part can be used to provide flow data: an annual survey of school-leavers undertaken for the Department of Labour. This is an example of a survey which would provide the answer to a number of policy questions if the basic data were accessible.

Policy Issues

In summary some of the policy issues which arise are:

- There is need for some form of coordination (with no coercion?) of non-CSO data with CSO data, in order to ensure quality control, i.e., some minimum standards and compatibility between individual data sets. There is also scope for some gatekeeping functions with regard to the increasing volume of surveys, including surveys of Irish firms, by those in the public sector. And there could be some rationalization of the existing surveys of Irish firms, leading to richer possibilities for data analysis.

- Should some minimum standards be set in relation to data which are collected as a by-product of government contracts, including their deposit in a central location?

- What forms of surrogates for the market system to identify user demands ("needs"?) can be developed - e.g., some mechanism for feedback and complaints, for wide consultation and receipt of suggestions at an early stage of the design of inquiries? For example, could there be greater use of charges in order to get an indication of the implicit value which different users put on different types of statistics?

- What forms of users' guide are required for the existing mass of data which public bodies hold? And what ground rules on access to public sector data by outside users should be established?

- There is need for public sector bodies to consult with CSO on the statistical possibilities of using administrative records. Related to this, how can the danger that the increased diffusion of computers among public bodies, including the local authorities, will lead to the onset of incompatible data sets, which are in turn incompatible with CSO data, be avoided?

- Despite the above remarks on the potential for using administrative records, there are many issues of economic and social policy where there is no substitute for survey or census statistics at the level of the individual or of the household, especially where it comes to establishing relationships between variables. This is especially true in view of the increasing

incidence of "non-standard" households such as lone-parent families and the many family policy issues which result. To what extent could existing surveys be used with additional questions being asked at marginal cost? In Ireland, there are many possible uses of a general household survey with a central base of questions and a varying sub-set of questions from year to year. This could provide information on education and on use of health services, for example. There is no possibility of adding to the Household Budget Survey (i.e., family - expenditure type), which is already overloaded.

- To what extent should a simpler Census of Population be conducted, with a shorter time lag for dissemination, leaving matters such as housing to sample surveys?

- To what extent should one try to force the development of intermediaries who would take primary industrial data and engage in "packaging" and marketing to industrial customers; are there economies of scale in this activity?

- Underlying the paper as a whole, there are the difficulties in deciding how much public resources to put into statistics. For example, non-CSO data will reflect producer preferences as to internal management needs or inter-departmental and inter-agency coordination: this may not be optimal from the point of view of society as a whole. ¤

Schematic outline of type of labour market flow data which are required

Initial State	Final State (numbered as in rows)
	1i) 1ii) 2 3a) 4 5 6 7
1. Employed	
a. full time	
b. part time	
2. Unemployed	
3. In partial retirement	
a. In the same job as when last fully employed	
b. In a different job	
4. Fully retired	
5. In labour force, sick	
6. Unable to work due to disability	
7. Not in labour force due to "home duties"	