Distr. GENERAL

8 October 2012

WP. 16
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Seminar on New Frontiers for Statistical Data Collection
(Geneva, Switzerland, 31 October-2 November 2012)

Topic (iii): Legal and institutional aspects of using new data sources

REALISING THE STATISTICAL POTENTIAL OF ADMINISTRATIVE DATA

Contributed Paper

Prepared by John Dunne and John Hayes, Central Statistics Office, Ireland

I. Introduction

1. The National Statistics Board (NSB), in its *Strategy for Statistics 2009-2014* report, laid out a strategy for achieving the board's vision of an Irish Statistical System, a holistic system based on the exploitation of administrative data. Recent changes in the economic and financial climate, with the concomitant reductions in public spending, have meant that the full exploitation of administrative data to support evidence-based policy making has now become critical.

2. This paper describes the Central Statistics Office (CSO) experiences and progress to date in positioning itself to achieve the NSB vision. In particular the paper will discuss the role of the Administrative Data Centre (ADC), a new set up in 2009 with the dual purpose of acting as a clearing house for administrative data received from public bodies, and to be a catalyst for the further development of the Irish Statistical System. This paper will cover different aspects of the journey to date: political, legal, organizational (internal and external) and technical.

II. Political environment

3. The NSB, established under the Statistics Act, 1993, promotes a whole-system approach to the production of Official Statistics in Ireland and the development of the Irish Statistical System.

4. The 2009 NSB paper *Strategy for Statistics 2009-2014*[1] laid out a strategy for achieving the board's vision of an Irish Statistical System. This strategy is built around five strategic priorities:

   (a) Ensuring that the Irish statistical system is coherent and that its potential to inform policy and fiscal decisions is fully realized;

   (b) Increasing value for money and efficient use of resources by prioritising investment in statistics across government bodies;

[1] http://www.nsb.ie/media/nsbie/pdfdocs/StrategyforStatistics2009-2014.pdf

(c) Developing systems to ensure that the burden of response on businesses, households, and individuals is minimized and that the statistical value of existing survey and administrative data is maximized;

(d) Ensuring that official statistics are fit for purpose by balancing timeliness, cost, and quality of data against existing and expected future demands;

(e) Improving access channels and promoting use of CSO statistics.

This paper also identified three critical infrastructural needs of the Irish Statistical System: a unique business identifier and a central business register; a unique personal identifier; and a spatial and geographic data capture.

5. The 2011 National Statistics Board's position papers, *The Irish Statistical System: The Way Forward* and *Joined Up Government Needs Joined Up Data*[2], elaborated on some of the core objectives of the earlier document, advocating:

(a) The development of the infrastructure to maximise the use of data sources – this includes the compilation of registers of persons, businesses, and buildings, with linkage between each such register – joined up data;

(b) The development of a professionally independent framework for the system;

(c) The addressing of data protection issues constraining the matching of data sources;

(d) The engagement of policy-makers.

Joined Up Government Needs Joined Up Data takes the idea of a national statistical infrastructure one step further and talks about the benefits of a national data infrastructure for not just statistical purposes but also policy and administration purposes.

6. Helping to realise the NSB vision for the Irish Statistical System is the fact that Ireland is a small country with typically centralized administrative and statistical functions. The language and examples presented in the NSB position papers with respect to Joined up Data and Joined up Government is playing a critical role in making the concept of the Irish Statistical System tangible to key decision makers in the wider Public Sector. Incorporation of this vision into the Public Sector reform plan and the requirement of 'joined up data' for administrative as well as statistical purposes facilitates the development of a partnership approach with other Public Sector bodies.

7. The government *Public Sector Reform Plan*[3] published in 2011 further supports the development of the Irish Statistical System with the following stated objectives:

(a) Improved sharing of data on businesses across the Public Service – this includes the development of business registers which are linkable to that of the Revenue Commissioners;

(b) Developing a code of practice for data gathering and its use for statistical purposes across the Public Service – this includes promoting consistent approaches to identifiers, classifications, and geo-spatial/postcode data;

(c) Assessment of the legislative environment with a view to identifying the scope for additional and greater uses of statistical data, including any potential legislative changes.

[2] http://www.nsb.ie/media/nsbie/pdfdocs/NSB%20ISS%20Position%20Papers.pdf
[3] http://per.gov.ie/wp-content/uploads/Public-Service-Reform-pdf3.pdf

8. Ongoing or completed projects that exploit administrative data for statistical purposes to address information gaps in key policy areas also play a significant role in promoting the development of the Irish Statistical System. For example, a project to track school leavers into further education, employment and or unemployment has significant value for policy analysts working in the education, children and labour market domains. It is also a tangible and concrete example of what joined up data can achieve.

III. Legal environment

9. The CSO was established statutorily under the Statistics Act, 1993[4]. This legislation contains articles that assign a number of key powers with respect to data held in public authorities to the Director General of the CSO:

   (a) The power to require a public body to provide copies of or extracts from any records in its charge for statistical purposes;

   (b) The public body shall consult and co-operate with the Director General on assessing the statistical potential of its records and in developing its recording methods and systems for statistical purposes;

   (c) The public body shall consult with and accept any recommendations the Director General may reasonably make if that public body proposes to introduce, revise or extend any system for the storage and retrieval of information or to make a statistical survey.

   The Statistics Act also includes articles on statistical confidentiality. These preclude the CSO from communicating to any person or body any information relating to an identifiable person or undertaking. The Statistics Act also positions the CSO in the context of the data protection legislation with respect to the collection of data. In summary, the critical pieces of statistical legislation are already in place to support the exploitation of administrative data for statistical purposes.

10. National legislation already incorporates the key elements in EU legislation with respect to exploiting administrative data for statistical purposes:

   (a) Regulation 223/2009[5] which contains articles relating to NSIs and Eurostat access to administrative records;

   (b) The European Statistics Code of Practice[6], which advocates the greater use of administrative data in order to reduce response burden.

11. Two other primary national pieces of legislation with respect to data held by Public Authorities are:

   (a) The Data Protection Acts, 1988[7] and 2003[8], which govern the uses of private information held by government bodies and other entities in Ireland;

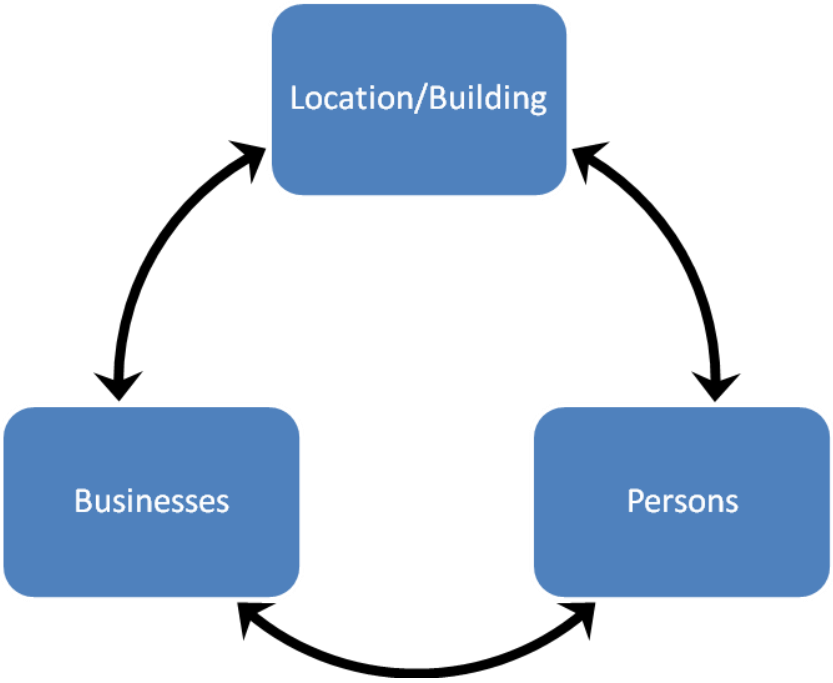
   (b) The Freedom of Information Acts, 1997[9] and 2003[10], which allow individuals to view any information held on them by government bodies.

[4] http://www.irishstatutebook.ie/1993/en/act/pub/0021/print.html
[5] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:0173:EN:PDF
[6] http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-32-11-955/EN/KS-32-11-955-EN.PDF
[7] http://www.irishstatutebook.ie/1988/en/act/pub/0025/print.html
[8] http://www.irishstatutebook.ie/pdf/2003/en.act.2003.0006.pdf
[9] http://www.irishstatutebook.ie/pdf/1997/en.act.1997.0013.pdf
[10] http://www.irishstatutebook.ie/pdf/2003/EN.ACT.2003.0009.pdf

IV. Organizational aspects

A. Joined-up data

12. The development of the Irish Statistical System has similarities with early chapters of the story describing the development of the Danish Statistical System. This story is available in *The importance of the archive statistical idea for the development of social statistics and population and housing censuses in Denmark (*Thygesen, Lars 2011)[11]. Joined up data envisages three comprehensive lists or registers (persons, businesses and buildings/dwellings) and the linkages between these registers. In summary, these lists should be able to link persons to their employers or school, link persons to other persons within the household, link a business to where it operates, and link a person to the location of their residence. The development or existence of these lists can facilitate the linking of data whether from surveys or administrative sources. Ideally, linking should occur using unique permanent official identifiers, however where there are gaps other record linkage techniques can be used.



13. The CSO Business Register is fully aligned to administrative sources from the Irish tax authorities.

14. The linkage between persons and business is available from the employer tax returns to the Revenue Commissioners for employer/employee relationships and from sources in Education authorities to identify where a person is studying. The former source, in conjunction with the Business Register and age/sex information from the Department of Social Protection, has been used to provide comprehensive statistical information about the flow of workers and jobs between firms[12] .

15. There exists a comprehensive buildings database for the state with relevant geospatial identifiers called the Geodirectory[13] . The database was established as a joint venture between An Post and Ordinance Survey Ireland. This database is available on a commercial basis. However this database does not contain any information on who lives in what building. It has some limited

[11] http://ww4.dst.dk/upload/nordbotten_and_denmark_final_draft_4.pdf
[12] http://www.cso.ie/en/surveysandmethodology/industry/jobchurn/
[13] http://www.geodirectory.ie/

information with respect to businesses, as a trading name is usually available as part of the address.

16. Ireland does not as yet have a post code system that can be used in administrative databases for compiling small area statistics. At present, the compilation of small area statistics from administrative sources involves matching address strings from administrative registers/sources against the Geodirectory to identify the geospatial element of the address. Typical challenges in this type of exercise are:

    (a) non standard address strings in data sources, and

    (b) non unique addresses in rural areas (typically a persons name has to be combined with a townland to allow mail to be delivered in a rural area).

17. The Department of Social Protection (DSP) is the primary source of person-based data and maintains the master list or Client Record System (CRS) of official Personal Public Service Numbers (PPSN) in the State. This list is available to the CSO and is the basis of an ongoing project to develop a Person Activity Register for statistical purposes.

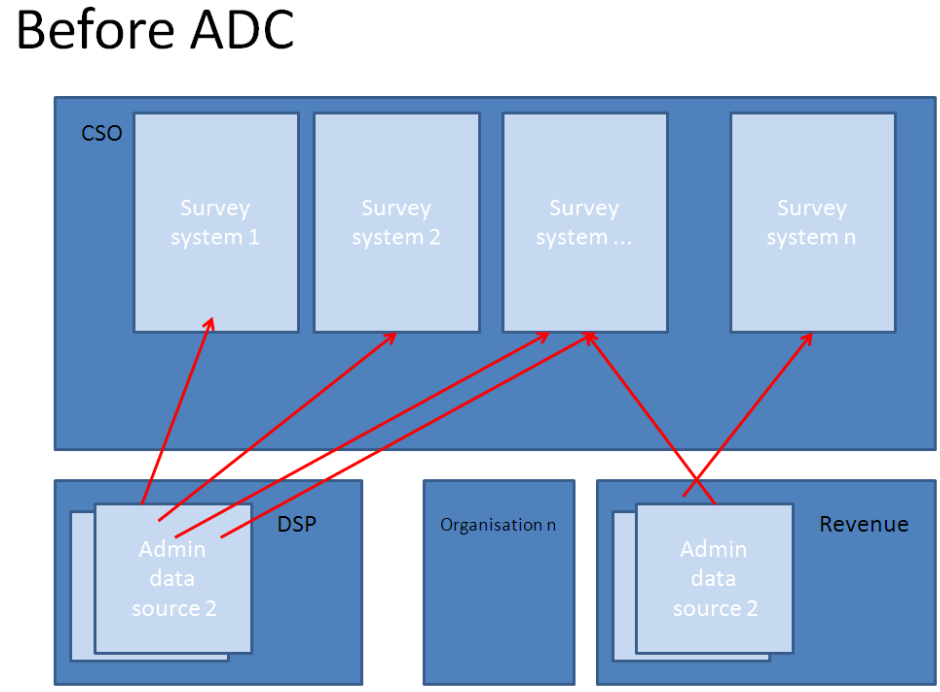
18. The objective of the Person Activity Register project is to build a statistical register that contains a summary of each person's engagement with key administrative systems (children's benefit, education, employment, unemployment, self-employed, pension, etc) on an annual basis. The primary objectives of this project are:

    (a) to provide general information on structural changes in the population over time;

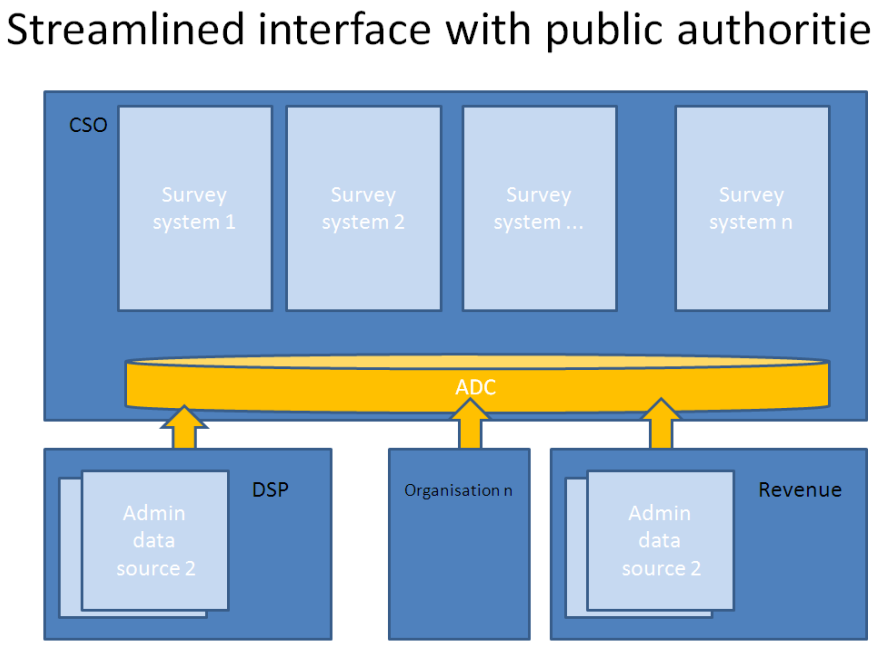
    (b) to provide aggregated information on policy or programme outcomes by tracking cohorts of the population over time and across systems.

B. Within the CSO

19. The Administrative Data Centre (ADC) is the CSO unit designated as the conduit for data transfers from other government bodies and is the central repository for received data from those bodies. This unit currently maintains over fifty different administrative data flows serving the statistical production systems in the CSO. ADC controls access to the data in accordance with confidentiality obligations under national and EU legislation, reinforced by international, national, and internal CSO protocols on data confidentiality. Subject to these criteria, ADC may also make anonymized data available as Research Micro Files (RMFs) to external researchers who demonstrate a legitimate research requirement to access the data.



And, following the setting-up of the ADC...



20. CSO can and does undertake data linkage exercises to further enhance the possibilities for statistical analysis. Such data linking may only be done with the prior written approval of the Director General of the CSO. Data linking is governed by the conditions of the CSO Data Protocol[14], and a public register of all such linking is maintained.

C. Interaction with other Public Sector bodies

21. The CSO is moving to a strategy of implementing institutional-level Memorandums of Understanding (MoUs) to underpin the flow of administrative data to CSO. This strategy replaces the existing process of underpinning each administrative data flow with a specific MoU. This

[14] http://www.cso.ie/en/aboutus/csodataprotocol/annex1/

strategy not only streamlines the administration governing such data flows, but also allows MoUs to address other items of interest relevant to both organisations. For example, in the case of the Office of the Revenue Commissioners, the memorandum of understanding[15] provides for a deeper, co-operative relationship. This relationship has allowed the CSO to adopt a business register which is based on the Revenue Commissioners' registration system and to use the Revenue Customer Number as a common business identifier between the two bodies.

22. The CSO is also involved in a number of senior level cross departmental groups addressing relevant data issues. This involvement is with a view to ensuring the best interests of the Irish Statistical System are considered. The ADC also leads the Statistician Liaison Group, a group setup in 2010 to act as a nexus for all matters of interest to the Irish Statistical System for the small number of statistical units that are operating in government departments. This latter group has been invaluable in working with the development of a code of practice for the Irish Statistical System.

VI. Technical aspects

A. Technical aspects

23. Data from other government bodies are first encrypted before being transferred using a secure transfer protocol to a CSO file server. Received data are converted to SAS format and held in a warehouse environment having source, analysis, and external researcher tiers. There are controls in place governing the levels of access by CSO users to the three tiers – usually users will not have access to the source tier, which contains identifiable information, without authorization from top management. Treatment of all data held by the CSO, whether collected directly or obtained from an administrative source, is subject to the CSO's Code of Practice on statistical confidentiality[16]. SAS is used as the storage medium for data in the ADC warehouse.

24. In the case of person-based administrative data, ADC anonymizes/depersonalises such files before making them available to CSO users as analysis files. Names and addresses are dropped. Date of birth is set to the first day of the month of birth. The Personal Public Service Number (PPSN, the official identification number for persons) is replaced with an alternative Protected Identifier Key (PIK) that, while serving to mask the original identifier, preserves the statistical value of the data in terms of linking across data sources and over time. The source for the Protected Identifier Key is a closely guarded secret within ADC.

25. All CSO staff have access, via a data portal, to core metadata and summary statistics on all administrative data held by ADC. Providing such information ensures all staff have the opportunity to consider and explore ideas to further exploit the statistical potential of administrative data. Core metadata is generated from SAS datasets and registered in a relational database.

26. The conceptual data model used to describe the administrative data coming into CSO can be briefly described as follows:

- **Administrative Data Flow** - It is assumed that data coming from a given system comes in on a periodic basis and as such is described as an Administrative Data Flow. When a flow is processed for availability in the analysis tier of the warehouse a new data flow is created.
    - **Data flow instance** - Each data flow can contain multiple instances – an instance refers to a time period usually relevant to some aspect of the data (i.e., date of registration, date of occurrence). An instance is typically described by its periodicity (monthly, quarterly etc).

[15] http://www.cso.ie/en/aboutus/descriptionsandfunctions/memorandumofunderstandingbetweenthecsoandrevenue/
[16] http://www.cso.ie/en/aboutus/statisticalinquiries/statisticalconfidentiality/codeofpractice/

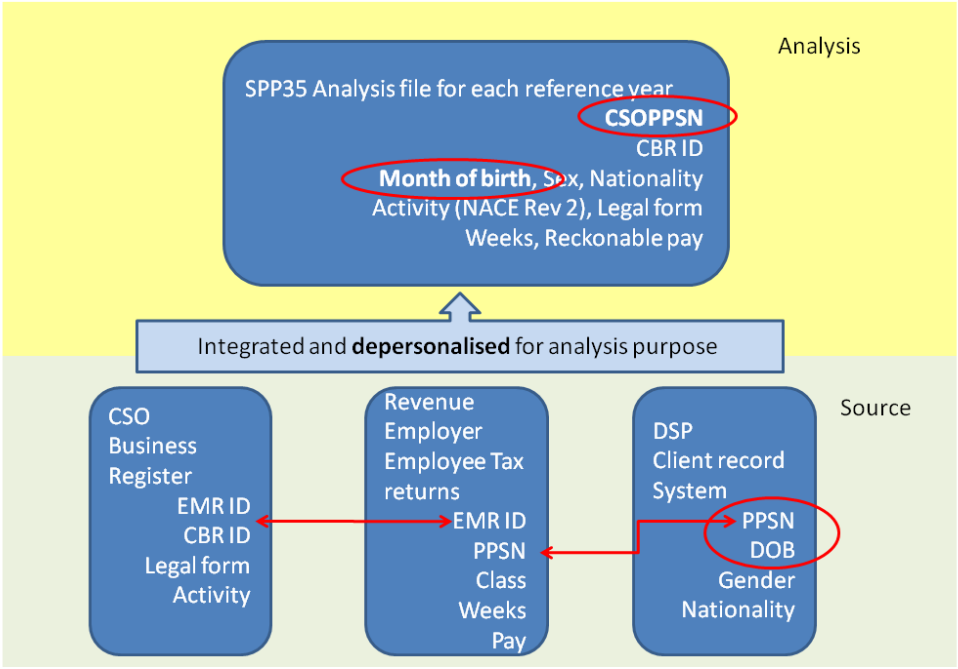
- **Instance version** - Each instance can contain multiple versions for that instance, for example CSO may receive multiple versions of the same tax record file with each subsequent version being more complete/comprehensive.

    - **Dataset Collections** - Each version then can contain a collection of datasets.

This model also underpins how the metadata and dataset summaries are presented to prospective users of the data.

27. The creation of analysis data files that serve multiple purposes is a primary consideration in bringing data onto the analysis tier from the source tier. Typically an analysis file will summarise records at the person/business unit level when being created rather than create an analysis file where there might be multiple transactions per person/business. The idea behind this approach is to create simple easy to use files that serve multiple purposes. For example, in developing the job churn statistical product to explore job and worker flows between enterprises, the underlying analysis dataset was created with many statistical purposes in mind. In particular, the analysis dataset provides the link between person- and business-based registers.

An example – creating the P35 analysis file



VII. Concluding remarks

28. Based on our experience to date, the following are some strengths and weaknesses in the CSO's approach to administrative data management. Amongst the strengths are:

(a) Having memoranda of understanding in place to define relationships;

(b) Having high-level liaison groups with each of the major providers of administrative data;

(c) Person to person relationships, in the form of liaison officers.

The weaknesses include:

(a) The resource focus of government bodies is administrative function, not statistical data collection;

(b) There are no conventions or standard methodologies to describe the data or how to use them;

(c) Data quality issues are sensitive and can be practically difficult to pursue;

(d) Due to resource constraints within ADC, the personnel best qualified to exploit administrative data do not necessarily have the time to do so;

(e) Person to person relationships (liaison officers).

29. Data sharing amongst government bodies in Ireland is still confined to a small number of organizations. The CSO is uniquely positioned through legislation to collect data from all other government bodies, while, at the same time, being restricted by the same legislation from sharing data on identifiable person or business entities. The key challenge for the CSO is to fully exploit the statistical value of administrative data (to mine the mine), especially by availing of increasing opportunities for joining up available data sources. Steps to complete fully joined-up data jigsaw might include:

(a) The implementation of a location or (x,y)-based identifier for buildings;

(b) The implementation in key public administration systems of a link between a person and a household / building, in particular where he or she normally resides, and where the household/building is identified by a unique identifier and can be associated with a small area code;

(c) The mandatory use of the PPSN number in the engagement of persons with the state through the different life stages: children's benefit, primary school, secondary school, third level, further education, employment, social welfare, retirement, death (i.e., cradle to grave approach)

(d) The implementation of a unique business identifier for businesses interacting with the state, which, when combined with a building identification number will facilitate the identification of the premises (and associated small area) attached to a specific business.

While there are gaps in the current joined up data jigsaw, there is still significant value to be gained. There may also be available statistical methods for bridging these gaps where they exist.

30. The Irish Statistical System continues to face significant challenges in the years ahead, however in the words of W. Edwards Deming, "It is not necessary to change. Survival is not mandatory."