NSB seminar 'Data as a Public Policy Issue'
October 25, 2016
Dublin Castle

Presentation: Building the National data Infrastructure (NDI) Part 2, Andrew O'Sullivan

Most European countries have some form of National Data Infrastructure to improve cross-referencing data for statistical and policy analysis purposes.

The term National Data infrastructure (or "NDI") was first developed in Nordic countries in the 1960s and has been widely adopted across Europe since then, especially in Northern and Eastern European countries. It is defined by a system of physical registers or lists, the three most important of which are people, businesses and locations. The maintenance of each register is the responsibility of a specific government department but it's the insight gained from cross referencing that holds the key to the extensive use of data driven policy, services and transparency in the Nordic countries. So, the NDI is usually represented by this diagram showing the interconnected registers. This gives the essential model for turning data into insight based on the three key identifiers.

The technology is available to build this kind of physical NDI requires but it also needs a legal and cultural readiness to implement, which is still evolving in Ireland. In the meanwhile, we plan to develop a virtual NDI that supports progressive register building over a longer period based on current legislation. It's less efficient, but it allows us to make progress on joining up the dots in the interests of better insights until a more formalised structure exists.

The public sector has a rapidly growing number of data sources – from digital services to less structured data that needs to be recorded as part of other service transactions with citizens. Each department or agency controls and stores its own data to enable its own operations and policy formation. The use of the three key identifiers varies from one department to another so, while the data itself is an important operational asset, it is usually difficult to cross reference it between departments which in turn makes it harder for data to really become a strategic asset . This means that each department's administrative data can become its own silo – so while we are often data rich for operations, we can also be insight poor at a wider Government level.

There are a growing number of other sources of data – such as traffic sensors, secondary data such as retail price surveys, business metrics and of course the census. One of the reasons the CSO needs to run a national census is to compensate for the absence of physical registers. Countries that have a full register based NDI do not need to run a full scale census since they already have most of the information they need and it is constantly up to date.

The CSO is the custodian of all information used for the purposes of official statistics in Ireland. Data is collected from multiple sources under the Official Statistics Act and is it highly confidential. We need to collect data to function but we take its protection very seriously. We will never share any information that would allow an individual to be identified. Confidentiality and independence are two of the core principles of the CSO and they form a wall around the data that we are entrusted with. Before the data can be used effectively, it needs to be linked together. In the absence of registers, we need to tie it together using a very inefficient and complex process of record linking before we can cross-reference multiple sources. This means after a lot of preparation work to make up for the lack of common identifiers, we can start to analyse the data and produce statistics and insights into the social and economic conditions in the country to Government, Businesses, NGOs and Citizens.

Government Departments also link data together from multiple sources so that they can analyse it to help inform and improve their services and policies. They need special data protection agreements and sometimes new legislation to access data outside their own department but that is a time consuming process so they need to know exactly what to ask for so that they create the right

shareable administrative data. The CSO provides analytics resources and expertise to a number of Departments already and we are expanding this Irish Government Statistical Service. Data scientists need data and especially linked data to be effective though. So while the CSO will never release individual level data to departments, we can answer their questions in aggregate and we can advise on analysis methodologies that will work.

Our approach is to building the virtual NDI is to identify high value pathfinder projects where the CSO can provide aggregate data and methodology expertise to identify what a department needs to collect and which of the three key identifiers it needs to cross-reference to assess its strategy or improve a specific service. For example, allocating an Eircode earlier in the planning cycle would make it easier to measure progress from rezoning to when a first time buyer turning the key in their new home. These so called longitudinal datasets (i.e. tracking the same kind of information at multiple points in time) are essential for policy evaluation. The insight gained from these projects is enormously valuable but every time we deliver one of them by going around this cycle, the quality of cross-referencing in a department improves and the use of identifiers in individual databases also improves. As we get more of these projects running we will create a form of virtuous cycle of better CSO data providing better high level insight and guidance to departments which in turn creates legitimate drivers to record the three key identifiers. With enough projects, we start to move towards a federated national data infrastructure. The data is still under the exclusive control of the department that needs it to provide essential services. The colours here don't refer to duplication of data across departments but to the usage of common identifiers. Using identifiers means it can be more readily cross-referenced if and when the necessary protection arrangements are in place. We are then starting to move from data silos to data nodes as part of a platform that lays the groundwork for joined-up data for joined-up government. The CSO will need to do spend less time record linking so we can concentrate on producing higher-level insights. The more data we have overall but even more importantly the better organised it is through this virtual National Data Infrastructure, the easier it becomes to create an informed society with quality open data, efficient researcher access and real insight based policy making. Better organised data is also safer data – you can better control and audit what you have, who accesses it and how you use it in the public interest.

To wrap up, our goal is to build up a virtual National Data Infrastructure through a series of pathfinder projects until we have a more formalised infrastructure for sharing information. The CSO and a number of Government departments are currently working on identifying more of the kinds of project that Paul mentioned that improve services and provide better policy insight, but all suggestions are welcome. The CSO is allocating staff to provide more data science skills to Departments. However, for this to work, it is critical that Departments and agencies expand the use identifiers to organise data, not least so we don't keep asking for the same information every time. One simple example is Eircodes – accurate location information is currently the weakest of the three identifiers but one of the easiest to collect from an implementation and legislation perspective. So, record the Eircode when you record the address.