Taking opportunities to plan quality - the development of a new Secondary data model for CSO Ireland.

Ken Moore

Central Statistics Office, Ireland

Abstract

A new CSO Secondary Data Model for CSO, Ireland has been developed to streamline, standardise and where possible automate the process of acquiring, receipting, ingesting and accessing data sources used across the Office. The model was developed in response to the increasing volumes of secondary data sources being received by the Office from the national statistical system, the introduction of a new data hub to manage this data and the strategic move to a “Secondary data first” principle. The model initially focuses on how we manage, govern and reuse administrative data, privately held data and publicly available data within the CSO.

As part of the development of the data model, Quality division placed a key role in ensuring that we took to opportunity to build and plan quality into each step of the new model from data acquisition to data access. This paper will outline the improved quality management elements introduced as the model moves from development to implementation including:

- Developing a quality assessment protocol for examining the quality of inbound secondary data from the statistical system
- Providing for improved engagement with our data providers so we have greater clarity on the data quality checks they carry out and the metadata they associate with the data prior to transmission to CSO
- The introduction of standardised quality checks on the data once received in CSO for ingestion into the Data Hub
- Increased feedback communications with data providers so that common data quality issues are discussed and where possible resolved
- The provision of centralised data services so that commonly used data flows are managed consistently at an initial data processing phase – centralised edits/data matching and linkages and improved metadata standards.

The paper will also outline progress made to date, the challenges being experienced with introducing these changes and the next steps planned for the implementation of the new model.

Keywords: standardisation, data quality, engagement, feedback, communications

1. Introduction & Context

The Central Statistics Office (CSO), Ireland has recently developed a Secondary Data Model to streamline, standardise and automate the process of acquiring, receipting, ingesting and accessing administrative and privately held data sources. In order to describe the variety and range of different data sources being used in CSO we use the term secondary data model in this paper. Secondary data sources form a central pillar of all statistical production in CSO, and it is estimated that secondary data is used in approximately 70% of our statistical outputs. A dedicated business division, the Administrative Data Centre (ADC), is responsible for managing, governing and providing access to secondary data sources in the CSO. The growth in access and availability to secondary data sources from other producers across the Irish Statistical System continues to expand with the ADC currently receiving over 240 statistical flows on a regular basis. Each statistical flow can contain many datasets with the total number of datasets being managed by the ADC estimated to be over 250,000. The frequency of these flows varies with data been received on a daily, weekly, monthly basis etc. Given the widespread use of secondary data use across the Office it is critical that the quality of this data is governed and managed in the best way possible in order to get the maximum value of the data to help the CSO inform policy makers and society. This paper will focus on the quality related elements of the model rather than on the technical elements behind the model.

2. Key Drivers

Prior to the development of the model, the process for acquired, ingesting and gaining access to secondary data in the CSO was somewhat fragmented. In general ingestion and acquisition was led by ADC and data was taken in through a dedicated ADC portal. However, some statistical production areas completed their own governance process and received data from suppliers directly independent of the ADC or the ADC portal. This raised the potential risk of inconsistency in the areas of governance, data transparency, coherence and coordination over time. In addition, in the past the primary focus for ADC was to act as a post box for getting as many statistical flows ingested for use which at times meant that the quality of the data and metadata was not managed in an effective or consistence way for certain flows. This in turn impacted on how the data was used by statistical producers in the Office. There were also challenges in how feedback was managed with data providers as the statistical producers did not always have a clear communication channel to follow up directly with the providers of the data flows they were using in their production with any queries regarding data quality.

3. Overview and key drivers of Secondary Data Model

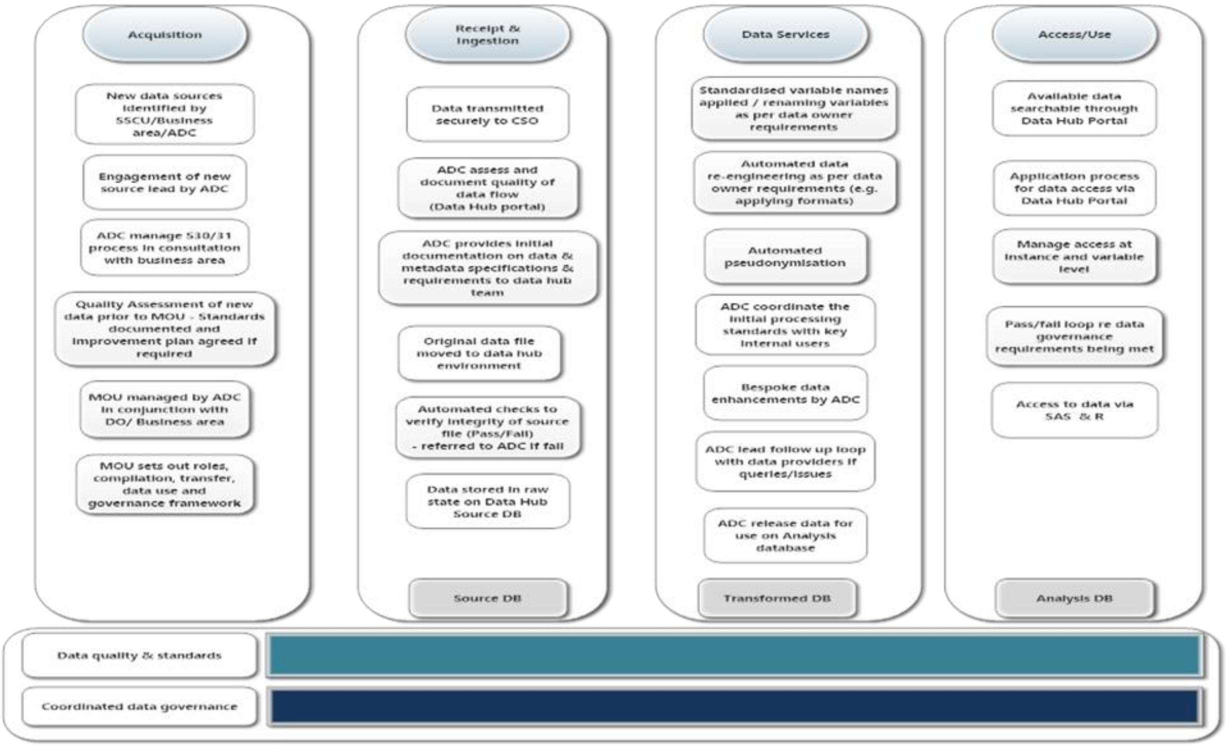
The overall purpose of the new Data Model was to provide clarity on who is responsible for which aspects of the process e.g., managing data suppliers, governance documentation etc. In order to successfully implement a secondary data first principle, there had to be a system to support that and to lay out the legal basis for our data acquisition and governance. The development of the CSO's Secondary Data Model responds to this challenge by clearly identifying and sequencing the process involved from data acquisition to access and use. It is focussed on streamlining, standardising and automating the process where possible. It places the Administrative Data Centre at the heart of the model as Business Partners proactively responding to user needs.

The key drivers to develop and introduce the secondary data model are:

- **Clarity** - lack of clarity on who is responsible for what, e.g., managing data suppliers, governance documentation etc.
- **Governance**–need to highlight and embed all aspects of Data Governance in all our data processing activity.
- **Support** – Provision of support to our internal and external users to further advance out strategic goals of secondary data first principle, the advance the implementation of the National Data Infrastructure and to deliver on our Data Stewardship role.
- **Quality –**need to embed quality into all aspects of secondary data (e.g., metadata, data standards, formats, data treatment, coherence in a consistent and transparent way
- **Use –** To enable us to increase the use of secondary data sources to support the principle of secondary data first.

A visual overview of the model is included in Figure 1.

Figure 1. CSO Secondary Data Model


Acquisition
New data sources identified by SSCU/Business area/ADC
Engagement of new source lead by ADC
ADC manage S30/31 process in consultation with business area
Quality Assessment of new data prior to MOU - Standards documented and improvement plan agreed if required
MOU managed by ADC in conjunction with DO/ Business area
MOU sets out roles, compilation, transfer, data use and governance framework
Receipt & Ingestion
Data transmitted securely to CSO
ADC assess and document quality of data flow (Data Hub portal)
ADC provides initial documentation on data & metadata specifications & requirements to data hub team
Original data file moved to data hub environment
Automated checks to verify integrity of source file (Pass/Fail) - referred to ADC if fail
Data stored in raw state on Data Hub Source DB
Source DB
Data Services
Standardised variable names applied / renaming variables as per data owner requirements
Automated data re-engineering as per data owner requirements (e.g. applying formats)
Automated pseudonymisation
ADC coordinate the initial processing standards with key internal users
Bespoke data enhancements by ADC
ADC lead follow up loop with data providers if queries/issues
ADC release data for use on Analysis database
Transformed DB
Access/Use
Available data searchable through Data Hub Portal
Application process for data access via Data Hub Portal
Manage access at instance and variable level
Pass/fail loop re data governance requirements being met
Access to data via SAS & R
Analysis DB
Data quality & standards
Coordinated data governance

4. Building quality into the Secondary Data Model

Given the review on how secondary data sources were being managed and governed, the Quality division of the CSO took the opportunity to collaborate with our colleagues in ADC and the CSO Data Office (responsible for data governance) to build as many quality improvements into the new secondary data model as possible and practical. These improvements include:

4.1 Centralisation of data transmissions and governance

It is intention of the model that over time all secondary data will be transmitted to the CSO Data Hub through the Secure File Transfer Protocol (SFTP). This means that those statistical business areas who currently receive their data flows directly will migrate to the centralised ADC ingestion over time. The Data Hub Team will receive the data and complete some initial data structure checks which are specified by the ADC Team. If the transmitted data fails the initial checks, the ADC team will be alerted to review the data and correspond with the data provider if necessary. In addition, all the data governance documentation will be coordinated and processed centrally in the ADC, thereby allowing for better consistency and coordination over time.

4.2 Quality Assessment Protocols

In the Data Model, the ADC is responsible for managing the quality of the files ingested before they are used across the Office. ADC will apply standard checks and edits to improve coherence and consistency of re-use. As part of this work ADC and the Quality team have worked on developing a quality assessment protocol. This assessment is carried out prior to transmission of the data to examine the data being transmitted to identify any quality challenges and potential improvements. If necessary, a quality improvement plan will be developed documenting the strengths and limitations of the data source. Initially, the CSO may have to identify gaps and apply changes. Over time, the ADC will work with data suppliers towards a minimum quality standard including but not limited to the use of labels, metadata, and supporting documentation.

4.3 Automation of data services

The Data Hub Team will work closely with the ADC team to automate the initial processing of the data. This includes, but is not limited to, the identification of required variables, file consistency re data layout and sequencing and consistent pseudonymisation of personal related identifiers. Over time, as the transmission becomes more consistent, further automation of the processing is expected. Once the initial processing of the data is completed, the ADC team will be notified that the files are available to the ADC for initial data services.

In addition to initial processing the ADC will provide a Data Linkage and Integration service to ensure that data is linked in a consistent manner. To this end, ADC have identified a spine of standardised variables which will support this work. This central spine includes the CSO PPSN (Personal identifier), Eircode (Location identifier) and Unique Business Identifier – these identifiers form the basis of Ireland’s National Data Infrastructure. Work is still ongoing on developing a standardised data linking and integration methodology.

4.4 **Common Editing & Automation**

The ADC will engage with internal data users or user groups e.g., PMOD User Group (Key Taxation data) to establish the required processing standards and will work to implement/automate these requirements. This will result in standardised editing or manipulation of data so that any changes to the secondary data is consistency across business areas and documented. ADC will also assess and document the quality of the data flow and will work with the Quality Team to establish minimum data and metadata standards will be applied along with standardised quality checks. This work, and that in the previous steps, will improve coherence. Again, these will be documented so that internal users are aware of the quality of the variables that are looking to use, including quality flags where there are potential issues with certain fields.

4.5 **Relationship building**

One of the key strengths of the data model is the focus on improved communications and the building of better working relationships with our data providers. As the data model matures, the ADC will have had extensive engagement and opportunities to strengthen relationships with their data providers either directly or through Joint Liaison Groups. Hence, the ADC in conjunction with users/liaison groups will follow up with data providers and will work with them to better meet the requirements of the data model and statistical production areas. This allows for greater clarity and oversight for both the providers and the users of data. The ADC will also look to our network of seconded statisticians and the Formal Statisticians Liaison Group (FSLG) for support in building these relationships. One good, practical example of this improved communications can be seen in the quarterly knowledge sharing sessions with our Tax Authority where they can explain which variables are key for them and describe the checks they carry out while CSO can describe how we use their data, what are our key variables and what consistency checks are used to review quality in statistical production.

5. Benefits of new approach

There are many benefits evident from the move to the new secondary data model which include:

- The centralised, standardised, transparent approach with all data being acquired, ingested, governed and managed by the ADC allows for greater clarity on the CSO's secondary data holdings. This provides reassurance to CSO's senior management team and to the data providers who are supplying data to CSO.
- The ingestion process has also become more efficient with elements of automation when data is received and initially processed in the CSO data hub.
- The quality of the metadata and associated documentation used to describe the secondary data and how it is created and managed externally by our data providers has also improved.

- By concentrating all the data governance requirements within ADC, the data governance process has been made more efficient and there is greater visibility of the controls and safeguards in place to meet our obligations under GDPR legislation. There is also greater evidence of controls for the Data Protection Commissioner as to how CSO manages and minimises personal information.
- The single point of contact between the ADC team and each data provider has allowed a better working relationship to develop. With the increased communications with data providers, there is now greater clarity on both sides of how the data is treated and managed externally by the data provider and internally by the statistical producers. There is also greater clarity on which variables are key and which variables may not have been a priority to the data provider.
- There is now greater collaboration internally between statistical producers using common data sources through improved communications and engagement with user groups. This had resulted in improved coherence across production/outputs for commonly used data files (e.g., PMOD) through the use of common edits and checks.

6. Challenges

As with all major change projects, there are many challenges that need to be overcome and managed. Some of these statistical business areas who received their secondary data directly from their data providers were obviously concerned with the planned changes, particularly as they would be required to change their processing systems and they feared they would lose control on the timeliness in receiving the data. While some were convinced that the new model would offer better quality, more consistent data, others are still resisting the shift to the new working arrangements and are adopting a “wait and see” approach. In addition, some data providers are slow to engage with the increased level of collaboration for a variety of reasons. Some producers are concerned about the quality of their processes and data but are reassured when we advise that we have dedicated quality supports and services available to them under the CSO’s data stewardship umbrella. Others are adopting a take it or leave it approach as their main concern is providing services to the general public and they are slow to change or update their administrative systems in case they performance of their systems are impacted. This will take time to manage but the information sharing sessions with them are convincing that there is value in improving the quality of their data for their own operations uses as we focused on discussing benefits of engaging with the new processes. By concentrating on the message that the benefits outweigh the effort required and that CSO will support them through this transition period, we are hopeful of getting full buy in and cooperation over time.

7. Conclusion

The CSO is a data-driven organisation that recognises that data is the life blood of an informed and democratic society. The increased use of secondary data sources, the modernisation of our statistical processes and systems, and overseeing, coordinating, and assuring the quality of official statistics are all key strategic priorities that must be achieved to deliver the CSO’s strategic goal of *Independent Insight for All*. The development and implementation of CSO Secondary Data Model is a key milestone on the journey to the achievement of these objectives.

Acknowledgment

The author wishes to thank all the staff of the CSO’s Administrative Data Centre and the Data Office for their work in developing the secondary data model and for their continued cooperation and collaboration as we now work together in the implementation phase of this project.