CSO’s new metadata portal – promoting standards, promoting consistency.

Ken Moore

Central Statistics Office, Ireland

Abstract

In order to promote the use of standardised metadata across the Irish Statistical system, CSO Ireland is developing a new, public facing metadata portal. This portal will contain resources to support users, researchers and the general public to make our standardised metadata more visible and accessible to allow all users to better understand the statistics we produce.

The portal is currently under development with a planned release date of Q2 2024. It will contain 2 key elements. Firstly a data standards resource hub where a number of data standards for key concepts will provide information on standard questions and response options, related classifications and code lists, general descriptions of the concepts being documented and API's to facilitate usage across the Irish Statistical system. Secondly it will also contain an external view of our internal metadata portal (Colectica) which allows users to see the key metadata elements for all of our Business and Social statistics inputs in a searchable format.

This paper will describe the journey to date in the development of the portal, highlight the content, outline the rationale for developing the portal and discuss potential next steps and future plans.

Keywords: metadata, standardisation, accessibility, harmonization

1. Introduction & Context

One of the key strategic goals of CSO Ireland is to provide a suite of data services to all producers across the national statistical system as part of its delivery on data stewardship. Embedded in this stewardship role is highlighting the value of data to policy makers and providing the system with support and access to a suite of data services to ensure that data is acknowledged and treated as a strategic asset. One of the core principles behind the provision of data services is that data must be supported by high quality metadata that is accessible and usable. This allows producers, users, policy makers and the general public to fully understand the potential of the statistical information being produced across the system and to move towards a more standardized approach to how data is captured and disseminated.

As CSO continues to promote data stewardship and the associated data services that are now available, there has been a welcome demand from producers for these services. The range of data services being offered is varied but the core message to the system is that the CSO has the data, the data processing, engineering and analytical skills and the legal basis to deliver data services. Included in the list of services being offered are some related to metadata which include:

1. Provide guidance on the correct use of standards and classifications.
2. Provide guidance on the use of standardised metadata including questions, response options and codes to get more usable data.

The broad engagement with the statistical system has resulted in a greater awareness by government departments of the value and potential in their data holdings and an acknowledgement that they must move away from the current siloed approach to a more harmonised solution using common classifications and standardised metadata. This has resulted in the demand for metadata services to be more transparent and useable. In order to meet this demand, the CSO is showing its commitment to making its metadata more accessible to users through the development of a new metadata portal.

2. Contents of Portal

There are two core elements to the CSO Metadata portal, the first being an external view of CSO's internal metadata system generated from our Colectica software, the second element being an under-development data standards content containing a suite of standards for core concepts under CSO's promotion of data stewardship services to the national statistical system.

2.1 External view of our Colectica content

The CSO has been using the Data Documentation Initiative (DDI) open data standard to document, describe, view and manage metadata for several years using software called Colectica. We currently use the tool for a variety of functions including for questionnaire design, quality reporting, the creation of codebooks for researchers and for documenting the concepts and variables used in the majority of our Social, Business and Environmental Statistics. The internal, working version of Colectica currently contains over 300 instruments (surveys), 14,000 questions and 4,500 variables. In order improve the re-use of this metadata and provide greater transparency to our key users and researchers, it was decided to create an external view of this metadata in the new CSO Metadata portal.

The Colectica content on the portal enables the documentation of the full data lifecycle adhering to the international open standard known as Data Documentation Initiative (DDI). It is used for "describing", "documenting" and "managing" both data and metadata across the entire data lifecycle. The content of all metadata items can be downloaded in multiple formats including PDF, XML and DDI.

The metadata contained in the Colectica portal is not finite or final as new data will continue to be added and expanded over time. In addition, the content of the external Colectica content will only contain final versions of the metadata packages generated internally during the statistical lifecycle.

The core uses of the Colectica content on the metadata portal include:

- It is a central question bank to enable the reuse of standard questions.
- It enables us to view all metadata items including for surveys, questionnaires and related documentation.

- Provides a platform to review the data across the GSBPM.
- Provides functionality to standardise content and improve metadata across all surveys.
- Enables the production of survey forms, quality reports and codebooks and other survey documentation.
- Enables version control and traceability of item changes over time.
- Provides a time series for all survey instances allowing users to see changes to survey content over time (e.g., The metadata content for the Labour Force Survey is updated on a quarterly basis).

The core DDI elements included in Colectica to describe the metadata for the core elements are detailed in Figure 1 below:

Figure 1: DDI elements and Metadata Definitions

Project	A group of related studies and resources
Series	A group of repeated studies
Study	A data-centric study
Instrument	A questionnaire/Survey
Metadata Package	A collection of reusable metadata definitions
Data Collection	Information about the data collection process
Question	A question that is asked to a survey respondent
Question Set	A set of related questions
Sequence	A reusable sequence of question and control constructs
Concept	Describe the basic ideas being explored by a study
Concept Set	A set of related concepts
Category	Generic term for items at any level within a classification
Category Set	A set of related categories
Code	Associates a numeric or alphanumeric value with a category
Code Set	A set of related codes
Universe	Describes the population being studied
Universe Set	A set of related universes
Organisation	An institution, company or other organisation
Organisation Set	A set of related organisations

2.2 New Data Standards content

As highlighted above the current silo-based situation lends to Public Sector Bodies measuring and defining concepts inconsistently, making it difficult to gain a coherent understanding of the data across the Irish Statistical System, thus limiting potential comparability. In order to address this challenge, the provision and successful implementation of common data standards across the Irish Statistical System will allow data to be structured to maximise comparability, leading to more reliable outcomes.

Each data standard will contain the following core metadata elements in a standardised format:

- General description of concept
- Standard question to be used to capture concept and associated response options and codes.
- Standard reference classification for concept
- Standard definitions on concepts/response options where available
- Details of where data standard is currently in use
- Links to related concepts
- Where available considering providing access to existing look up files to assist with coding in other administrative systems (e.g., look up file sitting behind a NACE/occupational coder)
- Summary technical specification containing concept name, standard label, version number,
- Release date, Standard owner, Contact details and review date
- Also looking at making an API available to allow reuse of standard in national administrative systems.

So far data standards have been developed and will shortly be released for the following concepts:

- Marital status,
- Full-time & part-time employment status,
- Household relationship status,
- Main activity status,

- Religion,
- Country,
- County.

Work is continuing on developing further data standards for the following concepts:

- Local Authority Regions (under development)
- NUTS Regions (under development)
- Sex (under development)
- Gender (under development)
- Sexual Orientation (under development)

Our decision to start with these statistical concepts, which are primarily Social Statistics based, is tied into the soon to be published National Equality Data Strategy which the CSO has developed. As part of this whole of government strategy the CSO has included a number of deliverables related to the development and implementation of core data standards and the launch of a new metadata portal.

3. Progress to date and Next Steps

While the external Colectica content is live, and content continues to be migrated across, it will be July 2024 until the Data Standards content for the portal is developed and completed. Work continues on the promotion of metadata portal to internal users within CSO and with our external users in government Departments and agencies across statistical system. As part of the promotion and ongoing engagement, we continue to seek feedback from users on the usability and potential improvements to the portal. As new standards and Colectica content is developed we will continue with the ongoing population and maintenance of the system.

So far engagement from data producers across the statistical system has been good. Several Departments are in the planning stages of upgrading their IT systems so timing has been good from that perspective. What we have found is that we have a number of different audiences that we need to communicate with regarding the availability of common metadata standards planned for the portal. These audiences include data experts, system development teams in technology and the policy people framing new policy initiatives. They all understand the motivation behind the need for common standards and high-quality metadata and welcome the dedicated support services being offered and the fact that the standards and metadata associated with our existing outputs are now available in an accessible and useable format.

Based on feedback received to date we are also examining the feasibility of adding a national data catalogue based on key data holding of CSO and other European and National producers in the Irish statistical system to the portal. This is based on engagement with statistical producers and policy makers who have raised the issue that there is a lack of data discoverability of what data exists outside of CSO in the national statistical system. While CSO hosts the statistical outputs for some national producers in an open data format on the CSO website, there is a lack of transparency of what other data assets exist to inform policy makers.