Implementing a Data Inventory Catalogue to Enhance Data Governance and GDPR Compliance in the Central Statistics Office (CSO), Ireland

Miriam O'Reilly

Central Statistics Office, Ireland

Abstract

The Central Statistics Office (CSO), Ireland, are implementing a data inventory application to effectively manage its data holdings, enhance data governance, and streamline business processes. This centralised metadata repository provides a comprehensive overview of data assets, including data inputs, outputs, programmes used and storage locations. The inventory also plays a crucial role in complying with the General Data Protection Regulation (GDPR) by facilitating data retention and ensuring data privacy adherence. Each piece of data in the inventory is classified according to a standardised scheme, aligning data storage and processing practices with GDPR requirements and other legal obligations.

The initial implementation of the data inventory has already yielded significant benefits, including increased transparency into data flows, enhanced data accessibility, and improved overall data quality.

This paper will document the progress made to date, the benefits achieved, and the challenges encountered. It will also highlight our future plans for the application including our aim of further integrating it with the CSO's data retention policy to assist in improved compliance with our GDPR requirements.

In conclusion, the data inventory application serves as an indispensable tool for cross-functional collaboration, data governance, data discoverability and GDPR compliance. It streamlines data usage and accessibility, ultimately contributing to enhanced data quality and overall organisational efficiency.

Keywords: Data inventory, GDPR, Governance, Accessibility

Introduction

Data is our paramount asset. Effective management of this resource is crucial, necessitating a delicate balance between data utility and regulatory compliance. Business continuity must be prioritised, ensuring minimal disruption to everyday workflows. Additionally, fostering public trust through transparency and responsible data practices is essential for maintaining societal buy-in.

This is increasingly challenged by the phenomenon of data sprawl. The ever-growing volume, diversification of sources, and expanding demand for access from researchers, creates significant hurdles to data visibility.

Data Governance supports data quality, security, and usefulness. It adapts as data practices and regulations evolve. One of those regulations that has had the greatest impact is the General Data Protection Regulation (GDPR). GDPR can be defined as "harmonising Data Protection practices across the EU and emphasises transparency, security and accountability by data controllers and processors, while at the same time standardising and strengthening the right of European citizens to privacy of their personal data".[1]

Current Challenges in Data Management

The Central Statistics Office is increasingly characterised by vast volumes of data dispersed across numerous systems. This presents significant challenges:

- Lack of transparency of data holdings.
- Identifying critical data assets and discerning their relative importance can be challenging.
- Unclear ownership structures make it difficult to determine accountability and access rights.
- Duplicate data collection efforts not only waste resources but also raise concerns regarding compliance with regulations like GDPR, particularly regarding data necessity and proportionality. Additionally, the prevailing "store forever" mentality for data retention necessitates re-evaluation.
- Mixed/poor metadata standards.
- Poor documentation to describe data holdings and associated statistical processes.

The Data Inventory is a proposed solution designed to address these challenges and facilitate a more efficient and compliant data management ecosystem.

Enhancing Data Governance with a Data Inventory Catalogue

The Data Inventory Catalogue functions as a centralised repository for metadata pertaining to our data assets. This comprehensive resource offers a holistic overview, encompassing data inputs, outputs, processing programs, and storage locations. Notably, the Inventory plays a crucial role in ensuring compliance with GDPR by documenting data retention periods and supporting adherence to data privacy principles.

[1] https://www.maynoothuniversity.ie/data-protection/introduction-gdpr-general-data-protection-regulation

The key features of the Inventory are:

- Records file locations and details of all the data sources for each iteration/survey period (Variable level information captured for Collect and Disseminate phases of GSBPM). This metadata is captured at a survey instance level (e.g., if a quarterly survey there is a new entry on the inventory for each quarter).
- Details which SAS/R programs/projects are used in statistical production.
- Describes, at a high level, the methodologies used, along with the data classifications and retention periods for all data sources for the entire statistical product life cycle.
- Ensures that any changes to data sources, flows or work responsibilities are recorded in the Data Inventory when prompted by the system when rolling over to a new survey period. This will ensure the Data Inventory is kept up to date and takes account of changes to data sources, programmes, and methodologies over time.
- Additionally, data ownership and accountability are transparent as all data assets are assigned to specific owners.
- Details regarding the data classifications utilised are also documented, along with retention periods for all data sources throughout the entire statistical product life cycle. The CSO's data classification scheme defines categories (and sub-categories) of data and sets out detailed rules on the treatment of data[2].

The Data Inventory Catalogue has proven to be an invaluable asset during our recent code migration from SAS to R. By readily identifying the number and location of SAS programs, we were able to streamline the migration process. Furthermore, the inventory enables us to conduct risk assessments by reviewing reports on personal data usage and access control.

Benefits

A. Promoting Transparency and Streamlining Data Management

The overarching objective of the Inventory is to establish complete transparency regarding our data holdings. This initiative offers several key advantages:

- Mitigating Knowledge Loss: By centralising information, the Inventory reduces the risk of valuable institutional knowledge being lost due to staff mobility.

[2] https://www.cso.ie/en/media/csoie/aboutus-new/legislationgovernancedatapolicies/csodatapolicies/CSO_Data_Management_Policy_Summary_2022.pdf

- Enhancing Collaboration and Reusability: The Inventory fosters clarity regarding how data from one section is utilised within other business areas, facilitating cross-departmental collaboration and data reusability.
- Quantifying Administrative Data Impact: The system enables the measurement of administrative data's contribution to various statistical outputs.
- Ensuring Repeatability: Over time, the Inventory facilitates the management of data assets, promoting the repeatability of statistical products generated by the Office.

B. Alignment with the GSBPM Model

The Data Inventory Catalogue aligns with the four core pillars of the Generic Statistical Business Process Model (GSBPM). It captures comprehensive information regarding data sources employed within the critical phases of production, encompassing data collection, processing, analysis, and dissemination.

C. Facilitating GDPR Compliance

The Data Inventory Catalogue fosters improved data governance by promoting:

- Increased Data Visibility: The centralised repository enhances data visibility, enabling informed decision-making based on a comprehensive understanding of available data assets.
- Facilitated Data Lineage Tracking: The catalogue facilitates data lineage tracking, allowing for more effective data quality management by establishing clear pathways for data throughout its lifecycle.

The Inventory simplifies compliance with GDPR by offering several functionalities:

- Efficient Personal Data Identification: The system streamlines the identification of personal data, enabling the implementation of robust data protection measures.
- Streamlined Subject Access Requests (SARs): The inventory provides a clear overview of data storage locations, expediting responses to subject access requests.
- Data Minimisation Support: The inventory facilitates data minimisation efforts by enabling the identification and elimination of unnecessary data collection practices. For instance, the CSO can leverage the inventory to determine if existing data holdings fulfil their needs before conducting new surveys.

D. Advanced Reporting Capabilities

The Inventory offers a robust reporting functionality, empowering users to generate insights into various data management aspects. Examples include:

- Data Retention Analysis: Reports can be generated to analyse data retention practices within the office, ensuring adherence to regulatory requirements.
- Data Flow Utilisation for Administrative Data: The inventory can be leveraged by the Administrative Data Centre (ADC) in the CSO to determine the utilisation of their data flows. Reports can identify which products utilise specific data flows, revealing the most and least popular offerings within the approximately 140 available options.
- Statistical Output Production: Reports can be generated to provide an overview of the number of statistical products being produced.
- Identification of Largest Personal Data User: The inventory can be queried to identify the area within the office that utilises the most personal data.
- Data Retention Lifecycle Management: Reports can be generated to identify product instances where the data retention period has expired. These reports can further detail the status (active, inactive, or archived) of such instances.

Implementation Considerations and Challenges

The Data Inventory Catalogue has undergone a successful development and population phase. The initial data loading process necessitated manual effort from statisticians, requiring them to document all processes involved in product generation and associated variables. However, the system is designed for streamlined upkeep. Statisticians can leverage an automated "rollover" function after each product release, assuming no changes were made to the underlying code. Where changes are made due to new data sources, changes in methodology etc these can be updated as part of the roll over process".

The Data Inventory is however resource heavy. The initial setup of a product requires one to one assistance and support from staff in the Quality team. Our team also oversees data quality and maintenance through regular product spot checks. This is a significant investment of time.

We continuously strive for system improvements, exemplified by the planned implementation of a notes feature. This feature will allow statisticians to document comments for future users, enhancing knowledge transfer and ongoing data management.

Encouraging User Adoption and Participation

Promoting user adoption and participation in the Data Inventory Catalogue is a critical aspect of ensuring its success. We are actively engaged in strategies to:

- Highlight System Value: Communicate the clear benefits of the Data Inventory Catalogue to users, emphasising its role in enhancing data governance, streamlining processes, and promoting data quality.
- Provide User Training: Offer comprehensive training sessions to equip users with the necessary skills and knowledge to effectively leverage the Data Inventory Catalogue's functionalities.
- Gather User Feedback: Solicit feedback from users through surveys and discussions to identify areas for improvement and ensure the system aligns with their needs.

Conclusion and Next Steps

The Data Inventory Catalogue creates clarity with processes, ownership, and usage. This supports data governance and GDPR compliance. To date we have concentrated on the implementation of the inventory, and data minimization and retention. Areas for future investigation include automating the production of the documentation required for GDPR compliance such as ROPA's and Transparency reports, alleviating the burden on our Data protection office and statisticians to produce them.

By implementing a well-designed data inventory catalogue, we are gaining a clear understanding of our data assets, improving data governance practices, and ensuring compliance with GDPR regulations.

References

https://www.maynoothuniversity.ie/data-protection/introduction-gdpr-general-data-protection-regulation

https://www.cso.ie/en/media/csoie/aboutus-new/legislationgovernancedatapolicies/csodatapolicies/CSO_Data_Management_Policy_Summary_2022.pdf