UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

ModernStats World Workshop

02 October 2024

21-22 October 2024, Geneva Switzerland

The designed governance for a central metadata management system: the Istat experience

Claudia Brunini (*ISTAT*)

brunini@istat.it

Abstract

A well-oriented governance is one of the essential components that has to be defined for achieving and supporting interoperability. The governance includes some crucial aspects such as legal and business policies, the active adoption of standards, and the roles and tasks that should be well identified, recognized and institutionalized. While projecting its central metadata management system (METAstat), Istat has also built a governance able to support the system in the central reference role for metadata.
The article illustrates the designed governance which specifies the essentials roles for a central maintenance of metadata. Because the metadata should be reused along all the phases and by all statistical processes, the standard GSBPM is used to identify all the lifecycle phases of the metadata.
For every phase and sub-process of GSBPM involved in metadata management, the roles are accurately identified. A detailed description of the tasks corresponds every role. The key element of the system is the statistical business process, that permit to connect referential, structural and terminological metadata.
The system will contain not only the metadata from the statistical production activity but also from other cross-cutting activities. This is convenient so that the system could update itself dynamically, efficiently and could provide complete documentation. The standard GAMSO is used in defining these segments.
With the aim to favourite the semantically interoperability, the central metadata management system is also equipped with a terminological component where every term has a proper cycle life and is connected to the structural metadata and referential metadata. The main references for the terminological component are ISO 1087-2019 and ISO 25964-2013. The ISO 1087-2019 supplies instructions on how to correctly manage the terminology. The ISO 25964-2013 helps in documenting the semantic connections.
The standard GSIM, since modelling the structural metadata, facilitates the communication between different processes and different phases within the same process.
The defined governance for METAstat focuses on roles, rules, interactions and processes in order to achieve FAIR data (Findable, Accessible, Interoperable and Reusable).

1 Introduction

A metadata management system (Istat named it METAstat) is defined as a common statistical infrastructure (UN, p. 520) that must be independent from each production process and capable to support them all. A metadata management system evolved as a common infrastructure could have many key benefits. Each element of the infrastructure can support all statistical processes and these can use the resources in order to gain

efficiency and reduce the costs. An additional advantage of a common statistical infrastructure is to promote harmonisation across statistical processes, through the use of common methods and standards.

In the central metadata management system METAstat, the process has a main role. In the first paragraph is explained how, thanks to it, the three major modules of the system are put in connection. The modules are referential metadata, structural metadata and terminology.

The core information of the central metadata management system is the referential, structural and terminological metadata. This information is captured and reorganised having standard GSBPM, GSIM and ISO as a reference. The standard GSBPM is used to identify the lifecycle phases for all the metadata, which allows to manage it in the different phases of the process. The standard GSIM, as structural metadata modelling supports, simplifies the communication between many processes and different phases within the same process. Taking glossaries into account, the ISO 1087-2019 supplies instructions on how to manage the terminology and the ISO 25964-2013 helps in documenting the semantic connections between each term. In the next two structural metadata and terminology paragraphs, it is explained how these elements are well linked.

The central metadata management system will contain standard concepts, variables and classifications used not only from the statistical production activity, but also from the other activities that support the principal one. This is appropriate in an active system, because the statistical information could be linked to other types of information such as administrative, legal, technological etc. The standard GAMSO is used to model the connection between the system and the cross-cutting activities. The last paragraph describes the topic.

2 The Statistical process: the core element of the architecture

Since the metadata should be reused along all the phases and by all statistical processes, the standard GSBPM is used to identify all the life cycle phases of the metadata in the processes. The roles are accurately identified for every GSBPM's phase and sub-process involved in the metadata management. A detailed task description accompanies each specific role. This is done for all types of statistical business process that is the cornerstone of the central system.

All the processes are uniquely identified by a code. The related sub-processes and products might be suitably documented starting from the design phase and in any case before their outcome is disseminated or used. Such documentation should be carried out actively through the IT services provided by the system, or via dynamic acquisition from other DB. All the metadata, referential, structural and terminological should be linked to an edition process.

Looking at metadata, processes are split in managers and users. The first ones generate new metadata and become responsible of it, the second group use already existing metadata.

The system supports the statistical production process in all phases metadata management through specific functionalities. Anyway, for metadata responsible processes that have their own management system, the central one should interact by dynamically acquiring the metadata, ensuring in this way their correct traceability and update.

The process manager is responsible for the metadata generated within the process itself. The responsibility concerns all metadata life cycle phases. Therefore, for example, the responsible of the metadata associated with the dissemination phase (both of microdata and aggregated data) is the process that generates the data to be disseminated and not the department that handles the dissemination phase.

The process manager should be an institutionalized role, i.e. he\she should be appointed by the directorate. The central metadata management system captures the start date of assignment from the administrative system. The process manager can nominate other appointed staff members to operate in the system. The relevant directorate formalizes the nominees.

All metadata activities are performed by the process manager and his team. He is responsible for entering the process into the system and completing all the information. He enters referential, structural and terminological metadata. He also takes charge of the changes and the definition of the states throughout the metadata life cycle.

In administrative source processes, the central metadata management system documents the statistical metadata regarding the outputs. It also provides an accurate description of the transformation processes that go from administrative metadata to statistical ones by establishing the connection between administrative source and output. Instead, the administrative source metadata documentation is under the responsibility of the data collection phase.

Each metadata operation is supervised by the structure that is responsible for the control and validation of the central metadata management system contents. This structure guarantees the availability of all metadata produced for managers and users inside and outside the Agency. It also ensures their correct update, their standardization, harmonization, consistency and integration.

The validation role is allocated to the centralized structure. This task has the goal to secure the standardization and consistency of the central metadata management system contents. It must be carried out in a simple and fast way, supported as much as possible by the application or IT services of the system. In cases of conflict between metadata, a decision-board is composed by the process managers who are involved and the representatives of the centralized metadata structure. The board has the duty to solve conflicts in short times. Metadata that has not yet been validated can only be used by the process that issued it. Only validated metadata can be visible and reusable by all institute processes.

3 The structural metadata

Each structural metadata is loaded into the system by a process, which becomes responsible for the metadata itself. This process is in charge of the management, i.e. its initial drafting and maintenance during all phases of the life cycle.

The process responsibility role on the metadata is the connecting element between the registry of processes and the referential metadata with the structural metadata. Therefore, between the GSBPM standard, used to model the process phases, and the GSIM standard used to model the structural metadata.

The single edition of the production process represents the minimum level of domain and so the main issuer of the structural metadata. Each process edition can have the role of responsible or simple user with respect to the structural metadata.

The processes distinction between managers and users allows to identify two profiles: the edition manager who formulates and issues the structural metadata and the user process manager, who cannot modify the metadata, but can propose changes to the responsible for the process that has generated the metadata.

It may happen that some metadata have more than one generating processes, in this case the responsible process is made up of a group of processes that will manage the metadata in harmony. They can eventually nominate a reference subject among themselves. Other two exceptions are metadata that arise within higher profile entities such as Commissions and Committees (for example the Ateco Committee) and metadata that arise to respond to transversal needs (for example the thematic reports, such as the Annual Report, or the ontologies). Metadata which originates in Committees are managed by the Committee itself, in the person of a specifically appointed thematic manager. If it is not possible to identify a thematic manager, the function of metadata manager is entrusted to the thematic structure where the Committee is based or, in its absence, to the methodological structure that manages the central metadata system. Metadata originating from cross-cutting needs are assigned to the most suitable processes; if no process can be identified, they will be assigned to the methodological structure.

The tasks of the structural metadata manager are: formulate the metadata and provide for any subsequent changes; act as an intermediary for the validate operation (which is under the responsibility of the manager of the structural metadata module of the central metadata management system); act as a contact for any modification proposals that come from the users of the metadata; define the dates relating to the life cycle of the metadata; fill in all the fields connected to the metadata; connect the metadata to the relevant data schemas, if they exist.

4 The terminology and the management of semantic resources

The central metadata management system is structured to support the semantically interoperability. For this reason, it is equipped with a terminological component where every term has a proper life cycle and is connected to the structural metadata, hence to the referential one.

The main references for the terminological component are ISO 1087-2019 and ISO 25964-2013. The ISO 1087-2019 supplies instructions on how to correctly manage the terminology, the ISO 25964-2013 helps in documenting the semantic connections.

The governance of the terminology is very similar to that of structural metadata. Each centralized term, like each centralized structural metadata, has a responsible process that takes charge of its management, i.e. its initial drafting and its maintenance during all phases of the life cycle.

The single edition of the production process represents the minimum level of domain and therefore the main emissary of the terms that forms the terminology collection of official statistics.

Looking at the terms, each process can act as responsible or user. The distinction of the processes between managers and users allows to identify two profiles: the person responsible for the process that formulates and issues the term, who takes charge of its management in every phase, eventually together with staff appointed by him; the person responsible for the process that is a user cannot modify the term but is able to propose changes to the term manager.

It may happen that some terms have more than one generating process, in this case the responsible is made up of a group of processes that will manage the term in harmony. They may nominate a reference subject among themselves. Other two exceptions are when the terms arise within higher profile entities such as Commissions and Committees or when they arise to respond to transversal needs (for example the drafting of thematic reports or ontologies).

The terms that arise in the Committees or similar have as responsible a thematic manager appositely appointed by the committee. If it is not possible to identify a thematic manager, the function of term management is entrusted to the thematic structure where the Committee is based or, if this does not exist, to the methodological structure that manages the central metadata system.

All other terms of cross-use that are not directly generated or defined by the production processes are taken care of by the methodological structure that is in charge of the central metadata management system. These terms have as responsible the terminology module manager. His task is to check the correct formulation and the right use of the term in the different phases of the process.

The tasks of the term manager are: formulating lemma and definition and providing for any subsequent changes; acting as intermediary for the validation step (which is in charge of the terminology module manager of the central metadata system); acting as a contact for each change proposal coming from term users; acting as a link for all semantic variant proposals, both about lemma and definition; defining the dates relating to the life cycle of the term; filling in all the system fields related to the term.

Instead, all the semantic relation between the terms, are defined by the terminology module manager.

5 The cross-cutting activities

A central metadata management system has the core mission to manage metadata from the statistical production. This kind of activity is one of the three overarching processes that in the GSBPM model has the goal to support the statistical production (the other two are quality management and data management). The activities that also support the statistical production, but are carried out at the level of the organisation, are modelled by the standard GAMSO. These activities are called cross-cutting activities.

A central metadata management system, in order to properly manage the metadata from the statistical production activity, needs to capture and contain data and metadata from the cross-cutting activities. This establishes the connection between the two standards GSBPM and GAMSO.

Below are given some relevant examples of connection between the core information of a central metadata management system and the data and metadata coming from the cross-cutting activities.

In the documentation of the process, all the information about its manager and the appointed staff that work with him must be acquired from the administrative management system that contains all data regarding the personnels working in the organization. This is valid also for the capture of data on organization itself, for example the organizational chart.

One more, when documenting the process, it is often necessary to report information on legislative sources that regulates the procedures, the definitions and the methods. The laws should be accurately described in a normative database held by the legislative sector. This information has also the function of supporting the documentation of the terminology, which often has a reference regulation.

A further example is connected to the terminological collection, that has the role to document the sectoral language of the official statistics. The collection includes not only the terms of the statistical production, that are the core information, but also the terms of the overarching processes (like quality) and of cross-cutting activities, such as IT or methods.

6 Conclusions

In a central metadata management system, the design of the governance is a key element. A well-oriented governance is a crucial component that has to be defined for achieving and supporting interoperability between the standards and, on a future note, with other components connected to metadata. The governance includes many key aspects such as legal and business policies, the active adoption of standards and roles with tasks which should be well identified, recognized and institutionalized.

References

1. COSMOS, The smart metadata manifesto, COSMOS (cosmos-conference.org), 2024.
2. ISO 25964-2:2013, Information and documentation Thesauri and interoperability with other vocabularies. Part 2: Interoperability with other vocabularies, 2013.
3. ISO 1087:2019 Terminology work and terminology science. Vocabulary, 2019.
4. UNECE, Generic Activity Model for Statistical Organisations (GAMSO), Version 1.2, January 2019.
5. UNECE, Generic Statistical Information Model (GSIM), Version 1.2, October 2020.
6. UNECE, Generic Statistical Business Process Model (GSBPM), Version 5.1, January 2019.
7. UNECE, Data Governance Framework for Statistical Interoperability (DAFI), Version 15 November 2023.
8. UN, Handbook on Management and Organization of National Statistical Systems, Version 2022/A.