The terminology module in a centralised metadata system: a crucial contribute to quality

Claudia Brunini[1]

[1]*National Institute of Statistic (Istat), Italy*

Abstract

An effective centralized metadata system is based on three key components: the referential metadata module, the structural metadata module and the terminology module. The referential metadata module describes statistical processes and provides all the elements for quality assessment; the structural metadata module describes the data involved in the processes assigning them the role and the relationships between them; the terminology module documents the semantic aspects. This work focuses on this last module. The terminology resource has the role of detailing the semantics of the metadata, allowing in this way its identification. The availability of accurate and strict terminological resources, following international models, guarantees an improvement in the quality of the statistical information. It also contributes to a clear identification of objects (units, populations, variables, etc.) and allows to overcome the constant problem of semantic overlap.

The outcome is eliminating the interpretative ambiguities of meanings. In particular, the terminological component represents a fundamental asset for preventing and managing the risk of specification error, which is relevant in the initial phase of the statistical production process. In this phase the concepts and dimensions, previously identified, are made operational in terms of detectable characteristics (variable), statistical units, target population and classifications as well as territorial and temporal dimensions. These aspects, if not carried out correctly, can cause severe consequences on some components of quality such as the relevance of the data and accuracy, in the end affecting the distortion of the estimates produced. These risks are more relevant in processes that use administrative sources, where it is more difficult to keep under control the relationships between the statistical definition adopted and the definitions of the original sources of data.

The contribution illustrates in details how the functionality of the terminological component can effectively support the phases of the statistical processes where the specification error is particularly risky. How it can facilitate the researcher's work and promote the construction of metadata internally coherent, harmonized with national and international sources and complete from the point of view of quality indicators.

Keywords: Metadata management, Glossary, Terminology, Semantic interoperability, Modernization

1. Introduction

METAstat, the new central system for metadata management currently under development in Istat, is designed as an open system (according to international standards) that offers functionalities useful to those who need to create and manage metadata. Taking into account that metadata are concepts that pervade the entire Institute and that there are already specific systems that manage metadata autonomously, the METAstat system has the task to facilitate and streamline the data production processes by providing exchange functions of information between different data systems. The system manages three types of metadata:

- Referential metadata, which describes the statistical processes and provides all the elements for quality assessment;
- Terminology, which describes the semantics of the metadata and allows its identification;
- Structural metadata, which describes the contents of the data involved in the processes and assigns the role of the data and the metadata as well as the relationships between them.

Within the system, the terminology module provides the functionalities for managing semantic resources. On the quality side, a single terminology collection has several advantages (UNECE, 2000). First of all it encourages the use of clear and unambiguous language, improving the information disseminated; it can also help the researcher to create well-defined concepts in the design phase. The article wants to illustrate how the functionalities of the module can help in preventing and managing the specification error.

2. The theoretical framework

The terminological component could represent a fundamental asset to prevent and manage the risk of specification error, which is not only relevant in the initial phase of the statistical production process, when the process is designed, but also during the subsequent phases of the analysis and dissemination, when the researcher may form new variables, indexes or aggregates. During the design phase the constructs, that are abstract ideas, are transformed into measurable concepts (variables), clearly identifiable dimensions or units (UNECE, 2019). The critical task before measuring is to design variables that reflect perfectly the constructs they should measure (i.e. income may include or not legal earnings, or marriages that may or may not include legally separated couples). Measurement is more concrete than constructs, because it consists in gathering information about construct itself. In traditional surveys measurements are questions posed to a respondent using words. In administrative sources

they have to be determined from pre-existing data, that have definitions usually focused on administrative aims and not always properly formulated. The nature of measurement, both in surveys or process from administrative data, determines that the response or transformation adopted cannot reflect exactly the variable. These aspects can cause severe consequences on components of quality such as the relevance of the data and accuracy, in the end affecting the distortion of the estimates produced (Istat, 2018).

Also the description of the units that form the target population involve a semantic aspect. The target population is the set of persons which has to be studied, the frame population is the set of target population members that has a chance to be selected into the survey sample. It is desirable that the frame population matches perfectly the target population. The description of the units that form the target population can in some cases be one of the determinants that origin the coverage error.

The specification error involves the construct validity, that is the degree to which the measures reflect the construct. In statistical terms validity is the correlation of the measurements, Y_i and the true value μ_i, measured over all possible trials and persons. When Y and μ covary, moving up and down in tandem, the measurements has high construct validity. A valid measure of an underlying construct is one that is perfectly correlated to the construct (Groves, R.M. et al, 2004). Regarding the coverage error, it produces a bias that can be described as a function of the proportion of the target population not covered by the frame one, and the difference between the covered and the noncovered population (Lepkowski J., 2005).

The management of the terminology resources can help to control and eventually reduce all the errors linked to a bad specification of constructs.

3. The governance and functionalities on the terms

The terminological component of a central metadata system can offer functionalities that help to manage the semantic contents in an organized dbase (Brunini C., 2021). The controlled terminology collection documenting terms and the semantic relationships between them constitutes one of the three key components of the new centralised metadata system (METAstat), currently being developed at Istat. The interconnection between the terminology collection and the other two components - the structural metadata system and the statistical process management system (referential metadata) - takes place according to three core governance principles:

1. each centralised term, like each centralised structural metadata, has a person in charge of its management, i.e. its initial drafting and its maintenance during all phases of the life cycle;
2. the production process represents the minimum domain level and therefore the main emissary of terms that form the terminology collection of the official statistics;
3. with respect to a term, each process may have the role of manager (creator of the term) or simple user. This distinction allows two profiles to be identified: the manager for the process in which the term is formed (he takes over its management in every aspect and at every phase and appoints any contact persons authorised to operate in METAstat); the manager of the process that is simply user (he cannot directly modify the term, but can propose modifications by interfacing with the person responsible for the process that generates the term).

The duties of a term manager are: formulating lemma and definition and making any subsequent changes; being the intermediary for the validation, that can assure harmonization; being the contact person for any proposed changes coming from users of the term; being the contact person for all semantic variant proposals, both on the lemmas and on the definitions; defining the term lifecycle dates.

When a new term is entered into the system and validated, it will be visible and usable by all production processes and at all phases, making semantic interoperability possible.

Harmonization is the activity in charge of the central structure that has the responsibility to control both the metadata and the management of the metadata central system. The harmonization is made after metadata are entered into the central system by the term manager, and it has the goal to integrate new metadata with the ones that are already used by the institute. Until the new metadata have not been checked, with the aim to verify the harmonization, they cannot be made available. They will be available to read and used by all other processes only after being harmonized. Harmonization is a phase crucial for the preparation of standard metadata. It is best carried out by the central structure that deals with metadata management, rather than by the thematic sectors, which have a vision mostly oriented towards specific contents. It is essential for the central metadata management system to offers all the functionality needed to achieve harmonization quickly and easily, therefore though important as an additional phase, harmonization must not slow down the production process.

A centralized metadata system should also offer adequate consultation and reuse functionalities. Every process manager should be able to navigate all the metadata connected

to the terms, so to verify if the term he needs yet exists. In this case he can adopt it without duplicate information. If the term does not exist, he can create a new one. The new term will be equipped with all the information necessary to allow its coherent reuse by the other processes.

One of the goals of a controlled terminology collection is also to highlight the semantic relationships between terms (Brunini C. 2023). If a certain concept is marked by multiple linguistic forms, a controlled collection denotes the form with the role of preferred term and associates with it the numerous other forms used to indicate it. This allows information about a certain concept to converge at a single point, namely the preferred term, through the links associated with it. Synonyms and equivalent semantic forms are adequately documented so that the user can attach his/her preferred term or alternatively one of the equivalents to the concept. A controlled terminology collection also documents polysemies, that are cases in which several concepts are marked by terms with the same lemma but different definition. This occurs with reference to different sub-domains and it is important that users have adequate documentation for each cases.

According to international standards (ANSI/NISO Z39-19:2005, 2010), three types of semantic relationships need to be made explicit: equivalency; hierarchy; association. The equivalence relationship allows the management of synonymy, quasi-synonymy and linguistic variants. The hierarchical relationships is based on degrees or levels of superordination/subordination, where the superordinate term represents a class or a whole, and subordinate terms refer to its elements, parts or individuals. The associative relationship covers associations between terms that are neither equivalent nor hierarchical, yet the terms are semantically or conceptually associated to such an extent that the link between them should be made explicit in a controlled terminology collection, on the grounds that it may suggest additional terms for use in indexing or retrieval (Folino A., 2013).

When a new term is input, it is linked to all the information related to the process that inserted it (referential metadata). It is also connected to the structural metadata set, since many terms are the linguistic identifiers of structural metadata. All these information can be inserted by the process manager or by a person by him or her appointed. Semantic relationships, on the other hand, can only be mapped at a later stage. The description of the semantic relations must be carried out by a subject of the central structure who has governance over the centralized metadata. The reason for this is that semantic connections have a transversal character and involve many themes simultaneously. The documentation of semantic relations has a dual thematic and technological aspect. From a thematic point of view, the documentation is

desirable to be created by sector specialists who know well the links between the entities underlying the linguistic descriptor. From a technological point of view, the use of ontologies is strategic. The documentation of semantic relationships allows to resolve many cases of terminological overlap and allows for a much more conscious reuse of resources. Consultation of semantic relations can certainly help the processes in choosing the best term to adopt.

With reference to user processes, the link is documented through the reuse function. Reuse of a terminology resource can occur during numerous stages of the statistical process. Undoubtedly, one of notable importance is dissemination, during which thematic glossaries are formed. A central terminology collection plays an important role in this context, because it allows the construction of coherent and integrated thematic glossaries. With support to the dissemination phase, not only reuse functionalities, but also organizational functionalities were imagined. Dissemination products manager can use the system not only to create their own glossaries, but also to manage them over time.

4. Conclusions and next developments

Centralized management of terminological resources allows specification error to be kept under control at different stages of the statistical process.

The availability of clear definitions is useful in preparing the conceptual framework of the survey, the terminological module allows to search for existing semantic resources, providing documentation of the statistical objects to which they are associated. It also allows to analyse which processes use them and with which specific definition. For a better analysis, additional information is also made available to researchers, such as the process responsible for its maintenance and updating, the dates of validation and any end of validity, the list of user processes, regulatory references and any other changes. In this way, the researcher can implement a conscious selection (or possibly a new formulation) of the semantic resources, proceeding with a correct attribution to its statistical objects.

References

Brunini C. et al. (2021). Verso un glossario unico per la statistica ufficiale italiana, in “Terminologie e vocabolari, lessici specialistici e thesauri, glossari e dizionari” a cura di C. Grimaldi, M. T. Zanola, Firenze University Press.

Brunini C., Recchini E. (2023). From a glossary to a controlled vocabulary: a preliminary Istat experience, Conference on New Techniques and Technologies for official Statistics, NTTS, Brussels, 7–9 March 2023.

Folino A. (2013). Tassonomie e thesauri, pp. 387-444. In Documenti Digitali, a cura di R. Guarasci, A. Folino, Iter, Milano.

Groves, R.M. et al. (2004). Survey Errors and Survey Costs, 50. John Wiley & Sons, INC., Hoboken, New Jersey. ISBN 0-471-48348-6.

Istat. (2018). Linee guida per la qualità delle statistiche del Sistema Statistico Nazionale, 16-20. https://www.istat.it/it/files/2018/08/Linee-Guida-2.5-agosto-2018.pdf

ANSI/NISO Z39-19:2005. (2010). (R2010), Guidelines for the construction, format, and management of monolingual controlled vocabularies.

ISO 25964-2:2013 (2013). Information and documentation — Thesauri and interoperability with other vocabularies — Part 2: Interoperability with other vocabularies.

ISO 1087:2019 (2019). Terminology work and terminology science — Vocabulary.

Lepkowski J. (2005). Household Sample Surveys in Developing and Transition Countries: Chapter VIII *Non-observation error in household surveys in developing countries*. 257. ISBN 92-1-161481-3. https://unstats.un.org/unsd/hhsurveys/pdf/household_surveys.pdf

UNECE. (2019). Generic Statistical Business Process Model, GSBPM (Version 5.1, January 2019), 11-15. https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1.

UNECE. (2000). Terminology on Statistical metadata, Conference of European Statisticians, Statistical standards and studies, no. 53, Geneva.