Metadata management systems, living spirals or flat line?

A. Portugal [1], S. Nunes[1], J. Poças[1]

antonio.portugal@ine.pt, susana.nunes@ine.pt, joao.pocas@ine.pt

[1]. Statistics Portugal, Portugal

Abstract

Metadata and quality management systems can support a consistent process through the identification of needs, design and build of statistical operations, proceed to the evaluation phase and restart of a new loop.

At Statistics Portugal Metadata Management System (SMI) the objects standardization, namely concepts, classifications, variables or collection instruments leads to a more optimized flow to launch a statistical operation in the field and disseminate its results. The multitude of analysis and results that can be produced is stimulus to a more challenging appropriation of data.

One of the challenges of today's information is the ephemeral meaning of data and the need of versioning. The different objects that make part of a metadata system rearranged with different versioning can lead to a multitude of information that multiplies the initial meaning of data. At SMI objects standardization is implemented to achieve a better understanding among all intervenients during the design and building phases. Reuse of objects is crucial to explore the potential of metadata management system. Good search tools minimize the creation of new objects.

Integration with other systems supplies information to the building phase of applications and also reports to statistical operation's analysis and evaluation of processes.

Information in metadata management systems can't be carved in stone. The user friendliness and users recognition as an added value to the process are essential to keep the system alive. Challenges are in place, new standards and procedures are continuously arising demanding an effective answer from the metadata management system.

Key words: Metadata, Management system, Quality, Standards.

1. Metadata and Statistics Operations

Metadata and quality management systems can support a consistent process trough the identification of needs, design and build of statistical operations, proceed to the evaluation phase and restart of a new loop.

At Statistics Portugal (SP) Metadata Management System (SMI) the objects standardization, namely concepts, classifications, variables or collection instruments leads to a more optimized flow to launch a statistical operation in the field and disseminate its results. The multitude of analysis and results that can be produced is stimulus to a more challenging appropriation of data.

2. Metadata

If data is the core communication of a piece of content, metadata is information about the content that provides structure, context and meaning. By structure we can refer a report or an xml structure, by context is identified who, where, when, and meaning is given by tags, legend and notes.

10374822	9869783	246353	258686
4923666	4681840	120758	121068
5451156	5187943	125595	137618
449799	426971	12533	10295
230243	218612	6395	5236
219556	208359	6138	5059
499768	472734	13876	13158
255773	241906	7060	6807
243995	230828	6816	6351
540674	509777	15072	15825
277470	261718	7715	8037
263204	248059	7357	7788
552373	519854	16406	16113
282068	265270	8376	8422
270305	254584	8030	7691
553108	518899	17626	16583
279030	261647	8952	8431
274078	257252	8674	8152
566505	532342	17898	16265

Fig. 1 *Data without metadata*

Identified types of metadata can be:

- Structural Metadata - Models the content types and attributes.
- Administrative Metadata - Indicates how, when and by whom the content was created, defines how it can and will be used, its status and who can access it.
- Descriptive Metadata - Describes the subject matter of the content.

Statistics metadata can be of any of previous identified types. The glue is the activity where they are generated.

Grupo etário	Sexo	População residente (N.º) por Local de residência (NUTS - 2013), Sexo e Grupo etário; Anual			
		Período de referência dos dados			
		2014			
		Local de residência (NUTS - 2013) (1)			
		Portugal	Continente	Região Autónoma dos Açores	Região Autónoma da Madeira
		N.º	N.º	N.º	N.º
Total	HM	10374822	9869783	246353	258686
	H	4923666	4681840	120758	121068
	M	5451156	5187943	125595	137618
0 - 4 anos	HM	449799	426971	12533	10295
	H	230243	218612	6395	5236
	M	219556	208359	6138	5059
5 - 9 anos	HM	499768	472734	13876	13158
	H	255773	241906	7060	6807
	M	243995	230828	6816	6351
10 - 14 anos	HM	540674	509777	15072	15825
	H	277470	261718	7715	8037
	M	263204	248059	7357	7788
15 - 19 anos	HM	552373	519854	16406	16113
	H	282068	265270	8376	8422
	M	270305	254584	8030	7691
20 - 24 anos	HM	553108	518899	17626	16583
	H	279030	261647	8952	8431
	M	274078	257252	8674	8152
25 - 29 anos	HM	566505	532342	17898	16265
	H	282792	265150	9268	8374
	M	283713	267192	8630	7891

Fig. 2 *Metadata, all the difference with significance*

Metadata in the context of statistical activity supports the processes described in GSBPM model as Metadata Management. GSBPM is implemented at Statistics Portugal through the Statistical Productive Process Model Handbook (MPPE 2010). This handbook will be entering in a process of revision. The process defined includes the mandatory and continuous use of SMI system.

3. Metadata management system

The Metadata Management System (SMI) consists of a repository of concepts, classifications, variables, data collection instruments and methodological documentation relating to statistical activities, mainly statistics operations, carried out in the National Statistical System (SEN), and disseminated data on the Portal of Official Statistics (www.ine.pt). The several components of this system are integrated, so its management is subject to strict rules of harmonization and integration.

Each subsystem has started to be designed and developed in 2003 until 2008 in observance of international norms. The integration in the subsystems had different levels, but were not complete. It was revamped and rebuild with the objective of a bigger integration and better performance in 2012. All the information was migrated to the new SMI system.

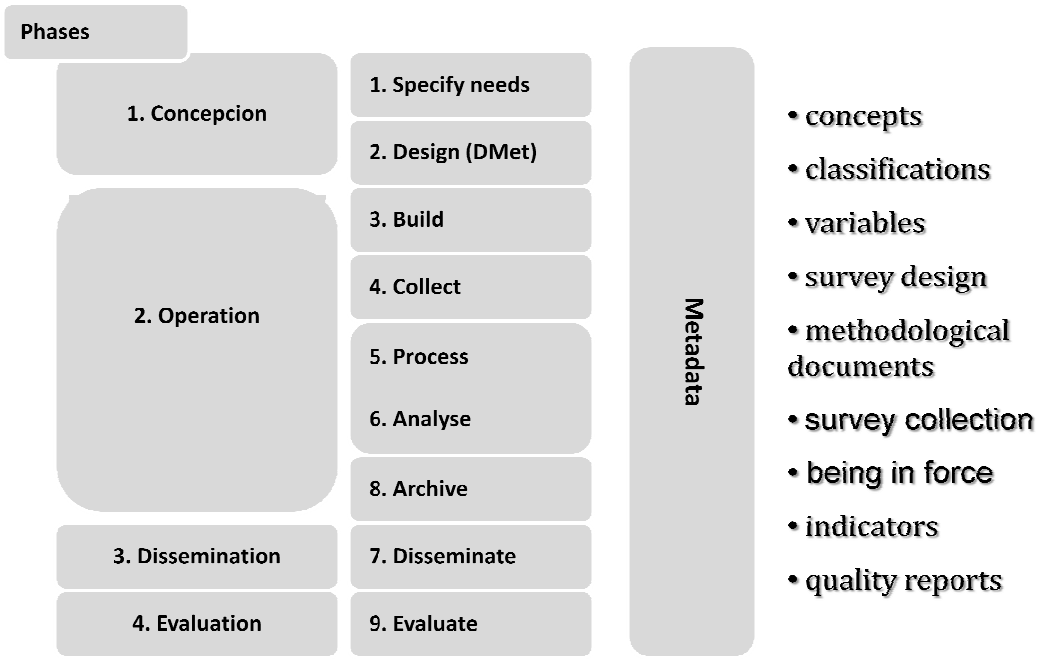

Fig. 3 *Metadata and the Statistical Production Handbook Procedures (≈ GSBPM model)*

The main goals of this repository are, a) to support the survey design process and b) support data dissemination and documenting indicators spreading through the dissemination database. It is intended that this system constitutes an instrument of coordination and harmonization within the National Statistical System (NSS).

3.1 Concepts

Concepts module contains terms and definitions used in the statistical activities conducted in the scope of the NSS.

This database has concepts that are in force but also those that are no longer valid, approved by High Statistical Council (CSE) or only in use but not yet approved.

It is intended to inform the users of statistical information about these concepts and promote their use aiming to obtain consistent and comparable statistical data.

The use of approved concepts in statistical operations contributes to ensure consistency, reliability and comparability at national and international level.

Statistical concepts are organized in conceptual systems that help the search in a theme.

The concept definition needs to be perceptible to the specialized user but also to the general public. A concept is a knowledge unit with origin on a unique combination of characteristics (ISO 1087-1:2000). Concepts are linked to questions in surveys, to indicators in data dissemination and influences data analysis.

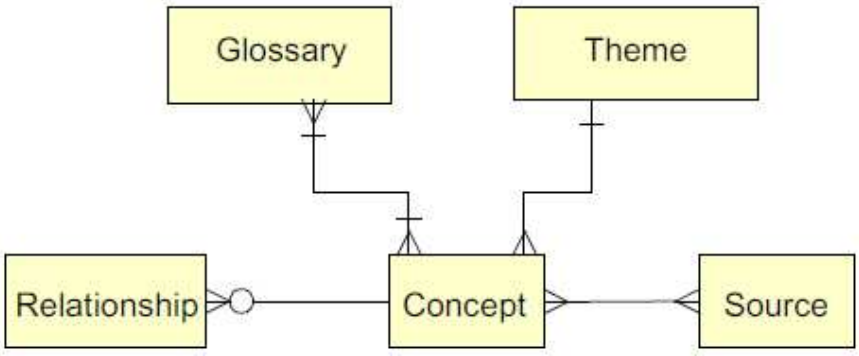

Fig. 4 *Concepts structure at SMI*

Regular updates in concepts and links with other metadata objects needs to be ensured.

Revisions or new versions have a great impact in variables and indicators.

3.2 Methodological documentation

Methodological Documentation module is the central subsystem, as it is related with all others subsystems. It contains all documents relating to the statistical activities in force in the NSS - Statistics Portugal and entities with delegated powers - and other entities intervening in national statistical production, as well as the data available to Eurostat and other organizations.

Fig. 5 *Methodological documentation search filter*

A NSS standard was produced in 2005, and a new version in 2011, for methodological documents that are enforced. It is composed by 12 items and 46 sub items. For each item filling instructions were made available to the users.

One of the mandatory activities of a statistical operation is the construction of the Methodological Documentation (DMET) at SMI, it corresponds to the GSBPM phase 2. The list of observation variables and classification versions are mandatory pieces to include in DMET. The information that is included in the chapter Data Collection Instruments can be used as input in the analysis and build phases of software application.

A survey is a statistical activity sustained on a predefined statistical methodology. Comprehends a complete statistical production cycle: data collection, data processing, analysis and data dissemination in respect to a certain statistical population or universe. Three types of statistical activities are differentiated: sample survey, census and statistical study. Administrative sources and derived statistics are included in the system. The concept "statistical operation" is registered in the system. Survey documentation is disseminated and classified by theme.

New outputs can be produced by grouping selected items of DMET. New templates can be defined with combined information from DMET and new items. This new templates can generate in the future quality reports like SIM 2.0.

3.3 Classifications

Classifications module contains national, European and International statistical classifications used in the statistical activity.

A classification is a structured list of discrete, exhaustive and mutually exclusive categories, identified by codes and names, each of which describes a possible value of the classification variable.

Classifications are organized in families and have versions. Each version of a classification is a set of categories that are valid for a given period of time.

Fig. 6 *Version search filters*

There are formal classifications but also non formal or decoding code lists to be linked to observation or dissemination variables and domain values. The non formal classifications can also be cumulative versions with valid codes along time used for time series dissemination, also known as floating versions.

The information structure is hierarchical, with levels, from higher to lower: family, classification, version, level and category. Version is an instantiation of a classification valid in a due date interval. Level is what make possible the aggregation. All categories of some level have code structure. A category is the value for one level of one version in a classification.

Two specific types of versions, or tables, the equivalence tables to link two versions of the same classification, p ex. different years, and the correspondence tables to link to different classifications, p.ex CPA and NACE.

The classification version is the entity that brings together most specific information, which is why the module is directly accessed by the list of existing versions. Each version of a classification is identified by a 6-digit code. All versions have relevant technical information

on the responsible entity, which owns the copyright on the classification, as well as information on levels that compose it, indexes, associated documentation, and correspondence with other versions and relationships with other modules of the system. All this information is accessible by consulting the detail on the version.

The access to a category item is done by choosing in the respective category list. Hierarchical and tabular views of the version are available. In the category list is possible to export (download) all or some categories of one version.

Exports are also accessible by direct access to families, versions, groups, correspondences and indexes.

3.4 Data collection instruments

Data Collection Instruments module contains all the data collection instruments that are registered in Statistics Portugal, questionnaires and when data is transmitted electronically the file structure is published and used in statistical activities performed in the NSS. The system contains information dating back to 1961 until today.

The objectives of this component are a) to provide information to characterize the data collection instruments, including respective image files, and b) to publicize information in the methodological documents related to the data collection instrument. The concepts of “Data collection instrument” and “Questionnaire” are registered in the system to search. Many types of data collection are registered, paper, electronic questionnaire, direct interview and telephonic interview. A registry number is produced for each data collection instrument. For each question a concept, not mandatory, a domain value and a variable need to identified. Each metadata object can be reused or build and proposed to adopt.

3.5 Variables

The main goals of this repository are, to support survey design and support data dissemination, documenting indicators exposed through the dissemination database.

Variables module contains a repository of all the variables observed and disseminated by NSS. This module includes observation variables and indicators provided by the statistical activities performed in the scope of the NSS, and the administrative sources used for statistical purposes.

A variable can be defined as a characteristic of a measure unit or population, can have different values group and a numeric measure or classification categories can be given.

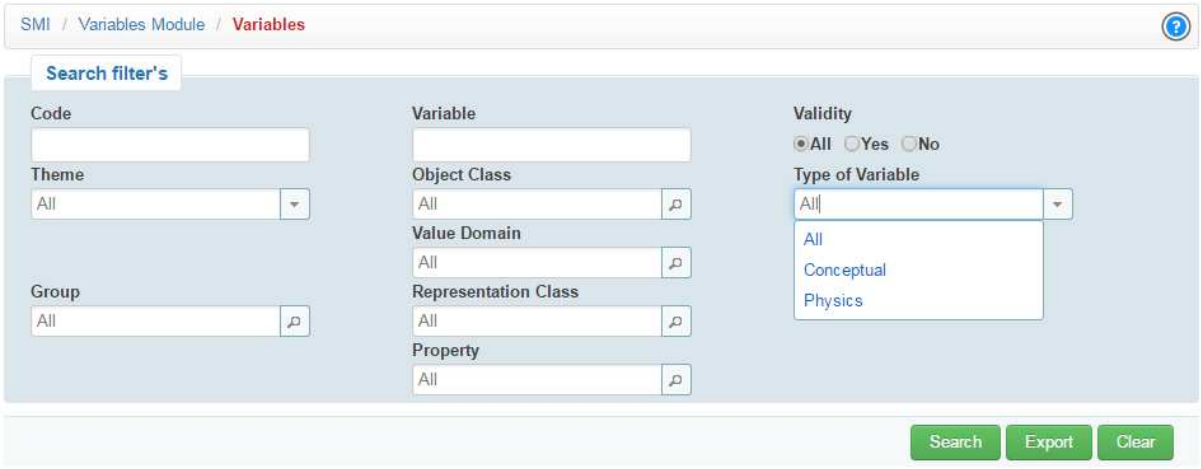

Fig. 7 *Variables search filters*

Variables, in this system, are organized in two levels: the first level presented conceptual variable and corresponds to the definition of the variable; each conceptual variable groups a set of variables (second level), depending on the way it is used in different contexts: statistical operations, questionnaires, statistics portal, etc.

Variables are linked to concepts and statistical sources and each one of the latest's are linked to a classification level.

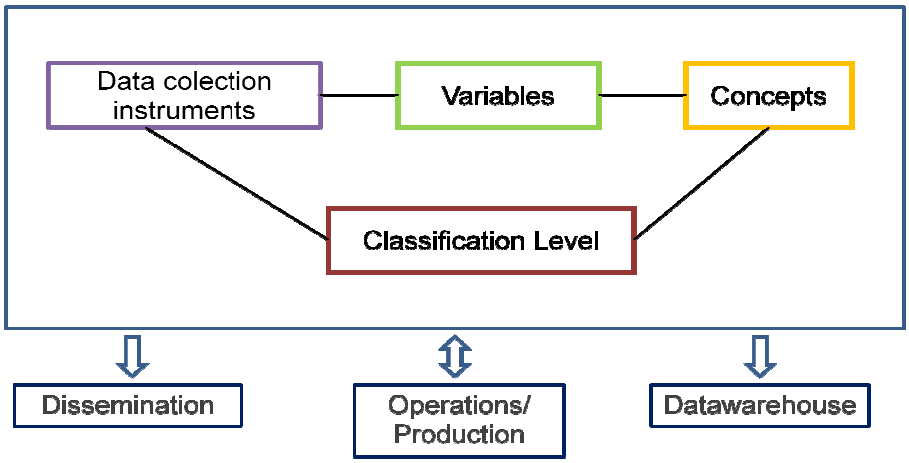

Fig. 8 *SMI subsystems relations*

Statistical indicators are composed by a set of variables, a variable measure that provides the data to search and analysis dimensions that allow data to break down by the criteria significant in each case.

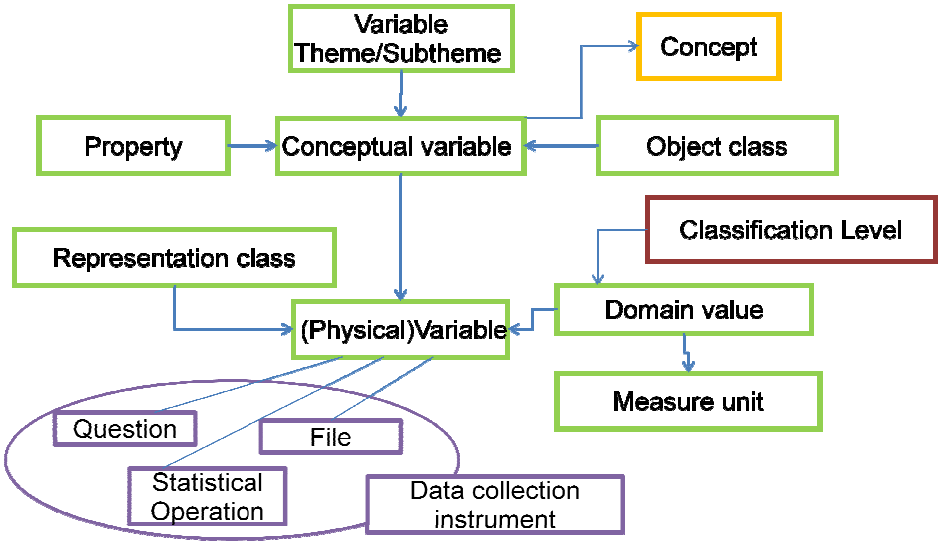

Fig. 9 *Variable metadata objects*

In general, the same variable can't be used for collection and for dissemination. Is collects a propriety of one kind of individual, and disseminate a characteristic of a population. This leads to increase the number of metadata objects that are difficult to use in a distinct way.
The data collection channel/mode can be a perturbing factor in the use of metadata objects as also the dissemination channel. Different versions of very similar metadata objects are needed, and poses difficulties in which should be used in each case.

4. Statistical Indicators

An indicator is the result of the combination of a measure variable, a time dimension, and geographical dimension (mandatory), normally other dimensions are concurring to the result.

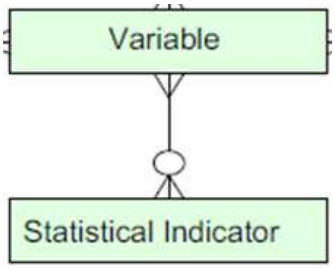

Fig. 10 *Indicators are variables*

All this metadata objects are variables ruled by the management rules previously defined.
Previously to data loading all metadata must be defined at SMI.

5. Timeline

One of the challenges of today's information is the ephemeral meaning of data and the need of versioning. The different objects that make part of a metadata system rearranged with different versioning can lead to a multitude of information that multiplies the initial meaning of data. At SMI objects standardization is implemented to achieve a better understanding among all intervenients during the design and building phases. Reuse of objects is crucial to explore the potential of metadata management system. Good search tools minimize the creation of new objects.

The creation of a new version or revision of a metadata object (methodological document, data capture support, concept, variable or domain value) needs to be ruled by strict rules. But the reality surprise us with new situations in which the rules don't fit and needed to be changed. The motive of evolution is the trigger to the action that needs to be done.

A methodological document can have new versions (major changes) or revisions (minor changes). A data collection instrument can be substituted (new observation variables and consequently new questions) or prorogated (just a new period, so is validation is extended).

Each metadata object is build in the system by the statistical operation responsible person (ROE) or subsystem manager, and starts with "Construction" status. When complete or a group of objects are available, the owner changes it to "Proposed" and it will be validated by metadata team, subsystems manager. At this phase adjustments can be made and "Approved" status is granted.

A new version variable can occur due to an end date in the concept. A new conceptual variable is created with a begin date set to the day after the end date of old version. The old conceptual variable gains a new end date.

A new version variable can occur due to an end in the classification version. A new physical variable is created. A new physical variable is created with a begin date set to the day after the end date of old version. The old physical variable gains a new end date.

A variable can be characterized depending on the role: Observation, Derived, Measure, and Dimension.

A new classification needs to be used due an administrative process or a new representation of reality, so a new version or variant enter in the system.

The conventions about the words that should be used in some context are tools that lead to a bigger standardization through the use of SMI. The change in this conventions, give space to different names for the same reality.

The evolution of concepts is triggered from a better adherence to “business” reality.

The evolution of wording in questions and answer possibilities, and classifications categories, has a big impact in metadata objects.

The evolution of wording in dissemination indicators affects all other metadata objects that compose an indicator. In a reverse way new versions of each metadata object that compose an indicator are affected by new wording in indicators.

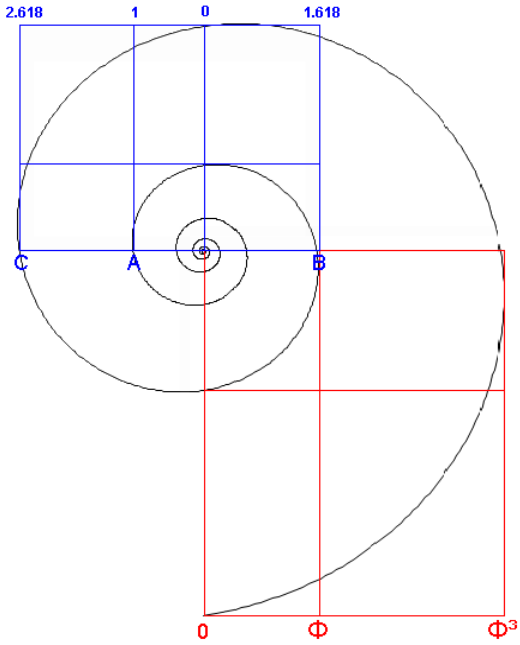

Fig. 11 *Evolution - discovers / reuse / adapt / versioning / new object*

The group of criteria to use some words in metadata objects in favor of another’s, even if they are synonymous, is not always simple and easy to determine. But one idea needs to be capital, consistency, otherwise the search and reuse of the system is quite difficult and time consuming to the user.

The group of criteria that is mandatory to break the continuity of a metadata object, with all the impact in data, mainly in long time series, needs to be established and observed continuously.

The perception of the user that the SMI have added value in the chain of statistics production is a great help for the managers of each subsystem and in the effort of good results in all process of metadata management.

The use in SMI of Portuguese and English for this technical tool increases the difficulty of daily operations in create and manage the metadata system. Sometimes grammar rules contradict technical metadata rules.

Integration with other systems supplies information to the building phase of applications and also reports to statistical operation's analysis and evaluation of processes.

Information in metadata management systems can't be carved in stone. The user friendliness and users recognition as an added value to the process are essential to keep the system alive. Challenges are in place, new standards and procedures are continuously arising demanding an effective answer from the metadata management system.

6. References

Pais, A. (2015) TSEE Course - SMI - Variables & Indicators Subsystem

Valente, I. (2015) TSEE Course - SMI - Classifications Subsystem

Guerra, A. (2015) TSEE Course - SMI - Methodological Documentation & Data Collection Support Subsystem

Saraiva, L. (2015) TSEE Course - SMI - Why Statistic Concepts? Concepts Subsystem

Morgado, I. (2013) Statistical Metadata (METIS) - Metadata Case Studies - Statistics Portugal

Portugal, A. and Poças, J. and Nunes S. (2015) SMI - Integrated Metadata Management System