ISTAT's New Strategy and Tools for Enhancing Statistical Utilization of the Available Administrative Databases

Giovanna D'Angiolini[1] | *Italian National Institute of Statistics, Rome, Italy*
Pierina De Salvo[2] | *Italian National Institute of Statistics, Rome, Italy*
Andrea Passacantilli[3] | *Italian National Institute of Statistics, Rome, Italy*

Abstract

This paper presents ISTAT's[4] new strategy to enhance the statistical utilization of administrative databases, aimed at: (i) disseminating information about the existence and accessibility of administrative databases to all potential statistical users; (ii) disseminating standard documentation about the information content and the quality of the available administrative databases to all potential statistical users.

Our purpose is to supply a proper utility to all institutions and people who wish to use administrative databases for statistics, including the owner institutions themselves, which may take advantage from the availability of tools to evaluate the quality and usability of their own databases.

This goal requires a massive activity of collecting information about the existence, the content and the quality of administrative databases. For this purpose we are currently undertaking a set of related activities, ranging from surveys of the administrative databases owned by local administration agencies to dedicated inquiries on the most important administrative databases, which are managed by central administration agencies.

Keywords

Administrative databases, quality documentation, quality database

JEL code

C10, Z19

INTRODUCTION

ISTAT is undertaking a general strategy that aims at making administrative data sources as serviceable as possible for statistical purposes by means of defining proper activities and dedicated tools, and levering forms of collaboration with owner institutions (UNECE, 2011).

In order to exploit administrative data sources for statistical purposes, the first step is to characterize their content in terms of collectives, which may be of interest to statisticians, and their features.

[1] Via Cesare Balbo 16, 00184 Rome, Italy. Email: dangioli@istat.it, phone: (+39)0646731.
[2] Via Cesare Balbo 16, 00184 Rome, Italy. Email: desalvo@istat.it, phone: (+39)0646731.
[3] Via Cesare Balbo 16, 00184 Rome, Italy. Email: passacantilli@istat.it, phone: (+39)0646731.
[4] Italian National Institute of Statistics.

For this purpose, it is necessary to examine their administrative procedures as data collecting tools and characterize their collected data as pieces of information about real-world observable items, which may assume a potential interest for statistical users. The aim is to determine what can be statistically used in a given administrative data source, independently of any current utilization, from a strictly documentation viewpoint. This requires an appropriate description of the administrative data source's information content, namely the definition of the administrative data source's ontology. This goal implies the capability of specifying the content of the administrative data source in terms of several collectives with their features, taking into account that an administrative data source generally observes both populations and sets of events that occur in time.

The more the statistical users exploit existing data sources, in particular administrative data sources, the more it becomes relevant for the description of the ontology of these sources to be standard and understandable, independently of any further particular statistical use. Moreover, statistical users need an accurate and comparable assessment of the data source's quality, which should exactly concern those collectives and features they are interested in.

ISTAT's strategy is carried out through (i) activities that aim at specifying the information content of each administrative data source, and analysing and measuring its quality, and (ii) ISTAT's supervision on those changes and innovation projects, which involve administrative data sources and administrative forms. In particular, the specification of each administrative data source's content and quality is attained by means of investigation on administrative data sources and their related administrative forms. An investigation is an analysis and documentation activity that employs standard tools, and is undertaken by ISTAT in collaboration with the owner institution. In order to support such activities, ISTAT is building methodological and information management tools, namely the DARCAP system and the Quality Assessment Framework for Administrative Data Sources.

DARCAP (Documenting ARChives of Public Administrations) is the web-based information management system to support investigations on administrative data sources and other documentation initiatives in order to provide potential users with structured documentation of their content and features. This tool also supports administration institutions in sending ISTAT their communications about innovation initiatives that concern administrative data sources or administrative forms. Furthermore, DARCAP supports ISTAT experts in producing structured documentation on the new information content of the administrative data sources that are involved in innovation projects, and in defining ISTAT's recommendations.

The *Quality Assessment Framework for Administrative Data Sources* is ISTAT's methodological tool to support statistical users in evaluating the quality of the available administrative data sources.

1 DOCUMENTING THE CONTENT AND THE QUALITY OF THE AVAILABLE ADMINISTRATIVE DATA SOURCES

The investigation on administrative data sources is performed by analyzing the available documentation and interviewing the source's experts belonging to the owner institution, as well as the source's users. The collected documentation is then organized according to DARCAP's database structure in order to be stored into such a database.

Firstly, we specify the denomination and the main characteristics of the administrative data source, the owner and the other managing institutions, the information flows and sets of administrative forms that are used to feed the administrative data source with data. Furthermore, the administrative data source's content is also documented; in particular, the main observed populations, which correspond to those collectives that are the target of the administrative procedure, and their related sets of events, each one with its definition. We also document the main characteristics of the single elements belonging to the specified collectives with their definitions, and the associated classifications (list of modalities) for qualitative characteristics. This work of conceptual description of the administrative data source's

content produces a specific source ontology that encompasses the following elements: the main collectives (which may be populations or sets of events), the main characteristics of these populations or sets of events, and the relationships that link populations and sets of events. The first result of this work is a network of main populations or sets of events, linked by 1-1 or 1-N relationships, in which every collective has its own definition and characteristics. A further analysis of the administrative data source helps determine more populations or sets of events that have associated their distinguished characteristics and relationships and are linked with the main collectives by means of subset relationships.

In terms of the investigation on administrative data sources, we also produce a first evaluation of their quality. More precisely, we ask the source's experts for information concerning each population or set of events. For each population, we document the entry and exit events, and the way by which their registration influence the population's coverage. For each set of events we document the way by which the single events are recorded into the source and the time distribution of events, as well as the coverage problems due to some of the following elements: registration scope (namely the capability of effectively registering all the single expected events) registration systematic distortion related to the purposes of the administrative registration procedure and registration timeliness (namely the time lag between the occurrence of the event and its registration). The main problems and the possible interventions concerning: the collectives' definitions, the suitability of the used classifications and their correspondence with standard classifications and the identification codes which may be used to link with other data sources are also evaluated. For the administrative data source as a whole, the main problems and the possible interventions concerning its statistical usability and its diffusion timeliness are also evaluated together with the related innovation strategies.

In order to perform a deeper analysis of the quality of administrative data sources it is useful and necessary to calculate standard quantitative indicators. As described in the next dedicated section, the Framework of quality indicators for administrative data sources defines concepts, methods and specific indicators for such an in-depth quality evaluation.

The investigation activity may suggest how to improve the content and the quality of the investigated sources. Moreover, in order to enable ISTAT to realize a more direct intervention on the existing administrative data sources we are now launching another activity, namely the supervision on changes and innovation projects related to the observed administrative data sources. To accomplish this task, ISTAT is asking a first group of administrative data sources' owner institutions to inform it about any kind of innovation project concerning their administrative forms and data sources in order to receive a technical and scientific evaluation. DARCAP provides a suitable environment to collect such communications (which may concern occasional as well as periodic changes), analyzing them to a certain extent, storing the extra documentation related to the communicated innovation projects and on the whole analysis process, and releasing opinions and recommendations.

All of the above activities are coordinated by a *Committee for Harmonizing Administrative Forms* whose members are nominated by ISTAT and the most important administrative data sources' owner institutions, which is supported by a *Network of experts*.

2 DOCUMENTING THE CONTENT OF ADMINISTRATIVE DATA SOURCES: THE CONCEPTUAL MODEL

The documentation activity aims to produce a standard, and therefore comparable, specification of the content of the available administrative data sources in terms of observed real-world objects, namely an ontology of the documented administrative data sources.

An ontology of an administrative data source is a structured description of its information content based on a standard conceptual model. In order to define such a conceptual model, we've analyzed the life-cycle of the administrative data and identified the different kinds of real-world objects to which they are referred. We've put such objects into correspondence with those objects to which any statistic

is currently referred, namely collectives and variables. Our conceptual model is oriented towards supporting the statistical exploitation of the administrative data sources, but it can be easily translated into other general-purpose conceptual models and languages for ontology specification (D'Angiolini, 2013). In the following section, we briefly introduce its main features.

Administrative data sources collect information about several kinds of real-world objects in order to support administrative activities (Brackstone, 1987, pp. 28–43). Firstly, any administrative activity entails collecting data about those entities which the activity addresses. Such entities are subsets of two general populations of persons, on one hand, and entities which perform economic activities, on the other hand. They may also be subsets of related populations such as households or territorial units. Moreover, information is collected about those particular sets of events that may involve these entities and are of interest for the purposes of the administrative activity. The observed *populations* and *sets of events* are linked by *relationships*. For both observed populations and sets of events proper information is collected about their characteristics, which may change in time. As an example, the Ministry for Public Education continuously collects information concerning students, schools and universities with their characteristics, as well as various sets of events such as degree course enrolments, examinations and degree earnings with their characteristics.

Therefore, inside an administrative data source we find two kinds of linked collectives: *populations* and *set of events*. Populations are subsets of the two most general populations of persons on one hand, and entities which perform economic activities on the other hand, or subsets of their related populations. Sets of events can be instantaneous (such as an examination) or durable (such as a degree course enrolment), and they may connect elements belonging to different populations. For example, any degree course enrolment event connects a student with a degree course. Each element of these collectives has *qualitative* or *quantitative characteristics*, such as a date of birth, residence, date of enrolment, examination score, as well as relationships with elements in other collectives.

According to a widespread ontology specification paradigm, in our conceptual model, a qualitative or quantitative characteristic is regarded as a relation that links an element belonging to a collective with an item belonging to a proper *classification*, or with a number in a numerical domain respectively. From a statistical viewpoint, quantitative characteristics and qualitative characteristics, together with their associated classifications, are regarded as variables. New variables can be defined as combinations of relationships and characteristics by means of logical and numerical operators. This is the reason why it is important to document the relationships among collectives. Finally, an administrative data source's ontology is a network of populations and sets of events that are linked by 1-1 or 1-n relationships and have associated quantitative or qualitative characteristics, the latter ones with their associated classifications.

Often some characteristics or relationships are associated with only a part of the elements of a collective. In this case, it is worth defining another collective that is a subset of the main collective, whose elements have associated such characteristics or relationships. More precisely, we distinguish between *subset relationships* and *partition relationships*. A subset relationship simply links two collectives when one gathers a part of the elements of the other. A partition relationship links a collective with many collectives which jointly partitions it, that is: each element of the partitioned collective belongs to one and only one of the partitioning collectives.

3 ASSESSING THE QUALITY OF ADMINISTRATIVE DATA SOURCES: BUILDING THE FRAMEWORK OF QUALITY INDICATORS FOR ADMINISTRATIVE DATA SOURCES

Trends such as the open data vision, the widespread development of data warehouses and the increasing usage of administrative data sources for statistical purposes not only by NSOs, but also by other organizations including their owner organizations' themselves, are all factors that are enlarging the scope of the quality assessment activity. In this scenario, NSOs should take responsibility for a new methodological

coordination task, namely to define rich and flexible sets of standards and repeatable quality assessment procedures for administrative data sources, as they currently do for surveys (UNECE, 2011). In order to meet such requirements, we have based our Framework on a careful analysis of the particular goals and features of the administrative data collection process and their effects on the quality of the collected data for each one of the different kinds of observed objects, which set up any data source's ontology (D'Angiolini et al., 2013).

Our Framework is organized according to the structure that has been proposed by Statistics Netherlands (Daas, 2009), which distinguishes three different views on quality, namely the Source view, the Metadata view, and the Data view. To each of these views, called "hyperdimension", is associated a number of dimensions, quality indicators and methods.

In the Source *hyperdimension*, the quality aspects relate to the administrative data source as a whole, the data set keeper, and the delivery conditions. The *Metadata hyperdimension* specifically focuses on the metadata related aspects of the administrative data source. It is concerned with the existence and the adequacy of the documentation, and the kind and structure of the identification codes. The *Data hyperdimension* focuses on the quality aspects of the data in the administrative data source. For the Source and Metadata hyperdimensions, we propose a set of qualitative indicators. As it has been mentioned above, in addition to requiring the administrative data owners to certify the availability of proper metadata, we also provide them with a standard tool for metadata specification: the DARCAP system.

As for the indicators in the Data hyperdimension, according to our approach, we aim to define a rich and well-reasoned quality indicators' frame in order to drive anyone outside or inside an NSO, particularly the administrative data source's owners themselves, in calculating and interpreting each indicator. Therefore, *the quality indicators are defined on the basis of the data set's ontology specification* and *the Data hyperdimension, which includes both qualitative and quantitative indicators.*

As we have seen, the qualitative indicators in the Data hyperdimension are specified by asking the data set experts a first qualitative assessment concerning some preliminary aspects of the data quality, such as the coverage and influence of registration delay on the coverage, distinctly for each collective (populations and set of events) in the administrative data source. As to the quantitative indicators, namely those indicators that are calculated from data and therefore require the availability of the data set, they must be calculable by the administrative data owner as well as by the NSO when it acquires the data set. The best scenario is when a collaborative calculation procedure is applied.

In order to define such quantitative indicators, first we have discriminated between possible errors, on one hand, and ways of checking them, on the other hand. The possible errors are defined in terms of those objects that may be present in an administrative data source's ontology in the following way.

For each object in an ontology, namely a collective, a characteristic or a relationship, we can build belonging statements concerning observed elements. More precisely, we can assert that a single observed element belongs to a set or that a couple of elements belongs to a characteristic or a relationship. In logical terms, statements concerning populations and sets of events correspond to a single variable predicate, statements concerning characteristics and relationships correspond to two variable predicates. Such statements will be true or false.

As an example, let us suppose we have an administrative data source whose ontology encompasses:

- Student (x), Degree_course (x), Examination (x) and Enrolment (x) which are collectives, more precisely Student (x) and Degree_course (x) are populations, Examination (x) and Enrolment (x) are set of events;
- Residence (x y), a characteristic that links each element x of the population Student (x) with an item y in the classification Town_codes (y);
- Examination_Student (x y) a relationship that links each element x of the set of events Examination (x) with an element y of the population Student (x);

- Enrolment_Student (x y) and Enrolment_Course (x y), two relationships that link each element of the set of events Enrolment (x) with an element y of the population Student (x), or Degree_course (x) respectively.

Examples of belonging statements, which involve observed elements are:

- the person identified by the fiscal code n is a student, namely Student (n);
- this person lives in Milan, namely Residence (n, Milan);
- there is an event i belonging to the set of the Examination events that concerns such a person, namely Examination (i), Examination_Student (i, n);
- there is an event i belonging to the set of the Enrolment events that concerns such a person and the degree-course Statistics, namely Enrolment (i), Enrolment_Student (i, n), Enrolment_Course (i, Statistics).

However, our conceptualization is more complex because the above statements have also time references as parameters, which are single moments for instantaneous events such as Examination, or (possibly open) periods for elements of populations such as Student, or durable events such as Enrolment.

The administrative data sources continuously collect and store data, which are in fact proper combinations of such belonging statements. Referring to the above example, for each new student, a new record is stored in the Student register which combines the statement Student (n) with the statement Residence (n, Milan) and other similar statements. Furthermore, another new record is stored in the Student register too, which combines the statement Enrolment (i), with the statements Enrolment_Student (i, n), Enrolment_Course (i, Statistics) and possibly other statements.

Each administrative data source has its own data collection procedure, which consists in accepting or not accepting such combinations of belonging statements inside the data source. As a result, at any time, any administrative data source stores a collection of belonging statements for each collective, characteristic or relationship in its ontology. It may happen that some of these statements are false, and that some true statements are not in the data set.

Therefore, at any given time we may have in the administrative data source:

- *Inclusion errors:* false statements (definitely or temporarily) accepted in the data source;
- *Exclusion errors:* true statements (definitely or temporarily) excluded from the data source.

Other errors may concern *wrong identification of the involved elements* due to problems in the identification code system, such as: syntactical errors in identifiers, identifiers for non-existing elements, lack of identifiers for existing elements, more than one identifier for each element and elements that share identifiers.

For *each collective* (population or set of events) the inclusion or exclusion errors correspond to the well-known *over-coverage* and *under-coverage* errors respectively. Therefore, when combining them with identification errors, we obtain a specification of all possible errors that concern belonging to the collective.

For *each mandatory characteristic* we may have an exclusion error, which corresponds to a nonresponse error. We may also have a combined exclusion and inclusion error if the element is linked with a wrong item in the classification or a wrong numerical value, which corresponds to a *measure error*. In terms of non-mandatory characteristics, we may have inclusion errors too. Identification errors may affect the observed characteristics when a change in a characteristic is registered for an element that is already in the data set, such as a new residence town for a student.

Errors that may concern relationships are specified in a similar way.

The available *quality check methods* are mainly: searching evident errors such as duplicate identification codes, linking with other data sources, using logical constraints (mandatory or incompatible combinations

between various belonging statements), calculating time lags between the moment of the events' occurrence and the moment of their registration.

Until now, we have defined a quality indicators' frame concerning the collectives' coverage and the elements' identification by means of properly combining possible errors and quality check methods. We are now analyzing the possible errors on characteristics and relationships in order to define two other quality indicators' frames concerning all kinds of nonresponses, measure errors and relationship errors. Obviously, in such indicators' frames, each quantitative indicator refers to several kinds of possible errors. Therefore, the quality evaluator will exploit the collected qualitative information about the nature and the respective relevance of the different kinds of errors in order to choose those quantitative indicators that are worth calculating and properly interpret them. Note that *our proposed indicators are distinctly calculable for each collective, characteristic and relationship in the administrative data source's ontology, in order to effectively support any statistical usage of the collected information.*

CONCLUSION

After a one year of testing activity, *we are now ready to start the investigation activity and the supervision activity on innovation projects concerning administrative data sources at operating speed*. Thanks to the work done in this first experimentation year, it has been possible to improve the investigations' supporting tools and the applied work procedures in collaboration with the administrative data sources' owner institutions.

We are now carrying out the work of specifying indicators in the Data hyperdimension on the basis of a careful analysis of possible errors anchored to the objects which may be present in an administrative data source's ontology. We plan to integrate the existing indicators, such as the BLUE-ETS indicators (Daas et al., 2011), in our indicators' framework as summaries of these more detailed indicators. Finally, the *Framework of Quality indicators for Administrative Data Sources* will contain *qualitative indicators for a preliminary quality assessment* in the Source and Metadata hyperdimensions together with *a rich set of both qualitative and quantitative indicators for an in-depth and customizable quality assessment* in the Data hyperdimension, which can be summarized in a more limited frame for a first overall certification task.

Moreover, in our opinion, an important advantage of our approach is that it gives foundations for *future research aimed at building a generalized probabilistic frame for the quality assessment activity* by means of properly reformulating the task of assessing the quality of any data collection as a problem of evaluating and composing the probabilities of all the possible errors.

References

UNECE *Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices*. Geneva: United Nations Economic Commission for Europe Press, 2011

D'ANGIOLINI, G. *Manuale per la documentazione di archivi, moduli e dataset nel sistema DARCAP* (Manual for the documentation of archives, modules and datasets in the DARCAP system). Rome: Italian National Institute of Statistics, 2013.

BRACKSTONE, G. J. Issues in the use of administrative records for statistical purposes. *Survey methodology*, 1987 (June), Vol. 13, No.1, pp. 28–43.

D'ANGIOLINI, G., DE SALVO, P., PASSACANTILLI, A., POGELLI, F. *Framework per la qualità degli archivi amministrativi* (Framework for the quality of administrative databases). Rome: Italian National Institute of Statistics, 2013.

DAAS, P., OSSEN, S., VIS-VISSCHERS, R., ARENDS-TÓTH, J. *Checklist for the Quality evaluation of Administrative Data Sources*, The Hague: Statistics Netherlands, 2009.

DAAS, P., OSSEN, S., TENNEKES, M., ZHANG, L., HENDRIKS, C., FOLDAL HAUGEN, K., CERRONI, F., DI BELLA, G., LAITILA, T., WALLGREN, A. *Report on methods preferred for the quality indicators of administrative data sources – Deliverable 4.2.* BLUE-Enterprise and Trade Statistics, 2011.