Towards a more efficient system of administrative data management and quality evaluation to support statistics production in Istat[[1]](#footnote-2)

Grazia Di Bella, Simone Ambroselli (Istat, Italy)

Abstract

Due to the increasing use of administrative data for producing statistics, management and methodological issues need to be better standardized within NSI and among EU member states. Istat is moving in this direction by setting up some centralized functions for the acquisition, storage, integration and administrative data quality evaluation.

**Keywords:** administrative data, microdata, quality.

1. Introduction

Dealing with an increase of the use of the Administrative Data (AD) for statistical purposes has become a common condition for the majority of the NSIs in the last decade. In Istat the administrative data sets acquired for statistical uses has increased from 90 in the 2009 to 230 in the 2013 as many statistical processes currently use them or are planning to review the production processes in this direction. In such conditions Istat is becoming more and more dependent on AD sources and the management of the AD acquisition and use inside the NSI required the formulation of a strategy to guarantee the quality and the timeliness in compliance with the rules on data security and privacy.

From the organizational point of view, Istat determined that some functions must be coordinated at a central level in order to avoid duplicate work among internal AD source users. From the operational perspective, a dedicated office named ADA (Administrative Data Acquisition and integration, under the Censuses and Statistical Registers Directorate) is responsible for the following tasks: acquiring, storing and integrating AD; evaluating AD quality; make AD and their metadata available to internal statistics producers (Figure 1).

**Figure 1 – ADA functions in Istat**

**Acquisition procedures**

**SIM**

**steps**

**AD quality evaluation**

Currently, these functions are in the implementation phase trying to automate the process as much as possible.

Regarding the performance of the Acquiring procedures, a strong coordination is required among internal statistics production processes and also with AD suppliers considering the changing from a system in which each process separately arranged for acquiring AD that it would use. The AD management process begins with the collection of AD requirements from the referents of Istat production processes (A). AD requests are then formulated ​​for each supplier and for each AD source following a standard form. The AD acquisition, storing and integrating functions up to the internal dissemination function (C-H) are performed in the new system SIM (Integrated System of Microdata), the core of ADA described in the next section. In addition to providing AD to all internal users, ADA is also involved in evaluating and documenting AD quality as input for statistical processes[[2]](#footnote-3) (I).

The centralization of the AD quality evaluation is an advantage for the efficiency of the system but also a guarantee for the quality of processes and products. In addition, a standardized procedure and coordinated approach at the central level allows activating those necessary feedbacks to improve AD quality, in collaboration with AD producers. AD quality issues are reported in section 3. The realignment of this process to the Generic Statistical Business Process Model (GSBPM) [1] may help to harmonize the existing infrastructure and to better clarify the ADA functions. To give a general idea, AD Acquisition procedures, compared to the eight phases of Level 1 and corresponding sub-processes, may be classified in the Phase 3. Build, sub-process 3.1 Build collection instrument and in Phase 4. Collect: 4.3 Run collection and 4.4 Finalize collection; while the SIM steps correspond to the 5. Process Phase: 5.1 Integrate data, 5.2 Classify and code. The ADA dissemination step is addressed to internal AD users and not to customers, than it is not classifiable in Phase 7. Disseminate. About the AD quality evaluation there is a close connection with the 8. Evaluation Phase as it is necessary to evaluate each sub-process; in the producer statistics perspective it is also important for the input quality management. It should also be noted that ADA functions also interact with: Phase 1. Specific Needs Phase, sub-process 1.1 Identify data needs, considering the potential of integrated AD with respect to survey data; 1.5 Check data availability, namely searching for currently available AD or, in case they not already available, AD for which a statistical availability and usability procedure may be started.

1. Integrated System of Microdata - SIM

SIM (Sistema Integrato di Microdati in italian) is the Istat Repository of integrated administrative microdata built with the aim of supporting the statistical production processes both for social and economic statistics. Integration means: (i) identifying each object in the Administrative Data Set[[3]](#footnote-4) (ADS) with a unique and stable (over time) ID number; (ii) to define, for each object, the logical and physical relationships among AD coming from different sources. According to the development of a unified view of administrative microdata, SIM supplies also metadata describing data and processes.

It includes social and economic data for:

* individuals and household characteristics; for example, demographic aspects, status in employment, educational level;
* places, for example, in terms of residence, work or school;
* the typology of the unit where people perform their activity/spend their time: houses, schools, places of work;
* the typology of the relationships among individuals, units and places.

*2.1 The process*

Observing the ADA process described in Figure 1, after completing the Acquiring procedures, ADS are stored in the centralized system (function C) and analyzed in order to identify the objects and their relationships useful for the development of the subsystems of integration (D). The ETL (extract, transform and load) process (E) for each source follows a set of standard rules:

1. each file is associated with an ID (key of the source) that identifies it in the system; the ID is stored in the metadata system;
2. the files are organized into one or more tables (relational structures);
3. each row of a table generally comprises: a serial number (internal key of the ADS); the reference dates; the IDs for the specific subsystems of integration (keys derived from data integration and identification step); the attributes of the objects in the ADS.

The step of integration (F) refers to the process of linkage and physical integration of microdata recorded in different ADS according to a strategy of integration and a set of algorithms to be applied. In this way each object in the system is identified with a unique and stable (over time) ID number. Data integration process feeds the development of the DBs for the integration of each subsystem.

The DBs for integration are warehouses of microdata useful to guarantee a unified view of the specific object under analysis showing information available in the different sources. Even though the physical structure of the DBs for integration depends on the different subsystem built, they generally includes:

* the ID of the ADSs;
* the serial number internal to all the ADSs in which the object is recognized;
* the ID of the object in the subsystem of integration;
* the variables used for the integration for all the sources in which the object is present;
* the time reference for linkage validity;
* the different kind of record linkage procedures used to enter in the DB.

The presence of a unique identifier (the ID of the object in the subsystem of integration) determines a *spider web* structure of relationships able to guarantee that every source is connected with the DB for integration and, at the same time, with all the others that are part of the same subsystem of integration.

In addition, the use of the SIM IDs allows to build data structures that could constitute the starting point for the statistical processes based on AD. In fact, it is possible to define a thematic DB linking data from the subsystems (or parts of them) and/or from the ADS. Using a syntax to define the rules of extraction, for example in relation to time or by statistical domains or by subsets of variables or objects, it will be possible to obtain a dynamic generation of data structures. Crossing the SIM border (H), the downstream statistical outputs will be automatically integrated because the objects (elementary units) have been submitted to the same process of integration and identification. The latter condition will be valid regardless the starting point used: one or more administrative DBs (traditional approach); subsystems (the DBs for integration); thematic DBs; a mix of them.

*2.2 Subsystems*

The subsystems identified are shown in Figure 2. The traditional socio-demographic domain is shown on the left (azure) while the economic domain is on the right (pale green). In the center is placed the subsystems able to link information both from individuals and economic units. Seven subsystems are under development; they can be grouped in: subsystems of the units; subsystems of the places; subsystems of the relationships.

The basic subsystems for the entire system are “SIM individuals” and “SIM economic units”. These two subsystems need to be developed before starting the procedures for all the others. Both from a logical and a technical point of view the two primary keys, respectively the *individual id* and the *economic unit id*, need to be assigned to all the basic objects of the system before the development of the subsystems of the places and the relationships.

**Figure 2 – SIM: the subsystems**



The second kind of subsystems developed is referred to the places of the economic units and individuals. “SIM economic units places” contains all the locations of the economic units identified in the administrative sources acquired. “SIM individual places” contains the information on residence, fiscal domicile (address), other addresses associated with the household (addresses for gas, electricity, water bills) and so on, for all persons recognized by the ADS stored in the system. Every address has its identifier (*individual places id* and *local units id*) and the presence of the primary keys (*individual id* and the *economic unit id*) allows to link data among different subsystems.

The subsystems developed for the relationships among different entities are: “SIM relationships among economic units”; “SIM relationships among individuals”; “SIM relationships between individuals and economic units”. In the first case, the kind of relationships considered are limited to the following: Enterprise/Enterprise (Events of transformation); Enterprise/Legal Units; Enterprise/Local Units; Control/ownership relations of the enterprise. The table contains the identifiers of the enterprises and the local units. In the second case the main purpose is the identification of the private households. In the latter case, the subsystem integrates information on the links among the individuals and the economic units where they perform their activity or spend their time. Three identifiers are necessary to tie everything together: *individual ID*; *economic unit ID*; *individual/units relationships ID*. The subsystem includes the LEED (Linked Employer Employee Data base) structures in which information from both enterprises (employer) and individuals, seen as workers, are broughttogether:. The flag for the kind of relationships allows to refine the research activities. At the moment, ten kinds of links attributable to the macro typologies “Job”, “Business role” and “Education” are admitted.

1. Administrative data quality evaluation

In recent years, Istat carried out various experiences in the AD quality evaluation (some of them are [2], [5], [6], [7]); other initiatives, in collaboration with the AD holder, have also been undertaken to improve the processes of generation of AD in case administrative forms changes, addition or new settings of the administrative information systems are expected. Finally, besides the Quality Guidelines for Statistical Processes already available [8], Istat is developing specific Quality Guidelines for processes using administrative data.

On the basis of these, a comprehensive strategy approach has been adopted considering quality aspects related to the AD management process from contacts with AD providers, until the dissemination of integrated AD to internal statistics producers that use them as input for the statistical processes.

In this paper, topics of statistical quality of the output produced by using AD will not be addressed, as they are tackled in the specific production processes and go beyond the tasks of ADA.

*3.1. The AD quality framework*

The AD quality framework adopted to evaluate the quality of ADSs delivered to Istat is based on that originally defined by Statistics Netherlands ([3], [4]) and then developed within the international BlueEts project, WP4 [5]. It considers a hierarchical and multidimensional approach including issues directly connected with the AD quality and those information for the AD management process aimed at improving the statistical AD quality/usability.

Three Hyperdimensions are considered: Source, Metadata and Data. In the *Source Hyperdimension* all aspects of quality related to the phase of AD acquisition are considered. Quality issues related to the AD generation process and the AD concepts are included in the *Metadata Hyperdimension*. Finally, the *Data Hyperdimension* addresses the specific quality of the AD supplied by source holders: the ADS. Each Hyperdimensions is described by quality Dimensions and quality indicators. At the last level there are the quality indicators measurements methods, that are specific measures appropriate to the AD specificity [5].

The Quality Report Card for AD (QRCA) is the document used for the dissemination of quality assessment in a statistics producer-oriented perspective.

Quality dimensions for each Hyperdimension are listed in Table 1.

**Table 1 – Dimensions and quality indicators**

|  |  |
| --- | --- |
| **SOURCE HYPERDIMENSION** | |
|  | Name and code of the administrative source holder |
| Supplier | Name and code of the administrative source |
|  | Name and code of the dataset (time reference of the data as defining attribute ...) |
|  | Referents for the supply |
|  | Istat source users |
|  | Istat statistical purposes |
| Relevance | AD holder possible statistical purposes |
|  | Importance of the source for Istat |
|  | Effects of the use in the reduction of the response burden |
|  | Meets the information demand? (Quality, Timeliness, Contents) |
|  | Legal provision |
| Privacy and security | Confidentiality |
|  | Security |
| Arrangements for the data | Costs |
| Delivery | Periodicity |
|  | Punctuality |
|  | Format |
|  | Dataset composition |
| Relationships and feedback with administrative source | Prior knowledge of any planned changes |
| holder | Inizio modulo  Feedback procedures in case of trouble |
| **METADATA HYPERDIMENSION** | |
|  | Administrative regulations generating the source |
|  | Description of modules / applications used for the acquisition / recording |
| Clarity/Interpretability | Objects description  Unique key |
|  | Variables description |
| Concepts comparability | Object definition comparison (vs. other AD source or statistical objects) |
|  | Variable definition comparison (vs. other AD source or statistical variables) |
| Data treatment (by data source holder) | Information on possible checks plan |
| **DATA HYPERDIMENSION** | |
|  | Readability/Convertibility |
| Technical checks | File declaration compliance |
| Integrability | Alignment of objects  Linking variable quality/Quality of the record linkage  Comparability of variables  Comparability of objects |
| Accuracy | Authenticity  Inconsistent objects  Dubious objects  Measurement error  Inconsistent values  Dubious values |
| Completeness | Coverage  Selectivity  Redundancy  Missing values  Imputed values |
| Time-related Dimension | Timeliness  Punctuality  Overall time lag  Delay  Dynamics of objects  Stability of variables |

Before adopting this AD framework, some experiments were performed in Istat with the aim to check first of all, the adaptability to Italian administrative data and secondly the possibility of implementation in the SIM. The first test took place as part of the Blue Ets - WP8 assessing Italian Social security Data [6], the second one (on some Education data) is in progress, but looking at the initial achievements it is possible to expect a gradual implementation in SIM, as will be described in the following paragraph.

*3.2. Producing QRCA through interoperability*

With the purpose of complying the appropriate efficiency and timeliness, a system that allows making the AD quality evaluation as automated as possible is being planning. Following the OECD Core principles for metadata management[[4]](#footnote-5), the strategy aims to take advantage of all the available metadata from the production process using AD (Figure 1), that is to make metadata “active” to the greatest extent possible for supporting the QRCA production.

The implementation of the *Source Hyperdimension* quality indicators takes advantage of all the information used for the AD acquisition procedures. The official AD requests formulated for each supplier and for each AD source is accompanied by: a) the date of delivery, previously agreed, necessary to comply with the timeliness; b) composition of the ADS to deliver as selection of the AD source (defined, where possible, according to the latest delivery sent net of changes known in advance c) the statistical uses that will be made (list of projects in the National Statistical Plan); d) the metadata request; e) other specific information such as agreements for data delivery, possible costs, …(function B). For the moment this information is managed in a not fully automatic way but Istat is proceeding in this direction storing and organizing it according with the AD quality framework in view of its reuse in the process of quality assessment. With respect to the Relevance quality Dimension, a specific Report for each main AD source is being finalized. It will provide information about all their statistical uses (derived automatically from the Acquisition procedures metadata, Figure 1) and about the compliance with the Istat requirements in terms of quality, timeliness, contents (derived from a very short questionnaire that could be submitted to AD source users).

In the *Metadata Hyperdimension*, the Clarity quality Dimension considers metadata that should be available from the data source holder.

In this regard, a procedure is in place which should allow the acquisition of the metadata together with data. The structural metadata which describe the data selection to deliver (ADS) are attached to the official letter sent to the AD source holder just described (function B in Figure 1).

The attachment is a worksheet produced automatically using the metadata of the last delivery. In the metadata acquisition procedure, this worksheet is used as a form for the AD source provider: the provider is required to include/update the following information: a) columns descriptions; b) possible changes with respect to the previous delivery in terms of new variables, variables no more available, changes in the population composition; c) date of data extraction from the source; d) other available free-form metadata. Metadata, when received, are stored in a DB and become available to automatically populate the QRCA.

About the Objects description (Clarity/Interpretability in the Metadata Hyperdimension, Table 1), the phase D. “Formal Concept Analysis/ identification of objects and relations” of the ADS, and the consequent data loading in the relational database (Figure 1), can allow us to automatically identify the set of objects/entities to be evaluated. Considering the possibility to automatically implement the Concepts comparability dimension (Metadata Hyperdimension), it should be possible to define a bridge linking the statistical units with the corresponding administrative units and the output statistical variables with the administrative ones used in the production process. Following the strategy of the interoperability among systems, new Istat *Unified Metadata System* (SUM) [9] could support the QRCA production. In case of processes already using AD, starting from the reference metadata, describing content and quality of statistical data produced and disseminated by Istat through the *I.Stat Dissemination System*, the traceability task (one of the SUM objectives), should be pursued retaining metadata processes. To work in this direction, a collaborative interchange of available metadata is being established. For the *Data Hyperdimension* too the description of the processes that take place in SIM may be reused as metadata to automatically define AD quality indicators. ETL process metadata (see § 2.1) may support the Technical checks dimension (dataset readability, convertibility and compliance with AD requested). The AD integration process is documented in SIM then, for the Integrability quality dimension, it is possible to reuse metadata to compute indicators describing the quality of the Linking variable and, in general, the quality of the record linkage procedures. Using unique and stable (over time) ID number for objects in SIM, Objects Alignment and Comparability indicators may be implemented comparing AD from different sources in SIM or AD with the main Statistical Registers. With respect to the latter, coverage indicators of the Completeness quality dimension, where possible, may be implemented too. Linking the same object in a dataset over time, using the SIM ID number, may produce Dynamic of objects indicators in the Time-related dimension.

These are some example of making metadata “active” for the QRCA production. If many indicators can be implemented automatically, however some others are not derivable from other metadata. It is the case of indicators of consistency checks (Accuracy dimension) for the implementation of which, an external intervention has to be considered for the check rules definition. As for the Relevance evaluation in the Metadata Hyperdimension already described, the AD source users may provide this information insofar as they have an interest to share with other AD source users checks performed on the data. Some “AD source users groups” are setting up in Istat for the most important data source holder (Tax administration, Social Security Institute, Ministry of Education, Universities and Research) in order to verify the possibility of sharing information or more specific analysis possibly useful to most users.

Essential references

[1] Australian Bureau of statistics. 2011. *Quality Management of Statistical Outputs Produced From Administrative Data*.

[2] Congia M. C., Rapiti F. (2010) Quality assessment and reporting in a short-term business survey based on administrative data, Documenti Istat n. 5, 2010

[3] Daas P.J.H., Ossen S.J.L., Vis-Visschers R.J.W.M. and Arend-Toth J. (2009). *Checklist for the Quality evaluation of AD Sources*. Discussion paper 09042, Statistics Netherlands.

[4] Daas, P.J.H., Ossen, S.J.L. (2011) *Metadata Quality Evaluation of Secondary Data Sources*. International Journal for Quality Research, 5 (2), 57-66.

[5] Daas P.J.H., Ossen S., Tennekes M., Zhang L. C. (CBS, Netherlands), Hendriks C., Foldal Haugen K. (SSB, Norway), Cerroni F., Di Bella G. (ISTAT, Italy), Laitila T., Wallgren A. and Wallgren B. (SCB, Sweden) (2011a). *Reports on methods preferred for the quality indicators of administrative data sources*, Deliverable 4.2 of Workpackage 4 of the BLUE-ETS project.

[6] Daas P.J.H., Tennekes M., Ossen S., Burger J., (CBS, Netherlands), Di Bella G., Galiè L., Bonardo D., Cerroni F., Talucci V., (ISTAT, Italy), Laitila T., Lennartsson D., Nilsson R., Wallgren A., Wallgren B. (SCB, Sweden), Hendriks C., Zhang L.C. and Foldal Haugen K. (SSB, Norway) (2013). *Guidelines on the use of the prototype of the computerized version of the QRCA, and Report on the overall evaluation results.* Deliverable 8.2 of Workpackage 8 of the BLUE-ETS project.

[7] ESSnet AdminData (2013a). *Final list of quality indicators and associated guidance.* Deliverable of Workpackage 6 of the ESSnet on Admin Data.

[8] Istat (2012) Quality Guidelines for Statistical Processes.

[9] Signore M., Scanu M., Brancato G. (2013) *Statistical metadata: a unified approach to management and dissemination.* NTTS - Conferences on New Techniques and Technologies for Statistics, Brussels, 5-7

[10] Unece (2007) Register-based statistics in the Nordic countries - Review of best practices with focus on population and social statistics.

1. The authors thank Carla Runci head of the ADA office. Although the paper is the result of the combined work of the authors, the Sections are to be awarded as it follows: Sections 1 and 3 to Grazia Di Bella; Sections 2 to Simone Ambroselli. The text published exclusively engages the authors, the views expressed do not imply any liability by Istat. [↑](#footnote-ref-2)
2. The meaning of “AD quality” is considered in relation to the AD reuse for statistical purposes, taking into account that the AD are not primarily produced for statistical purposes. For example, some AD may be very efficient in achieving administrative goals for which they were generated, while from the statistical point of view they may have a poor quality, that is, their statistical usability could be low. [↑](#footnote-ref-3)
3. The ADS is defined as the data set delivered by the AD holder. It is a selection of the AD source. [↑](#footnote-ref-4)
4. Common Metadata Framework Part A: Statistical Metadata in a Corporate Context), UNECE / Eurostat /OECD Group on Statistical Metadata (METIS) http://www.unece.org/stats/cmf/ [↑](#footnote-ref-5)