DARCAP: A tool for documenting the information content and the quality of the available administrative data sources

Giovanna D'Angiolini, Edoardo Patruno, Teresa Saccoccio – National Institute of Statistics (Italy), Carmine De Rosa, Enrico Valente – Top-Network (Italy)

Abstract Istat has developed a web-based system called DARCAP which manages standard documentation about the content and the quality of the available administrative data sources, in order to facilitate their utilization for statistical purposes. DARCAP encompasses three main subsystems, namely: DARCAP-Documenta, for documenting the information content and the quality of the administrative data sources and their related administrative forms, DARCAP-Innova, for enabling the administrative data sources' owner institutions to inform Istat about their innovation projects, DARCAP Consultazione, which allow the statisticians to navigate through the collected information. DARCAP is based on a proper conceptual model.

Introduction

In order to enhance the statistical utilization of available administrative data sources Istat is undertaking a general strategy aimed at surveying and analysing the existing administrative data sources in collaboration with their owner institutions [1]. Such a strategy is carried out through activities aimed at the specifying the information content of each administrative data source and analysing and measuring its quality, and through the 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 investigations on administrative data sources and their related administrative forms. An investigation is an analysis and documentation activity which employs standard tools and is undertaken by Istat in collaboration with the owner institution. In order to support such activities Istat has built a dedicated web-based system called DARCAP[2].

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

The DARCAP system' architecture

DARCAP system encompasses three main subsystems, namely:

- *Administrative data sources and administrative forms documentation system* **(***DARCAP-Documenta):* it provides ISTAT experts with functionalities for documenting the information content and the quality of the most important administrative data sources managed by central and local administration institutions with their related administrative forms, also through storing the results of interviews to experts of such data sources.
- *Innovation projects' communication supporting system (DARCAP-Innova)*: it provides the administration institutions with functionalities for informing ISTAT each time they plan any change or innovation in their managed administrative data sources and administrative forms. For the most important data sources, such a procedure allows ISTAT to give feedback and formulate recommendations.
- *Inquiry system (DARCAP-Consultazione)*: it provides the statisticians with the collected information about the available administrative data sources and administrative forms, which include a specification of their information content and a general assessment of their quality.

The three subsystems exploit a unique integrated database, in which we may distinguish three components: the objects' documentation component, the innovation projects' documentation component and the questionnaire.

Objects' documentation component: it is the part of the *DARCAP's database* which is dedicated to documenting the main features and the information content of the following objects:

- *Administrative data sources*
- *Administrative data sources' feeding administrative forms*

- *Administrative data sources' feeding datasets*
- *Administrative data sources' dissemination datasets*

More precisely, for each documented object the database stores several successive *versions* with their *validity period*, which is open for the latter version.

For each version of a documented object it is possible to specify a general description and other main features such as its managing institutions and, for the administrative data sources, those other objects' versions which are used to feed the documented administrative data source's version. Furthermore it is possible to describe the structure and the layout of the administrative forms' versions. Moreover for each version of a documented object it is possible to specify its information content according to a standard conceptual model to an extent which depends on the goals of documentation.

Note that the DARCAP's system objects' documentation component has been designed for storing an information content specification for both existing objects' versions and designed objects' versions. In such a way it supports all kinds of documentation activities, including the analysis of the information content of those designed administrative data sources and administrative forms which are involved in innovation projects.

Innovation projects' documentation component: it is the part of the *DARCAP's database* which is dedicated to documenting all the features of the innovation projects which are communicated to Istat as well as the content of the released Istat's recommendations.

Questionnaire component: it is the part of the *DARCAP's database* which is dedicated to storing, for each administrative data source's version, all the answers to the questions contained in the Istat's questionnaire on the administrative data source quality, which is submitted to the data source' experts during any investigation in order to collect information about several aspects such as the actual or potential use of the administrative data source's version, the information collecting procedures and the estimated coverage of the observed collectives.

The DARCAP subsystems

Administrative data sources and forms documentation system (DARCAP-Documenta)

At present the DARCAP's users in charge of documenting administrative data sources and administrative forms are only Istat experts. This subsystem provides them with three main environments: Registry, Information content specification, Questionnaire.

In the Registry environment the DARCAP's user documents the main features of a first or a new version of an administrative data source, an administrative form, a feeding or diffusion dataset, which are: the name and the validity period, the general description, the main owner institution and the other managing institutions, its related administrative procedures and regulating laws, and, for the administrative data sources, those versions of administrative forms, datasets, other administrative data sources which are used to feed the documented administrative data source's version.

In the Information content specification environment the DARCAP's user documents the information content of any version of an administrative data source, an administrative form, a feeding or diffusion dataset, as a network of populations and set of events with their associated characteristics, linked by means of 1-1 or 1-n relationships. Such information is referred to the particular update activity which generated it, which may be an investigation activity or another update activity such as loading the results of surveys concerning the many and various administrative data sources which are managed by local institutions. To make the documentation activity easier, all the information related to the previous version of an object is automatically associated with the new version of the object to be documented, so as to enable the DARCAP's user to work by deleting those pieces of information which are not valid anymore. Moreover it is possible to document the information content of a first or a new version of an administrative data source by means of importing the information content of its feeding administrative forms' versions, datasets' versions and administrative data sources' versions.

Such information becomes available to the end users in the Inquiry subsystem as a result of a specific validation operation which is launched by the DARCAP's user who specified it, or by a special supervisor user. As a consequence of the validation operation the system performs a set of checks and reports the results, moreover a preview of the graphic presentation of the information content is displayed and can be optimized.

In the Questionnaire environment for each administrative data source's version the DARCAP's user is led to fill an online questionnaire, in order to document the answers to the questions contained in the Istat’s questionnaire on the administrative data source quality which is submitted to the data source’ experts during the administrative data sources’ investigations. Non-responses are allowed.

Such information is presented to the end users by means of a downloadable pdf which contains the questions of the questionnaire with their collected answers and becomes available in the Inquiry subsystem as a result of a specific operation of closing the questionnaire which is launched by the DARCAP’s user who specified it, or by a special supervisor user. It is also possible to make visible to the end users only a part of the collected answers.

Two other documentation environments are available to the DARCAP’s user, the Structure documentation environment and the Information content’s origin documentation environment. In the Structure documentation environment the DARCAP’s user documents the structure of any administrative form’s version or dataset’s version, by means of singling out its different sections with their particular information content. For any administrative form’s version it is possible to specify the position of each section in the form’s layout, in order to show in the Inquiry subsystem each section with its particular information content by means of an anchorage to its position in the form’s layout. In the Information content’s origin environment the DARCAP’s users documents the origin of the information content of any administrative data source’s version, in terms of the information content of its feeding administrative forms’ versions, datasets’ versions, administrative data sources’ versions.

Innovation projects’ communication supporting system (DARCAP-Innova)

The innovation projects’ communication activity involves several kinds of DARCAP’s users. The administrative data sources’ owner institutions charge some experts with the task of communicating to Istat any innovation project or regular change which involve administrative data sources or administrative forms. Moreover such institutions appoint a supervisor of the innovation projects’ communication activity who also chooses those innovation projects which are submitted to Istat for recommendations. Istat experts receive and evaluate the innovation projects’ communications, they may document the information content of the designed administrative data source’s versions or administrative form’s versions, and document the released ISTAT’s recommendations.

DARCAP-Innova enables the experts belonging to the administrative data sources' owner institutions to specify a synthetic description for any innovation project which involves one or more versions of administrative data sources or administrative forms, and load into a dedicated repository the enclosed documentation, in various formats. Moreover such experts may enrich the description of the innovation project with the specification of the project's type, the project's thematic scope and the project's main features; such an enriched description of the innovation project can be also specified by their supervisor or by the Istat's experts when they receive the communication. Moreover DARCAP-Innova enables the supervisor belonging to the administrative data sources' owner institutions to choose among the innovation projects which have been communicated by experts, complete their description and finally send them to Istat.

DARCAP-Innova enables the Istat's experts to document the whole process of releasing recommendations, also by means of loading suitable documentation in the dedicated repository.

If necessary, the Istat's experts can analyze the innovation project in order to single out more specific projects that compose it, each one regarding only one of the involved designed administrative data source's versions or administrative form's versions. If necessary, they can also document the information content of each designed administrative data source's version or administrative form's version as well as the structure of each designed administrative form's version, by means of calls to the suitable functionalities of DARCAP-Documenta. When for a particular designed version of an administrative data source or an administrative form we want to closely follow and document each one of its design phases it is possible to specify several successive specific projects, each one corresponding to a more detailed description of the information content of such a particular designed version.

When the innovation project has been completed, the documented designed versions of the involved administrative data sources or administrative forms are transformed into new existing versions only by means of changing a suitable flag.

Inquiry system (DARCAP-Consultazione)

Such a DARCAP's subsystem provides the end user with two distinguished environments for accessing the documentation of the innovation projects or navigating through the documentation of the existing administrative data sources or administrative forms, respectively.

Accessing the documentation of the innovation projects: it is possible to search for an innovation project or a specific component of an innovation project by project's name and institution's name and display all the features of any innovation project, general or specific, including the documentation of the involved designed administrative data sources' versions or administrative forms' versions as well as the ISTAT's recommendations.

Navigating through the documentation of the existing administrative data sources or administrative forms. This environment provides the end user with two different search functions.

The first one is the s*earch for a version of an administrative data source or an administrative form by name and other criteria*, which depend on the kind of the owner institution. The search by name requires a string specification. For those administrative data sources or administrative forms which are owned by central institutions the other search criteria are: validity period, data source's type, managing institution's name. For those administrative data sources or administrative forms which are owned by local institutions the other search criteria are: validity period, managing institution's type and name, region, related administrative procedure's type, general thematic area and specific thematic area. Proper choice lists are displayed for each criterion. The system shows the list of those administrative data sources' versions or administrative forms' versions that satisfy the specified criteria, among which the end user can choose.

The second one is the s*earch for a version of an administrative data source or an administrative form by information content*: given a string specification, the system shows all the collectives, characteristics and classifications whose name contains the specified string, and for each of them its containing administrative data sources' versions or administrative forms' versions, among which the end user can choose.

Once the end users choose a particular data source's version or administrative form's version they can browse through its related documentation. More precisely they access:

- Name, description and validity period, and a simple list of the observed collectives, characteristics and classifications;
- A graphic presentation of the network of collectives and their relationships, with the possibility, for each collective, of viewing the list of the characteristics with their associated classifications and the network of those collectives which are its subsets.
- Other general features such as the owner institutions and the other managing institutions, the related administrative procedures and regulating laws, for the administrative data sources' versions their feeding administrative forms' versions, datasets' versions, administrative data sources' versions, and other information including enclosed documents and web sites' addresses.

Only for administrative data sources' versions, it is possible to download pdf documents which contain the filled Istat's questionnaire on the administrative data source quality, which gathers information about several aspects such as the actual or potential use of the administrative data source's version, the information collecting procedures and the estimated coverage of the observed collectives.

In the second version of DARCAP, for the administrative forms' versions it will be possible to view their information content referred to the different sections which set up their structure. It will be possible to highlight a section in the layout and open a window with the specification of its particular information content.

The DARCAP conceptual model

The documentation activity aims at producing a standard and therefore comparable specification of the content of the available administrative data sources' versions or administrative forms' versions in terms of observed real-world objects, namely an ontology of the documented administrative data sources' versions or administrative forms' versions. An ontology of an administrative data source's version is a structured description of its information content, based on a standard conceptual model. In order to define such a conceptual model, we have analyzed the life-cycle of the administrative data and singled out the different kinds of real-world objects to which they are referred, and we have put such objects into correspondence with those objects to which any statistic is currently referred, namely collectives and variables [3]. 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. In the following we briefly introduce its main features.

Administrative data sources collect information about several kinds of real world objects in order to support administrative activities [4]. First, any administrative activity entails collecting data about those entities which the activity addresses. Such entities are subsets of the two general populations of persons, on one side, and entities which perform economic activities, on the other side, or they are subsets of related populations such as households, territorial units. Moreover, information is collected about those particular sets of events which 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 r*elationships*. 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 about the students, the schools and the universities with their characteristics as well as about sets of events such as the degree course enrolments, the examinations, the degree earnings with their characteristics.

Therefore inside an administrative data source's version we find two kind of linked collectives: *populations* and *set of events*. Populations are subsets of the two most general populations of persons on one side, and entities which perform economic activities on the other side, or subsets of their related populations. Sets of events can be instantaneous (such as examination) or durable (such as degree course enrolment) and they may connect elements belonging to different populations, as an 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 date of birth, residence, date of the 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 which 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, the quantitative characteristics and the 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 the ontology of an administrative data source’s version is a network of populations and sets of events which 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 to define another collective which 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 partition it, that is: each element of the partitioned collective belongs to one and only one of the partitioning collectives. For each main collective its subset collectives, linked by several subset and partition relationships, may set up a possibly complex network. The DARCAP system checks the correct definition of such a network by means of the information content’s validation.

Finally we also document some other aspects which are important from the viewpoint of the statistical use of the documented administrative data source’s version: the identification codes which are adopted for the elements of each collective, the structure of such identification codes, the algebraic relationships which may connect different quantitative variables.

References

[1] United Nations Economic Commission for Europe (UNECE). (2011), Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices, United Nations Publication

[2] D’Angiolini G., Patruno E., Saccocci T. (2012), Specifiche del sistema DARCAP, Istat document

[3] D’Angiolini G. (2013) Manuale per la documentazione di archivi, moduli e dataset nel sistema DARCAP, Istat document

[4] Brackstone G.J. (1987), Issues in the use of administrative records for statistical purposes, Survey methodology