New experiences in the production of business statistics: the construction of the “Frame SBS” and SBS - data warehouse[1]

Francesco Altarocca[2], *Diego Bellisai*[3], *Antonio Laureti Palma*[4], *Roberto Sanzo*[5]

Abstract

This paper describes the first experience of data integration using administrative sources to estimate Structural Business Statistics (SBS) aggregates. It briefly shows the process that led to the construction of the “Frame”, the “integrated” dataset that is the base to derive SBS estimates from the year 2012. This approach represents an innovative way to produce business statistics: the use of direct surveys is limited in favour of the use of administrative data. Moreover, this paper presents an innovative IT proposal on how to combine the integrated production model and the warehouse approach for the production of Structural Business Statistics. This consists in a metadata-driven data warehouse of integrated SBS information which is well-suited for supporting the management of modules in generic workflows. Such a modular approach can improve the efficiency of collaboration among different statistical experts on common data.

Keywords: Frame, Business statistics, SBS, Administrative sources, Economic aggregates, Scientific workflow, Data Warehouse, Data integration, Process integration

Introduction

Structural Business Statistics (SBS) are a powerful instrument to obtain a lot of detailed information about most aspects of Italian enterprises. The Eurostat Regulation n.58/97 and SBS EU Council Regulation n. 295/2008 define the main economical aggregates that have to be sent to Eurostat; these aggregates have to be provided for a defined estimation set and these estimates have a statistical significance only if referred to a certain period and to a certain universe.

The Italian Statistical Institute (ISTAT) since 1998 has derived these estimates carrying

[1] This work is the results of the common effort of the authors. However the paragraphs can be attributed as follows: Introduction and par.1 Roberto Sanzo; Conclusions and par. 2 Antonio Laureti Palma; par. 2.1 Diego Bellisai; par. 2.2 Francesco Altarocca. The views expressed in this paper are solely those of the authors and do not involve the responsibility of Istat.

[2] Istat, e-mail: fraltaro@istat.it.

[3] Istat, e-mail: bellisai@istat.it.

[4] Istat, e-mail: lauretip@istat.it.

[5] Istat, e-mail: sanzo@istat.it.

out two different surveys depending on enterprise size[6]: the PMI survey (*Small and medium enterprise survey, including professional and artistic activities,* "*Rilevazione sulle Piccole e Medie Imprese e sull'esercizio di arti e professioni*"), a sample survey referred to enterprises with less than one hundred employees and involving about 120,000 enterprises, and SCI survey (*Survey on enterprise accounting system*, "*Sistema dei Conti delle Imprese*"), a total survey referred to enterprises with one hundred or more employees and involving about 10,000 enterprises.

In recent years, however, the increasing availability of information deriving from administrative sources has made possible to use them for statistical purposes both to make easier the analysis of respondents' data and to impute the statistical information when it was not available. In fact since the end of the '90s, the data about the Balance became an essential element of the production process of SCI, both during the editing step and during the integration of total non-response phase. In the 2000s, also the PMI survey began to use some administrative sources to try to estimate the information for non-responding firms.

The agreements between the ISTAT and the government (in particular the "*Agenzia delle entrate*" and "*Unioncamere*"), have ensured ISTAT to obtain an unprecedented amount of data that suggested the massive use of it for making statistics; the aim was to try to reduce the "statistical burden" for the respondents too.

The new approach for SBS was to obtain business information for each unit in Italian Business Register (ASIA, "*Archivio Statistico delle Imprese attive*") using administrative data: in this way, the aggregates could be simply derived using "sum of values" among different units instead of using sampling and calibration techniques. This approach was tested in 2010-11 on PMI Universe; for enterprises with 100 employees or more the SCI survey was preferred because it ensured a better quality of estimation. In fact, not all the SBS aggregates can be derived from administrative data: in an administrative source, the business information can be or not be present; moreover the data were not available for all units: some techniques of imputation or estimation are however needed to ensure a "complete sum of values" for all the units to obtain the main SBS aggregates.

The SBS Frame is the final result of this integration activity; it represents the first product completely derived from administrative data and used to obtain statistical aggregates: it provides a dataset including the values of main economical aggregates for each unit in SBS universe by Asia Register.

Finally, given the big amount of information to be managed and to increase its usability, the Frame has to be produced in an innovative application framework based on a data warehouse supported by a workflow engine. .

This paper is structured in two parts: in the first chapter, the steps to make the Frame will be synthetically described and some simple results will be showed; the second chapter presents an innovative IT proposal on how to combine the integrated production model and the warehouse approach for the production of Structural Business Statistics, consisting in a metadata-driven data warehouse of integrated SBS information.

[6] For more details, refer to: Curatolo S., De Giorgi V., Oropallo F., Puggioni A., Siesto G. 2015. "Quality analysis and harmonization issues in the context of the SBS frame". *Rivista di Statistica Ufficiale*. n.1/2016.

1. The “Frame”: the making of

In order to build an information structure able to estimate the SBS variables and the aggregates of national accounts (NA) using administrative sources, the following sources were considered :

- *Financial Statements* (hereafter BIL);
- *Sector Studies survey* (“*Studi di settore*”, hereafter SDS);
- *Tax returns* (“*Modello Unico*”, hereafter UNI);
- *Regional Tax on Productive Activities* (“*Modello IRAP*”, hereafter IRAP).

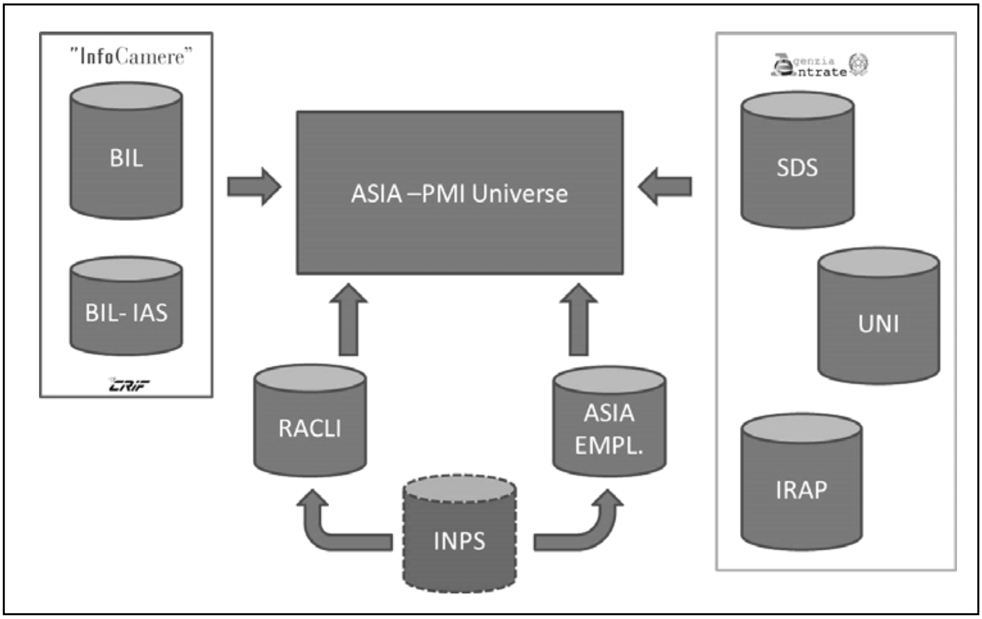
The Financial Statements (BIL) represents the main information document about enterprises dynamics: it includes the costs and revenues of the entire accounting period and includes the assets and the liabilities that a company has incurred during the year.

The SDS source is a tool that the Italian tax authorities used to detect the economical parameters for professionals, self-employed and small firms. The main aim is to identify the activity and the economic environment in which the enterprise operates, in order to assess its ability to produce real income.

The UNI source is the model of ordinary income tax return while IRAP is a local tax that applies to productive activities in each region: it must be paid only by those who carry out business activity and not by individuals.

As a reference universe is considered what is defined by the SBS Regulation and its identification is made through the Business Register (ASIA): in particular, we considered all the firms in industry and services (excluding financial firms and insurance) with less than 100 employees and active for more than six months in the concerned year (PMI Universe).

Figure 1: Logical scheme for the administrative data integration model.



Moreover, as support archives used in the different phases of construction of this information structure , from now on called Frame, are also used information from :

- Archive RACLI ("*Registro anagrafico del costo del lavoro per impresa*"), with regard to the cost of labour and its items[7]; it derives the Social Security Data from the Italian National Security Institute (INPS).
- ASIA-Employment, regarding the information about atypical workers and their remuneration.

The Figure 1 shows the logical scheme used for deriving statistical information from administrative source: the center of the scheme is the PMI universe as defined by ASIA according to SBS Regulation. The last two archives (RACLI and ASIA-Employment) entered the process to ensure a verification of information about personnel costs (see § 1.1).

1.1 Administrative items and statistical variables: general aspects and the personnel costs correction

The use for statistical purposes of information from administrative files, which, by definition, are kept for reasons other than those for which we will try to use them in this context, involves a whole series of preliminary activities to

- identify what items can be used to build the economic variables needed;
- evaluate the consistency with the statistical definitions;
- verify the information content of the items in the various administrative sources.

Regarding these three points, the different administrative forms were analyzed by experienced ISTAT personnel and the items useful for statistical purposes were identified. The next analysis was about the fitting of administrative and statistical values (in this case, comparing the values obtained by respondents to PMI survey) for identifying, for each administrative source, the different degree of fitting of the main economic aggregates used in SBS[8].

It was not easy to identify a systematic behaviour in compiling of the administrative forms by accountants or tax operators able to explain differences with surveys questionnaires. However, when the administrative item and the statistical variable should not be different according the definition, these differences could be considered "noise" and those could be treat using statistical methods.

An exception is represented by the variable "personnel costs" that has an additional source of comparison (RACLI): for enterprises using atypical workers the discrepancies between statistical and administrative information could be explained by the remuneration of the atypical workers; also these differences could be explained by independent workers even if the enterprise does not use atypical workers. In fact, the empirical analysis showed that sometimes the remunerations of the atypical or independent workers were wrongly

[7] For more details, refer to: Arnaldi S., Baldi C., Filippello R., Mastrantonio L., Pacini S., Sassaroli P., Tartamella F. 2016. "The labour cost variables in the building of the frame". *Rivista di Statistica Ufficiale*. N.1/2016.

[8] For more details, refer to: Curatolo S., De Giorgi V., Oropallo F., Puggioni A., Siesto G. 2016. "Quality analysis and harmonization issues in the context of the SBS frame". *Rivista di Statistica Ufficiale*. n.1/2016.

included in the personnel costs item instead of in “costs of services” or in profits [9].

For this reason, it was necessary to apply a correction procedure: however to ensure consistency with other economic variables, for constructing variable “personnel costs” in the Frame is used the information coming from administrative sources (except RACLI); RACLI is used in a procedure of "correction" (called COMBO) for separating the "noise" caused by atypical and/or independent workers from the administrative data.

Only in cases where the administrative data about labour cost were higher than the value in RACLI, the COMBO procedure tries to determine, according to the information derived by ASIA Employment, the best possible combination of remuneration values of atypical workers and independent to be deducted from the personnel costs, to get a value as close as possible to the RACLI value. Then the amounts deducted from the personnel costs were added to the costs of services (if related to atypical workers) or to the profit (if related to independents) to ensure consistency with other information.

The COMBO procedure was applied separately in the sources and has produced changes in a limited number of units (about 3.5% in BIL, less than 1% in the SDS and UNI) with variations respect the original data that, overall, ranging from -1.5% to -0.7% in terms of personnel costs and 0.8% to -0.2 % in terms of added value. In fact a conservative perspective was used: the correction is made only when there are a sufficient number of clues to ensure the successful application of the procedure. The problem did not concern for the source IRAP where the personnel costs is not used.

A similar correction, but in the opposite direction, was made (if COMBO did not act) for all units for which ASIA-Employment provided an indication of the presence of independent reclassified by ASIA Register as para-subordinates: their remuneration was deducted from the cost of services and moved in the profit.

1.2 Pre-treatment of administrative data: the data editing procedures

When an administrative item useful to statistical purpose was identified, at least other two steps were necessary:

- checking that the information relating to the same unit within each source is unique;
- checking and correcting all the inconsistencies that occur among items within each administrative source.

Regarding these points, it was necessary to implement some statistical procedures of data-editing to check, correct and/or eventually erase anomalies in the sources .

To ensure the “internal consistency”, the administrative data were treated directly, as those have been acquired, also verifying relationships with variables that are not useful for the construction of statistical aggregates. But because of the high amount of data (millions of records for hundreds of items), not all items were involved in the check activity: in fact, for each of these sources only items directly or indirectly useful to the construction of economic aggregates were tested.

Then, the main activities carried out were:

1. identification and elimination of "duplicate records",
2. identification and elimination of internal inconsistencies in each source,

[9] For more details, refer to: Arnaldi S., Baldi C., Filippello R., Mastrantonio L., Pacini S., Sassaroli P., Tartamella F. 2016. The labour cost variables in the building of the frame. *Rivista di Statistica Ufficiale*. n.1/2016.

3. exclusion of useless records

All these activities were performed independently on each of the administrative source.

In all considered sources were sometimes present more record referring to the same unit: these duplications were due to the fact that it is possible for the enterprises to re-present an administrative form already presented if they have to correct or to improve their declaration. But it is not simple for the statistician to understand what to do when there are more records for a single unit: the flags deriving from administrative form often are not satisfactory in all the cases because these flags seldom are nonzero. For this reason an interactive analysis of duplications was necessary: referring to the 2010 and 2011 occurrences, some rules were identified to help the choice. These rules made it possible to remove a lot of duplications in the 2012 and 2013 too; but since these rules are empirical, they are not exhaustive and then every year an interactive step is necessary to treat a residual number of cases.

The next step was to check and eventually correct inconsistent data. For this purpose for each source a consistency plan was built; a certain number of rules were considered: to accept a single record all of these rules had to be come true. In particular, domain edits, prorate edits or "less than" edits for "specification items" were the edits considered in the consistency plan: to correct inconsistencies, a deterministic method was used and the correction rules were based on the relationships among the variable.

At the moment, the check procedure does not provide comparisons with information deriving from the other sources: only for BIL is made a comparison with SDS data. In perspective we expect to use similar procedures for UNI, SDS and IRAP too.

When a record is formally exact, we had to decide if it was useful for the Frame: in fact if no item of the administrative form was used for derive a statistical variables, the related record was deleted from the dataset.

1.3 Harmonization of variables

The activity of "harmonization"[10] is crucial for binding the administrative items to statistical variables. It is carried out only after the elimination of the causes of "noise" previously described: to make the information recognizable and comparable as much as possible, the same names were used in the different sources, diversifying them only by a suffix indicating the origin of the source. These names come from names used in SBS.

All the steps described above must be driven by a set of metadata as comprehensive and clear as possible; unfortunately the greatest difficulty encountered in the first part of the work was just the availability of accurate metadata in electronic format. It is hoped, however, that for the next realizations of the Frame the metadata problem will be reduced through agreements with the government, the "owner" of the data.

The application of the procedures in the various steps of pre-treatment of data decreases the number of units available: this is the result of the "cleaning" of the data that includes not only the elimination of duplicate records but also of record with data incorrect or useless (e.g. all values equal to "1"); other deleted records are those records that do not have "significant" data for the purposes of this context (e.g. missing values for all item used for

[10] For more details, refer to: Curatolo S., De Giorgi V., Oropallo F., Puggioni A., Siesto G. 2016. "Quality analysis and harmonization issues in the context of the SBS frame". *Rivista di Statistica Ufficiale*. n.1/2016.

the construction of the economic variables here considered).

The Figure 2 shows a summary of the preliminary activities previously described for the construction of the Frame, starting from “Gross data” to arrive to “harmonized data”.

Figure 2: Steps of pre-treatment of administrative data for the construction of the Frame.

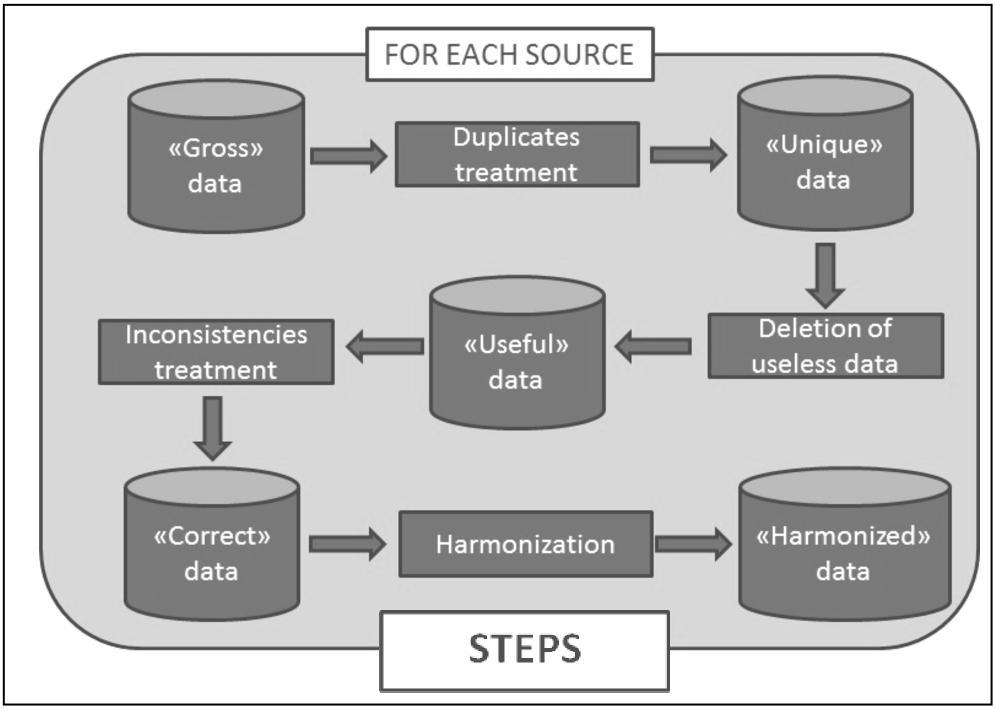


Table 1: Number of treated units in each source for steps of pre-treatment. Year 2012.

Steps	BIL	SDS	UNI	IRAP
Gross data	747,492	3,973,230	4,498,531	3,658,585
Duplicates deletion	746,953		4,329,326	3,532,179
Data Editing	743,737	3,973,230	4,319,252	3,495,617
Harmonized Data	743,737	3,580,745	4,313,632	3,494,493

The Table 1 shows the number of units treated in the different pre-treatment steps.

The harmonization step produced the table shown in Figure 3: this table is just a fictional representation of what really happened: the colored cells indicate the presence of SBS variable within a source. The Figure shows clearly that not all SBS variables are covered by administrative source and that the different administrative sources do not provide all the SBS variables: some sources provide more SBS variables then others. Moreover, also within the source, the same variable could be or not present, according to the type of enterprise and to compiled form.

In conclusion, paradoxically in some cases the information is too broad, in others it is virtually absent: for this reason the next step was to arrange this huge amount of data - with different degrees of quality - for identifying the best business information to assign to each unit in PMI universe.

Figure 3: Fictional representation of availability of SBS variables in administrative data.

		SDS		UNI									IRAP							
SBS Variab	BIL	F	G	PF-RG	PF-LM	SP-RG	PF-RE	SP-RE	PF-RS	SP-RS	SP-RS_IAS	SC-RS	SC-RS_IAS	IQ-I	IQ-II	IQ-V	IP-I	IP-II	IP-V	IC	IK
V1	■	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	■	■	■	■	■	■	■	■
V2	■	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	■	■	■	■	■	■	■	■
V3	■	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒		■	■	■	■	■	■	■	■
...	■	▒	▒	▒		▒	▒	▒	▒	▒	▒	▒		■	■		■	■		■	■
...	■	▒	▒	▒		▒	▒	▒	▒	▒	▒	▒									
...	■	▒	▒	▒		▒	▒	▒	▒	▒	▒	▒									
...	■	▒	▒	▒		▒	▒	▒	▒	▒	▒	▒									
...	■										▒	▒			■			■		■	■
...	■	▒		▒		▒									■			■		■	■
...	■	▒													■			■		■	■
...	■	▒	▒	▒		▒	▒	▒			▒	▒		■	■	■	■	■		■	■
...	■	▒	▒	▒		▒								■	■		■	■		■	■
...	■	▒		▒		▒								■	■		■	■		■	■
...	■	▒	▒	▒		▒	▒	▒						■	■		■	■		■	■
...	■	▒																			
...	■	▒	▒	▒		▒	▒	▒			▒	▒		■	■		■	■		■	■
...	■																	■		■	■
...	■	▒																			
...	■	▒	▒								▒	▒			■			■		■	■
...	■																				
...	■																				
...	■																				
...																					
...																					
...																					
Vn																					

1.4 Data integration: hierarchical order and source priorities

To assign a set of statistical variable to each unit of PMI universe; three situations were shown:

1. More sources for the same unit
2. Units with only a source
3. Units without information.

The choice is certain for point 2: if information is available only from a source, the SBS aggregates are derived from that source.

Regarding to point 3, it can be considered as a non-respondent case in a survey, then a statistical method of imputation of total response can be apply. In this context, the "predictive mean matching" and the "minimum distance donor" methods have been used[11].

The choice of source for units with more sources available was more complex: it was necessary a set of analysis for determining the accuracy between administrative and survey data in each source. Furthermore, these analysis showed different results depending on the variables used. For this reason not all economic variables derived in the different sources have been considered at the same level of accuracy[12]. The main reasons were:

- Definitions are very similar but not identical;
- "Noise" introduced by the administrative purposes of data collection from different sources (errors in the sources);
- "Noise" introduced by the statistical definitions that are not easily attributed to administrative definitions (errors in the survey).

Then, the choice was made taking into account two conditions:

a. The "importance" of variables with an high degree of accuracy;
b. The number of accurate variables.

Based on these two conditions, a list of source priorities was identified for deciding which source to use if more sources had data for the same unit: BIL showed a better accuracy to the survey data and this result was easily predictable because PMI variables definitions are based on Financial Statements legislation. In addition, BIL represents the richest source in terms of the number of variables.

Despite a smaller number of usable variables than BIL, SDS was a rich source considering the number of statistical units: more than 3.5 million units. The advantage of this source, compared to UNI, was that data are derivable from the same section (Section F) for all units, in spite of differences of firms activities; the only exception is the professionals who fill out the section G. Then, SDS was identified as an alternative source to BIL because of its uniformity and its relative richness of information.

The UNI source was used as third choice in case of lack of the first two sources: the reasons are to be found in the high heterogeneity of the variables, in a number quite limited of information and in the difficulty of standardizing the data because of the type of firm (individuals, partnerships and corporations), which filled different forms also because of the used accounting system.

For the year 2011 the IRAP source was acquired only when the making of the Frame was already nearing completion. Because the accuracy of it with PMI data has not been discussed in depth, it was used only as a last resort in case of absence of any other source and only for the corporations, because the analysis performed for the integration of non-response in the SBS survey about enterprises with 100 employees and over (SCI) showed a good accuracy with survey data.

In a second time, more analyses were conducted on other type of enterprises too: the results were the accuracy depends more on Section than on Form. In fact a very good accuracy for data coming from Section II of both forms IP and IQ, a good accuracy from Section I, a bad accuracy from Section V. These analysis permitted to change the order of

[11] For more details, refer to Di Zio M., Guarnera U., Varriale R. 2016. "The estimation of the main variables of the economic account of small and medium enterprises based on administrative sources". Rivista di Statistica Ufficiale. N.1/2016.

[12] For more details, refer again to: Curatolo S., De Giorgi V., Oropallo F., Puggioni A., Siesto G. 2016. "Quality analysis and harmonization issues in the context of the SBS frame". Rivista di Statistica Ufficiale. n.1/2016.

priority of sources for the next years, although a great problem of using IRAP source is that the stability of information is not ensured because it is more prone to policy decisions than other sources. However, in 2012 and 2013, IRAP source entered in the construction procedure as the others.

In conclusion, the order of priority used was the following:

1. BIL, if there is not then
2. SDS, if there is not then
3. UNI or IRAP, depending on sections and forms.

As previously said, for the enterprises as defined in the Business Register (ASIA) that are not present in any of the considered sources, it was needed a step of statistical integration of the missing information, relying on information from the RACLI source (for the personnel costs) and ASIA with regard to the turnover, derived from “Value Added Tax”, VAT, in particular.

1.5 The SBS Frame

After the correction for the remuneration of the atypical and independent workers (see § 1.1) and the harmonization of variables, the next step was the record linkage by the ASIA identification code to derive data from administrative sources (BIL, SDS, UNI, IRAP) referring to the SBS universe (in particular, PMI universe), discarding all information relating to non-active or not present in ASIA units.

The choice of data for every enterprise was made according to the order of priority as defined above. At the end, information from RACLI were linked too.

Some "exceptions to the priority" have been considered: the first exception regards the non-solvable inconsistencies: if the causes of inconsistencies are impossible to find, the involved source for that record was considered missing. Another exception was identified in BIL: records with an accounting period other than 12 months were discarded: in these cases the sources SDS, UNI or IRAP are considered more reliable, because those are referred to the calendar year.

Another exception came from a subsequent analysis of per-capita, in particular the personnel costs per employee. The cases of too high and/or too different from RACLI per-capita advised to waive the priority and to use the source following in hierarchical order: if no data was available or consistent, the unit was considered as a non-respondent and the information were imputed.

The Figure 4 shows the data integration step, as previously described: this figure is important because it shows the moment when the four different sources were linked to identify the statistical information to be used in all the following steps. Before step, the administrative datasets were treated independently and only in this time the different information are processed together.

The result of this step is a data set containing for each unit in PMI universe, all possible information related to a set of economic variables; the resulting dataset is exemplified in Figure 5, where the colored cells indicate the presence of information: for each unit one, more than one or none administrative source were available and the choice of data was made according to the rules previously described.

Figure 4: Administrative data integration: the Frame_base dataset.

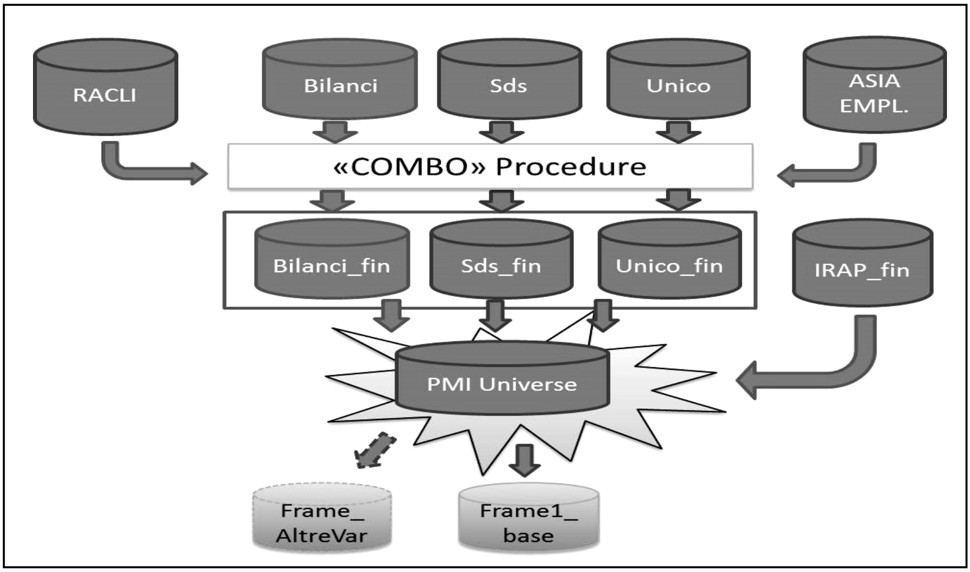
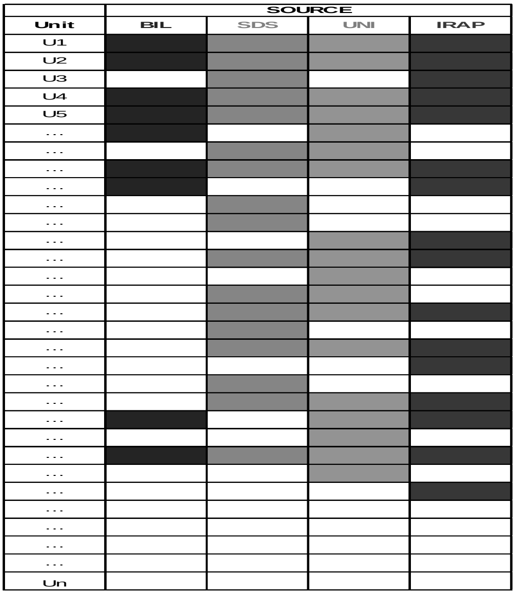


Figure 5: Exemplified representation of integrated dataset after linkage step.



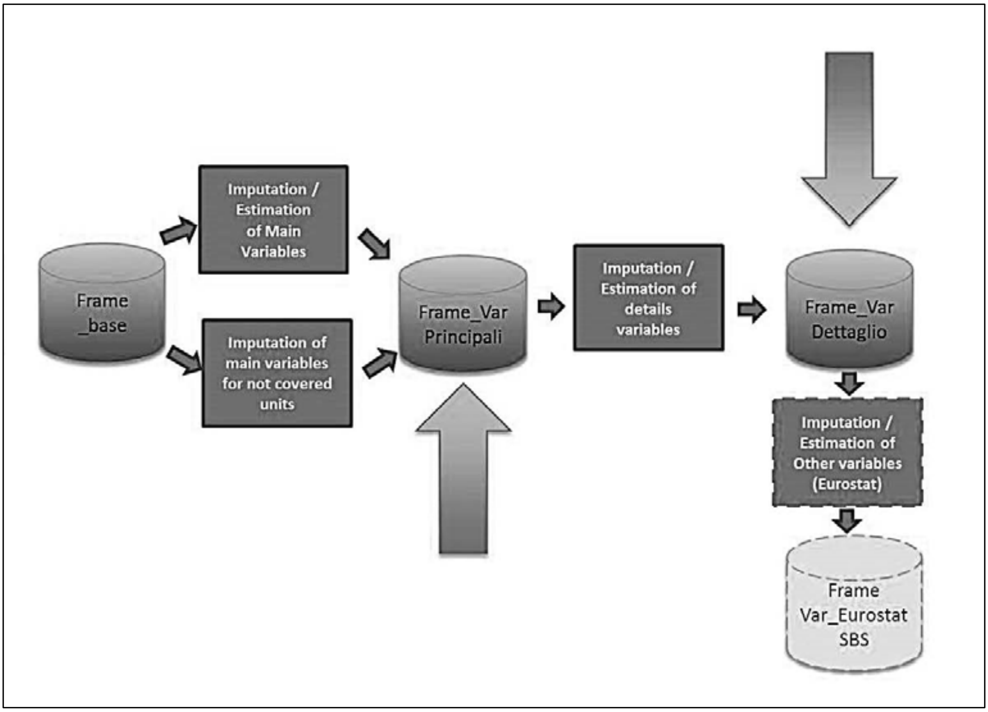
	SOURCE			
Unit	BIL	SDS	UNI	IRAP
U1	■	▨	▨	■
U2	■	▨	▨	■
U3		▨		■
U4	■	▨	▨	■
U5	■	▨	▨	■
...	■		▨	
...		▨	▨	
...	■	▨	▨	■
...	■			■
...		▨		
...		▨		
...			▨	■
...		▨	▨	■
...			▨	
...		▨	▨	
...		▨	▨	■
...		▨		
...		▨	▨	■
...				■
...		▨		
...		▨	▨	■
...	■		▨	■
...			▨	
...	■	▨	▨	■
...			▨	
...				■
...				
...				
...				
...				
Un				

At the end of this process for each unit - with at least one administrative source - was present a set of information provided by one and only one source; for the remaining units (without information available from administrative data) only the structural information deriving from ASIA and economic variable deriving from RACLI were considered: the resulting dataset is called *Frame_base*. This dataset represents the starting point for the construction of the SBS Frame, a complete set of information for the SBS estimates and for the National Accounts (CN) aggregates too.

A byproduct of the linking above is a dataset containing a set of variables useful for estimating the non-regular economy: on the contrary to the previous ones, the variables present in it, were not checked. The estimates of some other variables of detail for other revenue, made by the CN using the information from the "Notes" of BIL, are added too.

Starting from *Frame_base*, the next step is the imputation of missing values using both deterministic procedures and statistical ones, such as the predictive mean matching or the donor of minimum distance method[13]. The result was a new dataset called *Frame_VarPrincipali* where the values for all the main variables for all units of the Frame were available; it contained the main economic variables like turnover, costs, value added, personnel costs and so on. Also the details of personnel costs were present in it: those were estimated, if necessary, considering the structure provided from RACLI.

Figure 6: Final steps for the making of the Frame.


Frame _base
Imputation / Estimation of Main Variables
Imputation of main variables for not covered units
Frame_Var Principali
Imputation / Estimation of details variables
Frame_Var Dettaglio
Imputation / Estimation of Other variables (Eurostat)
Frame Var_Eurostat SBS

[13] For more details, refer to: Di Zio M., Guarnera U., Varriale R. 2016. “The estimation of the main variables of the economic account of small and medium enterprises based on administrative sources”. *Rivista di Statistica Ufficiale*. N.1/2016.

At the end, the last step was the estimation of the details of revenue or costs not previously calculated from *Frame_VarPrincipali*; the result was another dataset called *Frame_VarDettaglio*. The variables was estimated by the "projection estimator"[14] using survey data from PMI.

The Figure 6 shows the final steps of making of the Frame: the arrows indicate the datasets currently available to determine SBS statistics; they are the starting point for NA aggregates estimation. The Frame is available for years 2011, 2012 and 2013; for 2010 is available a prototype version.

The Table 2 shows the number of units in the Frame for the year 2012 according to the source of origin of the data: it is also shown the contribution that each source provides to the total of the number of employees and of some economic variables (turnover, value added and labour cost). The table has been derived from the elaboration of the Frame after the step of imputation and estimation of the main variables; in fact, also it shows the estimated values for not covered units. For 2013 the results are similar.

1.6 Future perspectives

Since 2012, the SBS Frame, with SCI survey, has been the core of SBS estimates sent to Eurostat: the main economical aggregates (e.g. turnover, added value, labour cost, etc.) are built simply using the "sum function"; instead to obtain other aggregates (e.g. detail of costs, detail of turnover, etc.) is necessary to apply a statistical method for estimation: also in these cases the SBS Frame provides data for each unit but they are significant just within the estimation domains of projection model.

Table 2 - Number of units and main economic aggregates of the Frame for source of data. Year 2012. (to be continued).

SOURCE	Number of enterprises	%	Number of employees	%	Turnover (mln €)	%	Value-Added (mln €)	%	Personnel costs (mln €)	%
BIL	707,167	16.0	4,572,661	37.8	1,138,480	64.4	221,203	53.2	139,348	66.4
SDS	2,985,929	67.5	6,064,848	50.2	433,273	24.5	153,962	37.0	51,612	24.6
- Section F	*2,280,386*	*76.4*	*5,093,983*	*84.0*	*378,115*	*87.3*	*113,779*	*73.9*	*46,952*	*91.0*
- Section G	*705,543*	*23.6*	*970,865*	*16.0*	*55,159*	*12.7*	*40,183*	*26.1*	*4,660*	*9.0*
UNI	542,721	12.3	722,050	6.0	34,369	1.9	13,648	3.3	3,190	1.5
- Professionals and self-employed										
- Form RE	*66,270*	*12.2*	*77,421*	*10.7*	*3,628*	*10.6*	*2,545*	*18.6*	*243*	*7.6*
- Form RF	*357*	*0.1*	*813*	*0.1*	*76*	*0.2*	*21*	*0.2*	*10*	*0.3*
- Form RG	*200,539*	*37.0*	*295,174*	*40.9*	*13,273*	*38.6*	*3,552*	*26.0*	*1,428*	*44.8*
- Form CM	*230,026*	*42.4*	*223,725*	*31.0*	*4,416*	*12.8*	*3,283*	*24.1*	*6*	*0.2*

[14] For more details, refer to: Righi P. 2016. "Estimation procedure and inference for component totals of the economic aggregates in the new Italian Business frame". *Rivista di Statistica Ufficiale*. N.1/2016.

Table 2 (continues): Number of units and main economic aggregates of the Frame for source of data. Year 2012.

SOURCE	Number of enterprises	%	Number of employees	%	Turnover (mln €)	%	Value-Added (mln €)	%	Labour costs (mln €)	%
UNI										
- Partnership										
- Form RE	*1,963*	*0.4*	*9,832*	*1.4*	*1,335*	*3.9*	*725*	*5.3*	*99*	*3.1*
- Form RF	*284*	*0.1*	*2,209*	*0.3*	*392*	*1.1*	*84*	*0.6*	*49*	*1.5*
- Form RG	*38,335*	*7.1*	*89,226*	*12.4*	*3,732*	*10.9*	*901*	*6.6*	*548*	*17.2*
- Company										
- IAS	*698*	*0.1*	*4,454*	*0.6*	*3,645*	*10.6*	*1,393*	*10.2*	*254*	*8.0*
- Others	*4,249*	*0.8*	*19,196*	*2.7*	*3,873*	*11.3*	*1,144*	*8.4*	*552*	*17.3*
IRAP	78,667	1.8	397,482	3.3	124,693	7.1	18,285	4.4	9,956	4.7
- Professionals and self-employed										
- Form IQ, Sect.I	*17,998*	*22.9*	*62,485*	*15.7*	*7,945*	*6.4*	*1,344*	*7.4*	*936*	*9.4*
- Form IQ, Sect.II	*2,467*	*3.1*	*10,893*	*2.7*	*1,658*	*1.3*	*341*	*1.9*	*199*	*2.0*
- Form IQ, Sect.V	*484*	*0.6*	*934*	*0.2*	*44*	*0.0*	*15*	*0.1*	*7*	*0.1*
- Partnership										
- Form IP, Sect.I	*15,984*	*20.3*	*87,989*	*22.1*	*15,638*	*12.5*	*2,205*	*12.1*	*1,599*	*16.1*
- Form IP, Sect.II	*3,437*	*4.4*	*27,432*	*6.9*	*7,667*	*6.1*	*1,184*	*6.5*	*658*	*6.6*
- Form IP, Sect.V	*60*	*0.1*	*242*	*0.1*	*12*	*0.0*	*4*	*0.0*	*2*	*0.0*
- Company (Form IC)	38,195	48.6	206,760	52.0	91,698	73.5	13,182	72.1	6,526	65.5
- Economic Public Corporation (Form IK)	42	0.1	746	0.2	29	0.0	10	0.1	29	0.3
NOT COVERED	109,489	3.9	328,388	5.1	37,038	6.2	8,559	5.3	5,725	6.3
- Professionals and self-employed	*47,219*	*43.1*	*89,488*	*27.3*	*6,393*	*17.3*	*1,594*	*18.6*	*807*	*14.1*
- Partnership	*13,145*	*12.0*	*40,311*	*12.3*	*3,357*	*9.1*	*846*	*9.9*	*443*	*7.7*
- Company	*40,787*	*37.3*	*148,503*	*45.2*	*20,709*	*55.9*	*4,645*	*54.3*	*3,273*	*57.2*
- Cooperative society	*5,667*	*5.2*	*40,701*	*12.4*	*3,528*	*9.5*	*921*	*10.8*	*735*	*12.8*
- Consortium	*1,196*	*1.1*	*3,019*	*0.9*	*667*	*1.8*	*127*	*1.5*	*106*	*1.8*
- Economic Public Corporation	*183*	*0.2*	*2,928*	*0.9*	*584*	*1.6*	*107.0*	*1.2*	*94*	*1.6*
- Foreign firm	*1,292*	*1.2*	*3,439*	*1.0*	*1,801*	*4.9*	*319*	*3.7*	*267*	*4.7*
TOTAL	**4,423,973**	**100**	**12,085,428**	**100.0**	**1,767,852**	**100.0**	**415,658**	**100.0**	**209,830**	**100.0**

As a new experience of massive use of administrative data for statistical purpose, the SBS Frame needs some years for adjusting the methods and the techniques; however, at this time the results are encouraging because the analysis of the differences between the Frame and the PMI estimates shows a similar behaviour between one year and the previous.

In recent years, ISTAT has invested a lot of skills (statisticians, IT resources etc.) to guarantee a long-term solution for the Frame construction process: some workgroups and task-forces have arranged to study different solutions for improving all the steps of the process, both about methodological and IT issues. This activity needs a continuous development and support to ensure an increasing quality of SBS estimates.

Also, the Frame will have to ensure the possibility of provide estimates for variables not available at this time: the use of other administrative sources (i.e. VAT or Financial Statements notes) could help to derive, for example, estimates about investments. The short-term statistics could be used too, for example to estimates the number of worked hours; this solution probably will be implemented in Frame 2014 while for the estimate of investments it is probably necessary at least one more year for properly studying the new sources.

Other activities are focused on improving the editing method introducing the use of probabilistic methods too, or on modifying the imputation techniques for units without administrative data (e.g. estimation of turnover by VAT using regression models in homogeneous domains); furthermore, a systematic selective editing procedure for the most influent units will soon be introduced in the process, as new analysis on increasing the quality of the choice of the administrative source when more sources are available will be further detailed.

Moreover, for the next Frame occurrences, it will be strategical to have a solid IT architecture to ensure the safety and the replicability of the process: this is a long-term activity but it will be one of the most important task in the future (in fact, it is currently in progress).

2. The warehouse system supporting Frame production

In statistical production the statistical burden is one of the main problems to be faced. To cope with the problem the two main directions are: the intensive use of administrative data and the reduction of the stovepipe production models. In fact, in a stovepipe model different surveys are totally independent from each other in almost every phase of statistical production processes.

Using administrative data often means to elaborate several large archives of data. In this case, an IT Data Warehousing (DWH) model[15] could be a suitable approach. In particular, we can consider a Statistical-Data Warehouse (S-DWH), i.e. a DWH specialized and optimized for producing statistical information and for data reuse (by storing data once but using it for multiple purposes). This means that we could manage several production processes for different topics in the same statistical domain, i.e. a data base for a single environment in which we could have data integration and sustain process integration. In fact, the data model underlying a S-DWH is not oriented to produce specific reports as well as for on line analytical processing. Instead of focusing on a process-oriented design, the underlying repository design is based on data inter-relationships that are fundamental for different processes of a common statistical domain.

[15] Inmon, Bill (1992).Building the Data Warehouse. Wiley. ISBN 0-471-56960-7

The S-DWH data model is based on the ability of realizing data integration at micro or macro data granularity levels: micro data integration is based on the combination of different data sources with a common unit of analysis, while macro data integration is based on integration of different aggregate information in a common estimation domain. In this paper, we will consider only micro data integration in order to support the complex procedure of the Frame production in which the integration of different sources of micro data is one of the crucial steps in the process.

In fact, the production process is articulated in a number of different phases or sub-processes. Schematically, each phase collects some input variables and produces some output variables. One way to find a common ground between different statistical actions is to focus on a generalized data input interface in which it is possible to identify and select the variables needed for data processing in each production phase. Adaptation of the data input and output interfaces of each phase of a process gives us the opportunity of managing the elaboration phases by using generic software components, i.e. using almost any statistical editing tool in a common application framework. The adaptability of the data input/output interfaces to the procedures are particularly helpful in statistical production based on administrative data when the input data layout and variable meanings are not under the direct control of the statistical producer, so that they can change for each supply due to national regulation changes.

In this way, the production of the Frame can be seen as a workflow of separated activities, which must be realized in a common environment where all the statistical experts involved in the different production phases can work. In such an environment the role of knowledge sharing is central and this is sustained by the S-DWH in which all information from the collaborative workflow is stored.

From an IT point of view this corresponds to a workflow management system able to sustain a data centric workflow of activities (scientific workflow[16]), i.e. a common software environment in which all the statistical experts (or data scientists), involved in the different production phases of the same process, work by testing hypotheses.

The Workflow management system then allows a controlled process through the standardization of working methods in a flexible environment. On the other hand, a scientific workflow based on a S-DWH can increase efficiency, reducing the risk of data loss and integration errors by eliminating any manual steps in data retrieval.

2.1 Data integration

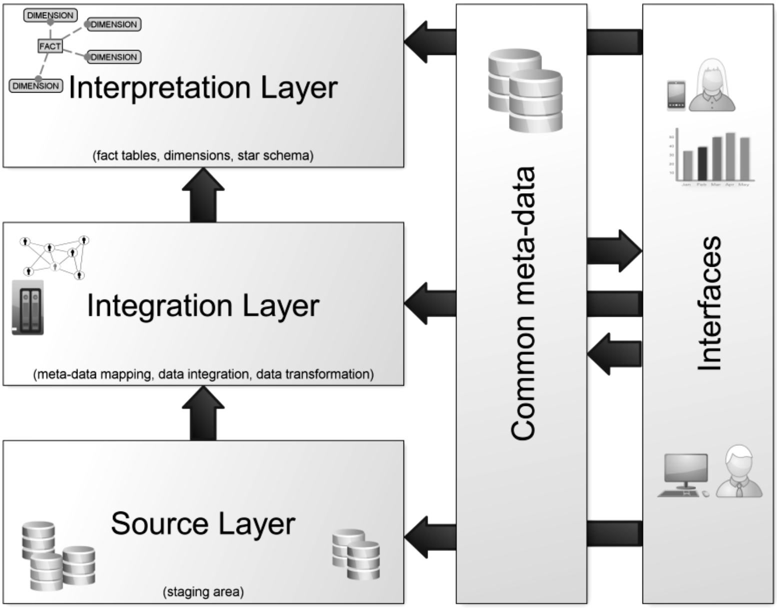
One of the most important activities in the process of building the SBS Frame is represented by *data integration*. This term refers to a set of activities aiming at creating a unifying view of data information coming from different and heterogeneous sources.

Often data integration problems are overcome through ad hoc approaches: each instance of the problem is treated case by case. In the Frame context, the key element to achieve this goal is given by a S-DWH.

[16] G. Scherp. A Framework for Model-Driven Scientific Workflow Engineering 05/2010 Procedia Computer Science. [17] N. Russell, A. Ter Hofstede, D. Edmond, W. van der Aalst. 2005. Workflow data patterns. In Proc. of 24th Int. Conf. on Conceptual Modeling Springer. Verlag: october

Statistical data warehousing systems extract, transform, and load data from different heterogeneous statistical data sources into a single schema so that data become compatible with each other and can be processed, compared, queried or analyzed regardless of the original data structure and semantics.

In classical DWH concept, it is first necessary to set up a database and a system that replicates data from production databases to the DWH's one. In addition to this, some tools to produce reports and view data are needed. All of the tools have to use and display metadata information.



The S-DWH is organized into three contiguous but essentially disjoint areas (the three boxes on the left in the picture above):

- the *source layer* contains the staging area. Here the supply of raw data from external sources takes place in the form of the *data provision* (which includes the set of elementary data, metadata, and classifications of a source for a given reference period). In this area also some preliminary consistency check operations are performed. This processing, called *provision acceptance*, allows to determine the completeness and the consistency of a given data supply;

  the *integration layer* deals with several Extraction Transformation and Loading (ETL) activities. Among them one can enumerate: finding and correcting inconsistent data, transforming data to standard formats in the DWH, classifying and coding, deriving new values, cleaning data and mapping the variables in the sources record layouts to the variables in the DWH dictionary. This layer, thanks to a suitable organization of metadata and thanks to the support of common metadata, allows decoupling the internal structure of the data sources from the interfaces. Thus, the analysts are able to perform queries on the data sources based on the dictionary of the concepts of the DWH regardless of the structure of the single sources;

- the *interpretation layer* contains, in the form of dimensions and fact tables, the elementary data coming from the ETL operations performed in the previous layer. Only data related to variables of data provisions, that have been associated to the S-DWH dictionary items in the integration layer, can be accessed by analysts for data mining purposes.

The three layers are accessible by suitable interfaces through a layer of common metadata. As far as this layer is concerned, the S-DWH metadata consist of many logical elements. The first one is the data provision component which represents the part that provides information on the data source available for the subsequent elaborations. It is defined by the collection of data, metadata such as classifications, record layout and other objects. The second element is the Statistical dictionary, also called dictionary of meta-attributes, which contains the definitions of the statistical objects to be used in order to perform data analysis or data mining. Another important component of the common metadata architecture is the meta-source abstraction. It represents associations (mappings) between meta-attributes (i.e. the interpretation layer metadata) and elementary variables collected by the sources. It thus helps giving a support for interpreting the information of the data supply. Two of the main advantages of the mapping component are given by a direct possible association between information and statistical data dictionary and by different interpretations of the same variable as the content of the variable itself changes in time. In particular, the interpretation layer can be accessed by an interface through the metadata represented by the S-DWH statistical dictionary.

Differently from a classical DWH, in which data flow only from one layer to the subsequent one, the S-DWH system can play a key role in circular processes of microdata re-use in the following way:

- it is employed (through the interpretation layer) to extract linked raw or cleaned micro data for specific variables from the internal and external sources to be used as input for a new editing and imputation statistical process. This constitutes a data transformation process which takes place by an asynchronous elaboration and uses the S-DWH as input/output data repository;
- part of the output of the new statistical process, i.e. the one corresponding to validated micro-data (with associated metadata and classifications), goes to feed the S-DWH through the source layer. The dictionary of S-DWH (in the metadata layer) is then enriched with meta-attributes created during the new statistical process.

This circular process, in the framework of a S-DWH, sustains the production of the SBS Frame. This type of process can be supported using a blackboard design pattern's paradigm; i.e. a shared area (the blackboard) which can be accessed by autonomous processes or actors in some coordinated and cooperative way. Generally, this working scheme allows to achieve complex goals that require multidisciplinary skills.

The first step consists in setting up a S-DWH on the domain of Structural Business Statistics: it contains data coming from different kinds of sources, administrative data, registers and surveys, both internal and external to the National Statistical Institute (NSI).

Subsequently, all available information are analyzed in the S-DWH and the variables, needed for the process of construction of the SBS frame, are chosen from the S-DWH and extracted for any required elaboration. In fact, the resulting data enter the statistical process of construction of the SBS frame and end up in an output, the Frame itself, which can be loaded in the S-DWH, becoming itself an internal source (besides being a byproduct of the integration of other sources).

More in detail, at the source layer level, one can find:

1. NSI surveys data;
2. NSI Statistical Business Register data;
3. Admin data.

In particular, the source layer of the S-DWH under implementation contains: the raw admin data coming from the Italian Revenue Agency (Income Tax Returns, "Studi di Settore") and from the Union of Chambers of Commerce (Financial Statements), the NSI Statistical Business Register (ASIA) data and the validated Small-Medium Enterprises survey data.

In the source layer some preliminary consistency checks between micro-data and record layout metadata are performed which make the data provisions ready to enter the subsequent layer for data integration. In the integration layer, after some ETL operations, data from the different sources are reconciled and mapped against common statistical concepts and subsequently loaded in fact and dimension tables in the interpretation layer. Here, data are thus ready for querying and analysis.

During the researchers' activities it is often necessary to inspect and view data in order to get detailed information on the survey units.

2.2 Process integration

In the previous paragraph some of the issues arisen in the construction of the SBS Frame have been discussed: in particular we focused on data and their integration. On the other hand, this paragraph deals with the integration of processes, tasks and all the elements needed for the Frame construction. Naturally, data integration and process integration are strongly intertwined.

The building process of SBS Frame can be defined as "scientific workflow" (SWF). This type of workflow is characterized by a strong dependence on data (data-centric WF), while a classic WF focuses more on processes than on the objects to manipulate (flow-centric WF). In fact, industrial (classic) WF is used when the process is not subject to substantial changes along its lifecycle and its number of instances (iterations) is large. Conversely, the scientific WF is characterized by frequently modified processes and by few instances.

One example of trial-and-error approach supported by SWF in the Frame construction, is the search of duplicated records in the UNICO source. Very often the sources are little documented or they are not documented at all (regarding to the statistics domain); therefore the only strategy researchers can use is to try some hypothesis on data. Subsequently it is necessary to test the hypothesis but effects could be observed only at the end of the sub-process or even afterwards. If the test succeeds, it is possible to move on; otherwise it is necessary to go back and try another strategy or modify some element of the existing one in order to remove all the duplicates.

Another great difficulty met in the production of the SBS Frame consists in collecting and organizing all the elements which concur to the results. As an example, every year some administrative sources metadata change and therefore several methodological adaptations to processes and procedures could be necessary.

Another example could be represented by a variable semantic change: this kind of event affects not only the mapping of the variables, but also check and correction rules.

These two simple examples are enough to show that the SBS Frame production involves different heterogeneous skills and tools (each sub-process, particularly statistical ones, uses its own instruments). Besides the issues outlined above, the huge amounts of data and strict time constraints involved, make the building process of the SBS Frame particularly complex and critical to manage.

In order to efficiently organize the WF[17] with the aim to support the production processes and improve quality, it is necessary to connect several entities such as the source variables and the related documentation. It is also important to gather and record the versions of any entity in order to fully document the process and guarantee its quality, reliability, replicability and reusability. A systematic collection of all the tested attempts could also contribute to the production efficiency because the researcher's team would be able to examine all past discarded hypotheses.

In the next paragraphs, it is described the design of an integrated environment for setting up, executing and documenting the Frame production process. This is articulated in following item functionalities:

- *design and management of a statistical workflow*. It allows designing, modifying and executing the main phases, sub-processes and elementary activities which constitute the statistical production process;
- *activities and processes schedule.* The activities, the remote processes and the procedures can be run by a scheduler in an automatic way. This is particularly useful when one deals with huge amounts of data. The scheduler's purpose is to translate the workflow design into a sequence of activities to be submitted to the distributed processing nodes. This sequence has to satisfy priority constraints planned during the design phase;
- *local and remote services call.* Each elementary activity can be either a native procedure (e.g. a SAS procedure, a *PL/SQL* program or an R procedure) or an external service, such as a web service encapsulating a high-level domain service (i.e. *BANFF*) that can be invoked from the platform. It is necessary to provide some mechanism of sharing information between systems;
- *integration of statistical abstractions*. A statistical production process has its own rules, constraints, methodologies and paradigms. The aim of the statistical abstraction layer is to supply a set of abstractions that make the researcher's work flexible, independent of technical details and more focused on research objectives. Among the possible abstractions there could be:
    - *meta-parameters*: the use of global parameters reduces the need to modify the scripts and variables necessary for other systems to operate correctly;
    - *partitioning or filtering units*: each type of record (unit) has its own processing path in the WF. The value of some variable could be used to filter units to the

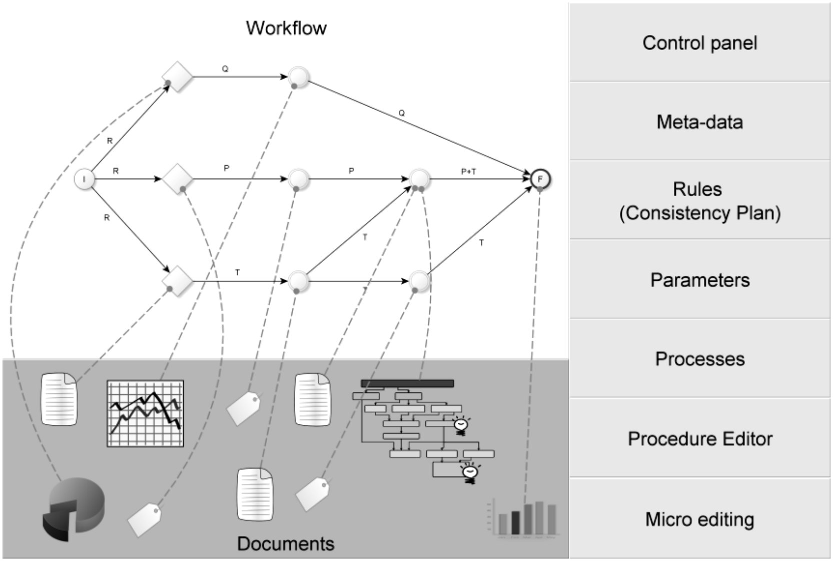
[17] N. Russell, A. Ter Hofstede, D. Edmond, W. van der Aalst. 2005. Workflow data patterns. In Proc. of 24th Int. Conf. on Conceptual Modeling Springer. Verlag: october

next processing step;

- *sampling test*: when the amount of data is very large, it is useful to test some hypothesis or programs on a subset of data in order to avoid loss of time and to early discover weak hypotheses;
- *rule checker*: a tool for finding inconsistencies in a formally defined set of rules and to manage efficiently semantic and definitional changes in sources;

- *documentation management and versioning. It is possible to associate one or more* documents and metadata to each WF element and, at any time, recall previous versions of the WF and all the elements connected.

The following picture presents a WF architecture that summarizes concepts and paradigms presented above.



The upper left box represents the key concept of the architecture. The WF is graphically represented by a custom graph and consists of sub-processes and domain modules. Each element can be linked to one or more documents (the bottom left box), for example: charts, reports, regulations, metadata and parameters, record layouts, text documents, etc. as well as information related to the execution of the sub-processes (actual and mean execution time, error and warning reporting, results, previous and subsequent sub-processes, etc.).

The right area contains “services” that the researchers could use to support most of their activities; in the following they are briefly described from top to bottom:

- *metadata* module implements a decoupling approach in data mapping. This type of abstraction introduces a new layer between data sources and statistical variables so that a semantic change in one administrative source does not affect statistical sub-processes that depend on the related statistical variables;
- *rules* module allow the researcher to write the consistency plan, check possible contradictions in the edits set, run the plan, log error and warnings and produce reports.

Moreover, this module assists the researcher in the activity of modifying an existing check plan in case some variables are introduced or deleted;

- *parameters* module is used to implement a basic form of parametric changes in all of the components of the WF. It can be thought as similar to a dashboard through which modifying thresholds, setting parameters, choosing elaborative units, switching on and off options, etc. For instance, suppose one parameter is shared by many sub-processes: a change in this value has an impact on all the sub-processes containing that parameter. The parameter is a placeholder that at runtime is set to the actual value (e.g. some sub-process can possibly change the parameters' value during processing);
- *processes* module provides information on actual state of active elaborations. It is possible to view the scheduled sequence of sub-processes and to recall the log of previous ones;
- *procedure editor* module is the development environment needed to create procedures or modify existing code. Such a module should support at least one statistical language (SAS, R) and one data manipulation language (PL/SQL). New languages can be added to this system in a modular and incremental way. The editor integrates a versioning system in order to restore a previous version of a procedure, document code changes and to monitor the improvements of the implemented functions;
- the *micro-editing* component is used in manual and interactive micro data editing activities. It can be a useful tool for statisticians to analyze some sample of micro-data.

Conclusions

In 2012 for the first time, ISTAT sent to Eurostat the SBS estimates using massively administrative data to derive the main SBS aggregates: this experience has shown that it is possible to have a lot of business information also without the use of direct surveys and if this information is appropriately supported by statistical methods it can be used to cover the most important aspects of enterprises' results. Of course the current process of Frame construction could get better: in particular, some steps could be improved using better and/or newer techniques, more efficient statistical methods or more generalized programs or software. But the confirmation of the quality of the results also for 2013 represents the most encouraging factor: accordingly, the SBS Frame is becoming the reference product for all the structural business statistics.

Furthermore, the analysis of the SBS Frame production system has shown the need to define a new informative infrastructure, which allows users to interact with a number of administrative data sources, and to yearly adapt and modify the Frame production procedures. The proposed infrastructure, that relies on concepts and paradigms of scientific workflows' management systems, involves the construction of an environment for process modelling, flexible and integrable with the standard statistical processing tools, and a data warehouse of microdata. The latter has been implemented for the specific domain of structural business statistics and allows to: manage the variable mapping, reconcile and follow definitional changes of the different sources in time, as well as build many customizable environments supporting the workflow process or data analysis.

References

Corsini V., T. Di Francescantonio, S. Filiberti, R. Sanzo. 2000. *Utilizzo integrato di fonti amministrative e fonti statistiche per la produzione di stime preliminari di alcuni principali aggregati economici previsti dal regolamento della Ue n.58/97 sulle statistiche strutturali*. ISTAT (Documento interno UDAS.10.00.3).

Dabbicco G. 2002. *Utilizzo dei bilanci aziendali civilistici ai fini del soddisfacimento del regolamento UE 58/97 sulle statistiche strutturali sulle imprese*. ISTAT (Documento interno UDAS.03.01.1).

De Carli R. 2003. *Integrazione tra dati statistici e dati amministrativi sui risultati economici delle imprese: prime evidenze dal confronto tra i dati individuali delle principali rilevazioni statistiche strutturali, di quelli desumibili dai bilanci civilistici e di quelli derivanti dalle dichiarazioni fiscali*. ISTAT (Documento interno UDAS).

Monducci R., G. Dabbicco, C.M. De Gregorio, T. Di Francescantonio, S. Filiberti, U. Sansone, R. Sanzo, A. Volpe Rinonapoli. 2003. "Prime esperienze sull'utilizzo integrato di fonti statistiche ed amministrative per la produzione di statistiche strutturali sui risultati economici delle imprese". In *Temi di ricerca ed esperienze sull'utilizzo a fini statistici di dati di fonte amministrativa*, a cura di P.D. Falorsi, A. Pallara, A. Russo. Milano: Franco Angeli.

Siesto G., F. Branchi, C. Casciano, T. Di Francescantonio, P.D. Falorsi, S. Filiberti, G. Marsigliesi, U. Sansone, E. Santi, R. Sanzo, A Zeli. 2006. *Valutazione delle possibilità di uso di dati fiscali a supporto della rilevazione PMI.* ISTAT (Documenti, 17).

Grazzi M., R. Sanzo, A. Secchi, A Zeli. 2009. *MICRO 3-ISTAT: A new integrated system of business micro-data 1989-2004*. ISTAT (Documenti, 11).

Grazzi M., R. Sanzo, A. Secchi, A Zeli. 2013. "*The building process of a new integrated system of business micro-data 1989-2004". Journal of Economic and Social Measurement,* 38: 291-324.

Sanzo R., D. Bellisai, T. Di Francescantonio. 2014. "Il processo di produzione del FRAME". Relazione presentata al Workshop scientifico: "Il nuovo «frame» delle statistiche sulle imprese: innovazioni metodologiche, uso delle fonti, risultati e potenziale informativo", Roma, 25 marzo 2014.

Buschmann F., R. Meunier, H. Rohnert, P. Sommerlad, M. Stal. *Pattern-Oriented Software Architecture, A System of Patterns*.

Inmon Bill. 1992. *Building the Data Warehouse*. Wiley.

Scherp G. *A Framework for Model-Driven Scientific Workflow Engineering.* 05/2010 Procedia Computer Science.

Russell N., A. Ter Hofstede, D. Edmond, W. van der Aalst. 2005. *Workflow data patterns.* In Proc. of 24th Int. Conf. on Conceptual Modeling Springer Verlag: October 2005.