INS integrated architecture: pilot application in external trade statistics and methodological improvements in data processing

M. Bruno, M. S. Causo, A. Najjar, G. Sindoni, T. Tkitek, C. Vaccari*

This paper describes one of the actions implemented in the framework of the twinning project "*Modernisation de l'appareil statistique tunisien*", namely the introduction of a standard IT architecture for statistical processes and its application in the external trade statistical pilot domain. The architecture covered all the Generic Statistical Business Process Model macro-phases, offering an opportunity to introduce methodological improvements in the INS external trade GSBPM Metadata and Quality Management phases. The new integrated IT architecture was designed with INS experts and focuses on the "core" production process, with the aim of standardizing and streamlining the data production phases by (i) enhancing the adoption of standardized metadata in the collection, processing and dissemination phases (ii) introducing a new methodological approach for selective data editing and automatic imputation based on robust statistical methods (iii) minimizing the need for manual intervention in data editing (iv) developing new IT procedures for outlier selection and imputation fully scalable to other statistical domains. The suggested methodological and architectural solutions are compliant with the standards adopted in the context of official statistics and scalable to different domains.

Introduction

The work described in this paper was performed in the context of the twinning project "*Modernisation de l'appareil statistique tunisien*" in which Insee, Istat and INS were involved for more than two years (2016-2018). The project was originally designed to respond to the need for reliable data, following three main lines of action:

1. Strengthen the governance of the statistical system;
- Improve the coordination of public statistical organizations;
- Establish a new legal and institutional framework;
- Adapt the regional organization of the statistical system to the new framework.

2. Reinforce data production by improving the quality of official statistics to meet international standards.

3. Ensure better data collection/dissemination by INS to support and monitor development policies.

To face these challenges, it is necessary to model the statistical processes according to official statistics international standards, such as Generic Statistical Business Process Model (GSBPM), Generic Statistical Information Model (GSIM) and Common Statistical Production Architecture (CSPA). Within the project Istituto nazionale di statistica (Istat) experts have designed a target architecture to support process standardization.

The proposed model has been tested in external trade statistics in order to harmonize scope, principles, concepts and definitions to the European framework (Eurostat, 2017).

Harmonization and standardization are the key elements for efficient production processes and allow

* Mauro Bruno, technologist, Istat, mbruno@istat.it, Maria Serena Causo, researcher, Istat, causo@istat.it, Anissa Najjar, statistician, Statistiques Tunisie, najjar.anissa@ins.tn, Giuseppe Sindoni, senior technologist, Istat, sindoni@istat.it, Tarek Tkitek, technologist, Statistiques Tunisie, tkitek.tarek@ins.tn, Carlo Vaccari, first level technologist, Istat, vaccari@istat.it

international comparability and coherence over time. In this project, specific actions have been carried out, such as:

1. Enhancing an effective inter-institutional cooperation with Customs Agency, in charge of external trade data collection;
2. Setting a system of standard commodity and geo-economic classification, to be reconciled as much as possible with the national Customs classifications;
3. Introducing standard metadata-driven processes based on standard IT tools;
4. Enriching the set of referential metadata, introducing editing parameters useful to streamline and automate data production;
5. Adopting procedures for automatic data imputation based on robust statistical methods.

More specifically, the first section provides an assessment of the current situation (AS-IS) at INS in relation to IT organization, software development and dissemination systems. The second one describes the proposed architecture (TO-BE) for each GSBPM macro-phase (Collect, Process and Disseminate) and a set of general principles and guidelines supporting INS in the transition towards the new architecture.

The next sections analyses in detail the External Trade use case, describing the improvements achieved in the external trade data production process and some open issues.

Finally, the roadmap for a complete implementation of the architecture is described. Short-term activities concern the extension of the pilot external trade system to other domains, while medium-term activities pertain to the following systems: i) metadata; ii) data collection; iii) data processing.

Towards an integrated architecture for statistical processes

State of the art at INS

To design an integrated architecture that fulfils the requirements of INS, a preliminary assessment of the current (AS-IS) scenario was performed. The technology scenario is quite heterogeneous, spanning from web development technologies like PHP and Java to .NET environments. Applications are based both on flat files and on relational databases. Different software technologies are used for statistical tasks such as data collection, classification, dissemination, etc. Concerning the External Trade statistical domain, the process was managed by a workflow engine, namely SQL Server Integration Services (SSIS). Important parts of the process, such as outlier detection and data editing, were performed offline by statisticians.

The main problems reported by the IT department are the short development cycles, lack of exhaustive definitions of requirements and lack of scope statements. Together with the above-mentioned heterogeneity, these challenges affect the quality and timeliness of the production processes. There are also planning issues in relation to software development, which is often disrupted by unforeseen but urgent activities. The heterogeneity of the software in use and the patchy IT skills of the staff concerned are additional concerns. There is currently no clear definition of user needs, and this leads to delays in software delivery and acceptance. The use of database technology in INS is generally low and the IT sector is not always involved in process analysis, in part due to the lack of an IT "culture" in the statistical domains.

Data dissemination is aimed at maximizing the quality and accessibility of the outputs. The global strategy is based on a web portal, built on top of data warehouses, which offers applications, services and open data to end users and international organizations. Data can be accessed by multiple devices and by machine to machine applications.

The various systems currently use different tools (Prognoz, Knoema, PHP), and clear strategic choices are needed in relation to dissemination tools and data warehouse technologies. There is also a need for standardized ETL (Extraction, Transformation & Loading) tools to manage the data flows, which are currently mainly managed by manual procedures.

Technical architecture for statistical production processes

The main objective of the new architecture is to standardize and improve the overall quality of data production processes by:

- Enhancing the set of indicators needed to guide and automatize the data editing phase;
- Minimizing the need for manual intervention in data editing;
- Introducing new IT procedures for editing and imputation that are fully extensible to other statistical domains.

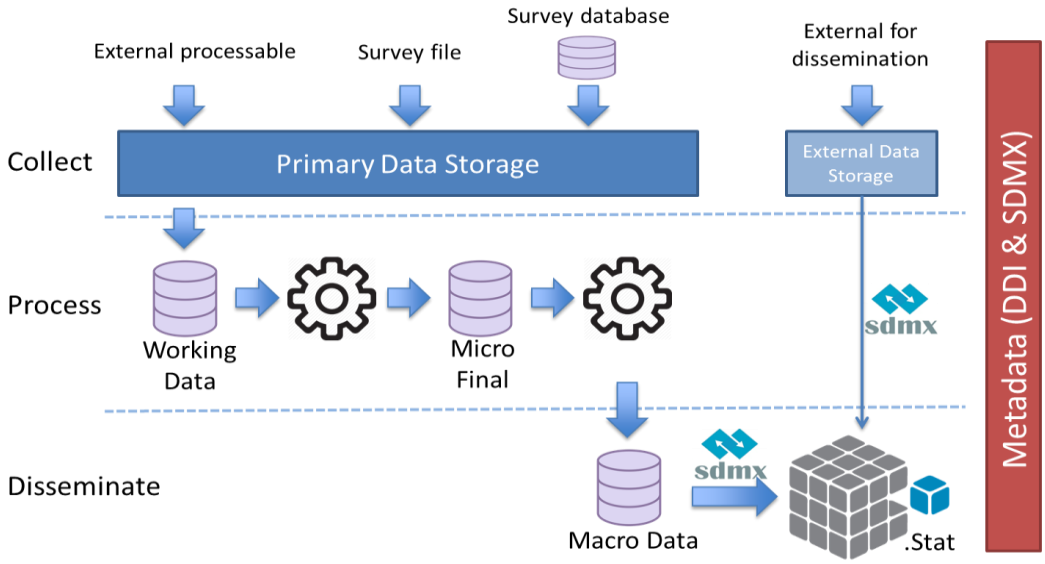
To implement a generalized architecture that fulfils the requirements of the different domains at INS, the following data repositories are needed:

- Raw data repository contains data provided to INS, whether through use of data collection capabilities or from external sources, such as administrative data;
- Working data repository: staging area needed for data processing;

- Metadata repository: metadata are the central element of the proposed integrated architecture and should be used in all phases of the production chain, from collection to dissemination (Signore and *al*. (2015) and Scanu and *al*. (2013));
- Dissemination repository: it was decided to use Statistical data and metadata exchange (SDMX) as the dissemination standard and T.STAT as the main dissemination portal, so the dissemination repository will be SDMX compliant by design.

A simplified model of the proposed architecture is shown below (figure 1). This framework covers three of the main phases of the statistical process, according to the standard GSBPM.

Figure 1:
Proposed integrated architecture



Collect [GSBPM Collect]

In the data collection phase different types of data are acquired. Depending on the data provider, data sources can be classified as:

- Internal data sources (e.g. direct survey collected using electronic questionnaires);
- External data sources (e.g. data from administrative sources, customs data for external trade).

Data can then be further divided, according to their treatment, into:

- Data for 'dissemination phase' (e.g. administrative data ready to be disseminated);
- Data for 'process phase' (e.g. customs data for external trade).

For each subset of data sources, the related building block can be identified[1]. Within this framework, the proposed building blocks are:

- Primary Data Storage: at a conceptual level this building block should store and manage raw data from different types of source. For example, data can be stored in relational databases to ensure consistency, efficiency and flexibility. The building block should also provide ETL functionalities to manage data transfer and capture from different channels. Preliminary treatments such as filtering, transcoding, normalization, translation and codification can be applied to facilitate data integration.
- External Data Storage: this building block should store data ingested from external sources and ready to be disseminated. It should also provide functionalities to integrate the ingested data with corporate INS metadata.

The described building blocks are closely related to the metadata building block (described below). In this stage, it is important for both elementary and aggregated data to capture metadata about the internal/external data sources (data provider, reference contact, data format, etc.).

Process [GSBPM Process & Analyse]

This phase includes all types of data treatment (e.g. cleansing, harmonization, validation) to transform the raw data stored in INS into statistical output for

[1] *According to the Enterprise Architecture Reference Framework (EARF), a building block is a potentially re-usable component that can be combined to build the information systems needed in ESS. Examples include: Metadata Management, Process Orchestrator, Primary Data Storage.*

statistical dissemination. The suggested building blocks in the proposed architecture are:

- Working Data Storage: at a conceptual level this building block should store and manage all the data transformations resulting from data treatment. It should also contain all auxiliary information (administrative data, benchmark data from other surveys, sample, etc.) needed to produce statistical outputs.
- Clean Data Storage: this building block should store and manage the 'final' clean microdata resulting from statistical operations performed in Working Data Storage.
- Aggregated Data Storage: this building block contains the data (e.g. indicators or multidimensional data) resulting from the aggregation of clean microdata.
- Workflow management: each operation performed in this phase is tied to a specific process step. This building block should therefore specify the sequence and the routing of the different process steps. The use of workflow management will ensure easy replication of statistical production across domains and hence minimize the cost of adjusting or expanding statistical production.
- Service catalogue: data treatment can be performed through the invocation of statistical services - programs implementing one or more statistical methods (extract sample, calculate weights, perform error checking, etc.) that can be invoked as a service. The service catalogue is the global repository that allows users to manage (search, insert, update) the available statistical services.

This framework covers three of the main phases of the statistical process, according to the standard GSBPM. In Figure 1 the cog  represents a statistical service, while the arrow symbolizes the process steps managed by the workflow management building block. In general terms, statistical services are available in the Service catalogue.

Disseminate [GSBPM Disseminate]

This phase is managed with the SDMX architecture and tools used to feed the ".STAT" dissemination portal. All data to be published, whether provided by external sources or produced as an output of the process phase, should be modeled in such a way to facilitate processing by the SDMX tools.

Architectural principles & guidelines

The following architectural principles should guide future activities according to Bruno and *al*. (2018) and Scannapieco and Vaccari (2011):

1. Metadata-driven approach: metadata are a key element in the proposed architecture. Standardized metadata should be used in questionnaire development, sampling, editing and imputation, etc. Whenever possible, codes should be dynamically generated from metadata;
2. Process and method standardization: harmonization of statistical methodology and IT components increases data quality;
3. Data storage: all data should be stored and managed in relational databases;
4. Data processing: data manipulation and flow control should be implemented using generalized software based on web applications working on databases;
5. Governance: shared governance across subject matter divisions ensures harmonization of concepts and minimizes survey-specific code.

Specific guidelines for each of the previously described building blocks and for metadata are provided below.

Data Storage

The building blocks used in relation to storage of the various types of data (primary/external/working data) should abide by the following guidelines:

- Each object stored in a data repository should be described in terms of structural/referential metadata;
- All the changes that have occurred to a data object over its lifecycle should be documented (data lineage).

Workflow management

The building blocks in the process and management standardization phase should be built to support the standardization and reuse of statistical methodologies (reduction of survey-specific procedures). They should:

- Permit scheduling of different process instances;
- Avoid overlapping of different user revisions;
- Provide a graphical user interface to (at least): (i) pass parameters to services; (ii) invoke services; (iii) monitor process execution; (iv) handle exceptions; (v) access (read/write) and visualize available data.

Service catalogue

The service catalogue should provide references to technical/ methodological guidelines and a link to IT services. Services in the catalogue should be Common Statistical Production Architecture (CSPA) compliant, meaning that they can be used as basic components in the production chain.

Metadata

The different phases of a statistical process should be described in terms of standardized metadata

(structural/referential)[2]. This cross-cutting building block should provide functionalities to model the inputs and outputs of the different services involved in statistical processes. Further metadata should be modelled according to the GSIM standard. This should facilitate the harmonization of the concepts and contents of the different statistical domains.

External trade statistical process in the INS Integrated Information System: methodological improvements in data processing

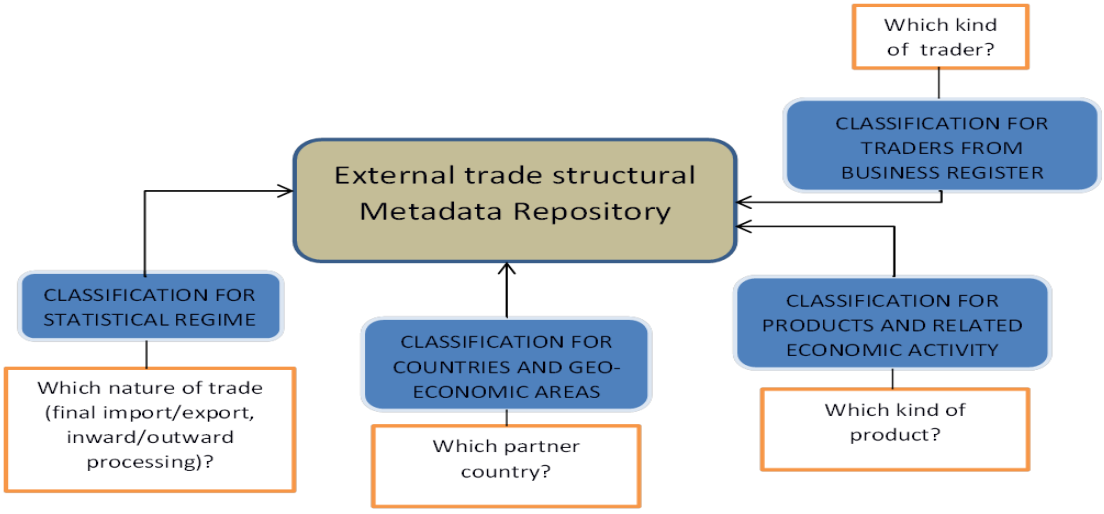
INS External trade statistics are based on Customs administrative data monthly transmitted by Customs Agency to INS. Raw data undergo different processing phases from data acquisition until dissemination of trade indicators. In the new data processing system developed during the twinning project, each process phase is metadata driven.

Structural metadata, such as variable definitions, identifiers, standard classifications, are exploited during each phase, from data acquisition until data dissemination, while referential metadata such as editing parameters are essential in data editing and error imputation phase.

The structural metadata classification scheme adopts, as much as possible, international harmonised classification systems, such as the Harmonized Commodity Description and Coding Systems (HS), the UN Central Product Classification (CPC) for products and Standard International Trade Classification (SITC), standard Geonomenclature based on ISO 3166 coding system for countries (Eurostat, 2017). However, a limit to the full adoption of international classifications is given by the administrative nature of trade data which serve both statistical and national Customs needs. For this reason, a strict cooperation between INS and Customs is needed to reconcile specific national definitions with international concepts and classifications, through appropriate correspondence relations.

Another important dimension which was proposed to be introduced as structural metadata is a trader register with traders' identifiers, to be matched with statistical business register. This is the key element to enlarge the scope of traditional trade statistics by enabling production of trade by enterprise characteristics (Eurostat, 2018). Such feature, not yet implemented, would require close collaboration with Customs and fiscal Authorities in order to enable appropriate linking of Customs traders tax identifiers to INS Business Register. The proposed structural classification metadata are represented in the figure 2.

Figure 2:
External trade structural classification metadata



Structural metadata undergo revisions in time. Therefore, the metadata repository requires both start and end date of validity defined for any classification modality, in order to easily apply appropriate metadata to trade flow time series.

[2] *Structural metadata describe the meaning of the data used, including the definition of a data element and a data set, variable names, variable types, unit identifier, classification identifier, etc. Referential metadata are the information objects necessary to run the process. These metadata contain the process flow and all parameters, rules and auxiliary data sets needed for the process steps involved.*

While structural metadata are updated with low frequency (standard classifications are typically revised every few years), referential metadata such as editing parameters need to be monthly updated, since they closely follow changes in trade characteristics. The way of calculating and updating these referential metadata is described in the section below.

Data acquisition phase

This is a metadata-driven process phase. Data received from Customs Authorities are stored in the INS Primary Data Storage. A preliminary treatment, consisting in formal validity checks, is performed while loading data in the Working Data Storage. Each variable is checked against the code list and format which is appropriate for the specific data field. The applied validation rules are based on a metadata system. Each concept type (dimension, attribute, measure) is required to be compliant with the expected trade data structure according to the external trade structural classification metadata repository.

Therefore, during the acquisition phase:

- Records which are not compliant with the structural metadata are not uploaded into the working database. They are eventually retrieved, after correction of the fields which failed to satisfy the validity rules. In case data rejection appears to be due to unexpected structural changes in Customs data (i.e. revision of Customs classifications for products or countries, or modifications in field formats), Customs are contacted and, eventually, the INS system of structural metadata is updated;
- Records are further filtered in order to process only statistically relevant Customs micro-data having statistically relevant regimes;
- Micro-data are connected with structural metadata needed to classify trade flows.

Preparation for data processing: referential metadata for data editing

In this sub-phase, parameters needed for the new outlier selection and error imputation are computed for the first time or refreshed considering the new uploaded data.

Three sets of editing parameters are computed:

- Editing parameters-based unit values "uv", i.e. ratios between traded value, expressed in Tunisian dinar, and net mass, in kg;
- Editing parameters for unit price "px", i.e. ratios between traded value, expressed in Tunisian dinar, and supplementary unit, defined for specific commodities (i.e. pieces, TJ, etc.);
- Editing parameters for unit weight "uw", i.e. ratios between net mass, expressed in kg, and supplementary unit, for commodities for which supplementary unit is required.

For each set of parameters, the same methodology is applied, based on a robust method for asymmetric, right skewed, distributions. The proposed methodology for outlier detection is based on a non-parametric method for asymmetric distributions (Tukey (1977), Thompson (1999) and Hubert and Van der Veeken (2008)), largely applied in several fields, and in external trade data editing (Narilli and Nuccitelli, 2018).

In the INS application, trade transactions micro-data over a set of 24 months from the current reference period are stratified by product and flow. For such strata, unit values, prices and weights distributions are considered. As several economic data distributions, the observed distributions are skewed to the right and need to be symmetrized by a logarithmic transformation before applying outlier detection methods. On the log-transformed distributions, robust position indicators are computed, namely median (q_2), first (q_1) and third (q_3) quartile[3]. The position indicators are used to compute editing parameters on the log-transformed distribution. After final exponential transformation, editing parameters are suitable to be used on the original asymmetric distribution of trade data.

For each commodity and flow stratum, minimum T1 and maximum T2 editing parameters are computed as:

$$T_1 = exp[q_1 - k(q_3 - q_1)]$$

$$T_2 = exp[q_3 + k(q_3 - q_1)]$$

"k" is a parameter which can be tuned appropriately by INS experts (typical values range from 1 to 3). Current setting is k=1.5.

At the end of this step, editing parameters for unit values (T_1^{uv}, T_2^{uv}), unit prices (T_1^{px}, T_2^{px}), and for unit weight (T_1^{uw}, T_2^{uw}), are used to update referential metadata related to each traded product, together with the exponential of the corresponding medians q_2, to be used for outlier imputation as described in the following section.

Data processing: outlier detection and imputation

Editing parameters computed as described in the previous section are used for detecting outliers in the appropriate product and flow strata of the current reference month, while the knowledge of the median of the distribution gives the possibility for automatic imputation for wrong variables.

Few outliers with large impact in value and quantity will be corrected manually, but the IT system provides

[3] Robustness is associated with high breakdown points, 50% for the median and 25% for quartiles. Such features make the outlier detection method robust even in presence of 25% outlier data.

the possibility of simply accepting the correction automatically proposed or manually performing a different correction.

The outlier detection and imputation phase is performed in three steps: at first the statistical value variable is inspected and corrected, then the net mass, and finally the supplementary unit. Since the three variables are correlated by unit values, unit prices and unit weights, the lack of coherence among them can be a reason not to allow automatic imputation for some records.

Check and imputation of statistical values

Even if statistical value is a variable which is checked at Customs level, there can be few cases of misreporting.

The selection of potential outliers is performed only on records with potential high positive impact, namely the ones satisfying the following conditions:

$value > 500\,000\ DIN$,

and

$uv > T_2^{uv}$

The above condition is not sufficient to state that the wrong variable is the statistical value, while the net mass is correct. The needed additional information is provided by supplementary unit, if available. In that case, potential errors for records with supplementary unit will be automatically corrected if both unit value and unit price are outliers, while unit weight is not an outlier, namely if all the following relations hold:

$uv > T_2^{uv}$,

$px > T_2^{px}$,

$(1-t)T_1^{uw} < uw < (1+t)T_2^{uw}$,

$T_1^{uw} \neq T_2^{uw}$,

where *t* is a tolerance parameter set to 0.3.

If the above condition is satisfied, the correction proposed for automatic correction by the system is:

$value_{corrected} = NetMass \cdot T_2^{uv}$.

If the conditions for automatic correction are not satisfied, the records are sent to manual revision, together with the above proposal for correction provided by the system, $value_{corrected}$.

For all outlier records sent to manual revision, the system provides statisticians with editing parameters and potential impact of the error in DIN (impact=proposed value – original value), useful to prioritize editing of records with very high impact.

The workflow for statistical value editing is summarized in Table 1.

Table 1:
Editing scheme for statistical value

conditions			VAL_corr
si VAL > 500 000 **et** uv > T2_uv	*corrections automatiques*	px > T2_px **et** 0,7 * T1_uw < uw < 1,3 * T2_uw **et** T1_uw != T2_uw	NetMass * T2_uv
	corrections manuelles	!= (px > T2_px **et** 0,7 * T1_uw < uw < 1,3 * T2_uw **et** T1_uw != T2_uw)	

By applying the new methodology to a typical reference month, about 0.04% of outlier records in statistical value were imputed for export and 0.02% for import.

Check and imputation for net mass and supplementary unit

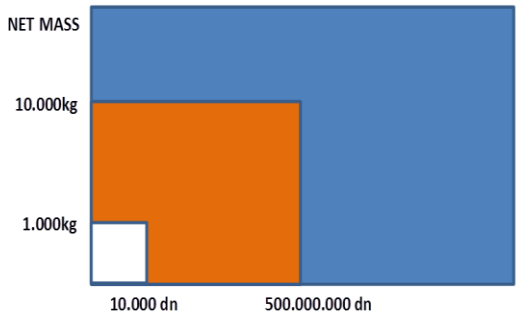
In this editing phase are selected as outliers:

- Records having unit value exceeding the maximum editing parameter or lower than the minimum editing parameter;
- Records having unit price exceeding the maximum editing parameter or lower than the minimum editing parameter.

Given the different potential impact on the aggregate figures of the different potential errors, the correction actions to be performed are differentiated as explained below (see diagram 1).

All potential errors in the “blue area”, i.e. having value > 500.000 dn or net mass > 10.000 kg are sent to manual revision, but the system provides the proposal for correction. Manual revision should be performed also on records having value < 500.000 dn or net mass < 10.000 kg for which there were less than 10 observations available for computing editing parameters. Indeed, for such records computed parameters are not robust enough;

Diagram 1:
External trade data revisions domains



All potential errors in the "orange area", i.e. having value between 10.000 dn and 500.000 dn and net mass between 1.000 kg and 10.000 kg are submitted to automatic revision. The automatic correction is not performed in case there were less than 10 records available to compute editing parameters;

Potential errors with low impact at micro-level in the "white area" are left uncorrected. Indeed, even if their impact can become relevant after aggregation, a further macro-editing performed at the end of the micro-editing process will detect potential residual anomalies.

For the proposals for correction or imputation, four different situations can be found.

1) The record has no supplementary unit: the system proposes a correction for net mass, according with the formula

$$NetMass_{corrected} = {value}/{\text{Median}(uv)}$$

2) The record has supplementary unit and both px and uw are outliers: the system proposes a correction for both net mass, as in point 1, and supplementary unit as follows

$$SuppUnit_{corrected} = {value}/{\text{Median}(px)}$$

if $(T_2^{uw} - T_1^{uw}) > (T_2^{px} - T_1^{px})$

or

$$SuppUnit_{corrected} = {NetMass_{corrected}}/{\text{Median}(uw)}$$

if $(T_2^{uw} - T_1^{uw}) \leq (T_2^{px} - T_1^{px})$

Indeed, the median of less broad distribution variable is chosen for imputation.

3) The record has supplementary unit, px is not an outlier, but both uv and uw are outliers: the system proposes a correction only for net mass

$$NetMass_{corrected} = {value}/{\text{Median}(uv)}$$

if $(T_2^{uw} - T_1^{uw}) > (T_2^{uv} - T_1^{uv})$

or

$$NetMass_{corrected} = SuppUnit \cdot Median(uw)$$

if $(T_2^{uw} - T_1^{uw}) \leq (T_2^{uv} - T_1^{uv})$

4) The record has supplementary unit, uv is not an outlier, but both px and uw are outliers: the system proposes a correction for supplementary unit

$$SuppUnit_{corrected} = {value}/{\text{Median}(px)}$$

if $(T_2^{uw} - T_1^{uw}) > (T_2^{px} - T_1^{px})$

or

$$SuppUnit_{corrected} = NetMass\ /Median(uw)$$

if $(T_2^{uw} - T_1^{uw}) \leq (T_2^{px} - T_1^{px})$

In case only uw is an outlier, the system cannot propose an automatic correction. The editing scheme is summarised in Table 2.

Table 2:
Editing scheme for net mass and supplementary unit

	conditions			NetMass_corr	SuppUnit_corr
1	si SuppUnit = NetMass			value/Median(uv)	
2	si SuppUnit != NetMass et	(px < T1_px ou px > T2_px) et (uv < T1_uv ou uv > T2_uv)	et T2_uw − T1_uw > T2_px − T1_px	value/Median(uv)	value/Median(px)
			et T2_uw − T1_uw =< T2_px − T1_px		NetMass_corr/Median(uw)
3		T1_px < px < T2_px et (uw < T1_uw ou uw > T2_uw) et (uv < T1_uv ou uv > T2_uv)	et T2_uw − T1_uw > T2_uv − T1_uv	value/Median(uv)	SuppUnit
			et T2_uw − T1_uw =< T2_uv − T1_uv	SuppUnit * Median(uw)	SuppUnit
4		T1_uv < uv < T2_uv et (px < T1_px ou px > T2_px) et (uw < T1_uw ou uw > T2_uw)	et T2_uw − T1_uw > T2_px − T1_px	NetMass	value/Median(px)
			et T2_uw − T1_uw =< T2_px − T1_px		NetMass_corr/Median(uw)
5		(T1_px < px < T2_px) et (T1_uv < uv < T2_uv) et (uw < T1_uw ou uw > T2_uw)		le système ne peut pas proposer des corrections automatiques	

By applying the new methodology to a typical reference month, about 0.5% of outlier quantity records were imputed for export, and 0.4% for import.

Finally, a macro-editing test is performed, looking for evidence of residual outliers present in the system. The year-to-year total growth rate (i.e. the percentage growth rate of the reference month "m" and year "y" with respect to the same month of the previous year) is decomposed in contributions associated to (product, partner country) cells, as follows:

$$contrib(cell = product, country) = 100 * \frac{value_{cell}(m,y) - value_{cell}(m,y-1)}{\sum_{cell} value_{cell}\,(m,y-1)}$$

Note that the sum of such contributions corresponds to the total year-to-year total growth rate.

Contribution are then sorted: the higher and lower ones are often associated to incorrect records to be further investigated.

However, many cells giving high positive or negative contributions can be associated to true economic phenomena. Therefore, the final macro-editing is also a tool for statisticians to have a final overview on the important trade components in data to be disseminated.

The same decomposition can be then performed for growth rate contributions in net mass and in supplementary unit.

Indicators for dissemination: trade balances and indices

After moving final edited micro-data in the final database, trade indicators are computed. At first, trade balances both at product and partner country breakdown are computed and moved to the dissemination Macro Database.

Finally, unit value and volume indices are computed.

Concerning unit value and volume indices calculation, first steps for introducing a new methodology of calculation in the new system were made. The proposed methodology has the advantage of producing robust indicators less sensitive to outliers (Anotori and Causo, 2008).

The proposed process consists in:

- No "a priory" selection of a specific basket of products[4];
- Production of chained indices where the base year is yearly refreshed;
- Exclusion from unit value calculation of transactions with outlier unit value;
- Imputation of outlier elementary indices;
- Estimation of base unit value and Laspeyres weights for products not traded in the previous year;
- Aggregation of Laspeyres, Paasche and Fisher indices;
- Chaining moving base indices to the fixed reference year.

In a hopefully future collaboration, the proposed methodology will be implemented and tested.

Results and perspectives in the pilot domain

The improvements achieved with the new system put in place in 2018 concern both statistical and IT management of the process. On the statistical side, streamlining the process gave the following advantages:

- The new integrated system is entirely managed by statisticians;
- The new integrated system allows to have an organized metadata repository and facilitate the update of his different files;
- The statistician has the possibility to define the parameter of control to be applied in the process and can also validate the correction proposed by the new system both manually and in automatic way;
- The proposed process offers more controls on data at elementary level which will improve the quality of data value, net mass and the supplementary unit;
- The automatization of all the controls steps can reduce the time to validate the data, so the new process is more efficient for quality and timeliness.

On the IT management side, the improvements involved several aspects:

- User Authentication and permissions: The application is designed with a built-in security system, which ensures a high level of data security and confidentiality, against unauthorized access or unintentional usage. System profiles are set for managing user accounts and roles to assign permissions that control access;
- Parametrizable application for data quality assurance: The system is based on several parameters requested and managed by the external

[4] *However, specific products which have "unique" quality content, so that they cannot be compared with any equivalent product traded in the base period, are excluded from the calculation (for example, Aircraft, Ships, Works of art, Jewellery).*

trade administrators. The management system of these parameters allows easy and fluid navigation to users in order to ensure the verification and correction of data;

- Data consistency, standardization, and user-friendly management system: The system is designed for non-IT specialist users as the subject unit users can easily access the system according to their special requests and generate the final reports. The centralized information system increases data integrity, security, support and storage capacity and it helps to improve the quality of data by ensuring consistency and traceability of corrections throughout the various processing steps.

Further future developments in the pilot domains are foreseen after the end of the project. An important aspect to be introduced in every statistical domain dealing with economic variables is linking data with business registers. At European level, External Trade data production divisions are in charge of the statistical production of trade data by enterprise characteristics (TEC data). TEC statistics are based on integration of International Trade in Goods Statistics data with Business Registers and allow reading trade data under business statistics perspective. The indicators produced in such relatively new statistical domain shed light on businesses behind trade flows, defining their characteristics in terms of business activity and enterprise size. These indicators are very useful for users and policy makers.

Moreover, having access to the national Business Register, can improve the quality of External Trade data production, by allowing to:

- Assess the coherence of traded values in data received by Customs with the size and economic activity of traders;
- Monitor on a monthly base the coherence in the time series of traded values at enterprise level;
- Assess coherence between traded products and economic activity of trader enterprises.

During the twinning project it was advised to take advantage from the integrated architecture for statistical process for integrating data from different statistical INS domain, both for data validation purposes and for producing innovative statistical products, such as TEC (Eurostat, 2018).

Implementation roadmap for an extension of the solutions developed for the External Trade pilot application to other domains

Short term roadmap

The first activities to be performed in the short term pertain to the extension of the solutions developed for the External Trade pilot application to other domains. The main principles in this pilot are:

- Data stored in relational database;
- Workflow management to implement transformation procedures for data processing;
- Simplified interface to help the statisticians to manage the statistical process independently of IT staff.

Medium term roadmap

While the pilot experience is being applied to other statistical domains, the following medium-term activities should be carried out:

Metadata

GSIM training: INS technicians should familiarize themselves with the GSIM standard. They could, for example, read the documentation and/or organize some training or study visits at national statistical institutes that are more advanced in GSIM use.

Analysis of case studies: The CSPro integration implemented in the ISTAT-Ethiopia cooperation project and the Serbian IST[5] solution must be studied to understand the pros and cons of both solutions.

Decision: after the above activity, INS should decide whether to use one existing solution (possibly with some modifications or adaptations) or to develop in-house solutions, which must however be compliant with the analysed alternatives.

Implementation: INS will develop or adapt the metadata system according to the following principles:

- GSIM-compliant statistical concepts, valid for every statistical domain and designed to be used as active metadata;
- Integration with the data collection system, including process controls and editing rules;
- Coherence with. STAT for dissemination.

[5] *Detailed documentation on IST project can be found at: http://webrzs.stat.gov.rs/ISTSite/IST_Home.aspx*

Data collection:

INS should develop/choose one or two metadata-compliant collection system(s) with the following characteristics:

- Metadata-driven: the structure of the data to be collected should be stored in the metadata system;
- Questionnaire design: the data collection system should provide an interface to design the questionnaire, starting from the contents (e.g. units, variables, classifications) available in the metadata system;
- Questionnaire microdata: the output should be stored in a relational database coherent with metadata.

Processing system

The processing system is implemented after the data collection system. It will have the following characteristics:

- Metadata-driven: deriving data flows from metadata;
- Apply editing rules: rules should also be stored in the Metadata system using SQL or VTL6 format;
- Manage workflows: the workflow should provide the following features: pass parameters to services, service invocation, process execution monitoring, exception handling, access and visualization of available data;
- Service catalogue: statistical services to be used in the workflow will be chosen from a catalogue where all services (CSPA compliant?) will be stored.

References

Anitori, P. and Causo, M.S. (2007), "Outlier detection and treatment: quality improvement in the Italian unit value indexes", Seminar on External Trade data production, December 2007, Eurostat.

Bruno, M., Luzi, O., Ruocco, G. and Scannapieco, M. (2018), "Standardization of Business Statistics Processes in Istat", Q2018 European Conference on Quality in Official Statistics.

Eurostat (2017), "Geonomenclature applicable to European statistics on international trade in goods", Manuel and guidelines, 2017 edition.

Eurostat (2017), "Compilers guide on European statistics on international trade in goods", Manuals and guidelines, 2017 edition.

Eurostat (2018), "Compilers guide on European statistics on international trade in goods by enterprise characteristics (TEC)", 2018 edition.

Hubert, M. and Van der Veeken, S. (2008), "Outlier detection for skewed data", *Journal of Chemometrics*, No. 22, pp. 235–246.

Narilli, M.G. and Nuccitelli, A. (2008), "Re-designing the editing process for the Italian data on Foreign trade statistics with countries outside the EU: some quality issues and results", European Conference on Quality in Official Statistics, July 2008, Rome.

Novkovska, B., Papazoska, H. and Ristevska-Karajovanovikj, B. (2012), "The GSBPM contribution to statistical business process standardization", European conference on quality in official statistics, Athens.

Scanu, M., Bergamasco, S., Cardacino, A., Rizzo, F. and Vignola, L. (2013), "A strategy on structural metadata management based on SDMX and the GSIM models", Conference Paper, Work session on statistical metadata (METIS), May, Geneva.

Scannapieco, M. and Vaccari, C. (2011), "Standardizing European Statistical processes: CORA and CORE projects", Conference MeTTeG 2011 - 5th International Conference on Methodologies, Technologies and Tools enabling e-Government, Camerino, Italy.

Signore M., Scanu M., and Brancato, G. (2015), "Statistical metadata: a unified approach to management and dissemination", *Journal of Official Statistics*, No. 31.2, pp. 325-347.

Thompson, K.J. (1999), Ratio edit tolerance development using variations of exploratory data analysis (EDA) resistant fences methods, 1999 FCSM Research Conference Papers.

Tukey, J.W. (1977), "Exploratory Data Analysis", Reading (Addison-Wesley), Massachusetts.

[6] *Documentation on the VTL standard can be found at: https://sdmx.org/?page_id=5096*