CONFERENCE OF EUROPEAN STATISTICIANS

ModernStats World Virtual Workshop 2020

27-30 October 2020 28 October 2020

ModernStats standards supporting the implementation and sharing of statistical services

Cotton Franck, Soulier Manuel (Insee, France)

Bruno Mauro, Amato Francesco, Ruocco Giuseppina (Istat, Italy)

franck.cotton@insee.fr, mbruno@istat.it

Abstract

Official statistical standards provide guidelines and recommendations to support both statistical production and auxiliary activities. In the last years, many National Statistical Institutes (NSIs) have increased the compliance to statistical standards, to gain efficiency and improve the quality of disseminated output. Statistical standards have also supported the implementation of the ESS Vision 2020 programme. Within this context, the project "Implementing Shared Statistical Services" (I3S) aims at developing and sharing reusable software solutions. More precisely, statistical standards have guided the implementation activities: GSBPM [1] has been the starting point for modelling statistical services, GSIM [2] concepts have been adopted to model data structures, while CSPA [3] principles have guided the development activities.

The paper describes I3S project design and development activities, focusing on the implementation of two statistical services, namely Relais (Record Linkage At Istat) and Insee's ARC (Accueil Réception Contrôle). Relais provides an integrated environment to solve record linkage problems, while ARC allows receiving data supplied by external providers, to control the compliance of the received files, and to transform administrative data to elementary statistical data.

1 Introduction

In the last years, the National Statistical Institutes (NSIs) have faced many challenges resulting from the considerable changes of the external and internal context. One of the main goals of official statistical standards is to support the modernisation process and enhance information and experiences sharing among countries. Within the ESS Vision 2020 programme, "Implementing Shared Statistical Services" [4] is one of the projects launched for developing and sharing generic software solutions and increasing the statistical services available in the CSPA (Common Statistical Production Architecture) catalogue[1].

The following paragraphs provide: i) a brief description of the project, ii) an overview of two statistical services implemented, iii) the analysis of how statistical standards have affected the service development and reuse, iv) a focus on the most relevant project results.

[1] CSPA catalogue is available at: https://www.statistical-services.org/

2 Implementing Shared Statistical Services (ESSnet I3S)

The ESSNet I3S aims at developing reusable and shareable statistical services, either from scratch or from existing components. Project activities are coherent both with CSPA principles and the ESS Enterprise Architecture Reference Framework [5] that is the reference framework for implementing the ESS Vision 2020. To foster service reuse and shareability, the use of free and open source software is a key requirement for service implementation. The project will also define architecture principles and guidelines for software deployment in the ESS context.

To achieve these goals the project has been structured in the following work packages (WP), having specific goals, deadlines, and deliverables:

- Work package 1: Develop new services. The deliverable of this WP is the release of three statistical services, selected from the list established in the ESSnet "Sharing Common Functionalities in the ESS" (SCFE) and available on the CROS portal[2].
- Work package 2: Define integration and architecture guidelines. The goal of the WP is to explore and describe the different architectures for service integration (e.g. orchestration, pub/sub, adapters, containers…). This study will also detail the minimum requirements that a service should fulfil for the best integration (e.g. authentication, authorization, auditing, logging).
- Work package 3: Build a sandbox and test available services. This WP provides a sandbox environment to install and test WP1 services implemented according to WP2 architectural principles. WP1 services will be advertised in the CSPA catalogue.
- Work package 4: Create and communicate success stories. This WP collects successful stories of sharing and reusing the statistical services implemented. Further, the WP describes the lessons learnt during the development activities, with the aim to promote service shareability principles.
- Work package 5: Communication and Dissemination of Results.

3 Statistical standards supporting service development

The design and implementation of a statistical service involves the analysis of several dimensions that can be grouped as follows: process, data, methods and tools.

- **Process**: concerning the process dimension, GSBPM allows referring a statistical service to a specific phase or sub-phase, avoiding overlapping between implemented solutions. GSBPM allows a standardized description of the process steps implemented by a statistical process (e.g. Probabilistic Record Linkage, Selective Editing, etc.). In other words, GSBPM helps defining, at a conceptual level, the main objectives of each process step (core logic).
- **Data and methods**: the data perspective is a crucial issue to manage. Within a statistical service, the data management includes input/output data and meta information related to the implemented methods (e.g. rules, parameters, thresholds). GSIM concepts help to standardize the structure of the different data objects involved in the service execution. Further, GSIM allows connecting each step executed by the service to a specific information object having a standardized structure. This feature enhances service reuse and shareability.

[2] The list of services is available at: https://ec.europa.eu/eurostat/cros/content/scfe-d4-1-initial-list-services-are-candidates-re-use-ess_en

- **Tools**: CSPA provides a reference framework in the design and implementation of statistical services. To achieve this goal, CSPA standard contains a set of principles and requirements that allow implementing shareable services.

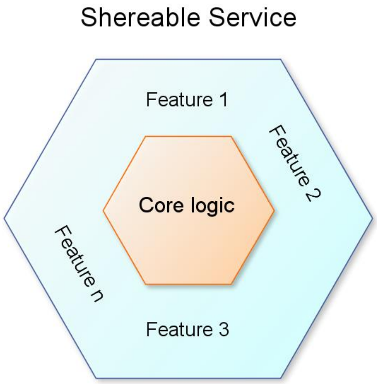

Figure 1: CSPA Statistical Service

According to CSPA the software components of a statistical service should be classified as follows (as shown in Figure 1):

- **Core logic:** the algorithm implemented by the service.
- **Features**: a set of components (not only software) that consider specific requirements from different stakeholders (e.g. IT, methodology, domain experts, etc.) and allow the execution of the service in several environments. Baseline features are, for example, documentation, internationalization, open source code, use of GSIM concepts.

Adopting statistical standards as reference frameworks for service implementation allows connecting the different dimensions explained above. More precisely, the relationship between statistical services and official statistics standards is summarized in Figure 2.

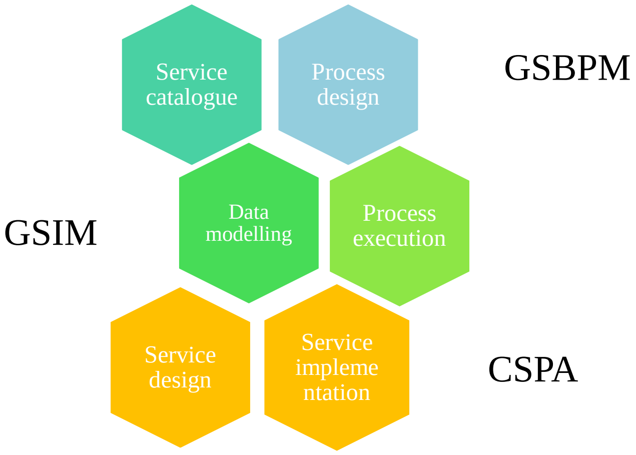

Figure 2: Statistical standards and their role in service implementation

In the following sections, these different subjects will be illustrated on concrete examples of statistical services.

3.1 *ARC implementation*

A statistical process generally includes a data acquisition phase. Multiple sources, and the more and more frequent combination of them, can provide these data (survey, administrative or private data). A conversion phase to statistical units and attributes is necessary before statistical treatments can be applied.

The ARC (from the French: Acquisition - Réception - Contrôles)[3] software allows to receive data supplied by the providers (several formats are supported, particularly XML), to control the compliance of the received files, and to transform input data into elementary statistical entities. The software enables the statistician to define and apply controls and mappings, to test them in a sandbox environment (linked to the software), and to put them into production without frequently calling on a developer (see Figure 3).

These functionalities/services aim at the statistician's independence and ability to adapt to the data evolutions, thereby avoiding impacts on the statistical chain.

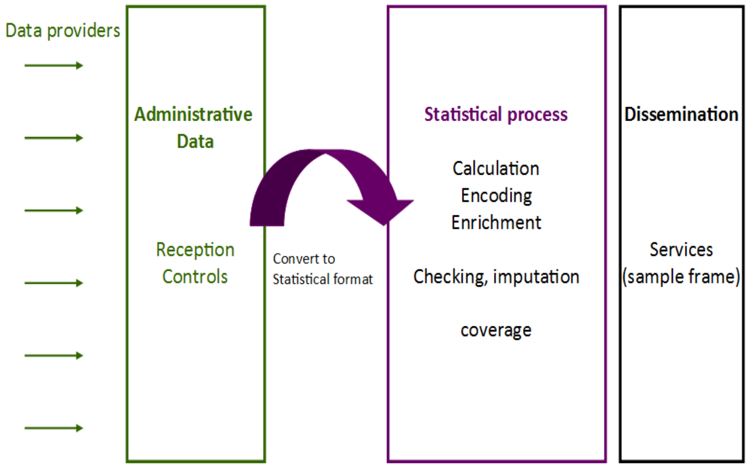

Figure 3: ARC Process Overview

The main functionalities of the ARC application related to the data processing life cycle are linked to the GSBPM "Process" step, and more precisely to elementary treatments necessary to integrate data, engage basic cleaning and editing tasks, but also transformation operations that are necessary to convert raw data into a statistical database compliant with statistical concepts and units. This is mainly related to sub-processes going from "5.1 - Integrate data" to "5.5 - Derive new variable and units", as schematized in Figure 4.

[3] ARC is an open source project. Source code is available from: https://github.com/InseeFr/ARC

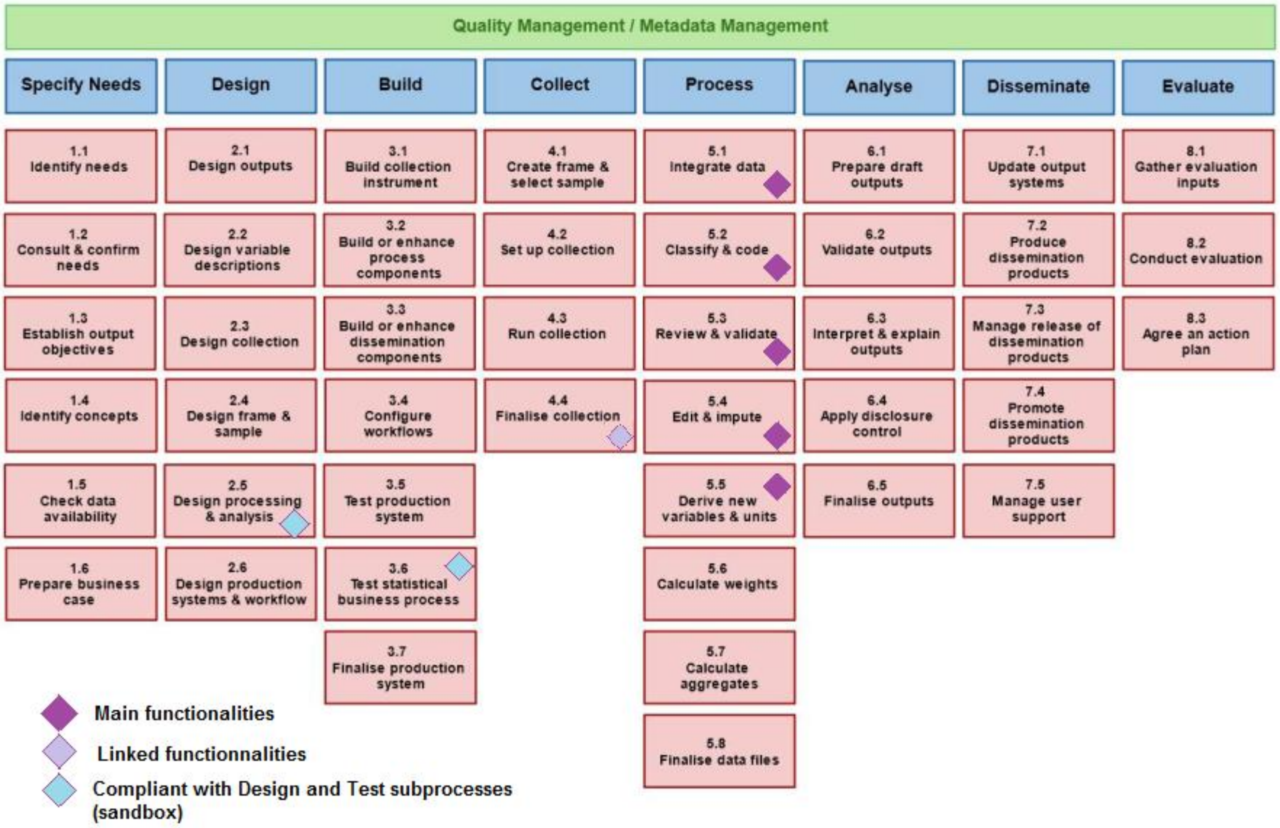
Quality Management / Metadata Management
Specify Needs
1.1 Identify needs
1.2 Consult & confirm needs
1.3 Establish output objectives
1.4 Identify concepts
1.5 Check data availability
1.6 Prepare business case
Design
2.1 Design outputs
2.2 Design variable descriptions
2.3 Design collection
2.4 Design frame & sample
2.5 Design processing & analysis ◈
2.6 Design production systems & workflow
Build
3.1 Build collection instrument
3.2 Build or enhance process components
3.3 Build or enhance dissemination components
3.4 Configure workflows
3.5 Test production system
3.6 Test statistical business process ◈
3.7 Finalise production system
Collect
4.1 Create frame & select sample
4.2 Set up collection
4.3 Run collection
4.4 Finalise collection ◇
Process
5.1 Integrate data ◆
5.2 Classify & code ◆
5.3 Review & validate ◆
5.4 Edit & impute ◆
5.5 Derive new variables & units ◆
5.6 Calculate weights
5.7 Calculate aggregates
5.8 Finalise data files
Analyse
6.1 Prepare draft outputs
6.2 Validate outputs
6.3 Interpret & explain outputs
6.4 Apply disclosure control
6.5 Finalise outputs
Disseminate
7.1 Update output systems
7.2 Produce dissemination products
7.3 Manage release of dissemination products
7.4 Promote dissemination products
7.5 Manage user support
Evaluate
8.1 Gather evaluation inputs
8.2 Conduct evaluation
8.3 Agree an action plan
◆ Main functionalities
◇ Linked functionalities
◈ Compliant with Design and Test subprocesses (sandbox)

Figure 4: ARC GSBPM Coverage

Sub-process “5.1 - Integrate data” is related to data integration from one or more sources. According to GSBPM, the input data can be from a mixture of external or internal data sources, and a variety of collection modes, including extracts of administrative data. The result is a set of linked data. ARC is dedicated to combining data from multiple sources or combining multiples files from a same kind of data source, in order to create an integrated database.

Sub-process “5.2 - Classify and code” is also covered by ARC, as it includes coding routines to convert initial variable codes to a pre-determined classification scheme.

Sub-process “5.3 - Review and validate” examines data to try to identify potential problems, errors and discrepancies such as outliers, item non-response and miscoding. The ARC user can create testing and validating rules, which will be applied to the data with detection of actual or potential errors. According to GBSPM, treatments in sub-process 5.3 are only dedicated to discrepancy and error detection, while effective data editing is done in sub-process 5.4.

Sub-process “5.4 - Edit and impute” convers a variety of updates, often using a rule-based approach. In ARC, data editing is mainly related to filtering and basic imputation. Functionalities include the determination of whether to add or change data, changing data values, and flagging data as changed in the process phase.

Sub-process “5.5 - Derive new variables and units” derives data for variables and units that are not explicitly provided in the collection, but are needed to deliver the required outputs. Through user defined rules, ARC may derive new variables by applying arithmetic formulae to one or more of the variables that are already present in the dataset, or applying different model assumptions.

Previously to these steps, ARC is also built to host raw data files that have been collected and deliver environment for storing large data collection before being processed. It thus can be used as a collection tool according to GBSPM sub-process “4.4 - Finalise collection”, loading the collected data and metadata into a suitable electronic environment for further processing.

Because it includes sandbox creation and testing procedures of user rules, ARC application is also compliant with GSBPM subprocesses dedicated to “2.5 - design processing” and “3.6 - test statistical business process”.

“2.5 - design processing” can include specification of routines for coding, editing, imputing, estimating, integrating, validating and finalizing data sets.

“3.6 - test statistical business process”, which includes activities to manage a pilot the statistical business process, including testing processing rules on data.

The last developments on ARC focus on expanding the GSBPM coverage of the service to the sub-processes 5.7 (“Calculate aggregates”).

Core logic implementation

From an implementation point of view, the ARC file reception process is divided into steps which may be individually parametrized and executed by the user, as shown in Figure 5.

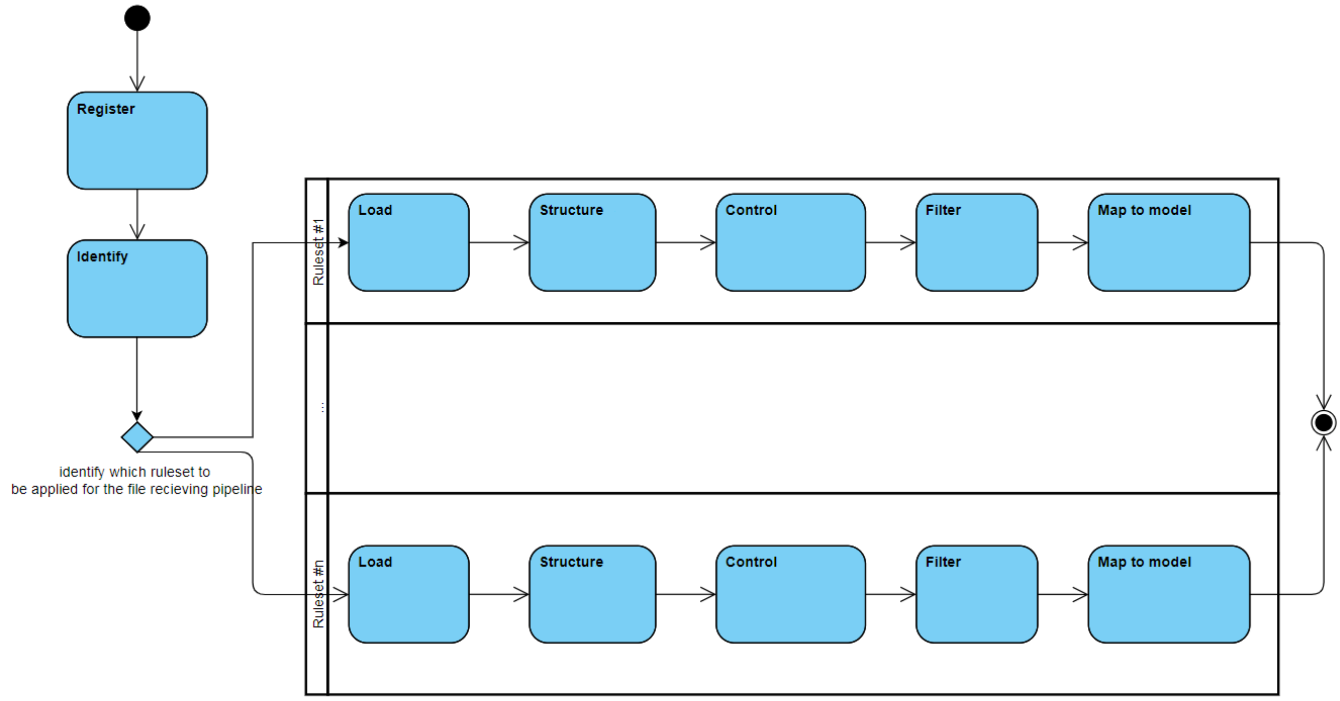

Figure 5: ARC pipeline

A short description of the different steps is given below.

Register	Register a set of input files and update the file system
Identify	Determine the "norm" of the file. The rules written by the users have a global input key called “norm”. This key identifies the different chains of treatments. The category and the sandbox determines which rule sets to apply to the file when processing the subsequent modules
Load	Apply the rules of the module to instantiate the proper file reader to load the data without transformation in the database.
Structure	Structure the previously loaded structured data consistently. Apply the rules of the module for advanced structuration features
Control	Apply the correction and control rules
Filter	Apply a filter rule to exclude records
Map to model	Transform and store loaded data in a standard user-defined data model

Table 1: ARC steps

ARC integration

During the course of the I3S project, important evolutions were made on ARC in order to integrate it with Istat’s Relais and Statistician Workbench (see below). In particular, we worked on the general architecture, the data model and the development of a layer of web services around the ARC engine. Now, the Statistician Workbench serves as a configuration interface: the engine retrieves the configuration from the workbench via web services, launches the execution accordingly and returns the results so they can flow further down the statisctical process.

3.2 RELAIS implementation

Relais (Record Linkag At IStat) provides an integrated environment to solve record linkage problems ([5], [7]). It has been implemented assuming that a record linkage process may result from the combination of different sub-phases, to achieve the best data integration solution. During the ESSnet, Relais has been re-designed according to CSPA principles. The design and implementation activities can be summarized as follows:

1. Analysis of current version of Relais software. The current version of the software, available in CSPA catalogue, is monolithic i.e. frontend, backend and database are strongly coupled. Such architecture has several drawbacks, e.g. the installation of Relais requires technical skills and this may represent a challenge for statistical users. Further the integration of monolithic software in modern IT environments (e.g. containers, cloud architecture, rest apis) is increasingly difficult.
2. Modelling of the AS-IS architecture. To re-design the software components according to official standard principles, an in-depth analysis of both the tool and the implemented methods has been performed. The results of this analysis have provided valuable inputs in the design of the target architecture, where each step of Relais has been connected to a re-designed software component.
3. Implementation of the new version of Relais, according to CSPA principles[4].

[4] The new version of Relais statistical service is available from: https://github.com/mecdcme/is2

Starting from the GSBPM sub-phase “5.1. Integrate data”, the tasks performed by the previous version of Relais have been analysed, in order to identify the service core logic and the auxiliary features. Relais core logic allows to execute both deterministic and probabilistic approach. As the software was already structured in a well-organized pipeline, the reengineering activities have concentrated on the refactoring of the subset of existing components related to the core logic. The operational steps that correspond to the core logic and allow to perform sub-phase “5.1. Integrate data” are reported in the figure below.

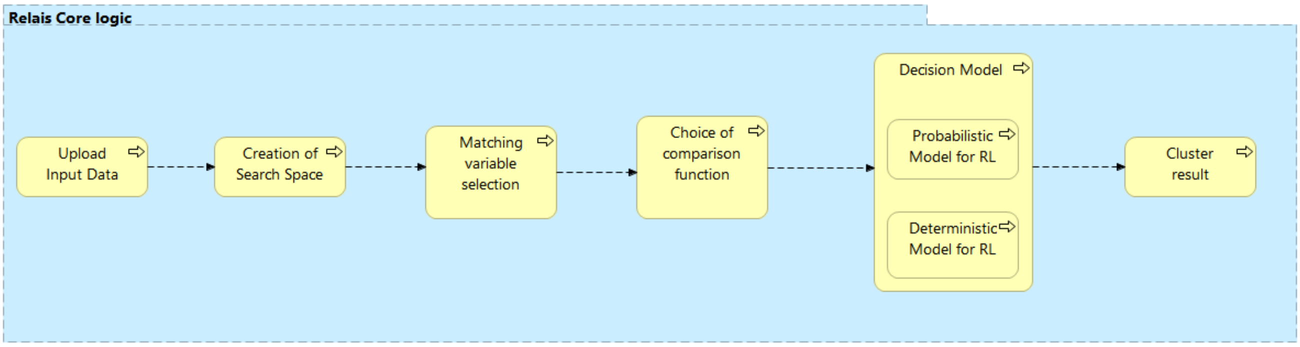
Relais Core logic
Upload Input Data
Creation of Search Space
Matching variable selection
Choice of comparison function
Decision Model
Probabilistic Model for RL
Deterministic Model for RL
Cluster result

Figure 6: Relais record linkage pipeline

The baseline features of Relais include a set of functionalities for data processing and output analysis. The more relevant functionalities to be finalized are: i) variable standardization (remove spaces, remove special chars, case conversion); ii) variable merge; iii) creation of new variables.

In the new version of Relais, data and metadata management are based on the following GSIM concepts: Business Function, Business process, Process step, Process Input/Output, Rule, Method, Data Set, Variable. These concepts have fostered the standardization of data structures involved in each task, and the service resilience in case of data or parameters changes (Figure 7).

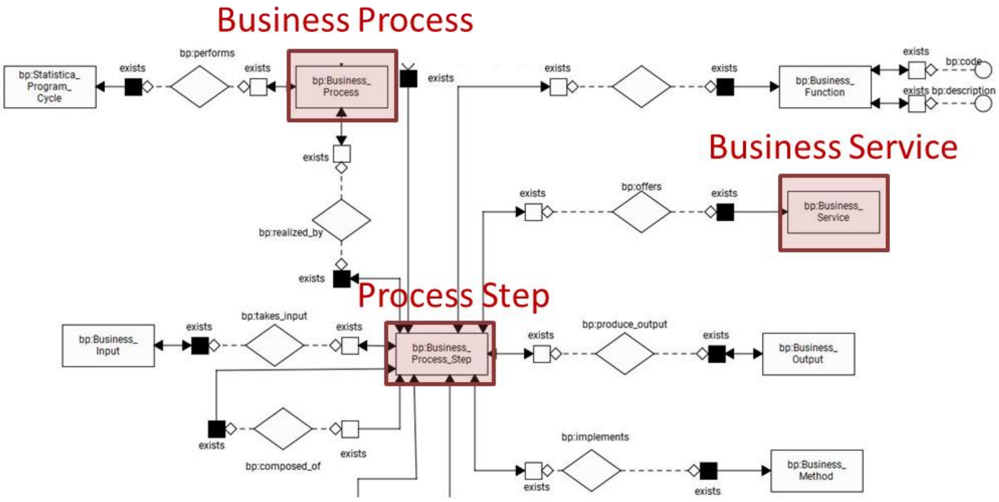

Figure 7: Subset of GSIM concepts used to re-design Relais

3.3 Statistical standards enhancing service reuse

The adoption of CSPA principles for the service implementation fosters the service reuse in different organizations and provides an assessment of service shareability. During the ESSnet, each service has been documented and implemented using open source software, according to CSPA recommendations.

One of the deliverables of WP1 consists in developing concrete reuse cases of statistical services between NSIs. Such an operation covers several steps. The service provider (or DO: Developing Organization) must first create a shareable service, or improve and package an existing service, in order to make it shareable. This implies in particular to work on documentation, internationalisation, modularization, abstraction of dependencies, etc. In the I3S context, it also means open-sourcing the code, which is viewed as a way to strengthen the trust relationship between the DO and reusing organizations (ROs). The RO must then define the reuse case, which is a project involving the implementation of the service in a local business context. It is crucial that subject-matter experts and methodologists be associated to the definition and management of the project, because experience shows that service reuse is more about strategy and organization then about IT. The actual reuse project should then be conducted in collaboration between the RO and the DO. Implementing the service in a new context requires methodological and technical support. It can also reveal opportunities of improvements or optimizations that can greatly benefit to the service. Thus, the DO should be ready to upgrade the code, or to review contributions from the RO. This type of collaboration has been set up between Istat and Insee as part of the I3S ESSnet. Insee was looking at enhancing the production process of its Permanent Database of Facilities[5], by reusing Istat's tool for record linkage Relais, and in return Istat proposed to define a reuse scenario for Insee's data acquisition software ARC. This double operation is still going on, but both partners already view it as a great success.

One of the lessons learnt during the reuse activities is that the adoption of statistical standards for service implementation also enhances the service reuse from other organizations. The compliance to common reference frameworks makes it easier to adapt the service features to different use cases, internal or external to the developer organisation. The harmonisation of the different dimensions achieved by the adoption of statistical standards is represented in Figure 8.

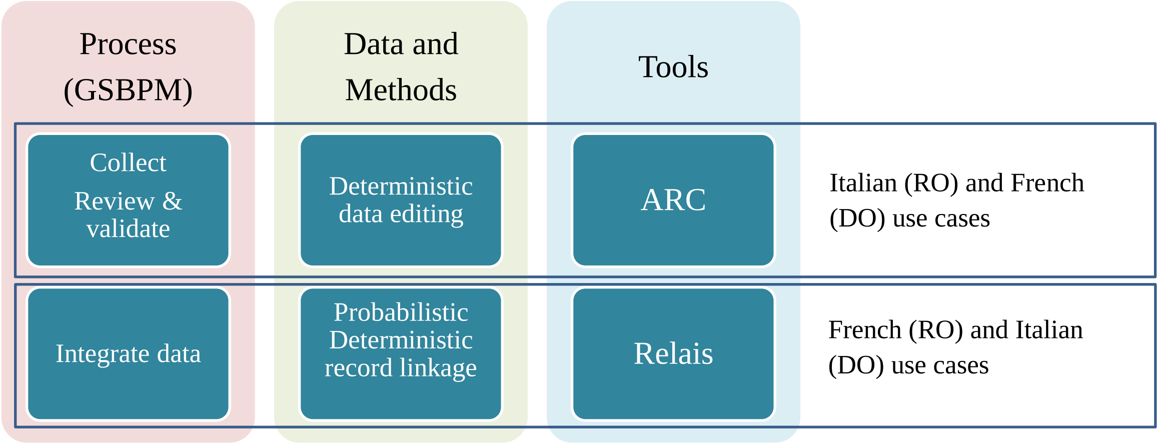
Process (GSBPM)
Data and Methods
Tools
Collect
Review & validate
Deterministic data editing
ARC
Italian (RO) and French (DO) use cases
Integrate data
Probabilistic Deterministic record linkage
Relais
French (RO) and Italian (DO) use cases

Figure 8: Statistical standards and their role in service reuse

[5] https://www.insee.fr/en/metadonnees/source/serie/s1161

3.4 From service reuse to service integration

The close cooperation between ARC and Relais teams has resulted in a much higher level of quality and functionality of both services. They are now associated in a common framework that could prefigure a future "statistician workbench" with shared user interface, process parameter definition, data access methods, etc. Istat and Insee are now convinced that common work has to go on after the ESSnet, and the question is now to define how that can be achieved. From the architectural perspective, the target result is the merge of ARC and Relais pipelines and functionalities in a unique environment. This objective is briefly represented in the figure below.

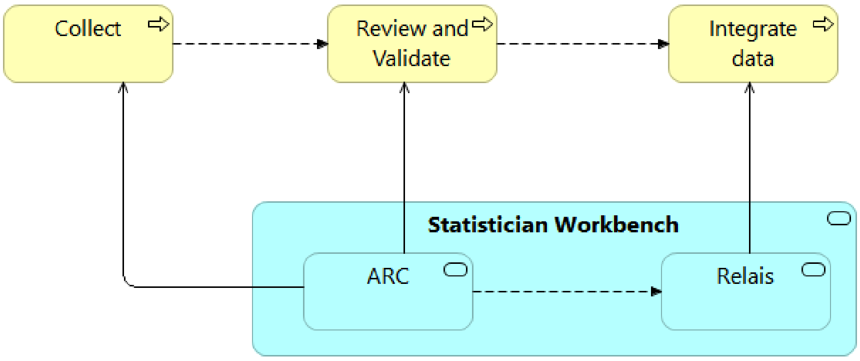

Figure 9: Integration of ARC and RELAIS in the statistical workbench

4 Conclusion

The adoption of statistical standards has enhanced the results of the activities carried out during the ESSnet. Particularly, the statistical standards have guided the architectural layers design (mostly the business and the information and application layers) and the application components refactoring. While GSBPM has been the starting point for the process chain analysis, GSIM has been the reference framework for modelling data structures. Further, GSIM concepts have been relevant to standardize input and output data of each process step, thus enhancing service reuse and shareability. The development of the application components has been driven by CSPA principles. In order to prioritize the development activities, the analysis of the service core logic and additional features has facilitated the iterative implementation. Considering the overall experience, one of the lessons learnt is that the alignment to statistical standards improves efficacy and effectiveness of a statistical process and enhances the cooperation between NSIs.

References

[1] Generic Statistical Business Process Model (GSBPM), version 5.1, available from: https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1

[2] Generic Statistical Information Model (GSIM), version 1.1, available from: https://statswiki.unece.org/display/gsim/Generic+Statistical+Information+Model

[3] Common Statistical Production Architecture (CSPA), available from: https://statswiki.unece.org/display/CSPA/Common+Statistical+Production+Architecture

[4] ESSNet Implementing Shareable Statistical Services (I3S) deliverables, available from: https://ec.europa.eu/eurostat/cros/content/projects-deliverables-0_en

[5] Enterprise Architecture Reference Framework (EAFR), available from: https://ec.europa.eu/eurostat/cros/content/ess-enterprise-architecture-reference-framework_en

[6] Cibella N., M. Fortini, M. Scannapieco, L. Tosco, T. Tuoto. 2007. RELAIS: Don’t Get Lost in a Record Linkage Project. In Proceedings of the FCSM 2007 Conference, Federal Committee on Statistical Methodology, Arlington, 5–7 November 2007.

[7] Fortini M., P.D. Falorsi, C. Vaccari, N. Cibella, T. Tuoto, M. Scannapieco, L. Tosco. 2006. Towards an Open Source Toolkit for Building Record Linkage Workflows. In Proceedings of the International Workshop on Information Quality in Information Systems (IQIS), Chicago, 30 June 2006.