**Implementing Shared Statistical Services - Reuse Report**

**Introduction**

In addition to service implementation, another objective of the ESSnet is the reuse of the implemented services. This task aims at providing a proof of concept of the statistical service shareability. This report describes the Istat’s experience in reusing ARC, a statistical service implemented by Insée.

**Reporting organization**

Istat (Francesco Amato, Mauro Bruno, Eleonora Ciocca, Giuseppina Ruocco, Simona Spirito)

**Unit**

Methodological Directorate, Data Collection Directorate

**Report version**

First version

**Report date**

January 2021

**Contact mail**

mbruno@istat.it

**Service reused**

**Publisher**

Insée

**Name**

ARC

**Version**

**Main functionalities**

ARC (from the French: Acquisition - Réception - Contrôles) allows receiving (administrative) data supplied by the providers (several formats are supported, particularly XML), to control the compliance of the received files, and to transform administrative data to elementary statistical data. 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.

**Links (code, documentation...)**

<https://github.com/InseeFr/ARC>

**Service reuse**

**Context**

The main steps performed by Data Collection Directorate, for the acquisition and treatment of supplies relating to the administrative sources, are as follows:

1. Annual planning of data supplies

2. Metadata update process and formalization of requests to data provider

3. Data transmission and monitoring of data supplies

4. Standardized loading process of the System of Integrated Microdata (SIM)

5. Technical checks

6. Release of data to final users and archiving.

Data providers receive by email the login credentials to upload and transfer data files using ARCAM. ARCAM is the ISTAT Portal for the acquisition of administrative sources. It consists of two interfaces, one for uploading and transferring data by AD owners (data providers interface) and the other for monitoring acquisitions by ISTAT administrator users (administrator interface). The access credentials are personal and comply with all the security rules. The data provider interface guides AD owners throughout the transmission phase. Data providers have a temporary area on DMZ (Demilitarized zone) available for data transmission, which they can access again to add or delete files. Once the data transmission is completed, ARCAM produces an automatic receipt of successful sending.

The monitoring phase consists of a daily verification of supplies acquired or those expected but not arrived. In the event that the acquisition is delayed with respect to the agreed dates, reminders are sent to data providers.

Once available at the Institute, ARCAM users release microdata to the staff who manage SIM (SIM users). SIM users are enabled for data access by a specific function in ARCAM.

SIM is a technological and organizational infrastructure that allows the use of administrative data in statistical processes in compliance with rules concerning the protection of personal data. The main purpose of SIM is to obtain a single system for pseudonymization and archiving of any AD source containing individual identification variables. When data are uploaded into SIM, some technical checks are carried out to verify the compliance of the data received with those required.

After the pseudonymization process SIM users release data to the staff involved in the statistical production processes (final users).

**Business case**

ISTAT acquires both administrative sources with personal data and sources without personal data.

Administrative sources with personal data are processed by SIM. When this type of data is uploaded into SIM, the following tasks are performed:

* Verify that the table acquired respects the established Record Layout
* Carry out the following technical checks for each supply:
* Number of records in the file acquired
* For each variable, percentage of non-missing values on the total of records
* For each classification variable, frequency distribution of its modalities, excluding missing values.
* Display the results of technical checks in time series.

Currently this type of checks are not carried out for administrative sources without personal data.

**Service reuse**

***General description***

The purpose of ARC reuse is to check administrative sources without personal data, acquired by ISTAT, before the release to production division for carrying out statistical surveys.

ARC reuse has involved some colleagues of the Data Collection Directorate. Based on their knowledge, they have chosen one of the administrative sources to test ARC functionalities: “Anagrafe dei corsi post Laurea” (Register of post graduate courses), year 2018. The extract of the file employed to test ARC is “PL\_CORSI\_csv”. It contains a table made up of 4 variables:

* Code\_un (university code): alphanumeric variable;
* Anno\_validita (year of validity): numeric variable;
* Id\_tipo\_laurea (degree code): alphanumeric variable;
* Id\_scuola (post graduate master code): alphanumeric variable.

***As-Is and To-Do architectures***

The As-Is architecture is currently based on two main systems:

1. ARCAM for upload all type of AD sources and release of administrative sources without personal data;
2. SIM for AD treatment to check, pseudonymize and release acquired personal data.

The following figures shows the As-Is architecture.



Figure 1: AS IS architecture

In the proposed To-Do architecture (Figure 2), administrative sources without personal data are managed by ARC application, to carry out the same type of technical checks performed by SIM.



Figure 2: TO BE architecture

It is necessary to implement new ad-hoc procedures to carry out the technical checks by ARC.

***Technical details***

Due to legacy infrastructure policy, ARC has been installed in a test environment.

For each type of check, new procedures have been implemented in the MAP MODEL, one of the main functionalities of ARC.

***Organizational impact***

ARC could enable to define and manage controls and mappings for administrative data that are not processed in SIM.

The following steps may precede the integration of the service in the process flow:

* Identify the subsets of AD to check using ARC functionalities;
* Analysis of the identified subsets, to specify technical checks and metadata requirements.

***Project management***

***Other aspects (e.g. financial)***

***Results achieved***

The results achieved are explained in the following report:

|  |  |  |
| --- | --- | --- |
| **Check** | **New functionalities to apply (Yes/No)** | **Activity status** |
| Verify that the table acquired respects the established Record Layout | Yes | To be analyzed and implemented |
| Technical check 1 - number of records in the file acquired | No | (already existing in the application) |
| Technical check 2 - for each variable, percentage of non-missing values on the total of records | Yes | Implemented |
| Technical check 3 - for each classification variable, frequency distribution of its modalities, excluding missing values | Yes | Implemented |
| Display the results of technical checks in time series | Yes | To be analyzed and implemented |

**Lessons learned**

**Problems encountered**

The definition of mapping rules may be quite difficult in some cases and may require the assistance of ARC IT experts.

**Missing functionalities**

It is necessary to work out new procedures to carry out the following checks that are not implemented yet:

* To check that the record layout of the supply acquired is as established
* To display in time series the reports of technical checks.

**General feedback**

To Do

**Conclusions and next steps**

To Do