CONFERENCE OF EUROPEAN STATISTICIANS
**Expert Meeting on Statistical Data Collection**
27-30 September 2021, Online

# A fully metadata-driven platform for the conception of survey questionnaires and the management of multimode data collection

Franck Cotton franck.cotton@insee.fr, Thomas Dubois thomas.dubois@insee.fr, Eric Sigaud eric.sigaud@insee.fr Benoit Werquin benoit.werquin@insee.fr
Insee (France)

## Abstract

For ten years now, Insee has been developing a survey data collection platform using international standards in a metadata-driven approach. Several components have been developed one after the other: a questionnaire generator (Eno, cf. https://www.insee.fr/en/information/5014703?sommaire=5014796), a platform for business surveys (Coltrane, cf. https://www.insee.fr/en/information/5014714 sommaire=5014796), a questionnaire designer (Pogues, cf. https://www.insee.fr/en/information/5014167?sommaire=5014796), a PAPI case management tool (Sting) and a survey data editing environment (Genesis). More recently, the Metallica project has set up a new platform serving household surveys, which require better support for multimode, additional functionalities for Pogues and Eno, and the possibility to handle CAPI in all its dimensions. Metallica pursues the long-term strategy of standardisation and industrialisation of questionnaires, processes, organisation and services that was is at the heart of the metadata-driven approach since its inception, and reaps its benefits.

Overall, the collection platform will run 30 surveys in 2021, covering nearly 1 million units, with much more to come in the next years. With this new collection platform, Insee has a tool that is both powerful and flexible, and ready for future evolutions (specific surveys with complex protocols, metadata-driven processes, survey designer).

In particular, flexibility was demonstrated at the start of the pandemic crisis, when a new Covid survey had to launched in urgency: it took 26 days from the idea to the publication of the results.

# A fully metadata-driven platform

***For ten years now, Insee has been developing a survey data collection platform using international standards in a metadata-driven approach.***

This paper describes this long-term strategic operation from a historical, then functional point of view.

## Vision and timeline so far

### The start of internet collection for Insee business surveys

In 2000, the SESSI, which was the French Ministry of Industry's Statistical Department at that time, started collecting data from businesses via the Internet with its monthly branch survey, before extending it to other surveys in the industrial sector. With the merger of part of SESSI and Insee in 2008, the data collection infrastructure became the property of Insee. Insee already had its own system, since 2004, for a dozen frequent and regular business surveys.

Gradually, the rapid spread of the Internet within businesses and the desire to reduce collection costs have encouraged other producer services, from Insee or ministerial statistical departments, to digitise their questionnaires. By seeking to respond as quickly as possible to the aspirations of producers and the pressing demand of businesses, we ended up with a juxtaposition of heterogeneous tools. This situation led Insee to seek to reduce the number of web data collection infrastructures in order to facilitate responses from businesses, while also reducing the IT maintenance costs of the various platforms.

### An Ever Increasing Demand for Digitization

At the same time, during the French national conferences on administrative simplification in April 2011, businesses raised their difficulties with regard to the statistical burden. Among the 80 measures resulting from these conferences was the objective of "Digitising 100% of Official Statistical Surveys" and centralising questionnaires on a single site from 2013. Everyone agreed – and the business interviews confirmed this – that this method of responding was less expensive for businesses than using paper questionnaires.

At that time, the availability of data collection by Internet was still insufficient: the surveys for which it was technically possible represented 73% of the volume of questionnaires sent by Insee to businesses, but only 51% of the surveys of the entire Official Statistical System. Therefore, the aim was less to reduce the statistical burden in the traditional sense of the term ("reduce the volume of surveys") than to reduce the response burden by simplifying collection systems for businesses.

### Birth of a New Project for Simplifying Business Surveys and Harmonising Designer Tools

This was the context in which the Coltrane (COLlecte TRANsversale d'Enquêtes – Transversal Collection of Surveys) project was started in 2010. Its main objectives were to:

- develop a "technical platform" for multimodal data collection;

- provide respondents and Insee survey managers, or even the Official Statistical System, with a set of functions that meet their needs (authentication of businesses on the platform, management hub for survey managers, etc.);
- provide a usage instructions for integrating surveys on that platform;
- provide a contact management infrastructure.

The project already had a broad scope in terms of its schedule. Its original objectives did not include being able to meet the needs of household surveys (or the population census). However, the idea remained that all potentially common functions between the two data collection areas developed during the project should eventually be shared.

### A New Logical Pathway Based on Active Metadata

One particular feature sets Coltrane apart from existing platforms: this feature is the close links that it has had, from the outset, with the RMéS statistical metadata repository project (Bonnans, 2019) and, in particular, the stated objective of using the metadata that describe the surveys questionnaires to directly generate the collection tools, hence the term "active" metadata to describe the system.

Insee has adopted the GSBPM to model its statistical processes and the DDI standard (Data Documentation Initiative) to formally describe the life cycle of the data, in particular the "questionnaire" objects. This structural choice made it impossible to use one of the collection software packages on the market, as they were not yet adapted to these standards.

### The Questionnaire Generator: Eno

The automatic generation of collection media, i.e. of the screens and programs that allow responses to be retrieved, based on their DDI description, forms the basis for the expected productivity gains. The Eno generator makes it possible, based on the specification of the questionnaire, to automate tasks previously carried out manually: development of the collection medium by a computer engineer and then testing and acceptance by the statistician; it makes it possible to ensure that several different collection methods (including paper) correspond to the same general description.

Based on this principle, this metadata was reused and, combined with additional controls, Eno enabled to automatically build a data editing tool. Thus, a survey clerk could treat the data based on the review and selection of data in an equivalent view of the questionnaire.

Therefore, unlike in the past, the designer no longer makes the same number of specifications as there are types of collection media. A single specification is sufficient and it guarantees the homogeneity of the questioning whatever the method.

After a prototype achieved in 2013 generating Open Document questionnaires for the Annual Structural Business Survey, a first release of Eno was put into production in 2015.

### Birth of the Questionnaire Design Tool (Pogues)

The writing of questionnaires in DDI in advance has quickly turned into a bottleneck, as the advantages of this metadata description language are paid for by a high level of verbosity, which increasingly overwhelmed the institute's scarce expert resources: the back-and-forths between survey designers and programmers had been shifted to the two or three DDI experts, but not eliminated.

To make designers autonomous and end-to-end masters of the collection media creation process, they had to be allowed to write DDI without knowing it. In addition, they needed to be able to quickly observe the results of their amendments, without having to resort to an expert. The idea was born: a questionnaire

design tool, within a “designer’s workshop”, making it possible to create the questionnaire templates and view them in one click, while directly feeding into the RMéS statistical metadata repository.

Pogues was put into production in 2018.

## Coltrane functionnalities to date: a Data Collection Platform

Using Pogues, the statisticians design the questionnaire template and view it in a “blank” and non-personalised version, then approve the final version.

They must then use the management application for their survey to provide the Coltrane platform with the survey sample and its characteristics (in XML format). For each surveyed unit, this so-called “personalisation” file contains:

- all of the identification data for the selected businesses (unit username, company name, etc.);
- possibly contacts: contact username if it already exists in Coltrane and if they want to share it with other survey managers, address, first name, last name, position, telephone number, email address, etc.);
- and other data relevant to collection: responses from the surveyed unit from a previous collection campaign, if available.

Coltrane generates and hosts all the web questionnaires for the units to be surveyed, based on the blank, non-personalised questionnaire template provided by Eno and the personalisation file provided by the external survey management application.

Once the collection period for a survey is open, Coltrane extracts the responses that have been collected since the previous extraction, with the frequency chosen by the survey manager (in practice, up to 4 times a day) then sends them (in XML format) to the survey data editing tool.

## Coltrane to Date: Services for the Surveyed Businesses

#### An Authentication Portal

When it is selected for the sample, the surveyed business receives a letter (or even an email) notifying it that it has been selected to answer a questionnaire. This letter, in addition to providing a brief presentation of the survey, contains the elements required to connect to the platform: link to open, username and password assigned to the natural person who will answer the survey and who is called the “Contact”.

#### A Response Portal

Once authenticated, the Contact is directed to a response portal and, more specifically, to a tab named “My Surveys”, a single access point for responding to all surveys managed by Coltrane. At the time of sending responses, the Contact is given the option of downloading a summary of their response in “PDF” format.

From the response portal, the Contact can of course enter or update their personal information (first name, last name, email address, telephone number, position, postal address, etc.) using the “My Account” tab. The Contact may also personalise their password and, in particular, if the Contact has received multiple usernames (one per survey, for example), they can “group them together” and have a single account to access and respond to questionnaires for several surveyed units or several surveys. On a practical level, this is a great advantage of the portal.

**A Help Desk**

Throughout their navigation on the collection platform, the Contact within the business may access frequently asked questions and online assistance via a form. This assistance is contextualised, i.e. depending on where the form comes from and can be sent to two types of stakeholders:

- the Insee-Contact hubs, for “technical” issues (how to access the collection platform, the reality of the survey, objectives, lost access code, etc.);
- the survey manager teams, for “business” issues directly related to the survey.

## Coltrane functionalities to Date: A Module to Manage Mail to Businesses

To successfully carry out the collection for a survey, it is essential to contact the businesses regularly, to remind them of their obligation to respond to labelled official statistical surveys, for example. Usually, there are four letter templates: start of the collection period, reminder, formal notice and non-response report. Coltrane also provides a letter template for additions to the panel and another to thank outgoing panellists.

During the Coltrane project, Insee developed a module to create “ready-to-print” letters. This "Mail Module"is now available for any other application with similar needs. The letters may be accompanied by a paper questionnaire, depending on the collection strategy adopted.

## Coltrane functionalities to Date: An Internal Application to Manage Contacts

Coltrane is more than just a user-friendly collection platform that facilitates the work of survey designers. It is also a contact repository that will be used by survey clerks who process responses and are in contact with surveyed businesses.

To that end, when a survey is integrated into the system, Coltrane imports massively the survey contacts already known (in particular when internet collection was already in use). Then, in order to bring this repository to life, the survey clerks use a “Contact Management” application: to create a contact, to amend its characteristics, to give the contact the right to respond to a survey or withdraw that right, etc. In addition, updates to the contact’s details, made by the respondents directly are recorded in real time in this repository.

Furthermore, the clerks of a survey can view the questionnaires of their respondents regardless of their status (already sent to Insee or not). In particular, this makes it possible to help businesses that have difficulty answering or to verify the sending or simply the saving of a questionnaire.

## A New Data Collection Project for households surveys

Overall, the collection platform will run 30 surveys in 2021, covering nearly 1 million units, with much more to come in the next years. With this new collection platform, Insee has a tool that is both powerful and flexible, and ready for future evolutions (specific surveys with complex protocols, metadata-driven processes, survey designer). In particular, flexibility was demonstrated at the start of the pandemic crisis, when a new Covid survey had to launched in urgency: it took 26 days from the idea to the publication of the results.

It represents a first achievement for data collection system at Insee. In addition to questionnaires, it offers services for surveyed businesses, a module to manage mail, an application to manage contact. At that time of achievement of Coltrane, demands to conduct household surveys on the Internet was growing and Coltrane offered opportunities to pool modules. The project called Metallica was born. Its first step was a household web collect platform, a “Coltrane’s Twin”.

### New functionnalities, new collect modes

If initially only web collections, with collection tracking and mailing services, were supported, the tools quickly expanded to support telephone and face-to-face collections with more detailed process monitoring and management functions centered on the interviewers' activity, with the target being multi-mode collection support, switching between modes and post-collection process automation (GSBPM Phase 5 - Process).

## Capabilities and tools (GSBPM projection)

***1 - Questionnaire Design***

### Survey design workshop (Conception)

Description: the survey design workshop is a collection of documentation (questionnaire design training, collection information system service offering documentation), processes (specification/receipt, deployment, etc.), and tools that enable stakeholders to specify and configure a survey collection process in the collection information system.

### Pogues (Conception)

Description: Pogues is a questionnaire template design application. Its user-friendly input interface allows a design by successively adding questions and organizing them into modules or sub-modules. Pogues is interactively linked with the Eno application for generating questionnaire models. This link between Pogues and Eno allows the designer to have an instantaneous view of the questionnaire model being built in the collection mode envisaged. Pogues is a tool attached to the statistical metadata repository RméS.

***2 - Generation and deployment of questionnaire models***

### Eno (Conception)

Description: Eno is a tool that generates survey questionnaires from their formal description in an international standard (DDI): put differently, Eno automates the production of collection instruments (self-administered web questionnaire or paper questionnaire, collection module for interviewers). These questionnaire models are then customized to produce the collection media, according to the different modes of collection (web, paper, telephone, face-to-face), within production collection infrastructures.

### Coleman Back office (Construct)

Description: Coleman's BackOffice services allow you to import into the Coleman web collection system the various elements needed to launch a survey: a sample, questionnaire templates, parameters, etc. This system allows you to prepare the collection and to launch the edition of the launching mailings.

### Sabiane Back office (Construct)

Description: Sabiane's BackOffice services allow you to import into the Coleman web collection system the various elements needed to launch an intermediated survey: a sample, questionnaire templates, interviewer assignment information, parameters, etc. This system allows you to prepare the collection and to launch the edition of the launching mailings.

### Protools (Construct)

Description: Protools is an application for administering the tools of the collection information system. It allows you to fill in and consolidate (from the ad-hoc repository) the metadata that make up a survey

(name, general information, specific FAQ, etc.) and those that allow you to configure its protocol (schedule, mode change strategy, statistical “scores” used during the protocol). It also allows you to manage user rights and roles on the various IS tools.

***3 - Online Collection***

## Coltrane (Collect)

The Coltrane portal is a “My Surveys” web-based application that displays a list of the different surveys, and the start/end dates of collection, in which a company respondent is participating. It provides secure access to web-based questionnaires and includes features that allow respondents to update their personal information. It also includes authentication and assistance functions for respondents.

## Coleman (Collect)

The Coleman portal is a website for promoting household surveys. It provides respondents with a display of information contextualized to the collection of a survey and allows secure access to Internet questionnaires. It integrates authentication and assistance functions for respondents.

## Stromae (Collect)

Stromae is a web-based data collection platform. It allows authenticated access to an internet questionnaire by a respondent, the storage of his or her answers, the provision of a proof of submission and manages the life cycle of internet questionnaires (response status, extraction, mirror site, etc.).

It allows the operation of internet questionnaires produced by a generation channel based on active metadata.

It is used for company surveys in version 1. A version 2, following a technological migration, will be used at the beginning of 2022 (milestone 2 Metallica) for household surveys. This version 2 offers a better accessibility of the internet questionnaires and a better support of the different response media (smartphone, tablet…) and a better resistance to load. It is intended to host all the internet, household and company collections, after migration of the latter.

***4 - Intervewer Data Collection***

## Sabiane & Queen (Collect)

Sabiane is a data collection application for interviewers. It offers data collection management functions for the interviewer, such as searching for and consulting information on the surveyed units, qualification functions for the identification and/or contact phases prior to data collection, and “interviewer questionnaire” (Queen) functions that allow telephone or face-to-face interviews to be conducted. It allows the operation of questionnaires produced by a generation process based on active metadata.

***5 - Surveys management***

## Coltrane (Collect)

The Coltrane Pilot application is designed for business survey managers to track collection and manage contacts. It is also intended for first and second level helpdesks to display contact information and useful information when interacting with a respondent.

### Moog (Collect)

Moog is an application for national and multi-mode monitoring of household survey collection. It offers functions for national pilots in their role of implementing the process (management of reminders, undelivered mail, etc.) and for support services in their mission of responding to respondents (information on respondents and on the survey).

### Sabiane (Collect)

Sabiane gestion is an application designed to monitor and manage a so-called local collection. As such, it offers a contextualized view according to defined organizations (they can, for example, represent the territorial distribution, such as a household survey division in a region), search/consultation functions for surveyed units (progress status, results of the identification/contact phases, etc.) and tables for monitoring the overall progress (within an organization) of the collection. It also allows you to manage manager/investigator workflows, such as questionnaire proofreading or validation of mail requests, and to close the collection of survey units that have not been finalized by an investigator.

***6 - Export data***

### Kraftwerk (Traitment)

Description: Kraftwerk is a post-collection processing engine. It allows the automatic execution of various post-collection treatments and controls on the basis of active metadata for the constitution of statistical databases for downstream processing.

## Next steps

### New functionalities

In the Insee active metadatas roadmap, the first next steps that one would consider are adding new functionalities beyond the building questionnaire process.
And natural ones are going deeper in the GSBPM process and add active metadatas in the '5 - Process' activity. Questionnaire metadatas could be easily re-use for building tools to review or correct the collected datas. Actually, a first implementation of this principle already exists at Insee as a questionnaire editor & validator for simple business survey. This first prototype of the principle would be extended to a questionnaire editor & validator for all surveys (household and complex business surveys) that need it.

But, others improvements inside the '4 - Collect' activities could be done. Using metadatas to describe the collect process would permit to pilot technical process based upon these 'active' metadatas which help to build process 'rules': Rules to choose which non-respondant that should be re-contact, which questionnaires should be considered finished even if they're not formally validated, which respondants should be switched to another collect mode (ie: from web to Capi). A set of metadata intended to describe the process as a whole and allow its traceability.

All these active metadatas are the 'description' of the surveys. And the Pogues tool should be extended to a real survey specificator tool. Where all the metadas related to the survey could be captured and used to generate technical tools and pilot the entire Collect & Process activities.

### Omnimode

But even in the already mature questionnaire generation process improvements should be made. The omnimode questionnaire is the goal. It's an unique specification for the questionnaire used for all the collect modes. It's a methodoligical concept to deal with the complexity of a multimode results' consolidation. But having a single specification for all the questionnaires offers other openings : same

variables, same interfaces, same process.
Particularly, pre-collect data's upload process and post-collect data's download process could be mainly automatized and standardized based upon this unique specification of the data variables used in a process collect.

## Conclusion

Insee has industrialized part of its collection processes, so 1 million questionnaires are operated by the web collection platform for business surveys and in 2022, the same number of questionnaires (with 3,000 telephone interviewer) will have been taken over by these platforms driven by metadata.

Efforts will continue to improve existing collection processes and associated tools.
Thus, if the collection and questionnaire applications, the tools allowing the specification and generation of questionnaires, are open-source. Progress remains to be made in this area, in order to improve the visibility offered on these tools so as to have a real open-source community around this work.

Other tools in the field, such as those for process monitoring in particular, would also benefit from being made open-source.

In the same way, if the metadata steering is a reality in the construction of questionnaires, more active metadatas can still be injected into the process, whether it is for its follow-up, its steering or for the post-collection phases of processing.

Finally, this industrialization is accompanied by a major effort to standardize practices, communications (paper mail or emails sent) and protocols. This effort must be continued in order to offer an industrialized, secure and metadata-driven collection process.