The French statistical Metadata Repository, **RMéS**:

*managing metadata throughout the whole statistical process*

**Marie Hélène Kérouanton, Insee, Quality Unit,** [**marie-helene.kerouanton@insee.fr**](mailto:marie-helene.kerouanton@insee.fr)

**Thomas Dubois, Insee, Quality Unit,** [**thomas.dubois@insee.fr**](mailto:thomas.dubois@insee.fr)

**Abstract**

*Since the 80s’ Insee has been developing a tool: DDS “Dispositif de Documentation structurée” (a structured documentation tool) used to describe and record structured metadata around statistical studies. The RMéS project (Repository for statistical metadata) intends to achieve similar goals all the while being compliant with the related international standards in order to enhance data exchange. The RMéS project will enable metadata to be shared and reused and will set up a metadata-driven production system.*

*Moreover, providing tools to manage metadata all along life cycle, the RMéS project will reduce the burden on statistical producers and improve the quality of metadata, by avoiding redundancies and inconsistencies. RMéS will deeply alter the way statistical operations are carried out. This necessary effort to change practices will benefit our entire institution and will act as a "quality catalyst" for our statistical products.*

**1. To start with…**

... **statistical metadata** are data that defines and describes other data:

* statistical concepts and classifications, used by many statistical operations;
* statistical metadata associated with each statistical production operation: statistical population, questions, questionnaires, variables, code lists, documentation on data collection process and methodology, quality report, etc.

 ...**a statistical operation** can be defined as all the phases of collecting, processing and disseminating data. It is year stamped and includes statistical surveys or administrative data (e.g.: the Enquête TIC 2017 [ICT survey], the Déclarations Annuelles de Données Sociales 2015 [Annual declaration of social data]).

  ...**a repository** is a single location for recording objects, with identification rules that ensure their uniqueness. In the RMéS context, the objects are metadata.

**2. Why overhaul our metadata management tool ?**

Several reasons made the implementation of a new metadata management tool at INSEE necessary:

* for technical reasons, *DDS “Dispositif de Documentation structurée” (a structured documentation too used to describe and record structured metadata around statistical studies)* is no longer intended to be maintained;
* DDS lacks reference to international norms and standards, while the volume of data exchanged is shooting up and their use must be automated;
* DDS offers few possibilities in terms of sharing and reuse metadata between several statistical operations, which generates redundancies and increases the risk of inconsistencies.

**3. The challenges of the RMéS project**

Setting up this repository is part of a process to improve the quality of Insee's statistical production. It will foster overall production consistency and will make it possible to automate certain operations using metadata.

There are three issues involved:

* Providing the best possible service to external users of Insee's statistical data, by ensuring that such data is clear, comparable and reusable, in accordance with the principle of "accessibility and clarity" of the European Statistics Code of Practice.
* Complying with international statistical production standards to support exchanges and comparisons between countries, in accordance with the principle of "coherence and comparability" of the European Statistics Code of Practice.
* Improving Insee’s public image by collectively working to produce quality data.

**4. The construction of RMéS**

*4.1. Storing metadata in a repository*

The RMéS is a repository for centralising all statistical metadata.

First, all the statistical concepts and classifications have been migrated from the current system to repository. Then, the first step in structuring metadata for statistical operations was to compile a list of statistical operations, to be shared by all actors involved throughout the statistical production process: Advisory committee, Committee on statistical confidentiality, Data remote access center, archive, insee.fr. Then the migrated metadata from the old system were assigned to each statistical operation.

*4.2. Designing models in accordance with international standards*

* Data collection instruments and variables are described according to the DDI standard and stored in a Colectica Repository©.
* Reference metadata are summarised in the SIMS (Single integrated metadata structure, European standard for quality reports adopted by the CSSE in November 2015)
* Concepts, classifications, list of opérations and quality reports are stored in a Datalift triplestore and published in open data linked format (RDF).
* A SDMX-RDF ontology has been created to implement quality reports (SIMS) in RDF
* Concepts and classifications are modeled according to SKOS and XKOS

These are being released in open source.

As a consequence, for technical reasons, the repository is structured in two bases (Colectica repository© and Datalift triplestore), but the uniqueness of the objects is guaranteed by their identifiers.

*4.3. Developing management tools and web services*

Respective owners of metadata, classification managers, survey designers, or the Quality Unit will be able to create or update their metadata using dedicated management applications.

Several management tools are being displayed:

* Colectica Designer© is being used to manage DDI metadata.
* Insee is developing a specific software to manage RDF metadata. A GUI to design questionnaire has also been developed, which provides a DDI file to be stored in Colectica repository© and used to generate questionnaires.

A set of web services allows client applications to use metadata stored in both metadata bases, Colectica repository and DataLift triplestore.

**5. Whom does RMéS concern?**

 As a structuring process for Insee's statistical production, the RMéS concerns everyone:

* Prescribers, who must ensure the process is adopted: the heads of the division in charge of supporting project owners, Comité des Investissements (Investments committee), Comité de suivi des projets (Projects monitoring committee), etc.
* The designers of statistical tools, who will have to use the RMéS metadata and services in their applications: project managers, project leaders, etc.
* Internal users, who will use it in their work – survey developers, disseminators, etc.
* Statistical operations users, who will benefit from better data quality: all those who use external statistical files, including Eurostat, statistical ministerial services, etc.

**6. What to expect from RMéS?**

In practical terms, RMéS will help lighten certain areas of the statistician's workload.

* Centralising and versioning[[1]](#footnote-1) the information will mean that metadata can be entered just once and then updated throughout the lifecycle of a statistical operation.
* Centralising and standardising will make it easier to pool efforts either internally or with our partners in a "shared" metadata approach.
* "Active" metadata will help automate the production of components in the statistical process (such as questionnaires for data collection platforms).

**7. The changes induced by RMéS**

* RMéS will deeply alter the way statistical operations are carried out. The necessary changes includes:
* Increased coordination, to ensure, right from the design phase of a new operation, consideration of the RMéS and its objectives.
* A new organisation that ensures that metadata is an integral part of all statistical work at Insee and that its management is effective at all stages of the process.
* This necessary effort to change practices will benefit our entire institution and will act as a "quality catalyst" for our statistical products.

**Terminology on RMéS**

API: an *Application Programming Interface* enables a software to offer services to another software.

CASD: the Remote Secure Access Centre (Centre d'Accès Sécurisé à Distance) is a facility that provides a secure tool for accessing individual data that is highly detailed, and therefore often subject to stringent security requirements, for researchers, data scientists, consultants, etc.

Colectica©: Colectica© is a software suite. Colectica Repository© will be the reference warehouse for storing the metadata from statistical operations (for example, variables, codes, questionnaires) described according to the DDI (Data Documentation Initiative) standard. Colectica Designer© is the graphic user interface designed for managing these objects. Colectica Portal© is is a web application, powered by Colectica Repository©, which enables data and metadata publication and discovery..

Coltrane (COLlecte TRANsversale d'Enquêtes): information system for collecting statistical survey data from businesses.

Datalift: the Datalift warehouse stores the metadata in more detailed format, as RDF (Resource Description Framework) triples, the basic Semantic Web language, makes it easier, for example, to automate the use and interlinkage of metadata and update the "definitions", "nomenclatures" and "sources" sections on Insee.fr.

DDI: *Data Documentation Initiative* is an international initiative launched in 1995. Its aim is to create and maintain a technical documentation standard for describing and preserving statistical information, particularly survey data. Standardising this documentation involves modelling the various statistical concepts (questions, questionnaires, variables, code lists, etc.) and their relationships in the form of XML documents.

DDS: The structured documentation system (Dispositif de Documentation Structur閑), a relational database, contains the documentation of an operation, and for more than thirty years has been used to manage and store documentation from a variety of statistical operations. Not all DDS databases fall within the scope of statistical metadata. Those that need to be migrated to the new system have been identified. For "occupation" knowledge bases, appropriate solutions have been implemented.

EDL(Local Data Warehouse): The EDL is a reference warehouse for data that is specific to the area, and even geographic units known as IRIS (*Ilots Regroupés pour l'Information Statistique*; aggregated units for statistical information), used for studies and advisory services, and for producing localised dissemination products made available on Insee.fr.

ENO: This questionnaire generation tool was developed as part of the Coltrane project.

Generic: This project aims to pool the initial processing of thematic surveys designed for businesses.

Pogues: This graphic questionnaire design interface, allows survey developers to design a questionnaire in DDI-compliant (Data Documentation Initiative) format, and to view its Web version (using ENO) at the push of a button. By the end of 2018, it will be connected to the RMéS repository.

Cnis portal (Conseil national de l'information statistique; National council for statistical Information): The Cnis website publishes practical information on the life of the Conseil National de l'Information Statistique, such as schedules of meetings of the various bodies, meeting announcements, agendas, and also descriptions of ongoing surveys that have been approved.

RDF*(Resource Description Framework)*, recommendation of the W3C (World Wide Web Consortium), makes it possible to disseminate data and metadata according to "Linked Data" principles. This graph model formally describes Web resources and their metadata. This initiative promotes the publication of structured data on the Web, not as isolated independent data, but by connecting such data to make a global information network.

SIMS (Single Integrated Metadata Structure): set of streamlined, quality, reference metadata, defined and maintained by Eurostat. Externally, it is already being used to provide quality reports at Eurostat. Internally, its expanded version will be used for documenting sources.

1. Versioning: storing and displaying successive, dated versions of an object, for example, in the event of revised definitions or nomenclatures. [↑](#footnote-ref-1)