Facilitating Access to INSEE Data

Cubes, Catalogue and Metadata

Jocelyne Mauguin * *and Nicolas Sagnes* **

INSEE (National Institute of Statistics and Economic Studies) publishes on its website a vast amount of data covering numerous economic and social themes such as demography, employment, national accounts, and price index. Given the richness of its website, INSEE must guide its users towards their desired data. Presenting all statistical data in a simple and coherent manner on the *insee.fr* website is therefore a significant challenge. A first level of dissemination, which is data visualisation, allows to comprehend a given subject through synthetic indicators presented as simple, clear, and easy-to-understand visuals. However, to go further in the analysis, more detailed data are made available. These are typically presented in an aggregated form: multidimensional cubes that cross-reference various variables of interest such as gender, age, or socio-professional category in household surveys. The challenge then becomes offering these informations in well-standardized and open-source formats, while also thoroughly documenting them, relying on international standards. These data must also be well-catalogued to facilitate discovery. To access them, INSEE's data consultation services are being updated to make it possible to navigate through these cubes. Finally, the data must be accessible both to internet users and to machines that harvest them: the use of the latter opens up new prospects for data consumption modes through artificial intelligence.

* Project Manager, INSEE, Direction de la diffusion et de l'action régionale (Directorate of Dissemination and Regional Action, DDAR).
jocelyne.mauguin@insee.fr

** Head of project, INSEE, DDAR.
nicolas.sagnes@insee.fr

INSEE publishes a vast amount of reference data for economic and statistical information on its website. With the development of data, one key challenge is to keep this offering as up-to-date, readable and accessible as possible. There are many challenges to address in order to achieve this, concerning not only the importance of data formats, documentation and its standards, but also services such as data visualisation, cataloguing or data navigation, not to mention the APIs[1], which are essential for the use of data by machines.

Presenting a Vast Amount of Data in a Simple Way

INSEE disseminates statistics on many topics, such as demography, employment, national accounts and price indices. These account for a large part of official statistics, with output on other topics being produced mainly by the Ministerial Statistical Offices (MSOs). These statistics are essential for the preparation of economic studies, in order to provide clarity on structural issues[2], the breakdown of inflation, factors driving poverty, etc.

This diversity is reflected in the vast amount of data following in the wake of the explosion of data in recent decades[3]. A few figures: INSEE publishes approximately 5,000 XLSX[4] files and 70,000 historical series on its website each year (for example, the gross domestic product series since 1949 or the monthly series of consumer price indices).

Given that it offers such a comprehensive range of data, INSEE must provide support to users of its website to help them find the data they require. The data must be easy to find and understand. To that end, INSEE strives to follow the main principles of the European Statistics Code of Practice[5], the cornerstone of the common quality framework for European statistical institutes. Consistency/comparability and accessibility/clarity are the key principles for the dissemination of statistics:

> ***INSEE must provide support to users of its website to help them find the data they require.***

- **Consistency and comparability**: comparisons of data over a reasonable period are possible; statistics are compiled on the basis of common standards for definitions, units and classifications in the various surveys and data sources.
- **Accessible and clear data**: statistics are presented with documentation for proper interpretation and useful comparisons; modern technologies, methods and platforms for information and communication are used; open data standards are offered, with access in a non-proprietary format (Ubaldi, 2013; Emilsson et al., 2020).

1 Application Programming Interface. The term web service is also used. The website insee.fr currently offers a web service, the results of which meet the international SDMX standard.
2 See (European Commission, 2015).
3 See for example *https://project.opendatamonitor.eu/*.
4 XLSX is a spreadsheet filename extension in Office Open XML format used by Microsoft Office from the 2007 version onwards.
5 *https://www.insee.fr/en/information/4249492*.

In order to apply these principles, the great diversity of the profiles and expectations of INSEE website users must also be taken into account. Here are a few examples: a student has to make a presentation on national accounts and only needs to consult a table of the main national accounts statistics (GDP, value added, etc.) on a web page; an individual rents out their apartment and wants to obtain the rent benchmark every year to reassess the rent; a researcher wants to analyse residential migration between municipalities and, to do so, they want to download the file containing granular population census data, etc. INSEE chooses to reach out to all these audiences and, as such, must offer different ways of accessing the data, starting with data visualisation (De Jonge and Ten Bosch, 2012).

▶ Figures to Facilitate Data Access

To let users discover the essential information for a topic, INSEE offers its key figures, which are often presented as infographics and summary tables: data visualisation, i.e. a set of summary indicators in the form of simple visuals that are easy to understand (Lagarenne et al., 2023). This is INSEE's preferred method of supporting its website users in reading the data and enabling them to understand the results of a study more easily. Thus, a time series[6] represented graphically in the form of a curve based on the available periods satisfies the needs of all audiences on most topics (consumer or production price indices, unemployment figures, employment, etc.).

Another example is the *Tableau de Bord de l'Économie Française* (French Economy Dashboard, TBEF), a multi-thematic data visualisation service provided on the insee.fr website. All the essential information on the various areas of public debate (economy, purchasing power, demography, society, wages, companies, sustainable development, etc.) is presented in accordance with three geographies (Europe, France and regions) (***Figure 1***). Statistics Denmark[7] provides a thematic navigation tree in the "Find statistics" section of its website: once the field is chosen, the statistical data are presented as figures with options for downloading this data and for a more granular analysis of the field.

▶ Downloading Data for Reuse

On the INSEE website, the data visualisation figures always include an option for downloading the data. This can be used by students to support their presentations or by economics teachers to prepare their courses. Journalists also take an interest in updates to these indicators or survey results to prepare an article; data journalists download the time series, in particular, to analyse a dataset in support of or in addition to a background article.

Beyond data visualisation, files with larger volumes of data are available, including in XLSX format. These files have more granular levels of detail or gather together all the available information on a given theme and not just an excerpt, as is the case for data visualisation figures. This provision of files to download is for INSEE website users who want to use the data directly for their own analysis, such as consultancy firms, researchers or certain

6 For example, this page grouping together the main indices and time series: *https://www.insee.fr/en/information/2868584*.
7 *https://www.dst.dk/en*.

▶ Figure 1 - View of the French Economy Dashboard

1. Entry by theme
2. Selection of regional, national or European view
3. Summary of the indicators of the theme, with a "Learn more" section to access additional datas
4. For downloading the data table
5. Data visualisation of summary indicators

> ***These downloadable files are for INSEE website users who want to use the data directly for their own analysis.***

local stakeholders. Thus, a Regional Council can study the economic activity of its region using the files on company creations aggregated in a highly granular manner on the basis of geographical location, activity, size and legal category of the companies. The level of detail can sometimes extend to personal data such as births, marriages or deaths taken from the civil register. The Regional Council can then perform its own aggregations and assess the needs for equipment and facilities based on the population of its region.

Organising the Supply of Datasets

Given the variety of themes and the diversity of applications, INSEE must organise its collection of files as effectively as possible, starting by defining their content. The challenge lies entirely in creating data files with relevant **axes of analysis** (also known as variables) for INSEE website users with diverse profiles. Considering wages as an example, if a journalist was interested in gender inequalities, they would compare wages by focusing on gender, while a researcher tracking changes in wages throughout a person's career would focus on age. It is therefore important to provide a dataset on average wages cross-referenced according to the axes of analysis "gender" and "age" to meet both needs.

The size of the data files is also important in this example. The files must not be too big (difficult to use for INSEE website users) or too small (need to consult in depth to analyse a topic). For example, a file containing data from the population census that includes all INSEE information on the French population would be far too large and the INSEE website user would easily get lost. It must be split based on themes, such as housing, family or the foreign and immigrant population. Decoupling can also be based on the degree of information: a file on housing with the main information to be determined, supplemented by a file containing additional information, intended for more advanced INSEE website users[8].

The Need to Standardise Datasets for Easy Use

In order to facilitate their use, the format of data files is generally standardised. So-called flat formats are used, first of all the CSV format or more recently Parquet (Dondon and Lamarche, 2023), because they are easily readable in a programming language[9] or even in a spreadsheet[10] if the file is not too large.

The statistical content of the files is also standardised. Firstly, each column of the file corresponds to a variable that is entered in accordance with its values. Next, the files do not contain labels in the column or row headings but codes, which are easier to use when a person wants to use the file: the column heading is a code relating to a variable (for example, the AGE code stands for age) and each cell in this column is a code relating to the values for this variable (for example code "Y35T39", which represents the age range of 35 to 39). Finally, the values in each column are in the same format. The main formats are date, character string or numeric format. Since the format of each column is fixed, their contents can be used more quickly by computer-based data analysis tools.

By way of support accompanying the file, the variable codes and their values are documented in a code dictionary where they are linked to their labels and grouped into code lists. For example, the variable code AGE has the label "Age" and has a code list consisting of codes such as Y35T54 (labelled "35 to 54") or Y_GE75 ("75 or over") (***Figure 2***). File variables can also be attached to well-defined semantic concepts. In the example, the code variable AGE will be attached to an age concept that specifies

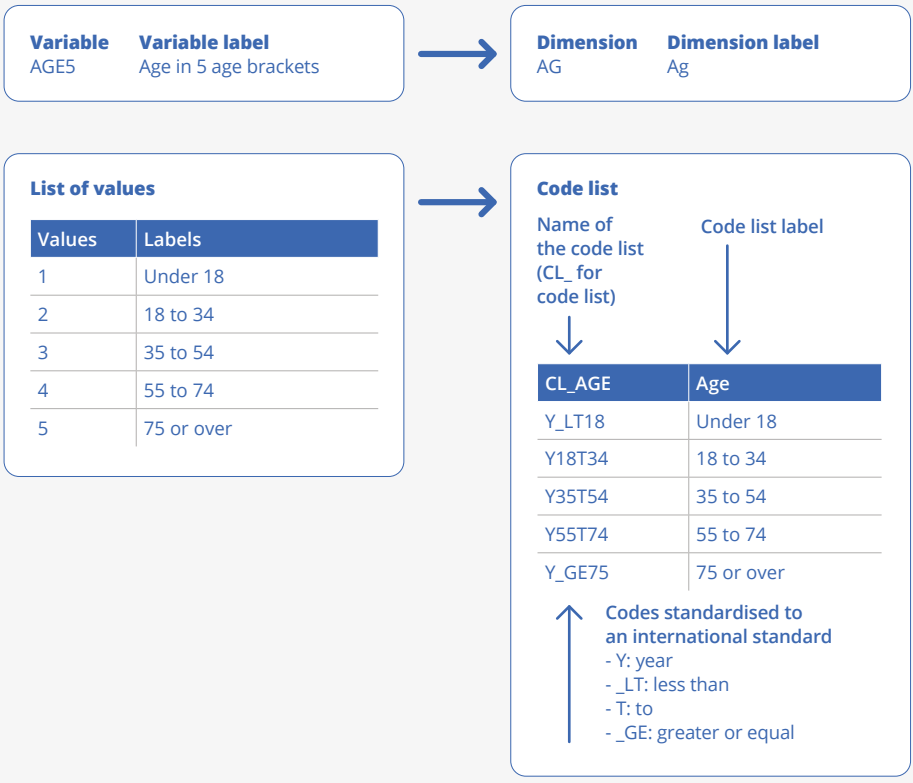
8 Such a decoupling reflects the way in which the population census is designed by INSEE: a primary use and a complementary use.
9 R or Python.
10 For example, Calc from the LibreOffice program suite.

> ***Variables must be comparable from one data file to another when they have the same meaning.***

whether it refers to age in terms of completed years or calendar years. Similarly, general information in the data tables such as details in the title, unit of measurement and the provisional or revised nature of the data is formalised and grouped into variables and code lists.

All of these data descriptions, known as structural metadata, are essential to understanding the data (Bonnans, 2019). When an INSEE website user is interested in a subject, they generally wish to obtain all the information available on it. Variables must therefore be comparable from one data file to another when they have the same meaning. Hence the importance, and even the necessity, of harmonising the identical metadata of the different data files, in order to ensure that the data are consistent across sources. To that end, a description standard that conforms to international standards is used by INSEE.

▶ Figure 2 - Modelling of the Age Variable



▶ Structured Data in the Form of Multidimensional Cubes

A fundamental concept of dissemination is that of a dataset, which refers to the information contained in the data file. Datasets must be considered to be distinct from files, since the same dataset may be present in multiple files of different formats.

> ***The datasets will be structured in the form of "multidimensional cubes" or "hypercubes".***

The datasets will be structured in the form of "multidimensional cubes" or "hypercubes", the dimensions of which are the axes of analysis. There are several hundred such axes throughout INSEE's output; the most common axes being age, gender, socio-economic group, industry and category of company (from a legal point of view or according to size). The points at which the dimensions of these cubes meet provide the values of the indicators, such as the number of inhabitants, the number of companies or income.

These multidimensional cubes are described using the international standard SDMX[11] [12], and, more specifically, its information model. This standard is used by the European statistical institute, Eurostat, in its data exchanges with Member States and by the United Nations for Sustainable Development Goal indicators[13]. These websites provide a clear explanation of this descriptive standard and its use format on dedicated pages of their data sections. The code dictionary of the Y cube is known as the Data Structure Definition (DSD). There are three types of cube variables: measures, dimensions and attributes. These components are defined as follows:

- **Measures**

The measures represent a phenomenon observed through statistics (population, accounting operations in business statistics or national accounts, overnight stays in hotels, indices of prices for consumption or industrial output, etc.).

- **Dimensions**

The dimensions correspond to the axes of analysis of the phenomenon being observed. If a person is interested in a population, it may be of interest to break down this measure using dimensions such as gender, age or employment status. Two dimensions have a particular status in dissemination: the time period (typically the year of reference for the data) and the geographical level (the region, for example).

- **Attributes**

The attributes provide information that is not essential to the value measured but is necessary to understand what is being measured. They can specify, for example, units of measure (number of persons or full-time equivalents), scale factors (units or thousands) and (final or provisional) value status.

11 SDMX stands for Statistical Data and Metadata eXchange. The SDMX initiative, launched in 2002, sets out standards to facilitate the exchange of statistical data and metadata between international organisations and their member countries, using modern information technologies. This format is sponsored by seven international organisations: the Bank for International Settlements (BIS), the European Central Bank (ECB), the statistical office of the European Union (Eurostat), the International Monetary Fund (IMF), the Organisation for Economic Cooperation and Development (OECD), the United Nations Statistics Division (UNSD) and the World Bank. For further details, see (SDMX, 2012).

12 Another standard based on the SDMX information model is the semantic Datacube web standard.

13 *https://unstats.un.org/sdgs/dataportal*.

With this modelling, a table of the population of Nantes broken down according to different axes corresponds to a multidimensional cube in which the measure is the population, with the dimensions of sex, age, socio-economic group, municipality and year, and in which the "number of people" attribute indicates that the population is measured in units and not in thousands of people, for example (***Figure 3***[14]). Another example: from the *Chiffres-clés* (Key figures) table of the *enquête Cadre de vie – sécurité* (Living environment and security Survey) on the number of victims of assault or theft outside the household by age and sex[15], the measure is the number of victims of assault or theft outside the household. It is measured on the basis of three dimensions: sex, age and type of violence. The units of measure (values in thousands of people, complaint rates as percentages) are informative and are therefore attributes.

The dimensions and their code lists can be reused from one dataset to another, which makes it possible to filter the datasets that contain the desired dimension (for example, age) or even to go further by more precisely filtering those that contain a certain code for that dimension (specifically, a particular age bracket). This is a very useful search function for a catalogue.

▶ A Catalogue to Discover Datasets

Designing and structuring data sets is not sufficient. INSEE website users must also know that they exist! This is why they are usually presented in a dedicated catalogue. INSEE website users can search this catalogue for a dataset of their choosing based on different criteria and then obtain the information and access associated data. The search criteria are very important for finding files efficiently, and the clearer the description of the datasets, the more accurate the search result will be.

In order to describe them well, international standards can be used, such as DCAT[16]. This standard describes the metadata used for cataloguing, that is, the relevant fields of a dataset that constitute the possible search criteria: for example, its creation date, its theme, its year of production, its geographical level (municipality, department, region, etc.) or its statistical programme[17]. The DCAT standard particularly helps to ensure that international comparisons are possible between the datasets of the various national statistics institutes (NSIs). Ultimately, a dataset will therefore have two types of metadata: its cataloguing metadata and its structural metadata (***Figure 4***).

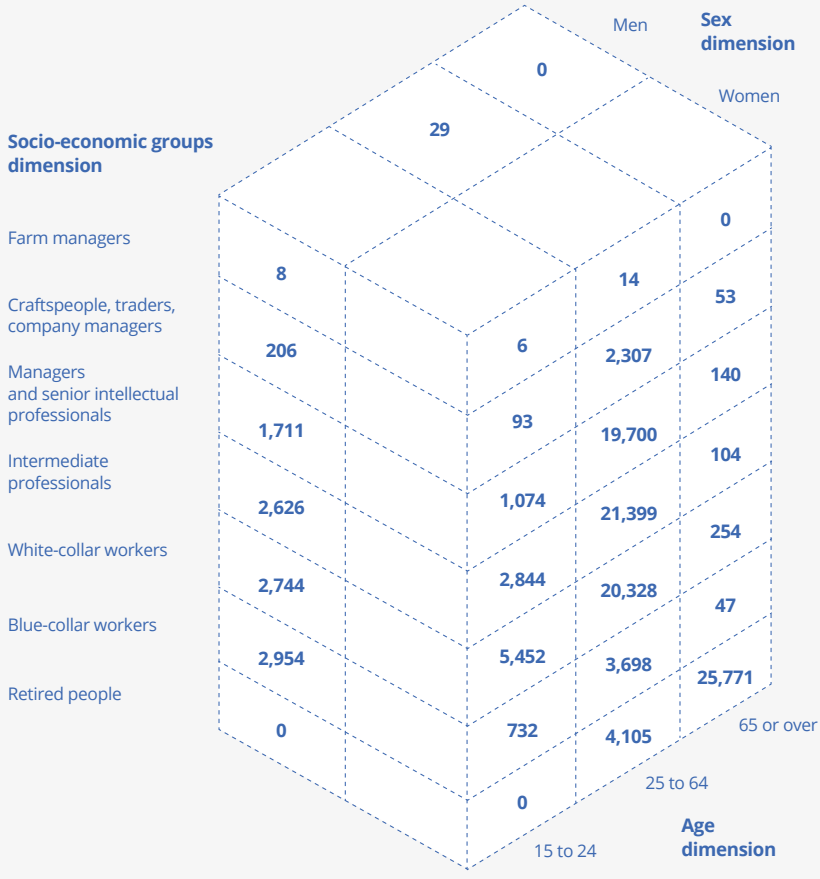
Once the datasets have been described, they can be presented in an online catalogue interface for easy access. This interface presents all the datasets and allows website users to filter them according to the search criteria. It also provides additional information about each dataset (a summary or the temporal coverage of the data).

14 In this case, it is possible to create a graphical representation of the cube, because it has only three dimensions.
15 *https://www.insee.fr/fr/statistiques/2525801*.
16 DCAT stands for Data Catalogue Vocabulary. The European Commission has endeavoured to describe a shared framework for cataloguing information, in the case of data catalogues. The data catalogues can equally come from the data providers (statistics institutes, government bodies, operators, etc.) or from aggregation websites providing aggregated information.
17 *https://www.insee.fr/en/metadonnees/sources*.

▶ Figure 3 - A Data Cube for the Labour Force Aged 15 or Over in Nantes in 2020, According to Sex, Age and Socio-Economic groups

Measure = Population
Geographical dimension = Nantes
Time dimension = 2020

Unit = Number of people
= Attribute

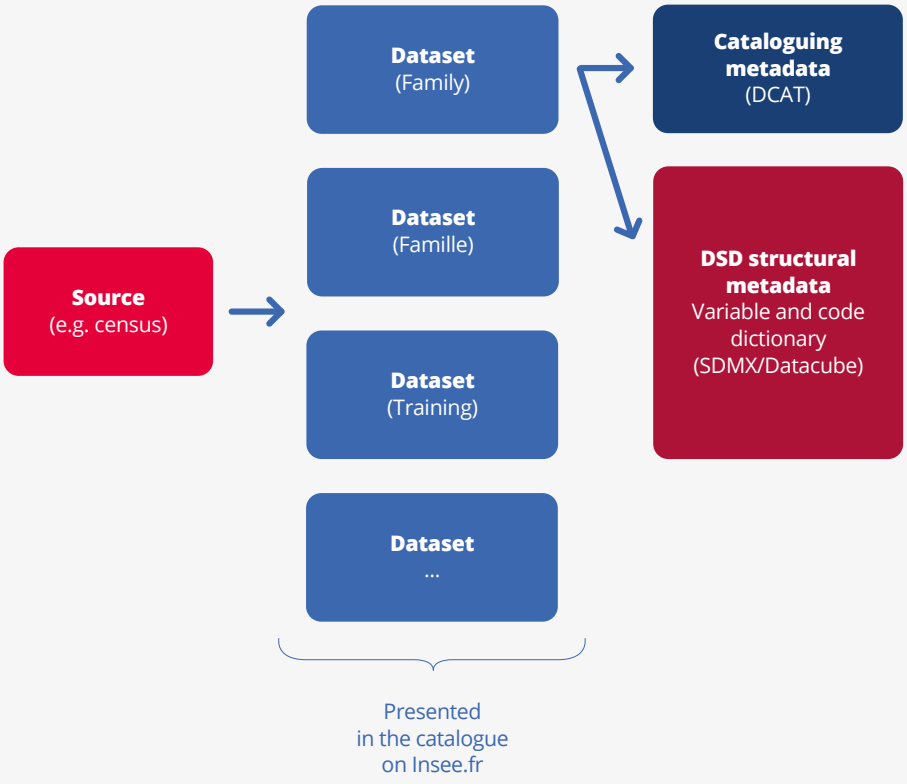


Sources: INSEE, 2020 population census.
Reading note: In Nantes in 2020, there were 732 women aged 15 to 24 who were blue-collar workers and 29 men aged 25 to 64 who were farm managers. The values are displayed for visible points where dimensions meet.

The catalogues available on official statistics websites mostly organise their datasets by theme (demography, employment, etc.) and then a more granular thematic navigation tree is created to obtain the desired dataset. There is a catalogue available on the Eurostat website[18] to consult the various European statistical datasets[19].

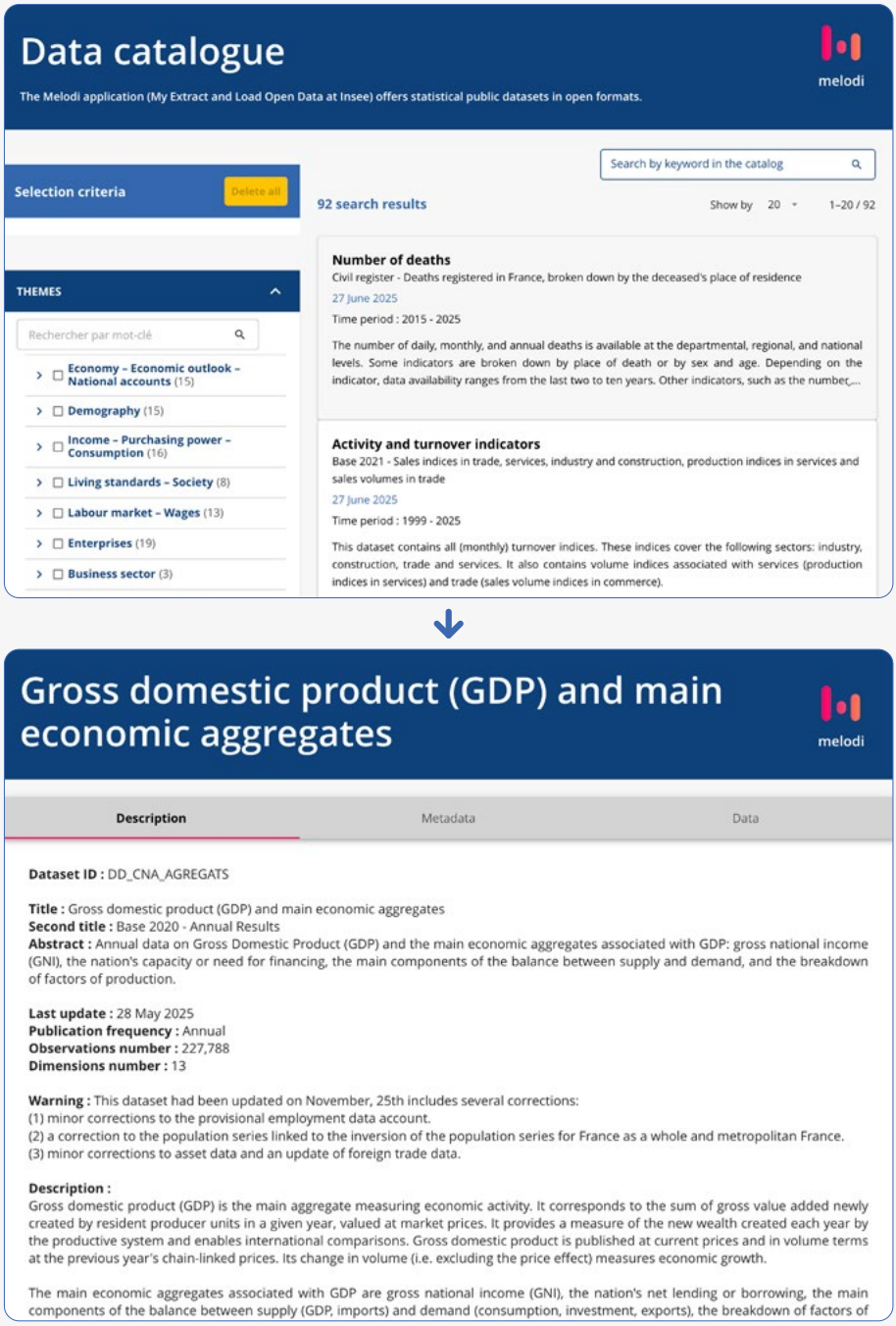
This is also the case for the German statistics institute, Destatis[20], which makes its statistical data available via its catalogue, Genesis. As often occurs, this website separates the catalogue from other statistical information (tables, publications, etc.). The datasets are accessible through the breakdowns of each theme. Selecting one allows users to view it before downloading it. The same is true for the Agreste website[21] of the statistical office of the French Ministry of Agriculture, which provides access to interactive tables in a thematic navigation tree in the "*Chiffres et analyses*" (Figures and Analyses) section.

Figure 4 - Metadata of a Dataset



18 *https://ec.europa.eu/eurostat/databrowser/explore/all/all_themes?subtheme=demo&display=list&sort=category.*
19 Another example of access is the open source platform for French official data: *https://www.data.gouv.fr/fr/pages/thematiques-a-la-une/.* Selecting a theme provides access to documentation providing details on the subject and datasets. Access can also be provided via a restricted navigation tree of more precise themes.
20 *https://www.destatis.de/EN/Home/_node.html.*
21 *https://agreste.agriculture.gouv.fr/agreste-web/disaron/!searchurl/4b54e171-2bf3-4c8b-93b9-06e41472066c!cda8b080-3e9e-4368-b41d-7a29c1da0be6/search/.*

Figure 5 - Catalogue of INSEE's Datasets
(*https://catalogue-donnees.insee.fr/en/catalogue/recherche*)



Reading note: The options on the left allow users to filter according to different criteria. The datasets are then displayed on the right. When selecting the required dataset, its description is provided. Datasets can be downloaded in CSV format and sometimes as XLSX files.

INSEE has such a catalogue[22] that allows more direct access to data via all facets of research (***figure 5***). Designed as part of a project to modernise dissemination, the purpose of this catalogue is to integrate all INSEE statistical data, while also meeting the two accessibility and clarity criteria of the European Statistics Code of Practice.

▶ Navigating through the Cubes to Analyse the Data

Once users have selected a dataset from the catalogue, they may find it useful to browse the dataset dynamically and create their own table extractions. Many NSIs offer this flexible method of browsing. On the website of the Italian statistics institute, ISTAT[23], the user selects a dataset, views it directly and accesses the documentation for each variable and value by clicking on the multiple information points. The user then personalises the dataset using a selection of variables and/or their values. Conversely, the website of the New Zealand statistics institute[24] offers immediate construction of its table before previewing and exporting it. The browsers of these two sites are particularly comprehensive in terms of the selection of data characteristics (units, presence or absence of rows or columns without values for the options selected, etc.) and offer several formats for exporting the selected data, which can include information on the data (provisional, revised, etc.).

In France, the Agreste website of the Ministry of Agriculture presents its cubes online and allows, for example, consultation of the cubes from the agricultural census on farms[25]. Likewise, the browser attached to the data catalogue on insee.fr is similar and allows users to extract part of a dataset by selecting the relevant options of the different axes. For example, an analyst working for the mayor's office studying local housing will be able to filter population census data for their municipality and the neighbouring municipalities.

These services allow different methods of browsing multidimensional cubes which can be summarised as follows[26]:

• **slicing**: one dimension is fixed to a value (a "slice") while the other dimensions are permitted to vary. Using the example of average wages by sex, age and socio-professional category[27] (***Figure 6a***), looking specifically at data on people aged 50 to 59 years: slicing is performed according to age by fixing the AGE dimension to the "50 to 59" value. The breakdown of the wages of people aged 50 to 59 years according to their socio-economic group is thus obtained (***Figure 6b***). If the user wants to examine the wage gap between men and women, a narrower slice is created according to sex by fixing the SEX dimension to the "Woman" value in order to obtain a cube for the average wages of women aged 50 to 59 years (***Figure 6c***).

22 *https://catalogue-donnees.insee.fr/en/catalogue/recherche.*
23 *https://www.istat.it/en/news/statbase-access-to-most-frequently-requested-data/.*
24 *https://infoshare.stats.govt.nz/.*
25 For example, the cube for farms by economic size and focus: *https://agreste.agriculture.gouv.fr/agreste-web/disaron/RA2020_001/detail/.*
26 This is explained in the data architecture literature under the term OLAP, an acronym that stands for "Online Analytical Processing". It is a database technology that is optimised for queries and reports, rather than for processing transactions (Codd et al., 1993).
27 *https://www.insee.fr/fr/outil-interactif/5369554/index.html.*

• **dicing**: this time several dimensions are cross-referenced on the basis of certain values ("dice"), in order to obtain a subset of the cube's data. Using this same example, the average wage of blue-collar worker women aged 50 to 59 years is extracted.

• **Drilling up or down**: it is possible to "drill up" and "drill down" in the data. This is particularly useful when there are nested aggregation levels, especially for nomenclatures, in order to study data with varying degrees of granularity. Thus, for the population of the age bracket of 50 to 59, we drill up on this age bracket to focus on the populations of the two age brackets of 50 to 54 and 55 to 59, or even of each age from 50 to 59. As another example, we drill down according to geographical levels, ranging from municipality level to country level.

These data browsing services are for all audiences, from individuals seeking personal information to professionals processing large amounts of data. Generally, however, the people who take an interest have the profile of a statistician or economist, while professionals who use the data regularly and systematically need other, more technical resources.

▶ The Harvesting of Data by the Machines

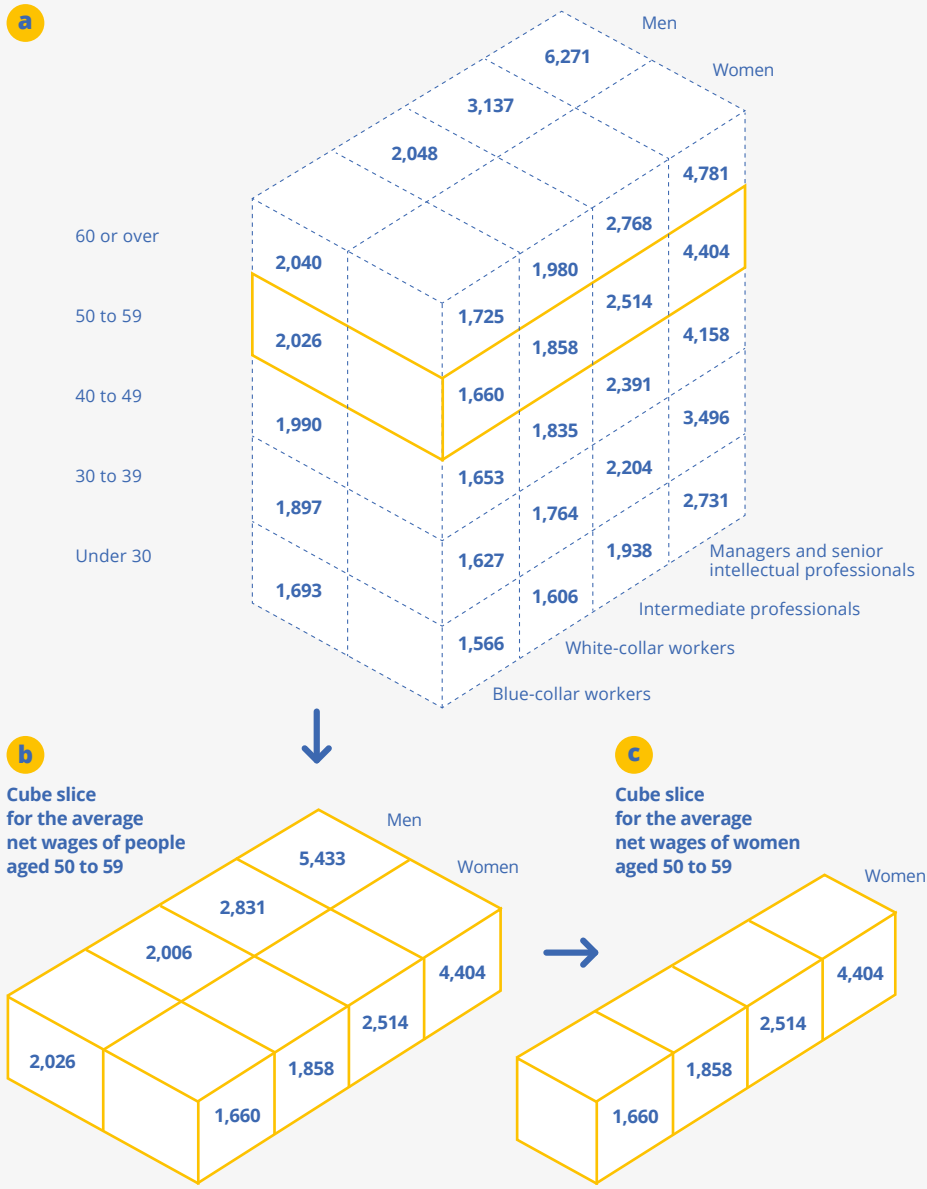
Data is increasingly being consumed on a machine-to-machine basis, with automated processing. This is the case, in particular, for companies that want to integrate INSEE data directly into their own information system. INSEE makes its data available via an API, a web service that can feed into client applications directly from its databases. The operating principle is as follows: the API client application is programmed to periodically contact the API to detect data updates and, if necessary, retrieve the latest information via a query (***box***). This form of consumption is particularly interesting because it avoids the need to manually download files from the insee.fr website while making it possible, through the configuration of the query, to retrieve only the data of interest (Jacobson et al., 2011). This is referred to as a machine-to-machine interface, because the retrieval is performed automatically by the client program, without any manual intervention.

Many organisations offer APIs (Boyd et al., 2020). For example, the *Caisse Nationale d'Assurance Vieillesse* (National Pension Fund for Employees, CNAV) offers an API[28]. From this API, it is possible to read data such as the number of pensions at 31 December broken down by gender, the total pension amount at 31 December broken down by type of entitlement, or the average monthly pension amount.

This is also the case for the OECD[29] or the Canadian statistical institute, StatCan[30]. Similarly, INSEE already offers APIs for various fields such as the *Banque de données macroéconomiques* (Macroeconomic Data Bank, BDM) for macroeconomic series or the *Diffusion de Données Locales* (Dissemination of Local Data, DDL) for local data: these will be replaced by a single API called Melodi[31] thanks to the modernisation of dissemination at INSEE.

28 *https://data.cnav.fr/api/explore/v2.1/console*. This example provides an illustration of "Swagger", with this web page presenting all possible queries and the format of the query result.
29 *https://www.oecd.org/en/data/insights/data-explainers/2024/09/api.html*.
30 *https://www.statcan.gc.ca/en/start*.
31 *Mon Espace de Livraison des données en Open Data de l'Insee* (My INSEE Space for Receiving Open Data).

▶ Figure 6 - Cube and Cube Slice for Average Net Wages in 2021 by Sex, Age and Socio-Professional Category



Sources: 2021 all employees database.
Reading note: From the cube for average net wages in 2021, a cube slice can be extracted for the average net wages of people aged 50 to 59; a cube slice can then be created for the average net wages of women aged 50 to 59.

▶ Box. How Do I Use an API?

Browsing via an API consists of using Internet addresses, also called URLs*, to query the dataset. The API sends the content directly (to the web browser page or to the client application) in a standard file format known as JSON**.

The URL structure is standardised as follows:

API name/Method/Name/Query filter.

The usual "methods" are DATA (to indicate that data is being retrieved) and STRUCTURE (to obtain the metadata details). The name is then that of the dataset (for the DATA method) or of its structural metadata (for the STRUCTURE method).

For example, the DS_TICM*** on the use of information and communication technologies by households provides the level of domestic Internet equipment ownership and the proportion of people with fixed or mobile domestic broadband. This information is sought by a company to assess the market for the production of electronic equipment.

To simplify matters, let's assume that the beginning of the URL is insee.api****. The company collects all data from the dataset in its web browser at the following URL: insee.api/DATA/DS_TICM.

The company can also extract a part of the dataset by filtering for its dimensions. If it is only looking for the rates of Internet equipment ownership among women in 2022, it will add the corresponding filter in the API query:

insee.api/DATA/DS_TICM?
MEASURE=EQUIP_INT&SEX='F'&YEAR=2022.

MEASURE is the measurement dimension fixed to the EQUIP_INT (equipment rate) code, SEX is the sex dimension fixed to the F (female) code and YEAR is the time period dimension fixed to 2022.

Finally, it should be noted that only data retrieval is possible; any and all calculations must be done at the customer's site using the data obtained by the API.

* *Uniform Resource Locator. An address that specifies the location of an Internet resource by specifying the protocol to be adopted, the machine name, the access pathway and the file name: https://www.insee.fr/en/accueil is a URL.*

** *JavaScript Object Notation (JSON) is a text data format derived from the notation of JavaScript objects.*

*** *DS for dataset and TICM for the enquête TIC ménages (Household ICT survey).*

**** *INSEE's future single API will be http://api-diffusion-catalogue-donnees-externe.insee.fr.*

> ***APIs multiplies the potential for the reuse of statistical data.***

This method for making data available is of particular interest for data dissemination, since it multiplies the potential for the reuse of statistical data. Data visualisation tools are generally based on APIs. Thus, the wage data visualisation tool on insee.fr[32] allows users to query salary data from different angles such as profession, socio-professional category or sex. When the website user selects a profession to discover the average wage, a request is submitted to INSEE's API, which finds the desired figure in INSEE's dissemination database and sends it to the tool, which displays it. The API sends the latest available value as it has direct access to the dissemination database. This tool on wages receives the value and displays it.

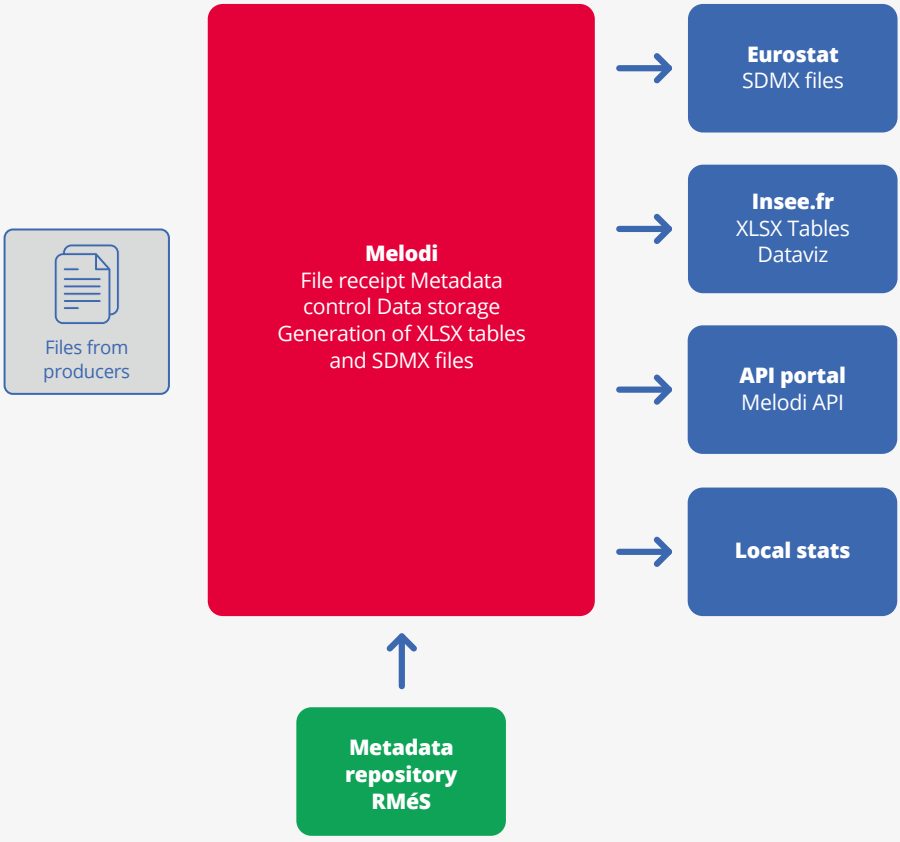
32 *https://www.insee.fr/fr/outil-interactif/5369554/index.html.*

▶ INSEE's New Dissemination Supply

To offer the services described above (catalogue, description of cubes, data browser, API, etc.), INSEE is involved in a modernisation project called Melodi[33]. This vast digital transformation project, which results in a standardisation of the data disseminated at INSEE, is based on several principles:

- the first principle is to centralise all the data to be disseminated in a single space called a statistical data warehouse[34] and to produce all the data products online (downloadable files, web tables, files sent to Eurostat, etc.). using the data in this warehouse, via shared tools for the entirety of the dissemination (***Figure 7***).

▶ Figure 7 - The New Process for Disseminating Data at INSEE



33 *Mon Espace de Livraison des données en Open Data de l'Insee* (My INSEE Space for Receiving Open Data).
34 See (Goossens, 2012) for a detailed presentation of a statistical data warehouse.

• the second principle is to describe these data on the basis of standardised metadata (SDMX/Datacube information model for structural metadata and DCAT for cataloguing metadata). In this respect, the Melodi process relies on INSEE's statistical metadata repository: RMéS. This organisation has major implications for INSEE's data production teams, who create the data to be disseminated and ensure they are delivered to the Melodi data warehouse. They must provide files that are in the expected format[35] and conform to the metadata previously described in the RMéS repository.

• a third principle, "Tell Us Once", prevents production teams from delivering the same data via different dissemination channels and greatly reduces the risk of data inconsistency.

Furthermore, the introduction of Melodi is a genuine opportunity to review the current data offering. First of all, this leads to a review of the statistical content being disseminated: to decide whether very rarely downloaded files are to be maintained and, conversely, to expand on frequently requested or new themes. Next, it is a question of redesigning the offering around the catalogue of datasets, which is a central access point, and its browser. The XLSX file offering can be reduced by refocusing it on the most requested indicators and INSEE website users looking for more specific or detailed data can be invited to consult the browser to build their own tables or download the files containing the entire dataset.

▶ What Next?

This need for the mass use of statistical data is becoming increasingly important and requires conceptual and technical developments to address it. We are thinking in particular of the Linked Open Data (LOD) technology. The principle is to structure data around metadata, which are resources that are used universally. For example, the Nouvelle-Aquitaine region would be referenced as a unique Internet "resource" and any data relating to that region would point to that resource. Unlike today, with every data producer being free to codify the region as they see fit, in the future they should refer to this universal codification. This use of universal metadata would ensure comparability between datasets.

Beyond the realm of Official Statistics, artificial intelligence (AI) is opening up the pathway to new services for querying data, to make the data even more accessible. The description of the metadata linked to the data makes them much easier for artificial intelligence algorithms to understand. This is particularly useful for chatbot/statbot tools where the user asks a question – for example, what is the latest unemployment rate? – the question is then interpreted by an AI algorithm to query the database and send the answer; the quality of the data description will then be a determining factor in the AI's ability to respond in a relevant manner.

35 Producers deliver much of their output in SAS or XLSX format in the current processes. Melodi imposes "flat" formats, such as CSV or Parquet, which are suitable for very large files.

▶ Bibliography

- BONNANS, Dominique, 2019. RMéS, INSEE's Statistical Metadata Repository. In: *Courrier des statistiques*. [online]. 27 June 2019. Insee. N° N2, pp. 46-57. [Accessed 6 February 2024]. Available at: *https://www.insee.fr/en/information/4195079?sommaire=4195125*.

- BOYD, Mark, GATTWINKEL, Dietmar, POSADA, Monica and VACCARI, Lorenzino, 2020. An Application Programming Interface (API) framework for digital government. In: *Publications Office of the European Union, Luxembourg*. ISBN: 978-92-76-18980-0. [online]. [Accessed 6 February 2024]. Available at: *https://publications.jrc.ec.europa.eu/repository/bitstream/JRC120715/final_version_pdf_version.pdf*.

- CODD, Edgar Franck, CODD, Sharon B. and SALLEY, Clynch T., 1993. Providing OLAP to User-Analysts: An IT Mandate. In: *E. F. Codd & Associates*. [online]. [Accessed 6 February 2024]. Available at: *http://www.estgv.ipv.pt/paginaspessoais/jloureiro/esi_aid2007_2008/fichas/codd.pdf*.

- DE JONGE, Edwin and TEN BOSCH, Olav, 2012. Visualising official statistics. In: *Statistics Netherlands website*. [online]. [Accessed 6 February 2024]. Available at: *https://www.researchgate.net/publication/267856653_Visualising_official_statistics*.

- DONDON, Alexis and LAMARCHE, Pierre, 2023. Quels formats pour quelles données ? In: *Courrier des statistiques*. [online]. 30 June 2023. Insee. N° N9, pp. 86-103. [Accessed 6 February 2024]. Available at: *https://www.insee.fr/fr/information/7635827?sommaire=7635842*.

- EMILSSON, Cecilia, RIVERA PÉREZ, Jacob A. and UBALDI, Barbara-Chiara, 2020. OECD Open, Useful and Re-usable data (OURdata) Index: 2019. In: *OCDE website*. [online]. [Accessed 6 February 2024]. Available at: *https://web-archive.oecd.org/2020-03-10/547558-ourdata-index-policy-paper-2020.pdf*.

- EUROPEAN COMMISSION, 2015. Creating Value through Open Data. In: *Official European data platform*. [online]. November 2015. [Accessed 6 February 2024]. Available at: *https://data.europa.eu/sites/default/files/edp_creating_value_through_open_data_0.pdf*.

- GOOSSENS, Harry, 2012. The statistical data warehouse: a central data hub, integrating new data sources and statistical output – Contributed Paper at the UNECE Conference of European Statisticians. In: *UNECE website*. [online]. 8 October 2012. [Accessed 6 February 2024]. Available at: *https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/mtg2/WP18.pdf*.

- JACOBSON, Daniel, BRAIL, Greg and WOODS, Dan, 2011. APIs: A Strategy Guide. In: O'Reilly Media, Inc. ISBN: 9781449308926.

- LAGARENNE, Christine, MINODIER, Frédéric and SAMSON, Odile, 2023. How should we present our data to communicate better? In: *Courrier des statistiques*. [online]. 11 December 2023. Insee. N° N10, pp. 7-29. [Accessed 6 February 2024]. Available at: *https://www.insee.fr/en/information/8325544?sommaire=8325635*.

- SDMX, 2012. SDMX 2.1 User Guide. In: *SDMX website*. [online]. 19 September 2012. [Accessed 6 February 2024]. Available at: *https://sdmx.org/wp-content/uploads/SDMX_2-1_User_Guide_draft_0-1.pdf*.

- UBALDI, Barbara, 2013. Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives. In: *OECD Working Papers on Public Governance*, N° 22, OECD Publishing, Paris. [online]. May 2013. [Accessed 6 February 2024]. Available at: *https://www.oecd.org/en/publications/open-government-data_5k46bj4f03s7-en.html*.