

SOCIAL SECURITY - 2023

METHODOLOGICAL NOTE

The Luxembourg Microdata Platform on Labour and Social Protection: **A service for scientific research**

Luxembourg
Microdata
Platform
on Labour
and Social
Protection



LE GOUVERNEMENT
DU GRAND-DUCHÉ DE LUXEMBOURG
Ministère de la Sécurité sociale

Inspection générale de la sécurité sociale

Table of contents

| | |
|---|-----------|
| ENSURE A HIGH LEVEL OF DATA PROTECTION..... | 7 |
| HIGHLY SECURE DATA ACCESS WHICH ENSURES DATA CONFINEMENT..... | 7 |
| ... WITHOUT HINDERING USE BY RESEARCHERS..... | 8 |
| HIGHLY SECURE DATA ACCESS WITH A SUITE OF MEASURES FOR ADDITIONAL PROTECTION . | 8 |
| Verifying the eligibility of users: show your credentials..... | 8 |
| Verifying project eligibility..... | 8 |
| Design of a data dictionary based on principles of privacy by design and privacy by default | 9 |
| Specific proportionality analysis for each request | 11 |
| Validation of treatment of the data request by a datateam | 12 |
| Contractual guarantees | 13 |
| Output checking | 13 |
| ENCOURAGE A DATA FOR RESEARCH APPROACH | 14 |
| QUALITY DATA COLLECTED AND STRUCTURED IN A DATA DICTIONARY | 14 |
| Quality data relevant to research | 14 |
| An evolving and increasingly rich dictionary | 14 |
| Data organised and supplied in the form of linked thematic registers | 15 |
| LINKING LMDP DATA WITH EXTERNAL DATA..... | 15 |
| A TEAM OF EXPERTS AVAILABLE FOR RESEARCHERS..... | 16 |
| THE LIFE CYCLE OF A DATA REQUEST INTEGRATED IN A DIGITAL SOLUTION: THE ASK4MDP APPLICATION | 18 |
| THE ASK4MDP APPLICATION IS DESIGNED TO ADAPT TO FUTURE DEVELOPMENTS IN DATA EXCHANGE | 23 |
| ADAPTABLE TO MORE REGISTERS AND MORE VARIABLES | 23 |
| ADAPTABLE TO MORE STAKEHOLDERS..... | 23 |
| ADAPTABLE TO AN EXTERNAL PSEUDONYMISATION SERVICE | 23 |
| ANNEX 1: PROJECTS SUPPORTED BY LMDP SINCE 2018..... | 24 |

Luxembourg Microdata Platform on Labour and Social Protection

The Luxembourg Microdata Platform on Labour and Social Protection (LMDP) is used to make personal data available for research in the areas of employment and social protection. This is administrative data, managed centrally by IGSS as part of its mandate, and collected for the most part by Luxembourg's social security institutions.

The constantly growing requirement for highly detailed data for research underlines the need for the technical and organisational safeguards necessary to preserve data confidentiality. This detailed personal data requires a high level of data protection in order to prevent any dissemination, which may be harmful to individuals, or any use by a non-authorised third party.

Hence the challenge for a service which provides access to this data consists, of getting the right balance between data protection, on one hand, and on the other, making it accessible to the world of research. In fact, data protection is not an end in itself, but an essential requirement to improve access to rich and relevant data and thus improve the effectiveness and efficiency of research

In order to respond to this dual need for data protection and accessibility, IGSS, in collaboration with the Ministry of Labour, Employment and the Social and Solidarity Economy, established the LMDP in 2018. Since its origin, it has developed in various directions, using the expertise gained in the projects supported by LMDP. The procedures and safeguards, which have been tested, refined and adapted to the needs of research, data protection and the requirements of this area of research, continue to grow and become more detailed. At the outset, with particular reference to social protection, LMDP restricted the data made available for research to information on cash benefits provided in the Luxembourg social security system. From now on, it is planned to integrate benefits in kind into the platform. This amplification of LMDP to include health benefits began at the start of the COVID-19 crisis, when the IGSS platform was used to provide information necessary to monitor the pandemic. The procedures and data protection measures proposed by the platform were adapted to ensure compliance with the provisions and requirements of the General Data Protection Regulation (GDPR).

Within the remit of LMDP, a group of measures guarantee data protection for all the elements and at all stages of a data request cycle.

This group of measures comply with GDPR by applying the basic principles of privacy by design, privacy by default and the management of risk of re-identification and disclosure of personal data. Hence, the data protection pillars formulated by LMDP are as follows:

1. Secure remote access which guarantees data confinement and traceability;
2. Eligibility criteria applied to users;
3. Eligibility criteria applied to data requests;
4. A data dictionary for protection by default of data made available for research;
5. Proportionality analysis specific to each data request, with rigorous application of the need-to-know principle and ad hoc protection measures;
6. Peer review of risks to privacy, related to data requests and protection measures;
7. Signature of a confidentiality agreement, which describes the user's obligations;
8. A systematic output checking procedure.

All these protection measures, which are described below, are applied at different points during the life cycle of a data request, which consists of several stages, from the initial request by a researcher, to the delivery of data. An application called ask4dmp has been created, as part of a process of digitalisation of procedures. This ensures the implementation of these procedures, and the smooth organisation and coordination of different stages in the life cycle of a data request, and centralises all the relevant information.

The ask4mdp application, whose main features are described in this document, has been designed with flexibility in mind, and to be able to adapt to future developments in data for research.

Finally, the principles and personal data protection measures embedded in the philosophy of the Luxembourg Microdata Platform, reflect the objective of the data Governance Act to provide a framework which reinforces confidence in data sharing.

ENSURE A HIGH LEVEL OF DATA PROTECTION

HIGHLY SECURE DATA ACCESS WHICH ENSURES DATA CONFINEMENT...

LMDP data are made available to researchers in a virtual desktop which they access remotely. This solution was implemented with the help of the CTIE, the State Centre for Information Technology, which means that LMDP can take advantage of the security measures developed by and for the Luxembourg State.

The technology used ensures strong authentication of users with a LUXTRUST certificate. This ensures verification of the researcher's identity¹.

The virtual offices are completely isolated from the outside world. It is not possible to extract or copy data from the LMDP or to import files, which are external to the platform. Only administrators of the virtual offices are authorised to upload external files, which the researcher has justified as necessary, and whose content has been verified (input checking).

Each research project is associated with its own virtual office, which is only accessible by researchers involved in the project. Hence, a researcher working on two projects must connect alternately to two different virtual offices, with different connection parameters.

Each research project has unique pseudonymised identifiers, which guarantee the interconnectivity of the registers made available to the researcher for their project.

Hence, to prevent any linkage of the data from different projects, the pseudonyms are different for each virtual office associated with each research project.

As soon as the project ends, access to the virtual desktop is removed.

Access to data in a virtual desktop ensures that data is confined in a secure space, which is an essential prerequisite for data traceability. Indeed, data provided outside a secure space, could be copied without limit at almost no cost, and disseminated to third parties. It would then be impossible to trace them.

Thus, securing access to data depends on the following principle:

one project ⇒ one virtual desktop ⇒ one pseudonymisation key ⇒ one login and password per researcher

IGSS decided to secure data by limiting it to remote access because it was deemed almost impossible to anonymise this data, although this is the only possible alternative for making it available in open source. This presumption is based on the results of preliminary analysis by IGSS, when LMDP was conceived. These results suggested that the small size of Luxembourg, together with the need to provide variables in varying quantity and level of detail to respond to research needs, meant that in some cases there might be a small risk of reidentification or disclosure of information, given that some combinations of variables could result in very small numbers of cases. Hence, for a genuine anonymization of IGSS data, it would be necessary to reduce the detail in the data to the point where it is no longer useful for research.

¹ More stringent authentication systems exist, particularly in France. Data access requires fingerprint recognition.

... WITHOUT HINDERING USE BY RESEARCHERS

Remote access should not create conditions for users which are so restrictive that they complicate or even prevent certain types of work².

Researchers must have at their disposal all the tools they need, as well as adequate computing power. LMPD was keen to ensure this, to make up for the fact that researchers cannot install software themselves.

Indeed, the partitioning of virtual desktops prevents this, which is very restrictive for researchers. To compensate for this, a range of software is made available, with the possibility of providing more at short notice if necessary³. This is also true for computing power, which may be adjusted according to analysis requirements and the volume of data.

HIGHLY SECURE DATA ACCESS WITH A SUITE OF MEASURES FOR ADDITIONAL PROTECTION

Data protection cannot be guaranteed only by securing access to data by means of remote access. While containment and ensuring the traceability of data greatly minimise the risk of unauthorised use by third parties, malicious or inexpert misuse of the data by persons authorised to access an LMDP virtual office is still possible. Hence virtual desktops constitute only one element of a comprehensive suite of complementary measures.

Verifying the eligibility of users: show your credentials

One of the best ways to protect data is to restrict access to qualified users, with experience of the statistical analysis of microdata. High quality statistical analysis will always result in the analysis of statistically robust groups, which are by definition large enough to greatly reduce the risk of re-identification present in statistics on smaller groups.

In addition, each user must be attached to a Luxembourg-based organisation, such that their research project is validated and supported by the management of that organisation.

Verifying project eligibility

Only projects of a statistical or scientific purpose are supported by LMDP, because these are included in those authorised for secondary data use by the GDPR⁴. Statistical purposes are well-suited to data pseudonymisation, this being a prerequisite for making microdata available under the GDPR.

² When LMDP was set up, access to data by “job submission” was investigated. This involves running programmes written by the researcher without access to the data on screen. This does little to reduce the risk of a malicious individual writing a programme to identify specific individuals, but greatly complicates the task of researchers. In particular, this method prevents detection of some types of error because they do not appear on the screen.

³ The software made available free of charge is : R, stata, Microsoft office,, latex, stattransfer.

⁴ Article 89


Design of a data dictionary based on principles of privacy by design and privacy by default

The cornerstone of the LMDP is its data dictionary (DD). This identifies the data available for research, ensures its visibility and enables researchers to narrow down the field of possibilities⁵.


Even if none of the variables in the DD may be used directly for identification (it does not contain names, addresses or social security numbers), some of variables may enable this indirectly, either through their specificity, or in combination with other variables in the database. Other variables, by virtue of their sensitivity, risk divulging personal information, in cases where individuals are re-identified.

The principles of privacy by design and privacy by default have taken precedence in designing the DD.

Two levels of protection by default...

Two levels of protection have been defined. The first is represented by the symbol  : variables preceded by this symbol are provided to the researcher at an aggregate level by default, where there would be a risk of re-identification at a greater level of granularity. The proposed level of granularity is always based on tests and preliminary reflections, with the aim of finding the best compromise between data protection and their value for research. In other words, the default level of granularity should correspond to a sufficiently large sample to avoid the risk of re-identification, while at the same time providing the modalities, which structure Luxembourg's society and are necessary to analyse how it functions.

For example, data on nationality is available at the most detailed level (all nationalities for all countries are available). However, providing this information to researchers at this level of detail carries a high risk of re-identification. For some rare nationalities, there may be very few, sometimes only one, individual resident in Luxembourg. To decide the default level of aggregation provided by the DD, IGSS defined which nationalities, or groups of nationalities, are necessary in order to understand the specific characteristics of countries. Hence, 7 modalities which consider the role of cross-border workers and of the Portuguese community in Luxembourg, were selected. The number of cases in each of these 7 modalities, have been tested, and are sufficiently large to ensure that no group with a small number of cases can be identified.

| | |
|----------------------|--|
| Name of the variable | I_nationality |
| Description | |
| Format | Character |
| Values | 0 Luxembourg |
| | 1 Germany |
| | 2 Belgium |
| | 3 France |
| |  4 Portugal |
| | 5 Other EU-28 |
| | 6 Other |
| Comments | This variable refers to the main nationality during the reference period. It can change from month to month for people who acquire another nationality. In cases of dual nationality, the provided nationality is the one considered as the first by the administration. |
| Source(s) | CCSS |



This work to define the level of aggregation focussed on those variables considered to be highly personally identifiable, and with a potentially high number of modalities: nationality, age (presented by default in 5-year age groups) and place of residence (presented by default at the level of cantons for Luxembourg). Although gender constitutes re-identifying information, this is not included because it has only two modalities.

Hence, if the researcher specifies a requirement for these four variables at the default level of granularity (and IGSS considers this to be proportionate) in their data request, the data will be supplied, because they carry a minimal risk of re-identification and disclosure.



⁵ The data dictionary may be consulted at: <https://igss.gouvernement.lu/dam-assets/microdata-platform/Data-dictionary-002-.pdf>



Nevertheless, making variables available at an adequate level of aggregation does not rule out any risk of re-identification. The accumulation of information, even in aggregate form, can result in the isolation and re-identification of an individual. Complementary tests were carried out in order to measure this risk. These are done simultaneously for age, nationality, place of residence and gender, and ensure that the combinations of the different modalities of these variables do not enable the re-identification of a very small group or a single individual. These tests highlighted the existence of some rare combinations producing very small groups of individuals, notwithstanding the high default level of aggregation. This residual risk, related to the small size of Luxembourg, could only be eliminated by greatly reducing the content of the data. It would be necessary to provide age, nationality, place of residence and gender, at even greater levels of aggregation, or not supply these variables at all, resulting in data unsuitable for most research projects, given their importance in explaining many socio-economic phenomena. Furthermore, by extending the simultaneous tests of records beyond the 4 re-identifying variables, other rare combinations may emerge, depending on the variables made available by LM DP.

Hence, there is a clear conclusion from the preliminary tests described above, and done before LM DP was established: the small size of the country means that LM DP must assume a small risk of re-identification in order to provide researchers with sufficiently rich data for quality research projects⁶. As a direct consequence, the alternative of anonymisation, which implies the elimination of any risk of re-identification, is not an option. Otherwise, it would have been possible to make a database available by open source, because in theory any risk of re-identification would have been excluded. This is the central element that explains the IGSS approach to privacy defined by a remote access to data to ensure data containment, and a set of complementary procedures to ensure data protection at different levels.

The second level of data protection is represented by the symbol   : variables preceded by this symbol are not, by default, supplied to the researcher. Given its sensitivity, this information is protected such that it does not enable disclosure of sensitive information in any case of re-identification. This protection applies to, for example, receipt of minimum income, or absences from work.

... which may be removed for research purposes but replaced by other protection measures

The two protection measures  and  are applied by default but may be removed if the researcher makes a case for this.

If the research project requires a more detailed level of granularity for variables protected by one padlock (), the researcher must show that the need-to-know principle will be respected in specifying their data requirements. If the project needs to access variables protected by two padlocks (), the researcher must show explicitly how they will respect the need-to-know principle.

As far as possible, removing a default protection measure is countered by another limitation to another aspect of the request, or an ad hoc protection which minimises the risk of supplementary re-identification generated by, for example, a more detailed level of granularity. This search for the best compromise between data protection and relevance is an integral part of request processing, in which the main challenge is to analyse the proportionality of the group of variables requested.

⁶ Algorithms exist for the identification of rare combinations, for specific groups of variables, which also propose solutions to protect individuals implicated by these combinations. This approach has not been used in Luxembourg, because it is inappropriate for a country of this size.

Specific proportionality analysis for each request

This stage is fundamental in data protection, and has two objectives: respecting the need-to-know principle, and minimising the risk of re-identification of individuals and disclosure of personal information. Proportionality analysis is specific to each demand. An ad hoc analysis is necessary for each request, resulting in specific measures adapted to the research project, which achieve the best compromise between protection and data relevance. In addition, each request is assigned to a single expert in IGSS, who is responsible for the entire request, and specifically for the proportionality analysis. Experts self-assign responsibility according to their field of expertise.

Application of the need- to-know principle addresses the following points:

- The relevance of using microdata: In their request, the researcher should describe the project's objective. In most cases, the planned analyses do need to be carried out at the individual level, but sometimes the request is reclassified as one for aggregated statistics because the platform experts do not consider it proportional to grant access to personal data. In these cases, the IGSS experts produce the statistics and provide them to the researcher ⁷.
- The field of study: In their request, the researcher indicates for example whether the research applies to the totality of Luxembourg's residents or the active population. The proportionality analysis here consists of adjusting the field of data to be provided in line with the research question. For example, the researcher states that they wish to work on the whole active population, whereas the description of the research project indicates clearly that the study is limited to private sector employees. In such a situation, LMDP will only provide information on private sector employees.
- The period covered by the data: in their request the researcher states the period to be covered by the data. For projects with a longitudinal dimension, it is common for a project to require data covering the whole period available in the LMDP (since January 2002). In some cases, IGSS experts consider the period requested to be longer than necessary, and therefore decide to reduce it.
- The different variables requested and their level of granularity: based on the DD, the researcher selects the variables necessary for their research and provides the justification. If the researcher is satisfied with the default level of granularity (which is strongly encouraged), no further justification is required. However, if the researcher wants a finer level of granularity, they must justify the importance of the variable, as well as the finer level of granularity. This proportionality analysis is carried out for each variable, independently of the others.

Minimising the risk of identification and disclosure is carried out in a second stage, using a global approach to the request that takes all the variables into account simultaneously. If the researcher requests certain variables at a higher level of granularity than the default, the experts dealing with the request evaluate the additional risks posed by this, and try to find measures to compensate this. These measures may be specific to the request, and endeavour to find the best compromise between data protection, and making data available for research. Box 1 gives several examples of compensation measures for minimising risks of re-identification and disclosure.

The IGSS approach to data protection, is adapted to all data, including health data, which is much more sensitive than, for example, employment data.

⁷ For example, one project consisted of generating structural indicators on the composition of private companies based in Luxembourg. The request for microdata was rejected, and changed into a request for aggregate statistics, for which IGSS itself produced a small number of clearly defined indicators.

Box 1 / Examples of ad hoc measures for minimising re-identification and disclosure risks**Example 1: Pseudonymisation of highly disaggregated data (zipcodes)**

Research question: impact of reform of parental leave on parental behaviour, including impact on take up of parental leave.

Researcher requirement: to measure parents' eligibility for parental leave. For this one needs to know whether the parent lives with the child, which determines eligibility for parental leave. However, LMDP does not contain a variable providing this information. Hence the researchers decide to use the zipcode (which is available but not proposed by default in the DD). It is hypothesised that a parent and child with the same zipcode lives in the same house, because zipcodes refer to streets in Luxembourg. The requirement was proportional, and use of zipcodes was the only possible proxy for households sharing the same house.

Ad hoc protection measure: a priori zipcodes contain information, which is highly re-identifying, because they relate to a very specific geographic location. The IGSS experts concluded that real zipcodes were not necessary to meet the researchers' needs, and that a pseudonymised zipcode was sufficient. Hence, with a pseudonymised zipcode, the researcher could associate parents with their child, and IGSS could guarantee a high level of data protection.

Example 2: Double pseudonymisation of sensitive data

Research question: evaluation of the effectiveness of job seeking measures

Researcher requirement: consider the possible existence of health problems to estimate the effect of an activity measure on the chances of finding a job. The researchers' idea was to use work absences to construct a composite health indicator. The need was proportional, and absences were the only possible proxy to determine the health status of individuals.

Ad hoc protection measure: work absences constitute very sensitive information, which is therefore made available to researchers very sparingly. Concerning the researchers' requirement, IGSS experts established that it was not necessary to connect the absence registers (which contain one row per absence) with other registers. It therefore proceeded in two stages: one file containing work absences and their characteristics was made available to the researcher with a first pseudonymisation key for the personal identifier. Hence the researcher could construct a composite indicator on individuals' health status. In the second stage, IGSS removed all the base variables from the file, retaining only the composite indicator, and pseudonymised the register with a second key in order to ensure the interconnectability of the registers. Hence, with a double pseudonymisation of the absence register, and by removing the base variables after calculating the composite indicator, IGSS reduced the risk of information disclosure.

Validation of treatment of the data request by a datateam

As shown by the examples described in Box 1, processing the data request requires both a good diagnosis of the risks to privacy associated with the request, and suggested measures to minimise the risks of re-identification and disclosure. All the points addressed in processing the request, and all the protection measures implemented, are recorded in a document.

When complete, the document is reviewed by an internal group of experts at IGSS who challenge the expert in charge of processing the data request, on the measures taken, or those, which should have been taken. This discussion with subject experts trained in data protection, ensures the quality and robustness of the proportionality analyses.

The validated document is archived for reference, in case of an audit by the National Commission for Data Protection (CNPDP).

Contractual guarantees

Making personal data available in the framework of the LMDP begins by signing a confidentiality agreement in which the researcher's obligations are specified. The agreement states that failure to respect these will result in the immediate closure of the virtual desktop even if the project involves several researchers. Although access to the virtual desktop is for a named individual, the contract is with the Luxembourgish organisation to which the researcher is attached, so that this organisation is aware of any possible risks and makes its researchers aware of the need for data privacy. It also ensures that Luxembourg's regulations are applied in case of litigation.

Output checking

As with any remote access system, LMDP enables researchers to retrieve project outputs at the end of the work. Unlike some other systems, this is the case throughout the project to ensure a degree of flexibility in working on the LMDP, and so that researchers can produce outputs at intermediate stages of the project. All results requested by the researcher are subject to output checking, to ensure that the results produced do not compromise the privacy of individuals⁸.

Diagram 1 / Summary of LMDP data protection measures

| Secure remote access | Eligibility of researchers and projects | Data dictionary combining protection by design and by default | Proportionality analysis | Contractual measures | Output checking |
|-----------------------------------|---|--|---|--|--|
| Strong luxtrust authentication | Researcher expertise in analysis of personal data | Two levels of default protection | Application of the need-to-know principle: ♦ to the field of study | Signed confidentiality agreement | Output checking (can include publications) |
| Data confinement and traceability | Affiliation to an institution based in Luxembourg Projects with a purely statistical purpose | ...which may be removed if necessary with compensation measures Tests on records based on combinations of re-identifying variables (age, gender, nationality, place of residence) | • to the period requested • to each variable (content and level of granularity) ⇒ ad hoc protection measures Peer validation of the analysis | Researchers made aware of data protection Access removed if protection rules not observed | |

⁸ No individual data can be removed from the virtual office.

ENCOURAGE A DATA FOR RESEARCH APPROACH

As mentioned in the introduction, data protection is not an end in itself. The LMDP philosophy considers that well-protected data is the strongest guarantee of openness to research in the best conditions possible. This is why it was designed to facilitate access to data.

QUALITY DATA COLLECTED AND STRUCTURED IN A DATA DICTIONARY

The data dictionary made available to researchers ensures that LMDP is transparent and attractive. Researchers can use it to validate the feasibility of their project, or to find new themes to explore.

Quality data relevant to research

The variables in the dictionary meet two criteria:

- they are relevant to research; some variables are not relevant because they are only useful for the administrative management of the social security institutions;
- they are of sufficiently robust quality to serve as a basis for research. Administrative data are compared with external sources to test their quality⁹.

In almost all cases, the variables obtained by IGSS from the social security institutions and retained for the DD must be transformed to make them suitable for simple and efficient statistical use. Indeed, they were collected for administrative purposes, the needs of which are not always compatible with statistical analysis.

Data transformation may be limited to renaming variables and their modalities to make them more intuitive and to meet international coding standards. In other cases, further transformation is required, such as constructing several variables to break down a range of information, which may be contained in one administrative variable.

An evolving and increasingly rich dictionary

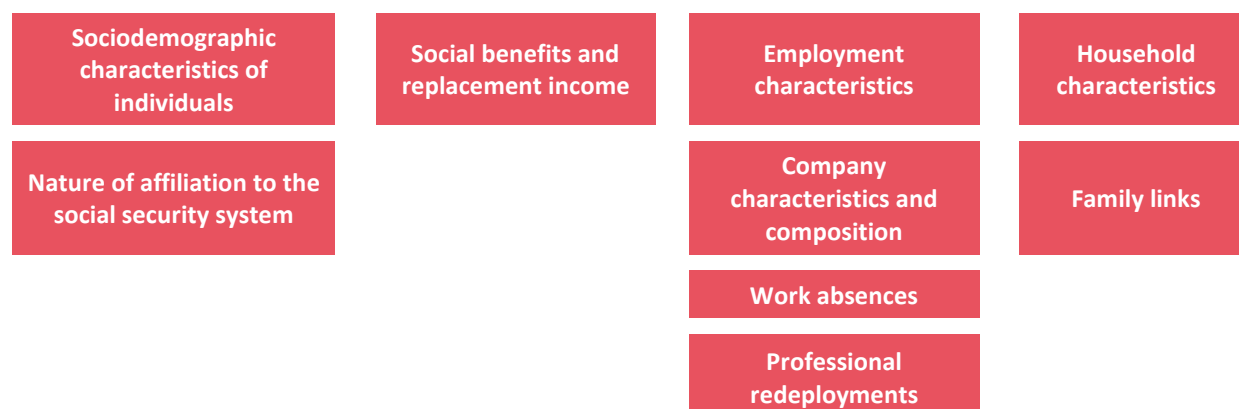
Since its first version, the DD has evolved by adding new variables. This has been driven by:

- the willingness of IGSS to respond to research developments and emerging concerns; hence IGSS foresees the inclusion of health data in the LMDP,
- a willingness to respond to a specific demand from a researcher resulting in the development of an ad hoc variable, which is then included in the dictionary,
- a willingness to make greater use of administrative data, by exploring databases not used so far.

⁹ Other means may be used to test the quality of information, where no external data is available: foreign referents, advice from database managers in the social security institutions.

Data organised and supplied in the form of linked thematic registers

To make the database more readable, data are organised in thematic registers. Currently 9 registers are proposed, covering the following themes:



All registers may be linked by different pseudonymised identifiers: a personal identifier, a company identifier and a job identifier.

Supplying files by thematic register enables researchers to link registers as needed and according to their chosen methodology.

Data are available by month from 2002 onwards, enabling analysis of long and detailed trajectories, as required by some research topics.

LINKING LMDP DATA WITH EXTERNAL DATA

External databases may be imported to the virtual desktop under LMDP procedures. These external databases may contain data which can be linked with LMDP data. For example, data from the employment development agency (ADEM) are regularly uploaded to the virtual desktop for projects analysing the career pathways of job seekers.

All administrative data containing the individual's personal matriculation number can be linked in this way, opening a wide and interesting range of research perspectives.

Clearly these linkages necessitate several protection rules, particularly relating to rigorous pseudonymisation procedure for the protection of data and individuals (see Box 2).

Data linking is often used to enhance surveys whose sampling plan was based on LMDP data. If this is the case, a survey may be linked to complementary data from the LMDP in a second stage. For example, a survey carried out in Luxembourg about child wellbeing was expanded by including LMDP data on parental sociodemographic and professional characteristics for the children involved in the survey.

This possibility of a posteriori linking made it possible to use a shorter questionnaire, thereby reducing the response time, and improving information quality because administrative data are more precise.

Extending a study in this way must be planned from the outset of the project. It is crucial to note that linking is only possible where explicit consent has been given for this. This consent is requested during the study. The organisation carrying out the survey must supply any information needed by the study subject, to make an informed decision about whether they consent to linking information they have just provided, with administrative data.

Box 2 / Procedure for linking LMDP data with external personal data

1. The researcher is responsible for requesting external personal data, and makes the request to the data owner.
2. The data owner is responsible for the proportionality analysis of external data.
3. However, LMDP requires a description of the file to be imported including the name, a description and the values of the variables. Hence the person processing the request within the platform can ensure that the file contents do not compromise LMDP protection measures. For example, if LMDP has specified that the place of residence at the level of the commune should not be provided because of an excessive risk of re-identification, it is essential to be sure that this information is not included in the external file. At the point where the file is physically imported, LMDP carries out an input check to ensure that the imported file complies with the description supplied. This additional verification was initiated after LMDP experts observed an involuntary mistake in which an imported file contained a non-declared and highly re-identifying variable.
4. Data transfer to the virtual desktop is always carried out by LMDP's IT team. This is now the procedure followed by all external data providers, ensuring the containment and traceability of all the data they make available for research.
5. In the absence of a trusted third party in Luxembourg, IGSS carries out the pseudonymisation procedure.

A TEAM OF EXPERTS AVAILABLE FOR RESEARCHERS

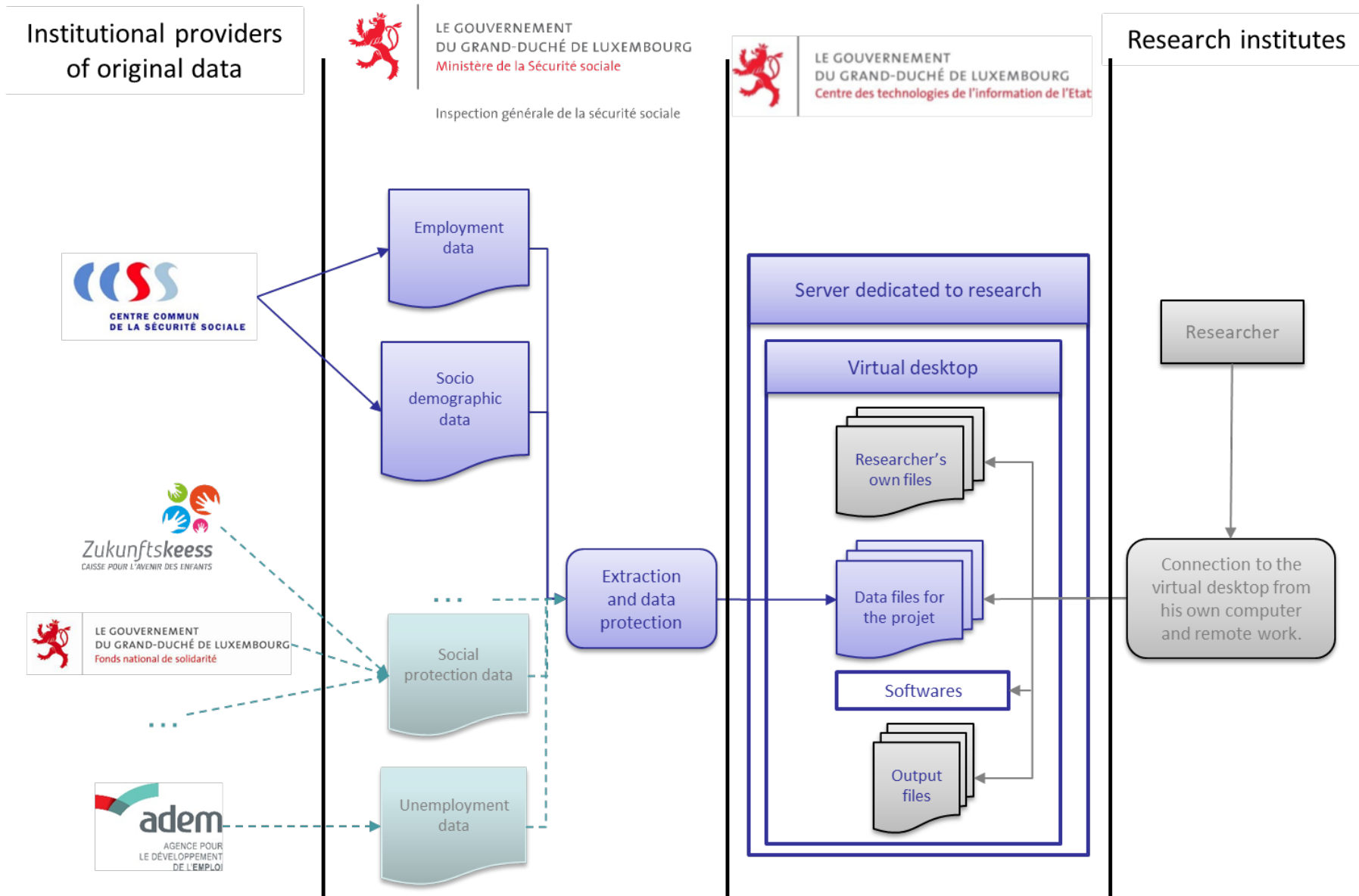
IGSS experts support researchers at each stage of their project:

- In specifying their request : given their extensive knowledge of the data available in the LMDP, IGSS experts are often able to develop the initial project of the researchers, by proposing variables with added value for their research.
- In carrying out their project : researchers sometimes need professional support to interpret specific results. The exchanges between the researcher and IGSS experts also enable the LMDP to identify specific problems. In addition, researchers frequently need technical support, which is provided by a team of responsive IT specialists.
- In exporting study results from the virtual desktop: researchers may be unaware of or lack training in privacy issues at the start of their project, and ask to export results from the platform that do not comply with the rules of statistical disclosure control¹⁰. IGSS experts make researchers aware of the problem and help them to formulate results appropriately.

IGSS willingness to provide a platform to support research means that it has assisted more around 50 projects since 2018 (see Annex 1).

¹⁰ [SDC Handbook \(cbs.nl\)](https://cbs.nl)

Diagram 2 / How the Luxembourg Microdata Platform works



THE LIFE CYCLE OF A DATA REQUEST INTEGRATED IN A DIGITAL SOLUTION: THE ASK4MDP APPLICATION

A data request involves several steps, in which several people with different roles and experience collaborate sequentially, or sometimes at the same time. As part of the Government's digitalisation initiative, an application called ask4mdp has been created. This ensures that processes are implemented, that the different steps in the life cycle of a data request are well-organised and coordinated, and that all information connected with it is centralised.

The main steps in the life cycle of a data request and their implementation in ask4mdp are presented below (Diagram 3). They include the different data protection pillars presented in the first part of this document:

1. **Creation of a user account:** a user account must be created to access ask4mdp.
2. **Preparation of the data request:** Once the account is activated, the researcher describes his request and its requirements using a request form available online. This form includes various sections which among other things, should allow IGSS experts to assess the eligibility criteria of the request and the researcher, and to get a sufficiently precise idea of the project objectives to enable the proportionality analysis and application of the need-to-know principle (Screenshot 1). If the project description is too imprecise for the LMDP experts to understand the researcher's needs, the IGSS expert responsible for the request can use the application to return the request to the researcher for further information. In another section of the form, the researcher selects the variables needed. The DD is integrated in the ask4mdp application. All available variables, their values, and any comments on them, are directly accessible in the application. It poses a series of questions for each variable selected by the researcher, who is thereby guided and obliged to justify their request: each variable must be justified in view of the research question; if a higher level of granularity than that available by default is requested, this must also be justified (screenshot 2).

Screenshot 1: Creating a data request for a research project on ask4mdp

The screenshot displays the 'Mes demandes' (My requests) section of the ask4mdp application. A sidebar on the left lists various sections: 'Membres du projet', 'Description du projet' (selected), 'Outputs prévus', 'Ressources externes validées', 'Logiciels', 'Remarques complémentaires', 'Variables nécessaires', 'Individual sociodemographic characteristics', 'Characteristics of individuals registered in the Luxembourgish social security', 'Characteristics of jobs', 'Characteristics of employers', 'Social benefits', 'Work absences', 'Child-parent relationships', 'Spouses relationships', 'Redeployments', 'Characteristics of households', 'Pièces jointes de la demande', and 'Confirmation de la demande'. The main content area is titled 'Description du projet' and contains several fields: 'Date de début du projet' (23 / 01 / 2023), 'Date de fin du projet' (31 / 12 / 2023), 'Description de la problématique de la recherche (600 mots maximum)', 'Description de la population concernée', and 'Période couverte par les données (mois, année)'. The 'Description de la problématique de la recherche' field contains text about evaluating the effectiveness of policies for employment in Luxembourg, specifically focusing on the transition from initial employment (CIE) to supported employment (CAE). The 'Description de la population concernée' field mentions the need for data on individuals who have benefited from CAE and/or CIE. The 'Période couverte par les données' field specifies the period from July 2007 to the most recent date possible. At the bottom right, there are two buttons: 'Enregistrer' (Save) and 'Enregistrer et continuer' (Save and continue).

Screenshot 2: Justifying the choice of variable and a request to change the level of granularity

Age at the end of the month

Character

CCSS

Aucun commentaire concernant cette variable

[Ne pas retenir la variable](#)

Justificatif de la variable *

Facteur explicatif de chômage de longue durée

Valeurs possibles

- 0 less than 20 years
- 1 20-24 years
- 2 25-29 years
- 3 30-34 years
- 4 35-39 years
- 5 40-44 years
- 6 45-49 years
- 7 50-54 years
- 8 55-59 years
- 9 60 years and more

[Masquer le détail](#)

☒ Demander une modification des valeurs proposées

Pour étudier la population active de 16 à 64 ans, j'aurais besoin d'une classe d'âge 60-64 ans. *

- Submission of the data request to LMDP** : after completing the request form, it is validated, and the researcher submits it to LMDP. The request may no longer be modified by the researcher. Automatic notifications notify the researcher of its submission, and the IGSS experts of its reception.
- Allocation of the request to an IGSS expert** : an auto-assignment system included in as4mdp allocates the request to an IGSS specialist in the relevant area (labour market, social protection or health). If the request covers two areas, the IGSS experts decide who will be responsible. As in any project, only one person should coordinate the request to ensure that it is followed up.
- Processing the request by an IGSS expert**: the expert assigned to the request must first determine the eligibility of the researcher and the request. As already indicated, they must answer three questions: Does the project need microdata? Is it legitimate for this researcher to use microdata? Is the project purely statistical? A positive answer to all three is necessary to continue processing the request and to initiate the proportionality analysis of the information requested. This analysis applies to the field of the data, the time period and variables requested. Based on the justifications supplied by the researcher, the application guides the IGSS expert in validating or invalidating each variable, and each request to modify the default level of granularity (screenshot 3). A messaging system is available to request further information if this is needed to make a need-to-know decision. Finally, the proportionality analysis generates the list of variables to be supplied to the researcher. This list is one of the key elements of the data request, resulting from a comparison of the needs specified by the researcher, and the need-to-know principle; as such it encapsulates all the protection measures used by IGSS to minimise risks to privacy. This list is also attached to the confidentiality agreement which is always associated with a closed list of variables. The application enables automatic generation of the list of variables.

Screenshot 3a: proportionality analysis of a variable without a request to modify the level of aggregation

La demande (v2)

Traitement

Validation DT

Confirm. du demandeur

Préparation du BV

Contrat

Finalisation

ANALYSE DE LA PROPORTIONNALITE DES VARIABLES

Contexte

Variables

Extracteurs ad hoc

Individual sociodemographic characteristics

reference_period

en attente de traitement

> Justificatif du demandeur

"Suivi des personnes dans le temps"

> Demande de modification du niveau d'agrégation

– Aucune demande de modification du niveau d'agrégation n'a été demandée –

[Poser une question au demandeur ?](#)

> Décisions IGSS

Justificatif

Niveau d'agrégation

DESCRIPTION DU PROJET

LEGITIMITE

FINALITE ET PERTINENCE DES DONNEES DISPONIBLES PAR RAPPORT AU BESOIN

ANALYSE DE LA PROPORTIONNALITE DES VARIABLES

DONNEES EXTERNES

MESURES DE PROTECTION GLOBALES

PROCEDURE DE PSEUDONYMISATION

CONSIDERATIONS RELATIVES AUX BUREAUX VIRTUELS

SOURCES POUR L'EXTRACTION DES DONNEES

Screenshot 3b: proportionality analysis of a variable with a request to modify the level of aggregation

i_age

Partiellement traitée

> Justificatif du demandeur

"Facteur explicatif de chômage de longue durée"

> Demande de modification du niveau d'agrégation

"Pour étudier la population active de 16 à 64 ans, j'aurais besoin d'une classe d'âge 60-64 ans."

[Poser une question au demandeur ?](#)

> Décisions IGSS

Justificatif

Niveau d'agrégation

Accepté Refusé

Saisissez votre commentaire...

Nouvelles modalités

Saisissez les nouvelles modalités...

ou [Annuler](#)

6. Validating the request processing: there is a double validation of the request processing.

- First by peers : the request processing is discussed by a team of IGSS experts (the datateam) to ensure that all risks to privacy have been identified and minimised.
- Secondly by the researcher : after validation by internal peers at IGSS, it is made available to the researcher via the application, who is thereby informed of the discussions held, the decisions taken and the final list of variables. Hence, they are assured of the relevance of any decisions taken by IGSS, and may flag up a decision which makes it difficult to carry out the project. Because the IGSS experts maintain close contact with the researchers, this is unlikely but cannot be ruled out entirely. Anticipating this allows for further discussion of any problematic aspects.

7. Allocating roles: the operational phase starts after validation of the processing, involving coordination of various persons with different roles. The expert assigned to the request, who has a complete overview, defines the necessary roles, and selects the colleagues to be assigned to each one (see screenshot 4: Table of roles). The notifications needed to inform and coordinate the various actors during the operational phase of the data request are sent based on this table.

Screenshot 4 : Table of roles

DEMO (en attente de traitement)

Demande suivie par **MZ Mireille Zanardelli** Échéance le **31/01/2023** Demandeur **VR Virginie Raymond (Ministère du Travail, de l'Emploi et de l'Économie sociale et solidaire)** Dates du **23/01/2023** au **31/12/2023**

La demande (v2) Traitement Validation DT Confirm. du demandeur Préparation du BV Contrat Finalisation

Intervenants de la demande

[Retourner à la fiche de la demande](#)

Experts thématiques * ☐ Sélectionner tous les experts thématiques

Extracteur principal *
Choisissez parmi ces utilisateurs l'extracteur principal pour cette demande

Extracteurs ad hoc

Pseudonymisateurs * ☐ Sélectionner tous les pseudonymisateurs

Input checkers * ☐ Sélectionner tous les input checkers

Préparateurs de BV * ☐ Sélectionner tous les préparateurs de BV

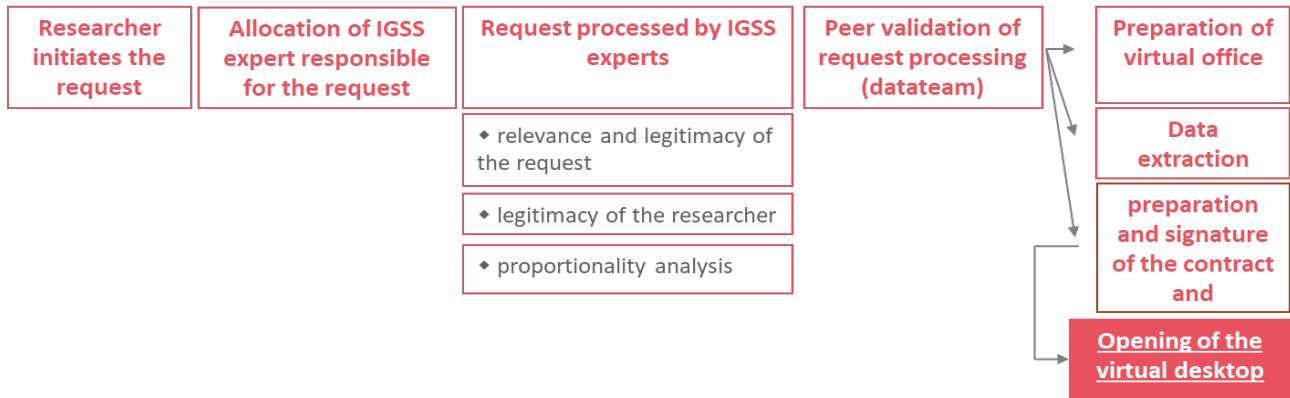
Enregistrer

8. Implementing the request: After processing and validating the request three steps are carried out simultaneously:

- Preparation of the virtual desktop, including creating the virtual desktop, the user accounts to access it, and installation of the software requested.
- Data extraction: this is carried out by the data extractors based on the list of variables. One feature of the application makes it very easy for the extractor to identify modalities which do not correspond to those proposed by default in the DD.
- Preparation of the confidentiality agreement which must be signed in order to open the virtual office. The application contains a screen which the researcher uses to provide all the information necessary for the contract. The application generates the contract automatically by publipostage (including the list of variables related to the request).

As the various tasks are completed, the application updates the status of the request, and shows this status on a dashboard. Dispatch of the contract is triggered by validation of the last task.

Diagram 3 / Steps in the workflow for an LMDP data request



THE ASK4MDP APPLICATION IS DESIGNED TO ADAPT TO FUTURE DEVELOPMENTS IN DATA EXCHANGE

The ask4mdp application is designed to adapt to future developments in data for research, thanks to its general parameters which will enable development of many dimensions of the application.

ADAPTABLE TO MORE REGISTERS AND MORE VARIABLES

Any number of registers may be added to the ask4mdp application, each containing as many variables as necessary. Hence, if other data providers wish to join the database, it will be very easy to do so by enlarging the DD.

ADAPTABLE TO MORE STAKEHOLDERS

If other experts outside IGSS wish to join the platform, it will be possible to enlarge the datateam, by adding further persons in the list of roles and involving them in request processing. Ask4mdp includes a feature to enable handover to a second expert during request processing.

ADAPTABLE TO AN EXTERNAL PSEUDONYMISATION SERVICE

Currently, there is no trusted third party in Luxembourg discharging the role of data pseudonymiser. If such a service were created at the national level, its integration in the application has already been planned, in accordance with secure and automated procedures, to ensure the efficient and fast exchanges between stakeholders inherent in a pseudonymisation procedure. The pseudonymisation service should be responsive, to avoid delaying or blocking projects.

Ask4mdp is therefore adaptable to a platform for data exchange, which goes beyond IGSS, with the aim of operating at the national level.

ANNEX 1: PROJECTS SUPPORTED BY LMDP SINCE 2018

| Year | Project acronym | Requesting entity | Subject | Sub-contractor |
|------|-----------------|------------------------|--|----------------|
| 2023 | EVAL_CAI | MFAMIGR | Analysis of the beneficiaries of a welcome and integration contract | LISER |
| 2023 | ChilDEV | LISER | Analysis of public and private investment in young childhood on wellbeing in children | |
| 2023 | FSE_REACT_EU_22 | MTEESS | Evaluation of short time employment policy implemented in 2020 | KPMG |
| 2022 | RACISM | MFAMIGR | Analysis of opinions on perception of racism in Luxembourg | LISER |
| 2022 | MET'HOOD | LISER | Impact of sociodemographic context on incidence of metabolic illnesses | |
| 2022 | OECD_LUX | Ministry of State | Evaluation of management of the COVID crisis in Luxembourg | |
| 2022 | CR_HOUSINQ | LISER | Relationship between workforce diversity and company performance | |
| 2022 | OSS | Town of Schiffflange | Socio-economic observatory in the town of Schiffflange | LISER |
| 2022 | A_HOUSE | Ministry of Housing | Sampling plan - housing survey | LISER |
| 2022 | Ind-FSE | MTEESS | Efficiency indicators of the European Social Fund (recurring project) | |
| 2022 | QoW | Chamber of Employees | Sampling plan - Quality of Work (recurring project) | INFAS |
| 2020 | TEVA | INFPC | School-Economically Active transition of young persons after technical secondary education (recurring project) | |
| 2021 | SURVEY-RACISM | MFAMIGR | Sampling plan – survey about of perception of racism | LISER |
| 2021 | EMP2021 | Ministry of Culture | Sampling plan - survey of museum practices | LISER |
| 2021 | VacciCovid | Directorate of Health | Analysis of COVID vaccine effectiveness | LIH |
| 2021 | Santé pour tous | Ministry of Health | Analysis of population health status | LISER STATEC |
| 2021 | BCL-survey | BCL | Sampling plan – income and wealth survey (recurrent project) | LISER |
| 2021 | SSMPOPCAR | Chamber of Employees | Analysis of SSM beneficiary characteristics | CSL |
| 2021 | INDEX2022 | MENJE | Analysis of socio-economic disparities in communes | LISER |
| 2021 | OVdL | City of Luxembourg | Socio-economic observatory of the city of Luxembourg (recurring project) | |
| 2021 | TRAJ_XB_FR | Region Grand Est | Career paths of French cross border workers | LISER |
| 2021 | FPE_CAE_CAI | MTEESS | Evaluation of the effectiveness of employment-promoting measures | |
| 2020 | Myenergy | Commune of Differdange | Socio-economic indicators of districts in Differdange | LISER |
| 2020 | OSE | Town of Esch/Alzette | Socio-economic observatory of the town of Esch/Alzette (recurring project) | LISER |
| 2020 | EvalFSE | MTEESS | Evaluation of measures financed by the European Social Fund | LISER |
| 2020 | COVID-TF-WP7 | LISER | Profiling of COVID cases | |
| 2020 | LST sampling | Ministry of Health | Sampling support for Large Scale Testing | |
| 2020 | MODVID-MICROSIM | LISER | Impact of covid on the financial situation of households | |
| 2020 | Commute_absent | LISER | Impact of length of journey to work on work absences | |

| Year | Project acronym | Requesting entity | Subject | Sub-contractor |
|------|-----------------|-----------------------|---|--------------------------|
| 2020 | S-Handicapes | MTEESS | Situation analysis of disabled employees on the labour market | |
| 2020 | MOBDET | LISER | Analysis of posted worker mobility | |
| 2020 | CASiNO | LISER | Impact of employer location on worker mobility | LISER |
| 2020 | COVID 19 - WP6 | MESR | Pandemic projections | University of Luxembourg |
| 2020 | ESICS | MENJE | Evaluation and improvement of the commune index | LISER |
| 2020 | StratVacc | Directorate of Health | Analysis of vaccine coverage | |
| 2019 | NETLUX | Chamber of Employees | Situation and development of the cleaning sector | LISER |
| 2019 | WISE | LISER | Determinants of exit from social minima | |
| 2019 | LuxChildWeB_2 | MENJE | Analysis of child wellbeing | LISER |
| 2020 | Workageing | LISER | Evaluation of measures to assist older workers | |
| 2019 | LuxChildWeB | MENJE | Sampling plan – survey of child wellbeing | LISER |
| 2018 | MIGAPE | LISER | Analysis of the gender pension gap | |
| 2018 | ISEC | MENJE | Socio-eco-cultural indices of pupils by commune | LISER |
| 2018 | Parent project | LISER | Analysis of joint parental leave choices | |
| 2018 | EVAL_CP | MFAMIGR | Evaluation of the effects of parental leave reform | LISER |
| 2018 | EVALAB4LUX | MTEESS | Evaluation of active employment policy | LISER |
| 2018 | FPL | LISER | Analysis of fathers' parental leave behaviour | |
| 2018 | CLD | MTEESS | Analysis of career choices of long-term unemployed | LISER |

MFAMIGR: Ministry of the Family, Integration and the Greater Region

LISER: Luxembourg Institute of Socio-Economic Research

MTEESS: Ministry of Labour, Employment and the Social and Solidarity Economy

INFPC: National Institute of Continuous Professional Development

MEN: Ministry of Education, Children and Youth

MESR: Ministry of Higher Education and Research