European Platform Undeclared Work

DATA MINING FOR MORE EFFICIENT ENFORCEMENT

A practitioner toolkit from the thematic workshop of the European Platform Undeclared Work

Frederic De Wispelaere and Jozef Pacolet
Research Institute for Work and Society – KU Leuven
Victor Rotaru
Labour Inspectorate - Romania
Steven Naylor
The Health and Safety Executive - United Kingdom
Dirk Gillis
International Research Institute on Social Fraud - UGent
Eleni Alogogianni
Hellenic Labour Inspectorate - Greece

This publication has received financial support from the European Union Programme for Employment and Social Innovation "EaSI" (2014-2020). For further information please consult: http://ec.europa.eu/social/easi

TABLE OF CONTENTS

LIST OF KEY TERMS AND ABBREVIATIONS

Compliance risk management process includes five consecutive steps. The first two steps relate to risk identification and analysis of risks. The next two relate to treatment planning (prioritisation and planning). The final step relates to evaluation.

Data collection is the process of gathering data from internal and external sources.

Data matching is the large scale comparison of records or files collected or held for different purposes, with a view to identifying matters of interest. With data matching, two or more sets of collected data are compared (comparison of records).

Data mining can be defined as a set of automated techniques used to extract buried or previously unknown pieces of information from large databases. By the use of data mining, correlations or patterns among dozens of fields in large relational databases will be identified.

Data sharing is the process of making data available to other users.

Data warehouse stores data from multiple sources that are required for analysis.

Knowledge discovery from databases is a process of data cleaning, data integration, data selection, data transformation, data matching/mining and knowledge representation.

Machine learning provides methods, techniques and tools which help to learn automatically and to make accurate predictions based on past observations.

Privacy by design is regarded as a multifaceted concept, involving various technological and organisational components, which implement privacy and data protection principles in systems and services.

INTRODUCTION

Extracting knowledge from databases by the use of data mining can support the ambitions of public administrations to tackle undeclared work by moving to a risk-based audit selection (i.e. selecting cases on the basis of the risk estimates provided). The use of data mining has become a powerful tool for risk management and risk analysis. It can support policy makers and enforcement bodies in making strategic decisions by monitoring and improving inspections. Moreover, it should result in a fairer and more selective approach and more efficient inspections. Currently, there are several initiatives within Member States on knowledge discovery from databases. Their importance has certainly grown over the last decade due to the increasing development and use of information technology and the growing availability of large databases. However, these initiatives are at different stages of development.

What is the toolkit about?

The aim of this toolkit is to enrich the mutual learning process on data sharing, data collection, data matching and data mining. The toolkit brings together findings from research and from exchanges taking place within the context of the European Platform Undeclared Work, on Member State experiences in data mining.

Data collection, data profiling, data matching or mining, and data sharing are intrinsically linked steps in the process of data exploitation, helping to improve efforts to tackle undeclared work, or even, in particular with regard to the last stage, in better predicting where undeclared work is more likely to take place. During this process concerns regarding data privacy and data quality (both as part of data governance) should be taken into account too. Data privacy needs to be built into any data gathering system from the very start. In addition, the continuous data quality assessment will assure the reliability of the system and will revise current data errors and prevent future data errors.

Different steps in the implementation process of a data analysis system should be identified as resources and ambitions may differ strongly among enforcement bodies.

Why is this important?

The importance of building an efficient data analysis system in order to predict, prevent and detect undeclared work cannot be ignored. The use of such a system, both for preventative and curative reasons, can help policy-makers and enforcement bodies to make strategic decisions, by enhancing monitoring and improving inspections. Moreover, the use of data matching or data mining might be an interesting method to estimate the size of undeclared work.

"Data mining as a tool for advice"

The use of a data analysis system makes it possible to:

- Provide inspectors with an initial and immediate level of information on enterprises, employers and employees to enable them to verify this kind of information during an inspection;
- Guide inspection activity towards those actors for whom inconsistencies are discovered through analyses that can be used to support the inspectors in their work;
- Estimate the size of undeclared work;
- Help policy makers and enforcement bodies with strategic decisions;
- Increase the perceived risk of detection.

Another important key factor for success is the data quality. Therefore, it is important that a data quality management system is developed and embedded within the data analysis system in order to assure a continuous quality assessment of data and data corrections.

How was the toolkit developed?

The toolkit is the result of the exchanges at the Thematic Review Workshop of the EU Undeclared Work Platform, held in Helsinki on 1-2 June 2017, the presentation of Member State experiences at this workshop, and the deliverables generated through these activities (i.e. Discussion Paper and Learning Resource Paper).

A Discussion Paper was written in preparation for the Thematic Review Workshop, which set the scene for discussions at the event[1]. The aim of the workshop was to provide an opportunity for participants to improve their knowledge and awareness of the importance of building efficient data mining systems in order to predict, prevent and detect undeclared work. The workshop provided an opportunity to exchange best practices, identify successful approaches to data matching/mining at national level as well as data sharing at cross-border level which have the potential to be transferred to other Member States. The workshop also explored the challenges that need to be overcome in order to develop efficient data analysis systems.

The Learning Resource Paper[2] presents a summary of the main elements of the discussion that took place during the workshop. This paper captures the key messages and information of value from the workshop and adds additional information based on other research/evidence wherever useful or necessary. It looks at the minimum requirements of an efficient data system and the challenges to overcome in order to develop a fully efficient data analysis system on the basis of a step-by-step approach.

Who is the toolkit for?

The toolkit is aimed at national enforcement bodies who already implement or wish to implement practices of data collection, data sharing, data profiling, data matching, data mining as well as data protection and data quality management.

How can the toolkit help?

Enforcement bodies that aim to implement a data analysis system have mainly three key questions to address.

The key questions are:

- What are the different steps and the minimum requirements for implementing an efficient data analysis tool?
- Which are the possible pitfalls and the challenges to overcome in order to develop a fully efficient data analysis tool?
- What about the return of investment and how can it be measured?

In order to meet these key questions the toolkit has the ambition to accomplish several objectives.

Objectives of the toolkit:

- to validate, support or boost efforts taken by enforcement bodies on data collection, data sharing, data matching and data mining;
- to assist enforcement bodies with the implementation of practices of data collection, data sharing, data matching and data mining;
- to assist enforcement bodies with the implementation of practices of reducing data errors through an effective data quality management system;
- to provide an overview of all major steps necessary for implementation;
- to provide checklists and step-by-step guides, as well as the main resources, elements of design, critical questions for implementation, and guidance on the design of monitoring and evaluation;
- to provide a practical information source on strategies and methods which furthers institutions' ability to implement these practices.

1 De Wispelaere, F. and Pacolet, J. (2017), Data Mining for More Efficient Enforcement – Discussion Paper, HIVA – KU Leuven, European Platform Undeclared Work.

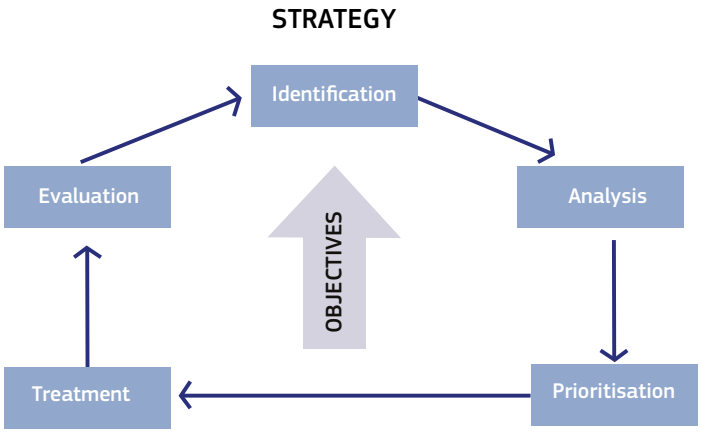
2 De Wispelaere, F. and Pacolet, J. (2017), Data Mining for More Efficient Enforcement – Learning Resource Paper, HIVA – KU Leuven, European Platform Undeclared Work.

The toolkit might be useful for those enforcement bodies who are at an early starting point. However, those who are advanced in the process can also still use the toolkit. Labour inspectorates might be using data analysis tools less frequently in their fight against undeclared work and other malpractices than tax administrations. However, there is no need to reinvent the wheel. Sharing good practices or working more closely together might be an efficient and effective way to increase knowledge.

1. FROM DATA COLLECTION TO DATA MINING: A STEP-BY-STEP APPROACH

The use of data analysis tools such as data mining is an example of how data and technology can support the compliance risk management process. By using data analysis tools, enforcement bodies could progress from data towards intelligence. They should not only be able to describe what happened or is currently happening (for instance, on the basis of the outcome of audits), but also to explain the current situation or even predict was is going to happen (for instance, on the basis of characteristics of audited companies or persons).

Figure 1. Compliance Risk Management Process



Source: EC (2010)

Data tends to be a reflection of past events, information can supply understanding of patterns and trends across data sources, and intelligence can add value by providing insight into the future (OECD, 2004). Different tools and techniques could be used by enforcement bodies depending on what level of 'knowledge' (data, information or intelligence) and 'technology' (low, middle or high) is present in the current situation or to what level they would like to evolve to. It is therefore important not to neglect the different realities that exist in enforcement bodies. Moreover, you cannot have everything and certainly not everything at the same time. One should always keep differences in resources and ambitions in mind.

The Practitioner's Toolkit does not propose a 'one-size-fits-all' approach aiming to implement a large-scale data mining project. There will be a strong focus on the different steps to be taken in the implementation process, rather than on the type of data analysis tool that should be implemented. For some enforcement bodies the collection of data that describes what happened / happens is already a major step forward. Other enforcement bodies might want to implement a data analysis tool that explains the current situation or even predicts was is going to happen. However, all of them will have to follow a rather similar implementation process described in this toolkit.

Three main consecutive steps are defined in the toolkit for the development of a data analysis system, including several intermediate steps:

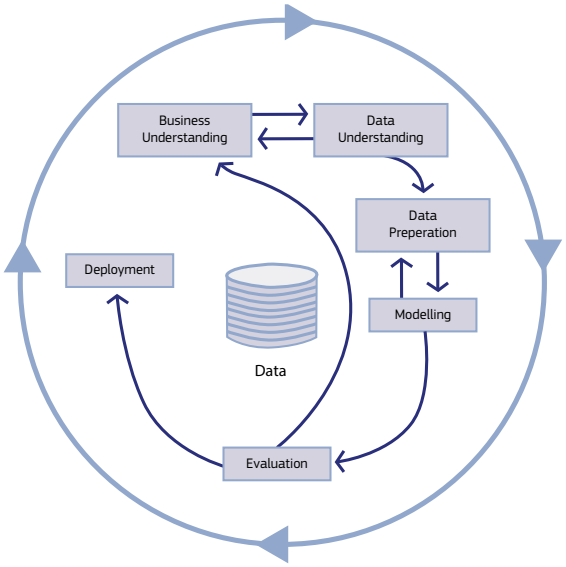
- **'Plan and design':** In this first phase, the idea for the development of a data analysis system is explored and elaborated.
- **'Implement':** After defining and planning the necessary steps, the data analysis system could be implemented.
- **'Monitor and evaluate':** The return of investment should be clearly monitored and evaluated. This could be an a priori assessment (before the implementation of the data analysis tool) as well as an a posteriori assessment (after implementation).

The steps defined in this toolkit are based, to a large extent, on the CRISP-DM model (Cross-industry Standard Process for Data Mining). This model, which was developed in 1996, organises the data mining process by six phases:

- Business Understanding;
- Data Preparation;
- Evaluation;
- Data Understanding;
- Modelling;
- Deployment

1. The first phase (business understanding) defines the business goals of the project (i.e. scope of the project: reference group, sectors of activity, type of fraud or error to be detected, etc.).
2. The second phase (data understanding) collects the data and assesses the source and quality of the data.
3. The data preparation phase covers all activities to construct the final dataset that will be used for building the model (for instance, data cleaning).
4. The modelling phase selects the data analysis tool and determines the algorithms.
5. The next phase (evaluation) focuses on the quality of the model in terms of achieving the project's business goals.
6. In the final phase (deployment), the enforcement bodies incorporate the results into the day-to-day decision-making process.

Figure 2. The Data Mining Process- CRISP DM



A summary of the content of the different steps is provided below.

Plan and design

Initiation phase

- Where are you in terms of data, technology and human/financial resources?
- What are the ambitions?
- What are the differences between the current and the desired situation? (i.e. gap analysis)
- What are the expected costs and benefits of an intervention? (i.e. cost benefit analysis)
- What are the boundaries?

Definition and planning phase

- What data will be collected/shared? (What do we have? What do we need? What are the partners involved? What are potential legal barriers?)
- What will be the 'privacy/data protection by design' strategy?
- What will be the data governance strategy?
- What will be the data quality strategy?
- What will be the data analysis tool?
- What will be the output?

Implement

- Preparation of the operational environment (IT infrastructure and operational procedures)
- Preparation of data for analysis
- Integrate Data Quality Rules into Data Integration Processes
- Train, test and implement the data analysis tool

Monitor and evaluate

- Feedback from front office to back office
- Feedback to the source of data
- Assess model outcome
- Assess compliance and risks

1.1. Plan and design

Three consecutive intermediate steps are defined: the initiation phase, the definition phase and finally the planning phase.

1.1.1. Initiation phase

The main goal of this first phase is to examine the feasibility of the project. The current situation, future ambitions, gaps between both as well as the boundaries, should be defined in this phase. In addition, by a cost benefit analysis, both potential costs and benefits will be identified and compared to each other.

The cost benefit analysis may differ depending on the ambitions one has (i.e. 'scenarios': only collecting data versus implementing a data mining tool). On the basis of the initiation phase the type of project will be defined: a pilot project or a large-scale data mining project or something in between.

Questions to be answered and tasks to be carried out in the initiation phase include the following:

Where are you in terms of data, technology and human/financial resources?

- Inventory of existing level of data collection, data sharing, data protection, data quality management, data matching and mining
- Assessment of your organisation's readiness to implement the project
  - Available hardware and software
  - Access to/availability of required data sources
  - Technical skills to carry on the implementation
  - Available financial resources to cover project costs
- Assessment of compliance with respect to General Data Protection Regulation

What are the ambitions?

- What are the most frequent fraud cases in your country you have to deal with? What forms of fraud should be addressed first to better support anti-fraud effectiveness in your country? (evasion,

avoidance, contribution fraud, social benefit fraud, bogus self-employment, under-declared work, errors, etc.)
- Where do you want to go in terms of data collection, data sharing, data protection, data quality management, data matching and mining?
- What are realistic short and long-term objectives and what should the result be?

Gap analysis
- What are realistic short and long-term objectives?
- What are realistic short and mid-term risks?

Cost benefit analysis
- What are the tangible short and mid-term benefits to implement the project?
- What are the total project costs (implementation and yearly maintenance costs)?
- What is the estimated timeframe for the project implementation?
- What are the expected costs versus the expected benefits (ROI)?

What are the boundaries?
- Should or could you have this ambition?
- Technical feasibility?
- Financial feasibility?
- Availability of human resources?
- What is the legislative framework on the protection of personal data?
- Any other constraints?

?

A) Gap analysis: current context versus ambitions

Data, technology and human/financial resources are the basis for the implementation process. Before starting the process it is important to determine the current and the desired situation in terms of data, technology and human/financial resources. The gap between the current and the desired situation is the starting point for determining the different steps to be taken. A marked difference across different countries and enforcement bodies in resources available may exist. The development of a data analysis system by enforcement bodies should therefore be defined in a very realistic way. However, circumstances might change ambitions and resources.

B) Cost benefit analysis

An intervention strategy entails indisputably high costs, both in terms of an increased numbers of persons engaged in monitoring and inspections and the high investment in technology. Before the development of a data analysis system, a cost benefit analysis will be required to assess the return on investment.

The return on investment should be more clearly estimated/measured in order to help enforcement bodies understand the impact of more advanced data usage on the outcome of inspections. Such evaluations also increase internal and external awareness of the capacities of public administrations and the potential value of investments in data collection/matching/mining in order to prevent and deter non-compliance with labour and tax rules.

It is important to define useful indicators to estimate/measure the performance and effectiveness of the data analysis tools and to ascertain the return on investment. These will be used to evaluate the total expected cost (i.e. in terms of additional human or financial resources) compared to the total expected benefits (direct benefits - results selected by data analysis tool – and indirect benefits - harmonization of the data, administrative gains, behavioral change) in order to determine whether the proposed implementation is worthwhile. The cost of persons engaged in monitoring and inspections (i.e. the resources of treatment) and the cost of the investment in the data analysis tool will be compared with the outcome of the audits selected by the analysis tool. The outcome could be expressed in terms of the amount recovered, in terms of efficiency but also in terms of the deterrent effect of it. Moreover, it could be compared with the return on investment for audits undertaken without the use of the data analysis tool (i.e. random selection). In the initiation stage, an a priori assessment of the costs and benefits will be required. However, also an a posteriori assessment will be useful. In this respect, some indicators like investigation time invested, investigation costs

and investigation quality can be used in an a priori assessment and in an after implementation assessment to reveal the level of investigation effectiveness. For an overview of indicators, we refer to the last phase 'monitor and evaluate'.

C) Potential boundaries

Several potential boundaries could be defined (in terms of availability of data, technology, human and financial resources; legislative framework applicable).

C.1) Availability of data

Before defining data needs, it should be clear which forms of fraud should be detected by the enforcement body. 'Undeclared work' defined by the European Commission is "any paid activities that are lawful as regards their nature but not declared". This is quite narrow and excludes other kinds of tax evasion and tax avoidance. Moreover, the borderline with criminal and illegal activities is sometimes very thin. Furthermore, a distinction between the detection of fraud on the one hand and error on the other should be made. Fraud can be defined as "any act or omission to act, in order to obtain or receive social security benefits or to avoid obligations to pay social security contributions, contrary to the law of a Member State" while error is defined as "an unintentional mistake or omission by officials or citizens".

An important requirement is that enforcement bodies have access to the most relevant information. A broad range of databases might be of interest when trying to detect undeclared work (social security, taxes, labour law, occupational safety and health, etc.). A key question is how to get accurate data. Enforcement bodies have sometimes inadequate access to databases and information in general. Moreover, in their fight against undeclared work, they often need to rely on administrative data from other administrative authorities. However, it is not always easy to gain access to this data.

A first step in this process is to look internally at which data are useful. Interesting data may be found within the enforcement body. For instance, the outcome of audits might be very useful to gather some evidence on the profile of undeclared work: which companies, employers, employees, benefit recipients are more vulnerable to fraud? Or within the administration in which the enforcement body is located. Afterwards, data might also be obtained from other public or private sources.

The data set required for the data analysis might cover:

- Internal data: data originating from the enforcement body itself or available within the administration in which the enforcement body is located;
- External data: data obtained by the enforcement body from other administrations or from other public or private sources.

C.2) Availability of technology

The use of data analysis tools, both in a preventative and curative manner, should help to maximise the audit benefits and to minimise the audit costs. The investment in technologies by enforcement bodies may range considerably: from the use of free open-source software to specially designed programmes and systems. Many advanced tools are available either as open-source or commercial software. However, the performance of the solution is a critical success factor for the project with direct impact on investigation duration. For this reason, particular attention should be paid to hardware sizing, to software requirements for hardware, and to the overall solution performance. Because each software platform for data analysis has particular requirements in terms of hardware resources, and because there are specific requirements in each country (in terms of data processing, volumes, number of users, etc.), it is impossible to generalise a set of requirements.

C.3) Availability of human and financial resources

Enforcement bodies might currently suffer a lack of staff and resources. As a result, the number of inspections and actions might be too limited to tackle undeclared work. It is therefore important to consolidate the limited resources towards easy, effective and efficient inspections. It could be answered by a further process of automatization and a changed strategy towards a more targeted approach on the basis of data matching or data mining tools.

There will be a need for specialist staff in case of the development of a data analysis system. Some enforcement bodies might have in-house expertise. However, some enforcement bodies may have difficulties

finding IT specialists. With regard to ensuring the required level of skills of the workforce employed in the data analytics service of the enforcement body, specific training provisions may need to be made. This could be solved by the support of skilled people who have developed the commercial data-mining tool (i.e. external staff).

C.4) Impact of the legislative framework on the protection of personal data

Almost all countries have national legislation in place that protects personal data and safeguards privacy. Enforcement bodies should be aware of these rules before starting the process of data collection, data sharing, data matching and data mining. Moreover, it raises a number of questions. One question is whether access to personal data is proportional to the enforcement body's objectives. Another question is which steps should be taken to ensure the personal data used is protected so that its misuse is avoided.

Moreover, in 2018 there will be a comprehensive reform of data protection rules in the EU (see following link). Individuals will have more information on how their data are processed and this information should be available in a clear and understandable way. Moreover, 'Data protection by design and by default' will become an essential principle. The Regulation promotes techniques such as anonymisation (removing personally identifiable information where it is not needed), pseudonymisation (replacing personally identifiable material with artificial identifiers), and encryption (encoding messages so only those authorised can read it) to protect personal data.

1.1.2. Definition phase

In this phase the steps will be defined that should be taken to implement the type of data analysis system that is concluded on the basis of the initiation phase.

A) With regard to data collection and data sharing

On the basis of the analysis of which data are needed and where they are available (internally or externally), it might be necessary to contact other public or private institutions.

List of questions to be answered:

- Who are the authorities/partners involved?
- What do we have?
- What do we need?
- What do the other authorities/partners need?
- What can we share?
- Are the other authorities/partners prepared to share data?
- Are we allowed to share these data?

?

1) Define and mobilise the stakeholders/partners involved

Effective data sharing, matching and mining is a step-by-step process which requires political will and trust between the different parties involved as well as a clear idea of what data needs to be shared. In general, authorities tend to not know each other, and tend to not speak to each other and thus do not know what kind of data other authorities have. For instance, there might be an urgent need for more collaboration at national level between labour inspectorates and tax authorities. Establishing a willingness to exchange data would benefit the authorities involved. It should be added value for all and result in the creation of a sense of community around data sharing. A first step in this process could be the design of a 'Memorandum of Understanding'.

Also, cross-national cooperation might be needed to fight undeclared work. Cross-national cooperation is possible by the use of the Internal Market Information System (IMI). It facilitates communication between national authorities involved in activities relevant to the internal market, such as the posting of workers and patients' rights. Furthermore, Council Directive 2011/16/EU lays down the procedures for exchange of information concerning all taxes related to income from employment. Finally, in the near future, the Electronic Exchange of Social Security Information (EESSI) will help social security administrations across the EU to exchange information more rapidly and securely.

Good practice example of cross-national cooperation: BENELUX (Belgium, the Netherlands and Luxembourg)

In 2015, a Recommendation of the committee of the ministers was signed which aims to enhance collaboration to fight social fraud and social dumping. Under this, Belgium, the Netherlands and Luxembourg will work together against social dumping, undeclared work and other types of social fraud. The recommendation contains three main objectives: a) dealing with unfair competition and social dumping by means of improved cooperation and data exchange; b) taking measures wherever necessary in order to find an answer to certain loopholes in the regulatory framework or to cooperation problems; c) looking for support, within Benelux and together with other countries, to take the appropriate measures at European level for the purpose of the provisions of a) above. These objectives should be accomplished by seven measures at Benelux level. One of the measures is multidisciplinary cooperation and exchange of data by means of the cross-border use of databases (red flags). The Benelux countries have two working groups on 'bogus construction' and 'benefit fraud', which aim at improving the cross-border administrative cooperation within the countries concerned (and eventually beyond). As part of these working groups, pilot projects on cross-border data sharing and matching of relevant data have been set up to try to tackle cross-border fraud and error.

For more information please see the good practice fiche here: http://ec.europa.eu/social/BlobServlet?-docId=18526&langId=en

2) Define data needs and sources

On the basis of a first consultation with potential partners, it should be clear which data from which sources could be obtained or shared.

3) Define potential legal barriers to data access

Possible questions are:

- Are we allowed to?
- How are we allowed to?
- Are we willing to do it?

One of the key steps is also to remove legal barriers to the exchange of data between authorities. There should be a legal basis to exchange data between administrations or between EU Member States and it should be verified that enforcement bodies have such authorisation at national or cross-national level.

B) With regard to data governance

The toolkit defines two crucial components in the data governance process. Steps should be taken to protect the content of the collected data. Secondly, it is important to monitor the quality of the collected data.

B.1) Data protection and data security

Data protection and data security are key issues which need to be built into any data gathering and mining system from the very start. 'Privacy by design', or its variation 'data protection by design', should therefore guarantee an effective protection of privacy and data (ENISA, 2014; 2015). It is regarded as a multifaceted concept, involving various technological and organisational components, which implement privacy and data protection principles in systems and services. The European Union Agency for Network and Information Security (ENISA) has defined eight privacy by design strategies, both data oriented and process oriented, aimed at preserving certain privacy goals (Table 1) (See also following link). It is essential to implement a coherent approach to data privacy protection, taking into account the complete lifecycle of the analytics (data collection, data storage, data analysis, data usage). One very important privacy principle directly related to the data collection phase is that of 'data minimisation'. The data needs should be precisely defined (i.e. what personal data are actually needed and what is not needed?). Furthermore, one of the most prominent techniques in the context of data analysis is that of anonymisation. Finally, a very important technique in privacy preserving analysis is encryption.

Table 1. Privacy by design strategies

	Privacy by design strategy	Description
1	Minimize	The amount of personal data should be restricted to the minimal amount possible (data minimalisation).
2	Hide	Personal data and their interrelations should be hidden from plain view.
3	Separate	Personal data should be processed in the distributed fashion, in separate compartments whenever possible.
4	Aggregate	Personal data should be processed at the highest level of aggregation and with the least possible detail in which it is (still) useful.
5	Inform	Data subjects should be adequately informed whenever processed (transparency).
6	Control	Data subjects should be provided agency over the processing of their personal data.
7	Enforce	A privacy policy compatible with legal requirements should be in place and should be enforced.
8	Demonstrate	Data controllers must be able to demonstrate compliance with privacy policy into force and any applicable legal requirements.

Source: ENISA (2015)

B.2) Data quality management

An important aspect is to have good, accurate, and well-structured data. Moreover, they should be comprehensible, manageable and meaningful. Poor quality data can create more problems than it solves. In fact if the data is inaccurate or contains errors, the analysis can be incorrect or even counterproductive (EC, 2010). To achieve high levels of data quality it is necessary to use filters that use many data integrity rules and thereafter to carry out checks relating to the quality of the data. It is therefore crucial to devote time and resources to clean the data. Clarifying terminology between agencies at national level is another important step. Finally, it is essential to have a robust data referencing system with good descriptions of the data explaining what they are and identifying the sources (i.e. a well-established data library). It is notably important not to lose track of the origin, the primary source, or the exact definition since 'the devil is in the detail'.

Some of the data might not yet be available electronically. For instance, detailed information on audits and their outcomes might only be available on paper. A first essential step is therefore the digitalisation of the inspection data.

The quality of the data could be assessed according to different quality components (relevance, accuracy, timeliness and punctuality, comparability, coherence, accessibility and clarity). These quality components are defined by Eurostat for the assessment of data in statistics. However, they may also be used as a base to assess the quality of the collected public or private data.

Relevance	Relevance is the degree to which data meet current and potential user needs. It refers to whether all data that are needed are produced and the extent to which concepts (definitions, classifications etc.) reflect user needs.
Accuracy	The extent to which the data are free of identifiable errors. Accuracy is the closeness of results of observations to the true values or values accepted as being true. To be correct, a data values must be the right value and must be represented in a consistent and unambiguous form.

Timeliness and punctuality	Timeliness of information reflects the length of time between its availability and the event or phenomenon it describes.
Comparability	Comparability aims at measuring the impact of differences in applied data when several databases are compared or over time.
Coherence	Coherence of data variables is their adequacy to be reliably combined in different ways and for various uses. When originating from different sources, data variables may not be completely coherent in the sense that they may be based on different approaches, classifications and methodological standards.
Accessibility and clarity	Accessibility refers to the conditions under which users can obtain data. Clarity refers to the data's information environment whether data are accompanied with appropriate documentation, whether information on their quality is also available (including limitation in use etc.) and the extent to which additional assistance is provided.

In addition to this list, other indicators to measure the data quality are, among others, completeness, validity, stability and continuity.

B.3) Data storage

Enforcement bodies could store their data in a centralised data warehouse. The data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading and periodic data refreshing. Holding separate database systems is more costly and less efficient than using a combined system which is accessible to all users.

C) With regard to data technology

Some key lessons when developing a data analysis tool

- It is important to share experiences/good practices. Look at the data analysis tools used by other enforcement bodies at national or cross-national level.
- You could start it as a pilot project (i.e. a small and manageable project) and if successful then move slowly forward.
- The design should be clear as it is difficult to correct it afterwards.
- Incomplete and incorrect data records should be reported to the data source holder, who is expected to take action in order to correct it. Thus, we can expect to reduce, in time, 'dirty data'.

Different tools and techniques can be used by enforcement bodies depending on what level of 'knowledge' and 'technology' is available. Moreover, enforcement bodies have to take into account constraints on the available financial and human resources. Finally, it is much more difficult to predict future cases of undeclared work (i.e. a predictive model) than selecting cases on the basis of past and present data.

Data mining has a wide number of applications and as a consequence a large number of data mining tools have been developed over decades. This toolkit provides an overview of the existing data-mining tools without making any final conclusion about the best choice for enforcement bodies. Most important is that enforcement bodies know the existing data-mining tools and are aware about their features, advantages and limitations.

Table 2. List of popular commercial and open-source data-mining tools

Popular commercial tools	
ADAPA (Zementis)	www.zementis.com
CART	www.salford-systems.com
IBM SPSS Modeler	www.spss.com
MATLAB	www.mathworks.com
Oracle Data Mining (ODM)	www.oracle.com
SAP	www.sap.com
SAS Enterpriser Miner	www.sas.com
SQL Server Analysis Services (SSAS)	www.microsoft.com
Teradata Database	www.teradata.com
TIBCO Spotfire / Statistica	https://spotfire.tibco.com
Popular open-source tools	
ADAMS	https://adams.cms.waikato.ac.nz
KEEL	www.keel.es
KNIME	www.knime.org
ORANGE	https://orange.biolab.si
Rattle (R)	https://www.r-project.org/
R Analytic Flow	http://r.analyticflow.com/en/
RAPIDMINER	www.rapidminer.com
WEKA	www.cs.waikato.ac.nz/ml/weka

Source: Mikut and Reischl (2011); Altahi Abdulrahman et al. (2017)

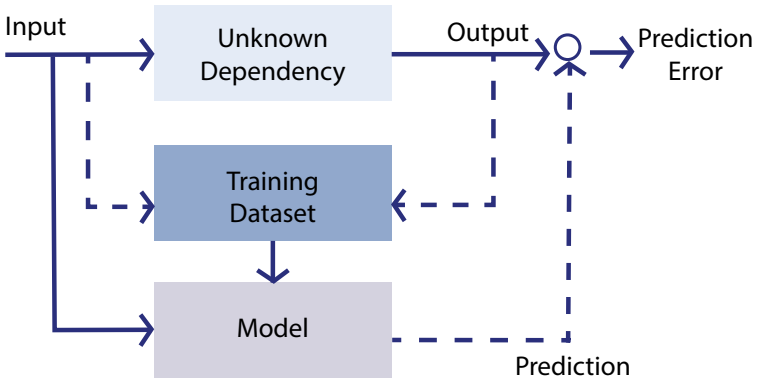
D) With regard to data methodology (i.e. risk analysis process)

Several methodologies can be used to detect (possible) fraudulent cases of undeclared work through the use of data-matching or data-mining tools.

In case of data matching two or more datasets are compared. For example, enforcement bodies can verify, amongst other practices, if social benefits are cumulated with declared paid work.

In case of data mining, the primary challenge is to build a model that has the ability to accurately predict whether or not a company or a person is compliant. It supposes the identification of patterns in a set of data by an algorithm. It is a well-defined procedure that takes data as input and produces output in the form of a model.

Figure 3. Risk analysis process



Two approaches can be used. Deduction, which is the first approach, begins with an expected pattern/theory/hypothesis that is tested. For instance, in some cases the development of the data mining tool is based on the input received by inspectors. They define certain risks that could be captured/detected by the use of data-mining techniques. The second, induction begins with the observations/data and seeks to find a pattern within them and theories are proposed as a result of the observations. For instance, on the basis of the outcome of previous audits, the characteristics of employers who have committed undeclared work could be compared with the characteristics of employers who made no infringements.

Detecting outliers is the way in which some Member States go about data mining. It is primarily about detecting anomalies in normal behaviour. It determines the difference between normal and suspicious/new behaviour, identifies anomalies, categorises and prioritises risks for further investigation.

Examples of data-mining techniques are (Khwaja et. al, 2011):

- 'Decision trees': this technique identifies groups of individuals or businesses that are as homogeneous as possible based on a set of predefined variables. It is based on an algorithm using separation criteria to identify the groups;
- 'Neural networks': this technique is similar to decision trees in the sense that it seeks to identify homogeneous groups based on a set of variables and criteria. However, it does not require a hierarchy in the variables;
- 'Clustering': this is another segmentation technique that allows for the simultaneous analysis of several possible explanatory variables during the segmentation process.

Furthermore, convinced that fraudsters are often connected to each other (for example, via the same accountant, managing directors, clients, suppliers, etc.) or that they may have many things in common with other fraudsters, Member States started network analytics to rank and profile cases.

A behaviour model is used to identify behaviour on the basis of logistic regression analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable (e.g. non-compliance) and one or more nominal, ordinal, interval or ratio-level independent variables. Dynamic benchmarking spots different patterns between similar businesses in close proximity. This benchmark will differ among regions. For example, one ratio compares declared turnover to the use of credit cards. A too high share of payments by credit cards in total declared turnover, in comparison to the benchmark determined for the specific area, will raise an alarm.

E) With regard to the output

The use of the data analysis tool could result in different types of output.

E.1) Individual or aggregated risk-based audit selection: risk ranking and prioritisation

The central goal of risk assessment is to choose which risks are to be treated, given the available human and financial resources (EC, 2010). The identified and analysed risks must be ranked and prioritised, because they differ from each other in several ways, for example financially, the appropriate treatment (capacity or timing) or the effect. However, the perceived likelihood of control and thus the deterrent effect might be very low when relying mainly on these practices. Therefore also random audits are still useful. This ensures that every company or person has an equal chance of treatment. The results of the random activities can also identify new risks. And not least, the results can be used as an indicator for the effectiveness of the risk management process which is in use.

Some indicators which play a role in this process include:

- The amount which is involved directly or indirectly;
- The resources for treatment;
- Deterrent effect: the treatment of some risks can have positive deterrent effect on others;
- Social and compliance effects: the tackling of some risks may have a greater effect on the social acceptance and/or the compliance level of companies/persons than others.

The reason why the employer is considered to be vulnerable to fraud should be clear for the inspector. Some of the technology being used is user-friendly and presents the results visually, making it easy for inspectors to access and use.

E.2) Revealing patterns of undeclared work

Companies, employers, employees or benefit recipients who have committed undeclared work might show similar characteristics. Moreover, data could reveal that sectors of activity or regions are more sensitive to undeclared work. Patterns detected by data analysis tools could be reported in the annual reports of the enforcement bodies.

E.3) Measuring undeclared work

The use of these practices is also a way to measure the size of the committed undeclared work in terms of persons and amount involved. Revealing and measuring what is hidden is a very difficult task. Both direct and indirect methods can be used. Indirect methods are often based on the comparison of macroeconomic aggregates while direct methods are based on statistical surveys. Both type of methods have their merits and limitations. However, the use of data from the enforcement bodies might also be relevant (for instance, on the basis of random audits).

1.1.3. Planning phase

In the last intermediate step some final decisions should be taken.

List of key questions to be answered:

- What data will be collected/shared?
- What will be the 'privacy/data protection by design' strategy?
- What will be the data governance strategy?
- What will be the data quality strategy?
- What will be the data analysis tool?
- What will be the output?

In this phase some good practices are defined which might be useful for enforcement bodies. They are mainly based on the presentations of practices and experiences of data collection, data sharing, data matching and data mining by the participants to the Thematic Review Workshop organised in Helsinki.

A) What data will be collected or shared?

Two good practices of data sharing are presented below. The Crossroads Bank for Social Security (CBSS) in Belgium is an example of how data, with attention to data security and privacy protection. The Grey Economy Information Unit in Finland shows how challenges of sharing data between agencies could be overcome.

Good practice example of data sharing: The Crossroads Bank for Social Security (CBSS) – Belgium

The CBSS elaborates an E-government strategy within the Belgian social sector. The mission of the CBSS is to stimulate and to support the actors in the Belgian social sector to grant more effective and efficient services with a minimum of administrative formalities and costs for all those involved. It also promotes the information security and the privacy protection of the actors in the Belgian social sector so that all those involved can have confidence in the system.

All the social security institutions are connected to a network for the electronic data traffic managed by the CBSS and have the legal obligation to electronically ask one another for all information available in the network. The CBSS regulates the data exchanges. Every socially insured person is identified throughout the whole social security system by a common and unique identification key and has an electronically readable identity card containing this identification number. In 2016 some 1.1 billion electronic data exchanges took place with a response time for the online messages of less than 4 seconds in 99.27 % of the cases. The CBSS provides labour inspectorates with information useful to plan action and investigate cases. The main Belgium labour inspectorates have access on a permanent basis to interesting databases via the use of the CBSS.

For more information please see the good practice fiche here: https://www.eurofound.europa.eu/data/tackling-undeclared-work-in-europe/database/e-government-in-social-security-sector-belgium

Good practice example of data sharing: Grey Economy Information Unit (GEIU) - Finland

The Grey Economy Information Unit (Harmaan talouden selvitysyksikkö), a specialist unit within the Finnish tax administration, gathers information and conducts investigations on undeclared work. The unit is authorised to keep a database within the meaning of the Data Protection Directive (95/46/EC, 1995), containing information necessary for the preparation of reports (at individual, sectoral or national level). This permanent structure creates a better cooperation and coordination between authorities and removes barriers and limits to the exchange of information. Challenges of sharing data between agencies were addressed early on. Authorities with the appropriate legal permissions can become a client of the GEIU and access the different reports available. The Finnish parliament has been responsive in removing legal barriers along the way.

For more information please see the good practice fiche here: http://ec.europa.eu/social/BlobServlet?docId=18511&langId=en

B) What will be the 'privacy/data protection by design' strategy?

Following the aforementioned description of the privacy by design strategies by ENISA (2014; 2015), the table below provides an overview of the possible implementation measures in each of the phases of the process of knowledge discovery from databases.

Table 3. Implementation of the privacy by design strategies

Phase	Privacy by design strategy	Implementation
Data Acquisition/ Collection	Minimize	Define what data are needed before collection (reduce data field, define relevant controls, delete unwanted information, etc.). Privacy Impact Assessments.
	Aggregate	Local anonymisation (at source).
	Hide	Privacy enhancing end-user tools, e.g. anti-tracking tools, encryption tools, identity masking tools, secure sharing etc.
	Inform	Provide appropriate notice to individuals- Transparency mechanisms.
	Control	Appropriate mechanisms for expressing consent. Out-out mechanisms. Mechanisms for expressing privacy preferences, sticky policies, personal data stores.
Data Analysis & Data Curation	Aggregate	Anonymisation techniques (k-anonymity family and differential privacy).
	Hide	Searchable encryption, privacy preserving computations.
Data Storage	Hide	Encryption of data and rest. Authentication and access control mechanisms. Other measures for secure data storage.
	Separate	Distributed/de-centralised storage and analytics facilities.
Data Usage	Aggregate	Anonymisation techniques. Data quality, data provenance.
All Phases	Enforce/ Demonstrate	Automated policy definition, enforcement, accountability and compliance work.

Source: ENISA (2015)

However, a privacy be design strategy could also be defined at a macro level.

Good practice of data protection: The Commission for the Protection of Privacy (CPP) - Belgium

The Commission for the Protection of Privacy (CPP), better known as the Privacy Commission, is an independent body ensuring the protection of privacy when personal data are processed. The Privacy Commission was established by the so-called 'Privacy Act'. The Privacy Act is intended to protect citizens against the abuse of their personal data. The rights and obligations of the individual whose data are processed as well as the rights and obligations of the processor have been laid down in this act. Several sector committees have been established within the Privacy Commission. These committees ensure that privacy is protected when personal data are processed in a specific sector. For example, the sector committee 'Social Security and Health' protects the privacy of beneficiaries of the Belgian social security network, and ensures particular supervision of the communication of health-related data. It consists of two sections: the Social Security Section and the Health Section. The Social Security Section has the power to grant authorisations for the communication of social personal data by the Crossroads Bank for Social Security and deals with complaints and resolves disputes regarding the violation of the legal rules it supervises.

C) What will be the data quality strategy?

Different steps could be taken to assure the quality of the collected data. By data quality management the data quality could be assessed, reported and if needed improved on the basis of the quality components listed in the 'definition phase' (i.e. relevance, accuracy, timeliness and punctuality, comparability, coherence, accessibility and clarity).

Data quality management process:

- Define data quality: define the quality components and standards.
- Plan and implement: develop and implement a set of procedures to produce, check, and ensure data of acceptable quality.
- Perform acceptance tests and evaluate results: perform tests to compare delivered data to acceptability metrics.
- Take corrective action: take steps to clean, correct, re-collect or reprocess data as needed to achieve data acceptance standards.
- Report on data quality: document the data quality standards, protocols, processing methods, acceptance tests, and results. Report inappropriate data records to the data source holder, who is expected to take action in order to correct it.
- Improve the process: use the knowledge and experience gained to modify processes as needed to improve data quality.

D) What type of data analysis tool?

Different types of data analysis tools could be implemented. Below some good practices of the Federal Public Service (FPS) Social Security in Belgium and of the HRMC (Her Majesty's Revenue and Customs) in the UK are provided. Moreover, a follow-up visit on this topic took place on 27 and 28 September 2017, hosted by the Federal Public Service Social Security in Brussels. The event provided the opportunity to learn more about the data mining approaches used in Belgium to tackle undeclared work.

Good practice example of a well-functioning analysis tool: Connect Tool, HMRC, UK

Within HMRC (Her Majesty's Revenue and Customs) the vast majority of data is cross-matched using the data-analysis tool, Connect. This ingests over three billion data items and looks towards matching them and producing connected entities. The data collected is a mixture of own data from tax returns and 'third party data' from other public and private institutions. In total there are around some 40

data sets amounting to 22 billion lines of data and 600 million documents. There are some 250 data analysts and 4,000 users. The tool uses information from all HMRC data systems related to tax declarations for self-employed individuals, employees and employers, companies and business, property and land taxes, and indirect and consumption taxes and makes connections between the data to identify all data related to individuals and businesses. In this way HMRC is able to see a comprehensive picture of its taxpayers and the data relating to them. Recently HMRC has been using the data within Connect to create maps of undeclared working, overlaying their data onto mapping software to provide a detailed visual map of undeclared work down to street and property level. They aim to use this approach to better target their compliance resource into risky locations.

For more information please see the good practice fiche here: http://ec.europa.eu/social/BlobServlet?docId=18525&langId=en

Good practice example of a well-functioning analysis tool: FPS Social Security, Belgium

For several reasons the practice of data analysis by the Federal Public Service Social Security in Belgium might be interesting for other labour inspectorates. Firstly, with a limited staff a lot of useful input to the inspectors is provided. Secondly, the data are collected on the basis of the Crossroads Bank for Social Security (see above) and stored in a data warehouse. Thirdly, the investment cost is relatively low through the use of open-source software. Fourthly, inspectors are strongly involved as they provide input and validate the results. Finally, results are visual and therefore user-friendly for inspectors. Currently, three data analysis tools are used: 1) on the basis of machine learning; 2) on the basis of network analysis; 3) on the basis of anomaly/novelty detection.

For more information please see the good practice fiche here: http://ec.europa.eu/social/BlobServlet?docId=18372&langId=en

E) What will be the output?

E.1) Individual or aggregated risk-based audit selection

Technology should be user-friendly and should present the results visually, making it easy for inspectors to access and use (for instance, by the use of colour-coded results, flagging potential fraudsters). Examples of good practices are presented below from the FPS Social Security in Belgium and HRMC in the UK.

Good practice example of a well-functioning reporting tool: MiningWatch, FPS Social Security, Belgium

Since the beginning of 2015 MiningWatch is available for all inspectors of the FPS Social Security in Belgium. It can be considered as the (missing) link between the data mining experts and the labour inspectors. By this tool the results of data mining are shown and classified by the risk levels of enterprises (by using specific colours).

For more information please see the good practice fiche here: http://ec.europa.eu/social/BlobServlet?docId=18372&langId=en

Good practice example of a well-functioning reporting tool: the Integrated Compliance Environment (ICE), HRMC, UK

Spider diagrams show the entity and the associated data links such as addresses and telephone numbers, bank interest, income and lifestyle. Users can quickly identify potential compliance and fraud risks and present complex relationships in a simple format, ready for further investigation.

E.2) Revealing patterns of undeclared work

Good practice example of a well-functioning reporting tool: reports published by the Grey Economy Information Unit, Finland

The unit produces three types of reports

- **Compliance reports:** Investigate specific organisations and persons suspected of engaging in undeclared work at the request of other organisations, such as the police, customs bureau and Finnish Centre for Pensions as well as authorities dealing with work safety, debt recovery and bankruptcies. The report describes the operations and finances of an organisation or an associated person and the management of obligations related to taxes, statutory pension, accident or unemployment insurance contributions, or the fees charged by Finnish Customs. A compliance report is also available in Excel. During 2015, the Grey Economy Information Unit prepared a total of 202,184 compliance reports.
- **Classification reports:** These are highly standardized anonymous reports. Some 100 classification reports are published every year (for instance, restaurants in a specific area). Reports should be interesting for decision makers.
- **Grey economy reports:** Some 10 to 15 reports every year mainly interesting for policy makers.

For more information please see the good practice fiche here: http://ec.europa.eu/social/BlobServlet?docId=18511&langId=en

1.2. Implement

An automated process of knowledge discovery from databases (KDD) could be followed. Steps to be taken in the implementation phase are:

1. Preparation of data for analysis:
   - Store collected data;
   - Profile the data;
   - Establish metrics and define expectations/targets;
   - Implement Data Quality Rules;
   - Data Quality Assessment;
   - Explore and modify: data cleaning, data integration, data selection, data transformation;
   - Address identified data errors;
   - Review exceptions and refine quality rules;
   - Monitor data quality versus expectations/targets;
2. Integrate Data Quality Rules into Data Integration Processes
3. Train, test and implement the data analysis tool

1.3. Monitor and evaluate

1) A crucial role for the inspector

For several reasons it is important to involve inspectors when constructing the data analysis tool. Input from inspectors might be needed to define and assess the potential alarms. Involving inspectors at all stages of the data gathering and matching/mining process is important to gain their trust. Moreover, in cases where inspectors are involved, the chance is greater that they will endorse and promote the tool.

2) Measuring the return on investment

The return on investment (ROI) should be clearly measured in order to help enforcement bodies understand the impact of more advanced data usage on the outcome of inspections. Such evaluations also increase internal and external awareness of the capacities of public administrations and the potential value of investments in data sharing/matching/mining in order to prevent and deter non-compliance with labour and tax rules.

The effectiveness of the data matching or data mining tool could be assessed (a priori and a posteriori) by looking at the result of the audit. Four possible outcomes are thinkable:

- True positive (TP): When data correctly predict someone is engaging in undeclared work.
- True negative (TN): When data correctly predict that undeclared work is not taking place.
- False positive (FP): When data falsely predict someone is engaging in undeclared work, whilst in fact s/he is not.
- False negative (FN): When data do not alert that undeclared work is taking place.

Table 4. Outcome of the prediction

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

The total number of inspections based on the use of the data analysis tool is: $T_{insp.} = FP + TP$

Several indicators could be defined to measure the effectiveness and performance.

The outcome of audits selected by the data analysis tool could be compared with the outcome of random selected entities.

Accuracy (TP+TN)/(TP+TN+FP+FN): This indicator measures the percentage of cases predicted correctly by the model. The ambition should be to minimise the number of FN and FP cases. One of the key goals of data mining is to reduce false positives in order to avoid one wasting time on false positives every day as valuable time and resources will be lost. Alarms should be set at optimum levels to reduce under/over linking of data creating false positives. Good models will reduce false positives, but even the very best of models will not eliminate them. Moreover, in fraud detection, misclassification costs (false positive and false negative error costs) are uncertain, can differ and can change over time.

Efficiency or positive predictive value TP/(TP + FP): This indicator measures the percentage of noncompliant cases likely to be detected if predicted evading cases are audited. The percentage of true positive cases will be counted. Notably, how regularly does the data mining correctly identify undeclared work which is then proved by the inspector? The audits provided for the FP-cases are lost efforts. It also relates to the feasibility of proving undeclared work in practice by the labour inspector. It is not always possible to prove undeclared work in practice, despite obtaining a true positive result. Finally, also the time that is required to detect the fraud could be used as a variable to measure the efficiency.

Negative predictive value TN/(TN + FN);
True positive rate or prediction efficiency: TP/(TP + FN): This indicator measures the percentage of noncompliant cases correctly predicted by the model.

True negative rate: TN/(TN + FP);
Proficiency: The amount which is involved. The detection of higher fraud amounts might have priority. Moreover, the additional tax identified and collected should be measured.

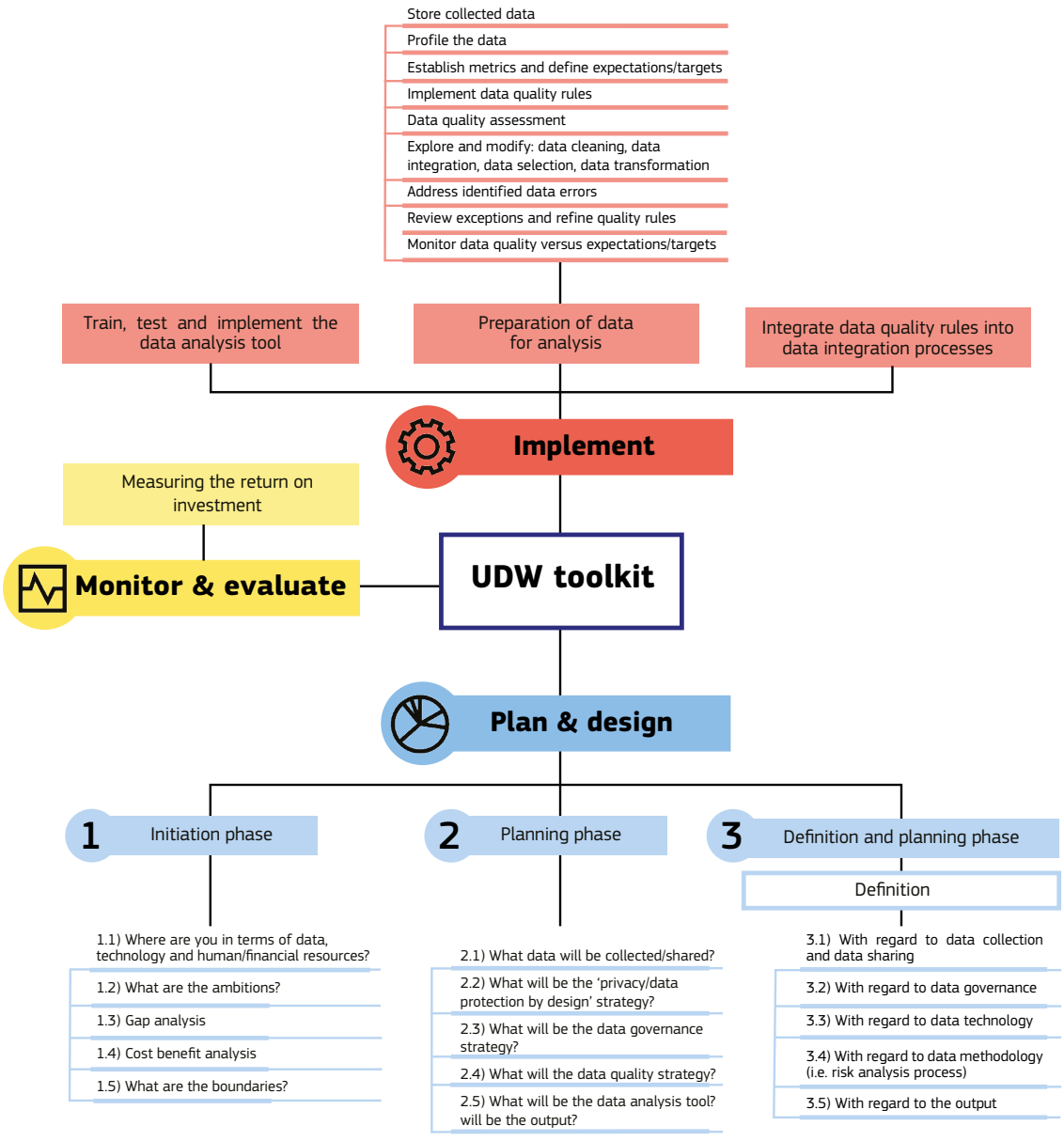
Measuring the behavioural change (impacts over time): The evaluation should not be limited to the direct impact but extended to measuring its ongoing impact on behaviour. For instance, it might have a more positive effect on general compliance. Therefore, some compliance indicators could be defined. However, the perceived likelihood of control and thus the deterrent effect might be very low when relying mainly on these practices. Moreover, audits have not only the aim to detect undeclared work but certainly also to reduce the size of it as a result of their deterrent effect.

Follow-up: What is happening with this company six months later?; Has it changed its behaviour?

Coverage (width and depth): Total data set of companies, employers, employees, self-employed, benefit recipients, etc. included in the data matching / data mining tool.

2. PRACTITIONER'S TOOLKIT MIND MAP

The following mind map is a visual representation of the topics presented in this document. Its purpose is to help the reader to build an intuitive framework with the presented guidelines



1.1

a) Inventory of existing level of data collection, data sharing, data protection, data quality management, data matching and mining

b) Assessment of your organization's readiness to implement the project

- Available hardware and software
- Access to/availability of required data sources
- Technical skills to carry-on the implementation
- Available financial resources to support project cost

c) Assessment of compliance with respect to GDPR regulation

1.2

a) What forms of fraud should be addressed first to better support anti-fraud effectiveness in your country?
b) Where do you want to go in terms of data collection, data sharing, data protection, data quality management, data matching and mining?
c) What are realistic short and long-term objectives and what should the result be?

1.3

a) What are realistic short and long-term objectives?
b) What are realistic short and mid-term risks?

1.4

a) What is the estimate timeframe for the project implementation?
b) What are the expected costs versus the expected benefits (ROI)?

1.5

a) Should or could you have this ambition?
b) Technical feasibility?
c) Financial feasibility?
d) Availability of human resources?
e) What is the legislative framework on the protection of personal data?
f) Any other constraints?

2.4

Data quality management
a) Define data quality
b) Define Critical To Quality Elements
c) Plan and implement
d) Perform acceptance tests and evaluate results
e) Take corrective action
f) Report on data quality
g) Improve the process

2.5

a) Individual or aggregated risk-based audit selection
b) Revealing patterns of undeclared work

3.1

a) Who are the authorities/partners involved?
b) What do we have?
c) What do we need?
d) What do the other authorities/partners need?
e) What can we share?
f) Are the other authorities/partners prepared to share data?
g) Are we allowed to share these data?
h) How are we allowed to share these data?
i) Are we willing to share these data?
j) Process
- Define and mobilise the stakeholders/partners involved
- Define data needs and sources
- Define potential legal barriers to data access

3.2

a) Process
- Data protection and data security
- Data quality management
- Data storage

3.5

a) Process
- Individual or aggregated risk-based audit selection: risk ranking and prioritisation
- Revealing patterns of undeclared work
- Measuring undeclared work

FURTHER INFORMATION AND RESOURCES

Altalhi Abdulrahman, H., Luna, J. M., Vallejo, M. A., Ventura, S. (2017), 'Evaluation and comparison of open-source software suites for data mining and knowledge discovery', *WIREs Data Mining Knowl Discov,* Vol. 7.

EC – DG TAXUD (2010), *Compliance Risk Management Guide for tax administrations.*

Mikut, R. and Reischl, M. (2011), 'Data mining tools', *WIREs Data Mining Knowl Discov*, Vol. 1.

Khwaja, M.; Awasthi, R.; Loeprick, J. (2011), *Risk-Based Tax Audits. Approaches and Country Experiences*, The World Bank.

OECD (2004), *Compliance Risk Management: Audit Case Selection Systems.*

ENISA (2014), *Privacy and Data Protection by Design*

ENISA (2015), *Privacy by design in big data*

EUROSTAT (2007), *Handbook on Data Quality Assessment Methods and Tools*

European Platform Undeclared Work (2017), *Practitioners Toolkit: Drafting, Implementing, Reviewing and Improving Bilateral Agreements and Memoranda of Understanding to Tackle Undeclared Work*

FEEDBACK NOTE

We hope that you found this toolkit useful. If you have any feedback or comments, please do not hesitate to contact us on: EU-UDW-PLATFORM@icf.com