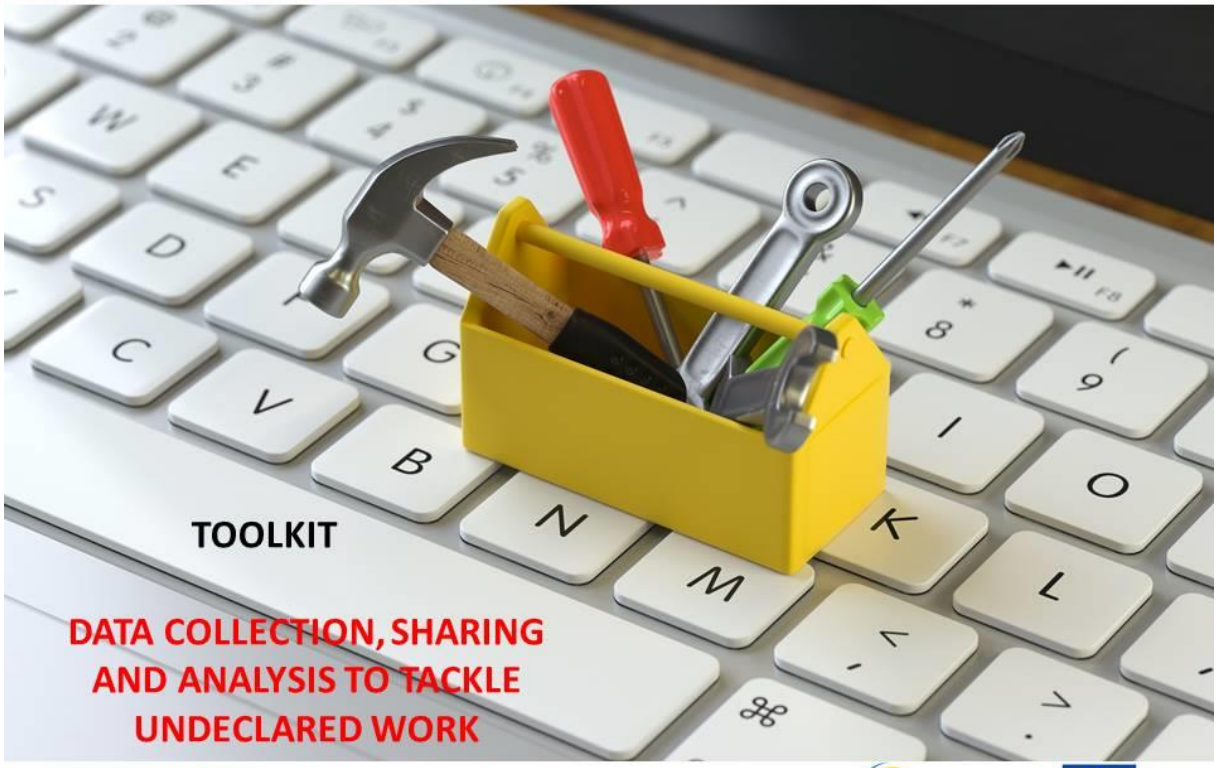
Employment & Social Affairs Platform 2 Project #ESAP2



RegionalCooperationCouncil

This project is funded by the European Union

RegionalCooperationCouncil

This project is funded by the European Union

Western Balkan Network tackling undeclared work

DATA COLLECTION, SHARING AND ANALYSIS TO TACKLE UNDECLARED WORK: A TOOLKIT

Sarajevo, June 2021

By Colin C Williams

University of Sheffield

This document has been prepared for the Regional Cooperation Council (RCC) within the framework of its Employment and Social Affairs Platform 2 Project (ESAP 2), funded by the EU, and reflects only the views of the author. Responsibility for the contents lies solely with the author and RCC and/or EU cannot be held responsible for use which may be of any information contained herein.

Contents

Executive Summary

Collecting, sharing and analysing data can improve enforcement authority performance not only in detecting undeclared work but also preventing non-compliance. The aim of this toolkit is to support enforcement authorities in the Western Balkans to improve their knowledge and awareness of how to develop efficient databases to detect, prevent and predict undeclared work. The objectives of this toolkit are (i) to discuss the challenges faced around data collection, sharing and analysis and (ii) to offer tips on how data collection, sharing and analysis could be improved, not least by reporting good practices.

Data collection

- **Data collection** is the process of gathering data from internal and external sources.
- There are large variations in the **maturity levels** of economies in relation to data collection, with some economies in the very early stages of data gathering whilst others have more developed databases.
- For enforcement authorities to be effective, they require electronic access to data on inspections/audit, businesses and employment to enable potential undeclared work to be identified. This requires:
    - the existence of databases (e.g., inspection records, businesses and employment registers), and
    - the development of an IT system that collects and stores comprehensive and high-quality up-to-date individual-level data, based on a robust data referencing system with good descriptions of the data explaining what they are and the sources, that can be made available to all relevant levels of the organisation, including inspectors.
- This collection of data is not only for the purpose of selecting workplaces to inspect but also for preventative actions such as selecting businesses and workers to whom notification letters and educational and awareness raising materials can be sent.
- Rather than simply use existing data to detect and prevent undeclared work, a **strategic approach to data collection** would start by asking "what data/information does my enforcement authority need to be able to identify undeclared work?". Sources are then identified that could provide such data.
- **Data protection and data security** are key issues which need to be build into any data gathering system from the very start.

Data Sharing

- **Data sharing** is the process of making data available to other users.
- There are large variations in the **maturity levels** of economies in relation to data sharing, with some economies in the very early stages of data sharing between authorities whilst others have a single central unit that collates all databases in a central warehouse. The Grey Economy Information Unit (GEIU) was set up in Finland to address this problem. The unit is a central point for producing and sharing information on the grey economy and its control.

- To implement effective data sharing systems, **a first key step is to remove legal barriers to the exchange of information** between agencies. This may require legislative action, bilateral agreements or Memoranda of Understanding (MoUs).
- It also requires that **data protection** and **data security** is built into any data sharing system from the outset. **Privacy by design**, or its variation "data protection by design", is a multifaceted concept, involving various technological and organisational components, which implement privacy and data protection principles in systems and services**.**
- Another barrier to the exchange of data is the degree of **political will** and **trust**. Data sharing requires political will and trust between the different parties involved as well as a clear idea of what data needs to be shared.
- **Interoperability of the data shared by agencies is crucial**. This requires a cross-government information technology infrastructure that actively supports the implementation of standardised processes. The design and the architecture of the information technology infrastructure will need to reflect the operational needs of the enforcement bodies and be capable of being updated without prohibitively high effort and cost. This can be achieved, for example, by making it mandatory for employers to register and de-register electronically their employees by their first day of starting work and the end of their last day of employment. Responsibilities for all these activities can be clearly defined to ensure accountability.

Data Analysis

- Data analysis involves either **data matching** (i.e., the large-scale comparison of records or files collected or held for different purposes, with a view to matching two or more sets of collected data) or **data mining** (i.e., a set of automated techniques used to extract buried or previously unknown pieces of information on potential instances of undeclared work from large databases, records or files collected for other purposes).
- There are large variations in the **maturity levels** of economies in relation to data analysis, with some economies in the very early stages of data gathering whilst others are using sophisticated techniques and technologies.
- To be effective in detecting and preventing undeclared work, there is a need for:
    - **Up-to-date robust data** to be available.
    - The databases containing the data to be **inter-operable**.
    - The appointment of **specialised staff** to administer and produce intelligence.
    - A well-functioning **data analysis tool**.
    - Measurement of "proof of concept" and "**returns on investment**" to secure resources.
    - Analyses to be **easily available to inspectors** to help them in detecting and preventing undeclared work, and for inspectors to be involved in all stages of developing the data analysis tool.
- Good practices are outlined of where this has been achieved.

1. Introduction

Modern enforcement authorities are increasingly engaging in data collection, sharing and analysis to improve their performance in detecting and preventing undeclared work, and to complement the qualitative judgements based on the experience and local knowledge of inspectors (European Platform Tackling Undeclared Work, 2017).

Collecting, sharing and analysing data can improve enforcement authority performance not only in detecting undeclared work but also preventing non-compliance. Databases containing information on businesses, employment, inspection visits and other "third party" data (e.g., bank account data) are fundamental tools for detecting and preventing undeclared work. Most enforcement authorities have databases, although at different levels of sophistication and completeness. There are also variations in the extent to which these databases are shared, and available to, others outside their enforcement authority. So too do the analytical methods used to identify potential instances of non-compliance and select targets vary across enforcement authorities.[1]

The aim of this toolkit is to support enforcement authorities in the Western Balkans seeking to improve their knowledge and awareness of building efficient data mining systems to detect, prevent and predict undeclared work. The objectives of this toolkit are (i) to discuss the challenges around data collection, sharing and analysis and (ii) to offer tips on how data collection, sharing and analysis could be improved, not least by reporting good practices.

To quote The European Commission's Directorate-General for Taxation and Customs Union (DG TAXUD):

> "There is no need to reinvent the wheel. Within … tax administrations there is a lot of knowledge and experience. Sharing good practices, joining forces and working closely together have proven to be efficient and effective ways of making tax administrations stronger".[2]

Therefore, the starting point of this toolkit is that labour inspectorates and tax authorities in different economies face common problems and have much to learn from each other. Sharing information on the challenges they have faced/face and how they have, or are seeking to, overcome them is valuable. It prevents enforcement authorities from having to "reinvent the wheel".

To examine this, the toolkit defines data collection, sharing and analysis as follows:

- **Data collection** is the process of gathering data from internal and external sources.
- **Data sharing** is the process of making data available to other users (De Wispelaere and Pacolet, 2017).
- **Data analysis** involves either **data matching** (i.e., the large-scale comparison of records or files collected or held for different purposes, with a view to matching two

[1] Historically, tax authorities have had more advanced databases and have used more sophisticated data analysis to detect and prevent undeclared work than labour inspectorates. See EC–DG TAXUD (2010), *Compliance Risk Management Guide for tax administrations*, European Union; Khwaja, M.S., Awasthi, R. and Loeprick, J. (2011), *Risk-Based Tax Audits. Approaches and Country Experiences*, World Bank; OECD (2004), *Compliance Risk Management: Audit Case Selection Systems*; OECD (2004), *Compliance Risk Management: Managing and Improving Tax Compliance*.

[2] https://ec.europa.eu/taxation_customs/business/tax-cooperation-control/administrative-cooperation-mutual-assistance-overview/tax-administration-tax-compliance_en

or more sets of collected data) or **data mining** (i.e., a set of automated techniques used to extract buried or previously unknown pieces of information on potential instances of undeclared work from large databases, records or files collected for other purposes) (De Wispelaere and Pacolet, 2017).

This toolkit does not focus upon **risk assessment**. Risk assessment occurs after data collection, sharing and analysis and will be discussed in a separate report later in the year. The focus in this toolkit is on the collection, sharing and analysis of data, which is the precursor to allowing effective risk assessment to be conducted.

In section 2, a review is undertaken of the challenges around **data collection** and good practice examples of how enforcement authorities have overcome these challenges. Section 3 examines the issue of **data sharing** again in terms of the challenges often faced by enforcement authorities and good practice examples of how these have been overcome and section 4 reviews the challenges around **data analysis** and good practice examples of how other enforcement authorities have overcome them. Section 5 then draws together some conclusions and recommendations.

2. Data Collection

Enforcement authorities, if they are to be effective, need to ensure they have access to data that can enable potential undeclared work to be identified. This can involve establishing for example:

- a case management database of audit/inspection outcomes;
- a business register, and/or
- employment register so that for example, real-time data is collected on the first day (and last day) of work of employees and/or their working time.

It can also include:

- "web scraping" to gather data on specific businesses or individuals.

It may also involve seeking access to "third party" data, including:

- individual or business bank account information from banks;
- information from telecommunications providers on individual businesses or employees, or
- information on internet service providers (ISPs) on the activity of businesses or employees.

This collection of data to identify undeclared work is undertaken for the following reasons:

- to enable data-driven risk assessment to be undertaken to select businesses for workplace inspections;
- to enable the data-driven preventative action of risk-based selection of businesses and workers to whom notification letters can be sent, and
- to enable the data-driven preventative action of selecting businesses, workers and citizens to target with educational and awareness raising materials.

To ensure that enforcement authorities have access to data that can enable undeclared work to be identified, this requires:

- the development of databases (e.g., employment registers, business registers) that can be analysed using data mining and matching for the purpose of identifying potential instances of undeclared work, and
- the development of an IT system that collects and stores comprehensive and high-quality up-to-date individual-level data about customers. Some enforcement authorities are more advanced than others on (1) developing these datasets, (2) having up-to-date real-time data which is accessible to all who need it, and (3) their ability to use the data collected for efficient enforcement.

These data bases need to:

- collect the data/fields/variables required to detect undeclared work;
- have access to these data on a real-time/up-to-date basis;
- be available to all relevant levels of the organisation who need these data, including inspectors.

The major **challenges** preventing the development of data bases are that enforcement authorities need:

- the financial resources available to develop these databases;
- the technical skills available to develop these databases;
- appropriate legislation on personal data and privacy safeguards that enables access to such data for enforcement authorities, and
- the "political support" to develop these databases.

At the most basic level, enforcement authorities need to develop a **case management database** reporting the records of inspections/audits undertaken and the results. Traditionally, these were kept in written form. In modern enforcement authorities, they are electronic records. They are also electronic records that are inter-operable which are up-to-date and fully accessible to those who need to use them in the organisation (and those outside who have permission).

In Belgium, for example, the results of inspections used to be recorded on paper rather than electronically. This was very resource intensive. However, an extensive programme of e-government has seen the introduction of electronic systems. This digitalisation of inspection records now allows these data to be shared with other authorities as well as used in data analysis. Similar digitalisation processes have occurred in many economies.

The second type of data collected relates to establishing employment registers, taxpayer registers, business registers, etc that are collated for purposes other than purely to detect and prevent undeclared work but can be used for this purpose. Box 1 provides an example of the Register of Employment in Estonia, Box 2 of the Revisal Employment Register in Romania and Box 3 the Incomes Register in Finland.

Box 1. Register of Employment, Estonia

Aims: the objectives of the introduction of the Employment Register were:

- to reduce the use of illegal labour;
- to improve the protection of employees' social rights;

- to simplify and streamline the work of the tax authorities;
- to increase the availability of electronic data and bring information relating to employment into a single system, and
- to reduce the administrative burden on employers and various public sector stakeholders, to simplify the operating principles of the social guarantee system.

Description: In Estonia, employers are required to register their employees in the electronic Employment Register before the employee starts work. The Register contains up-to-date data on employment in one place, enabling them to input and track their details in one single registration system. The Register of Employment is the responsibility of the Estonian Tax and Customs Board (ETCB). Registration is required for all employees, regardless of the form of contract, and also people working on a voluntary basis. Registration can be carried out by: E-Channel (E-Tax / E-Customs), at the ETCB, by phone or SMS. By offering three ways of registering and eliminating paper submissions, the process is easy for employers. Furthermore, employers can access their "live" employee data and make their own changes to it, at any point in time. Employees also receive automatic notification when their employment is registered and can check to ensure it is correct. Data from the Employment Register is used by several stakeholders. It is used to: determine health insurance; determine unemployment benefits (on termination of employment); monitor the working conditions of migrant workers; monitor and investigate accidents at work and verify tax compliance (labour taxes).

Alongside the electronic register, new IT tools have been introduced for the ETCB tax auditors. For example, mobile offices have been introduced, which tax officials can use during on-site inspections to perform and record all operations. Their new mobile working environment means that they can access all tax and employment-related data they need for inspections on-the-spot. A mobile app has also been developed, which enables officials carrying out inspections to use smartphones to gain direct access to the Employment Register, including a photo of the employee, alongside his / her employment data.

The initial cost to set up the Employment Register was €403,200. The annual maintenance cost is around €33,000.

Evaluation: With the introduction of the Employment Register, joint inspections are now carried out by the labour inspectorate and police and border guard. The local contact points of these authorities exchange operational information on a weekly basis and through this decide which employers should be subject to joint inspections. Through the joint inspections, a media campaign and press releases, the ETCB drew attention to the Register and its importance. These awareness-raising activities are necessary to encourage/inform people to use the database. Undeclared work has been reduced due to the Employment Register.

The impact in terms of additional tax revenue generated shows that an additional €11.8 million was collected in 2014 – this was linked to the registration of an additional 21,000 workers following the new requirements to register employment in the online Register. The introduction of the Employment Register also made it possible to measure the size of undeclared work in Estonia - by finding the proportion of undeclared workers identified during inspections as a percentage of the workers employed in the firms where inspections were carried out (controlled employees). The share of undeclared work according to this measure has decreased from 9.99% in 2014 to 6.28% in the first quarter of 2016.

Source: https://ec.europa.eu/social/BlobServlet?docId=17227&langId=en

Box 2. REVISAL Employment Register, Romania

Aim: The aim of this system was to:

- reduce the bureaucratic burden of registering labour contracts;
- increase the capacity of the labour inspectorate to reduce undeclared work by (i) increasing the transparency of employers' obligations towards the employees (ii) facilitating inspections and the detection of undeclared and under-declared work by providing labour inspectors with substantial information before going into the field**.**

Description: Since 1 January 2011, registration of labour contracts has been done exclusively through the REVISAL system. It provides information about all individual employee work contracts. All employers are obliged by law to fill in the database, using a desktop application which is provided free of charge by the Labour Inspectorate. Employers must send a complete record of employment for each new employee to the territorial Labour Inspectorate not later than the last working day before the employee's start date. Information in the database includes the following:

- Details about the employee, e.g., name, citizenship, personal identification code;
- Starting date of the individual employment contract;
- Job title, type of contract, working hours, salary and bonuses;
- The period and reasons of suspension of the individual employment contract, as well as the termination date.

Termination of employment contracts must also be sent to REVISAL which is then sent to the Labour Inspectorate within 20 working days (maximum) from when the work terminates. This means that if a person is working while their contract has been terminated, it is undeclared work. Furthermore, if a person is working more hours than what is included in REVISAL, it is under-declared work. The database can also help detect tax evasion (for example, if the salary included in REVISAL is lower than the salary the employee actual receives).

Employers must keep a paper file for each employee (the "personal file") in their offices, which must include all the paperwork necessary for the employment process. The employer must provide these files to the labour inspectors, if requested. At the written request of an employee or ex-employee, the employer must provide copies of all the documents included in their personal file, as well as copies of the pages included in the REVISAL registry.

The legislation clearly stipulates that if a company does not register a labour contract before the employee's start date, it is considered undeclared work and a fine can be imposed.

Evaluation: REVISAL has considerably reduced the bureaucratic burden of the Labour Inspectorate therefore increasing its capacity. Furthermore, the system is an essential part of planning inspections, as it gives labour inspectors access to key details about employee labour contracts, which helps to detect undeclared and under-declared work.

Success factors are: availability to all companies of a desktop application which has been critical to its success; provision of comprehensive guidelines on how to install and use the system has contributed to its success; and ability to integrate the system into payroll software is a key success factor. The system could be easily transferred to any interested economy, as the mandatory elements required for registration can be taken from the labour contract.

Source: https://ec.europa.eu/social/BlobServlet?docId=21643&langId=en

Box 3. The Incomes Register, Finland

Aims: The Incomes Register is an online database. It contains comprehensive information on individuals' wages, pensions and benefits. Data providers report individuals' earnings to the Incomes Register in real time, whenever a payment is made. The overall aim is to reduce undeclared work by showing in real-time which payment obligations have been met, to clarify to employers and employees what they need to do next, and to enable quick and easy detection of inconsistent information by authorities. The specific objectives are:

- To simplify and clarify employers' reporting obligations and simultaneously reduce the administrative burden involved;
- To tackle undeclared work by increasing the real-time transparency of the fulfilment of employers' obligations and enable efficient detection of omissions.
- To make income-related reports available automatically;
- To create direct communication between private payroll systems and the Incomes Register;
- To provide a real-time user interface for the citizens concerning their own earnings, pensions, and benefits, and
- To provide reports for different authorities according to each authority's mandate.

Description: The Finnish Incomes Register provides an up-to-date, comprehensive repository of individual earnings, social insurance contributions, benefits and pensions data which various authorities and all employers are obliged to report. It deters undeclared work by allowing the Finnish Tax Administration to establish whether all required payments have been made, reducing the likelihood of companies not reporting all incomes or reporting inconsistent information to different authorities.

The Incomes Register development project (KATRE) was introduced by the government in 2014 to lighten companies' reporting burden and simplify earnings-related reporting. The project was part of a wider digital infrastructure development programme. The law on the Income Information System (53/2018) obliged the relevant authorities and employers to use the new electronic system as a tool for reporting all payments by employers, benefits and pension providers. The first phase came into force on 1 January 2019 starting with the reporting of salaries and earnings. The number of parties using the information will increase in 2020. Phase two came into force 1 January 2021 and introduced the reporting of benefits and pensions.

The main activities include the following:

- The Register interface provides two user roles: those reporting on payments (e.g., employers) and those viewing payments made (e.g., citizens, government authorities);
- Employers as well as pension and benefit providers report all payments through the same electronic system, which conveys the information to each data user according to their legal entitlement;
- The reports include details of social, health, pension, accident and occupational disease contributions, and unemployment insurance contributions of varying types;
- The deadline for reports is 5 calendar days after a payment is made;
- Failure to report results in a fine of €135 for a delay up to 45 days, and after that, a

maximum fine of €15 000 per month is imposed, and

- The system allows citizens to access their own up-to-date payments and produce different reports from the data for various uses.

Evaluation: Digital and automatic reporting has achieved high coverage among employers with the help of supporting materials. Real-time reviewing of individual payments has become possible. Cooperation with the developers and many stakeholders has been a key success factor.

Outputs include the following:

- 250 000 employers submit reports to the Incomes Register;
- 65 instruction videos, 10 webinars, 2 web courses, marketing in newspapers, webpages, tv and outdoors marketing (budget €250 000).

Outcomes so far include:

- 87% of employers report digitally and automatically. Only 0.04% submit paper reports;
- Successful cooperation with private systems providers to build technical interfaces from the private payroll systems to link with the government Incomes Register;
- There is real-time data and transparency for data users, employers and citizens
- Challenges in the setup of payment processes and systems for the development and implementation of the project;
- Good practice is to provide ready-to-use materials for communications and training;
- The software companies' readiness to apply and build the interfaces should be monitored carefully. They should be given enough time to prepare;
- Adequate provision of resources for customer service is important in the beginning (first few months) before employers etc. are familiar with the new practice.

Sources: https://ec.europa.eu/social/BlobServlet?docId=21459&langId=en

Incomes Register web page: https://www.vero.fi/en/incomes-register/

Law on Incomes Information System (53/2018): https://www.finlex.fi/fi/laki/alkup/2018/20180053

English education material on Register: https://www.youtube.com/watch?v=dGdmL3n34JE

Usually, when developing databases to detect and prevent undeclared work, the decision is made to use data that has been collected for other reasons. This is the practical approach to data collection. Indeed, many innovative methods can be adopted, such as using big data and social media to detect and prevent undeclared work. However, when developing databases to detect and prevent undeclared work, enforcement authorities could adopt a **strategic approach to data collection** (see Box 4).

Box 4. Adopting a strategic approach to data collection

The practical approach to data collection uses the available data that has been collected for other purposes to detect and prevent undeclared work. A strategic approach to data collection

starts by asking the following question:

> What data/information does the enforcement authority need to identify instances of undeclared work?

To answer this, a meeting bringing together enforcement authority staff at various levels could be organised. At the outset of this meeting, it would be made clear that the aim is not to think about what data currently exists. Instead, it is to examine in an ideal world what data/information would enable instances of undeclared work to be identified (e.g., every unregistered worker would automatically report themselves to the authority or their employment would be easily viewable; all envelope wages would be reported by under-declared workers; all undeclared transactions by those in self-employment would be instantly viewable or reported by the self-employed or by their customers).

Having identified the ideal information/data required to enable instances of undeclared work to be known, the second step is then to consider whether and how such information/data could be made available/collected. This will require those attending the meeting to feel free to explore ideas without fear of retribution by colleagues. The outcome might be that new sources of information, or initiatives that could generate this information, will emerge.

3. Data Sharing

When using data mining and matching to tackle undeclared work, two types of data can be used by an enforcement authority:

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

Most enforcement authorities first look internally at which data are useful for detecting undeclared work. For example, audit/inspection data and their results might be used as well as data located within the enforcement body (e.g., declarations on the work performed by, and salaries of, employees).[3]

However, a broad range of databases might be of interest when trying to detect undeclared work (e.g., social security, taxes, labour law, occupational safety and health, bank account data). For this reason, external data is also often sought from other administrations.[4]

This requires the sharing of data by and with other bodies, especially administrative data from and with other authorities. However, access to these data is not always an easy task. Despite the development of databases across the enforcement bodies responsible for tax, social security and labour law compliance, there currently often remains a lack of a fully coordinated approach to data sharing.

[3] For instance, internal data used by the Financial Administration of the Republic of Slovenia are: data of tax payers, VAT returns data, social contributions data, corporate income tax data, personal income tax data, cash registers data, import and export data, etc.

[4] For instance, the sources of data used by the Lithuanian state labour inspectorate are: SLI information system (DSS IS), SLI e-service system for employers (EPDS), register of legal entities (JAR), Interinstitutional data system (TDS), Board of State Social Security Fund (VSDFV), Department of Standardization, State Tax Inspectorate.

Sharing data across government departments enables all individual enforcement authorities to improve the data at their disposal. If this is to be shared electronically, it requires a cross-government information technology infrastructure that actively supports the implementation of standardised processes. The design and the architecture of the information technology infrastructure needs to reflect the operational needs of the enforcement bodies and be capable of being updated without prohibitively high effort and cost.

When seeking to share data, bilateral agreements or Memoranda of Understanding (MoUs) on data sharing between authorities are often developed. When this is the case, the following questions need to be asked:

- What do we have?
- What do we need?
- What do the other authorities need?
- What can we share?
- Are the other authorities prepared to share data?
- Are we allowed to share these data?

When sharing data, attention needs to be paid to the legislation to protect personal data and safeguard privacy and whether this limits data exchange or the use of the data. Therefore, the following questions need to be asked regarding legal barriers to data access:

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

Almost all economies have legislation in place that protect personal data and safeguards privacy. In Finland, for example, there is a data protection law in general but there are also special laws in different fields which govern the rights of authorities to obtain information and process it. In the UK, there are various pieces of legislation, primarily covered by the Data Protection Act. Within the taxation sphere, the UK is covered by Commissioners for Revenue and Customs Act 2005. This means that data cannot be shared unless this is covered by specific Memorandum of Understanding (MOU) for very specific compliance purposes. Several laws in Norway[5] have rules about limiting the use of reported data. In Slovenia there are certain limitations for exchange of personal data according to the Personal Data Protection Act, but there are also special provisions concerning data, relevant for detecting undeclared work, which enable administrations access to it.

In all economies where legislation exists that protects personal data and safeguards privacy, enforcement bodies should be aware of these rules before starting the process of sharing data. A key question is usually whether the access to personal data is proportional to the enforcement body's objectives. Furthermore, persons should have the right to have access to the information on where and how the data are processed and they should be able to react on it. Finally, steps should be taken to ensure the personal data used is protected so that its misuse is avoided.

Therefore, **data protection** and **data security** are key issues that need to be built into any data gathering and sharing system from the very start. **Privacy by design**, or its variation

[5] "Skattebetalingsloven", "Ligningsloven", "Skatteforvaltningsloven" and "Personvernregisterloven".

"**data protection by design**", is a multifaceted concept, involving various technological and organisational components, which implement privacy and data protection principles in systems and services. "Privacy by design" should guarantee an effective protection of privacy and data. The European Union Agency for Network and Information Security (ENISA) has published two reports, namely a report on Privacy and Data Protection by Design (2014) and a Report on Privacy by Design in Big Data (2015).[6] In its 2015 report, ENISA presented eight privacy by design strategies, both data oriented and process oriented, aimed at preserving certain privacy goals (see Table 1).

Table 1 Privacy by design strategies

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

Source: ENISA (2015)

For enforcement authorities, it is essential to implement a coherent approach to data privacy protection over the whole lifecycle of the analytics (data collection, data storage, data analysis, data usage). An important privacy principle in the data collection phase is that of "data minimisation". The data needs should be precisely defined (i.e., what personal data are 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. It is essential, therefore, that there is a **legal basis to exchange data** between administrations.

A second barrier to the exchange of data relates to **political will** and **trust**. Data sharing requires political will and trust between the different parties involved as well as a clear idea of what data needs to be shared. Often, authorities do not know each other very well, and often do not know what kind of data other authorities have or what they might want from each other. Establishing a willingness to exchange data requires an understanding of the added value for all to create a sense of community around data sharing instead of a forced integration. A first step in this process could be the design of a "Memorandum of Understanding" or bilateral agreement.

Looking at the current situation, despite the development of databases in the enforcement authorities responsible for tax, social security and labour law compliance, there is presently a

[6] See https://www.enisa.europa.eu/topics/data-protection/privacy-by-design

lack of a fully coordination approach to data sharing. Many enforcement authorities have difficulties in accessing data from other enforcement authorities. Sometimes they have access. Sometimes they do not. This can be for many reasons. For example, one enforcement authority may not share data with another or not share it electronically. However, even if access to data from other authorities is accessible, it can be the case that their databases are not inter-operable with the databases of the receiving authority.

For this reason, some economies might decide to adopt a fully coordinated cross-government approach with one central unit collating the various datasets and acting as a warehouse for all relevant authorities. Box 5 provides an example from Finland of how the traditional problems with sharing data have been overcome by creating one central unit that provides a data mining and analysis service for all government ministries involved in tackling undeclared work.

Box 5: Grey Economy Information Unit (GEIU), Finland

Aim: To join up the previously fragmented function of data analysis and transcend the need for data sharing by establishing one central unit to produce and share information on undeclared work to all interested public bodies.

Description: The Grey Economy Information Unit (GEIU) was established in 2011. It produces and shares information on undeclared work. Through its service, the unit provides a single point of access for permitted public authorities to gain information on organisations and individuals within organisations suspected of engaging in undeclared work. The GEIU is responsible for gathering and disseminating information on the grey economy. The authorities permitted to request compliance investigations are defined in the enacting legislation, as are the purposes for which a compliance report can be prepared. The GEIU produces three types of report:

- *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 geographical area). Reports should be interesting for decision makers.
- *Grey economy reports*: Some 10 to 15 reports every year mainly interesting for policy makers.

The service is fully automated with a full web interface which means, for the most part, that compliance reports are produced automatically and delivered to the information system of the requesting authority. The GEIU does not charge for the preparation of reports. It is also entitled to obtain, free of charge, the information it needs to prepare the reports. The GEIU also operates a public website on the Grey Economy and Economic Crime for public agencies, companies and individuals providing an overall picture, and topical information on the phenomena of the shadow economy. There are 24 employees.

Evaluation: The GEIU has produced 2 million compliance reports since it was established in

2011. From receipt of a request for a compliance report, it takes the GEIU about one day to complete. Currently there are 21 authorities with permission to request compliance reports from the GEIU. The fully automated online web interface which allows compliance reports to be delivered automatically to the requesting authority helps improve efficiency, giving those authorities more time to tackle the grey economy.

The public website content is produced in collaboration with 21 authorities and ministries involved and is published in three languages including Finnish, English and Swedish. This provides statistical information on the impacts of action taken against undeclared work, as well as providing companies and citizens with information on how to act or protect themselves against such harm.

Source: https://ec.europa.eu/social/BlobServlet?docId=18511&langId=en

However, the GEIU in Finland does not store data in a centralised data warehouse. The data are extracted directly from the information systems of the various authorities. This means that for every new audit, a request to the relevant providers (i.e., authorities) is sent. So, the slowest authority determines the response time to a request. In general, nevertheless, a request that is sent in the morning will be answered by the evening of the same day. This creation of a single unit, therefore, is one response to the problems involved in data sharing.

Another interesting example of a solution to the problem of authorities gaining access to information from other institutions can be found in Belgium. In this economy, the Crossroads Bank for Social Security (CBSS) has been developed. The CBSS is not a database but an application that grants or denies access to databases of different administrative authorities. As a result, in most cases all contact between administrative authorities regarding the exchange of personal data will be channeled through the CBSS. The administrative authorities can use the data they have been granted access to directly, without the intervention of the body to whom the data belongs. Box 6 provides a summary of this initiative in Belgium.

Box 6. The Crossroads Bank for Social Security (CBSS), Belgium

Aim: The aim of the Crossroads Bank for Social Security (CBSS) is to provide a gateway improve service delivery to the socially insured people and the companies involved. Social benefits are automatically granted without citizens or their employers having to make declarations anymore and the administrative burden for citizens and companies has been drastically reduced.

Description: The Belgian social security consists on the one hand of 3 insurance systems (workers, self-employed workers and civil servants), that cover maximum 7 social risks (incapacity for work, industrial accident, occupational disease, unemployment, old age, child care and holiday pay - the so-called branches of social security), and on the other hand of 4 assistance systems (subsidies for the handicapped, guaranteed family allowance, minimum income and income guarantee for the elderly), that grant people specific minimum services after checking their subsistence resources. In total about 3.000 institutions are responsible for the execution of the Belgian social security. More than 10.000.000 socially insured persons and 220.000 employers have very regular contacts with those institutions to assert their rights, to furnish information therefore or to pay contributions.

The Crossroads Bank for Social Security (CBSS) is a gateway for data from 14 social security institutions, and offers electronic services for citizens. The CBSS, despite its name, is not

itself a databank, it is a network for data flows from different institutions. Each institution holds its own data, are the authentic source of the data and there are conventions about the treatment of the data, agreed through regular meetings and continuous collaboration. The CBSS initiative started in the early 1990s and has been developing ever since. The legislative changes needed for the CBSS to be created, included the legal translation of a common vision on information management and on information security and privacy protection and the obligation for each institution participating in the CBSS to use unique identification keys for their data.

Evaluation: 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 advantages of creating this system include efficiency gains. The Belgian Planning Bureau calculated that the information exchange processes implied an annual saving of €1.7 billion per year. That was an enormous stimulation in continuing the process. There were also gains in speed and in effectiveness, as the CBSS made possible the provision of services of better quality as well as the provision of new types of services, such as personalised simulation environments and a push system of automated granting of subsidies.

In terms of success factors, the CBSS was able to gain support thanks to a clear long-term vision which was also combined with some quick wins. A key success factor in gaining support included the fact that a small team consisting of experienced civil servants, scientific experts and political advisors worked closely with the Federal Minister of Social Affairs. Critical success factors included a common vision on electronic service delivery, support by policy makers at the highest level, the trust of all stakeholders, and respect for the legal allocation of competences. This top-level political support and the gradual involvement of the general managers of all public social security institutions, the social partners, and the general managers of the private social security institutions was also significant. Finally, it was important to ensure that electronic service delivery included a multi-disciplinary approach including legal, ICT, communication, coaching, training and change management. Additional success factors included adaptability to an ever changing societal and legal environment, the availability of sufficient financial means and most importantly, a radical cultural change within government, from the hierarchy to the operators, who made the CBSS possible.

Source: www.ksz-bcss.fgov.be/nl/information-english

4. Data Analysis

Data analysis takes two forms:

- **Data matching** is the large-scale comparison of records or files collected or held for different purposes, with a view to identifying potential instances of undeclared work. With data matching, two or more sets of collected data are compared.
- **Data mining** is a set of automated techniques used to extract buried or previously unknown pieces of information from large databases. Correlations or patterns among dozens of fields in large relational databases are identified. Two approaches can be used. Firstly, deduction begins with an expected pattern/theory/hypothesis that is tested. Secondly, induction begins with the observations/data and seeks to find a pattern within them, and theories proposed resulting from the observations.

Most authorities focus their efforts on data matching but there is growing interest in data mining, especially in tax authorities.

Although some economies have a fully coordinated cross-government approach to data analysis, with a central unit providing a common data analysis function on detecting undeclared work to all relevant authorities (e.g., the Grey Economy Information Unit in Finland as described in Box 5), this is exceptional.

In most economies, data analysis is conducted at the level of the different administrative authorities. For this to be effective in detecting and preventing undeclared work, there is a need for:

- Up-to-date data to be available.
- The databases containing the data to be inter-operable.
- Specialised staff who can administer and produce intelligence using data mining and matching.[7]
- A well-functioning data analysis tool.
- Resources to be available to fund such analysis, which requires "proof of concept" in terms of the "returns on investment".
- Such analyses to be made easily available to inspectors to help them in the field detect and prevent undeclared work.

Each is here considered in turn.

Up-to-date data availability

To effectively detect and prevent undeclared work, it is necessary to have up-to-date data and for that data to be accurate, and well-structured, and electronically available. Poor quality out-of-date data can create more problems than it solves. In fact, if the data is not up-to-date, inaccurate or contains errors, the analysis is likely to be incorrect or even counterproductive. It is therefore crucial to devote time and resources to clean the data and ensure that up-to-date data is being used.

It is also necessary 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, and the exact definition used for each variable.

The databases containing the data need to be inter-operable.

If internal and external databases are to be merged, or even two internal databases, then it is necessary to ensure that this can be done electronically and that they are compatible. To do this, it is necessary to ensure that the definitions of each variable are clear.

The Crossroads Bank for Social Security (CBSS) in Belgium is part of the e-government strategy to stimulate and support the actors in the Belgian social sector to develop 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

[7] International Labour Organization (2013), *Labour Inspection and Undeclared Work in the EU*, ILO, Geneva.

and unique identification key and has an electronically readable identity card containing this identification number. The data shared by the agencies is inter-operable.

However, this inter-operability of data in Belgium required a lot of effort. Clarifying the terminology between agencies has been important. Belgium's Crossroads Bank started off with 120 definitions of "wages" which was eventually narrowed down to 12 definitions. A weakness of the Crossroads Bank for Social Security (CBSS) is that no data from the Belgian tax administration is included, largely due to data protection issues.[8]

Specialized staff who can administer and produce intelligence

In many economies, a major barrier to effective data analysis is the lack of specialist staff in inspectorates able to conduct such analyses. Decisions therefore need to be made by inspectorates about whether to divert resources away from inspections for example and towards developing a specialist data analysis team that can conduct data analysis to select risk-based targets for inspection and provide information on targets to be selected for notification letters and awareness raising campaigns.

As the work becomes more technical on data analysis and shifts beyond simple data matching to data mining, the need for specialist staff grows. This leads some enforcement bodies to employ in-house expertise. However, it is often difficult for enforcement bodies to find IT specialists. This has been sometimes solved by seeking external support of skilled people who have developed the commercial data-mining tool (i.e., external staff). In other cases, internal staff who have the requisite skills are re-deployed to focus more upon data analysis. Unless the specialist staff are employed to administer and produce intelligence, the data available is unlikely to be used to its full capacities due to the lack of data extrapolation and shortage of automated intelligence options.[9]

Unless specialist staff are employed, it might also be difficult to elaborate a set of parameters which can provide "alarms" for potential undeclared work. However, the building and testing of these models requires technical IT staff to create a custom-built data mining tool.

A well-functioning data analysis tool

The use of data analysis tools, both in a preventative and curative manner, should help to maximise the audit benefits and to minimise audit costs. Different tools and techniques can be used by enforcement bodies depending on what level of "knowledge" and "technology" is available. The key lessons when developing a data analysis tool are that:

- It is important to share experiences/good practices with other authorities. An enforcement authority should examine the data analysis tools used by other enforcement bodies.
- It is useful to start it as a pilot project (i.e., a small and manageable project) and if successful then move slowly forward.[10]
- The design should be clear because it is difficult to correct it afterwards.

Data mining has a wide number of applications and therefore, many data mining tools have been developed over decades. The aim here is to provide some information on the existing data-mining tools and some awareness about their features, advantages, and limitations.

[8] www.ksz-bcss.fgov.be/nl/information-english
[9] International Labour Organization (2013), *Labour Inspection and Undeclared Work in the EU*, ILO, Geneva.
[10] For instance, HRMC in the UK started *Connect* as a pilot project.

Many advanced tools for data mining are available either as open-source or commercial software. Table 2 provides a list of the most popular commercial and open-source data mining tools. Some of these data mining tools are offered free of charge by using an open-source licence. Their features are shown in Table 3. Furthermore, it is important to know the advantages and disadvantages of the available data-mining tools. Several articles are published which assess the popular open-source data-mining tools.[11] In Table 4, the advantages and limitations of some popular open-source tools are listed.

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

TOOLS	LINK
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)	www.r-projects.org
RAPIDMINER	www.rapidminer.com
WEKA	www.cs.waikato.ac.nz/ml/weka

Source: Mikut and Reischl (2011); Altahi Abdulrahman, Luna, Vallejo and Ventura (2017)

Table 3. Features of the popular open-source data mining tools

Tool	Type	Features
Rattle (R)	Statistical Computing	• Data exploration, Outlier detection, Clustering, Text Mining, Time Series Analysis, Social Network Analysis, Parallel Computing, Graphics, Visualisation of geo spatial data, Web Application Big data; • Data and error handling, requires array language, poor mining.
ORANGE	Machine learning, Data mining, Data	• Visual Programming, Visualisation; • Interaction and Data analytics;

[11] See for instance, 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.; Mikut, R. and Reischl, M. (2011), "Data mining tools", *WIREs Data Mining Knowl Discov*, Vol. 1.; Rangra, K. and Bansal, K. L. (2014), "Comparative Study of Data Mining Tools", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 4, No. 6.

	visualisation	• Large toolbox, Scripting interface; • Extendable documentation.
RAPID MINER	Statistical analysis, data mining, predictive analytics	• More than 20 new functions for analysis and data handling, including multiple new aggregation functions; • File operators to operate directly from RapidMiner; • A macro viewer that shows macros and their values in real time during process execution.
KNIME	Enterprise Reporting, Business Intelligence, Data mining	• Scalability, high extensibility; • Sophisticated data handling, intelligent automatic caching of data, Data visualisation; • Import/export of workflows.
WEKA	Machine Learning	• Forty nine data pre-processing tools, seventy six classification/regression algorithms, eight clustering algorithms, fifteen attribute/subset evaluators, ten search algorithms for feature selection; • Three algorithms for finding association rules; • Three graphical user interfaces; • Poor documentation.
KEEL	Machine Learning	• Classification Discovery, Cluster Discovery, Regression Discovery, Association Discovery, Data Visualisation, a user-friendly graphical interface, evolutionary learning.

Source: Rangra and Bansal (2014)

Table 4. Advantages and limitation of the popular open-source data mining tools

Tool	Advantages	Limitations
Rattle (R)	Purely statistical	Less specialised for data mining, requires knowledge of array language
ORANGE	Better debugger, shortest scripts, poor statistics	Big installation, limited reporting capabilities
RAPID MINER	Visualisation, Statistical, Attribute Selection, Outlier detection, parameter optimisation	Requires prominent knowledge of database handling
KNIME	Molecular analysis, Mass spectrometry	Limited error measurements, no wrapper methods for descriptor selection, poor parameter optimisation
WEKA	Ease of use, can be extended in RM	Poor documentation, weak classical statistics, poor parameter optimisation
KEEL	Evolutionary algorithms, fuzzy systems	Limited algorithms

Source: Rangra and Bansal (2014)

The investment in technologies by enforcement bodies ranges considerably from the use of free open-source software to specially designed programmes and systems. Some enforcement bodies make use of commercial software[12] while others make use of open-source software[13].

Turning to the **analysis process**, either data matching or data mining can be used. In the case of data matching, two or more datasets are compared. In this way, for example, enforcement bodies can verify, among other practices, for example, if a person is claiming social benefits they are not entitled to whilst they are also in paid work.

Examining data mining, the primary challenge is to build a model that accurately predicts whether a business or worker is compliant. It supposes the identification of patterns in a set of data using an algorithm. The development of such a data analysis tool requires feedback from inspectors on whether risks are captured/detected using the data-mining techniques. The intention is to identify variables (i.e., determinants of non-compliance) that can be used in data mining to successfully predict non-compliance.

Examples of data-mining techniques include:

- "**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, because it does not require a hierarchy in the variables it is more powerful"; and
- "**Clustering**": "this is another segmentation technique that allows for the simultaneous analysis of several possible explanatory variables during the segmentation process" (Khwaja et al, 2011).

Based on these techniques, the objective is to detect **outliers**. An example is here provided from the UK of data analysis using the Connect system. Data analysis was undertaken of the hotel sector in the UK where turnover to credit card transaction ratios were used to identify outlier hotels where turnover to total credit card transactions deviates from the norm. This **dynamic benchmarking** of the hotel sector occurred on an individual city level and for types of accommodation provider (e.g., small hotels), since credit card to turnover ratios are higher in cities and larger hotels than in smaller hotels and in smaller towns. A too high share of payments by credit cards in total declared turnover, in comparison to the benchmark determined for the specific area, resulted in an alarm/red flag. This data mining initiative required that third party data were available from banks on credit card transactions, which were compared with reported turnover on tax returns, to identify "outlier" hotels who deviate from the norm. This could be similarly applied in the restaurant sector to identify "outlier" restaurant businesses and many other tourist industries (e.g., tour guides, tour operators) who appear to have turnover to total credit card transactions that deviate from the norm.

Hence, data mining is primarily about detecting anomalies to the norm. It determines the difference between normal and suspicious/new behaviour, identifies anomalies, categorises and prioritises risks for further investigation. **Machine learning** provides methods, techniques and tools which help to learn automatically and to make accurate predictions based on past observations. For instance, the Federal Public Service (FPS) Social Security in Belgium uses

[12] Slovenia: SAP and QlikView; Norway: Oracle Business Intelligence; UK: SAS enterprise guide.
[13] Belgium: R.

machine learning to detect differences between employers. Based on the outcome of previous audits, the characteristics of employers who have committed undeclared work are compared with the characteristics of employers who made no infringements. By the application of machine learning, employers are afterwards ranked by their vulnerability to fraud. The reason why the employer is considered vulnerable to fraud will also be clear for the inspector (i.e., red flag for a certain alarm).

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, the FPS Social Security Belgium has also recently started network analytics to rank and profile cases. This application of **network analytics** follows its earlier successful application to tackle VAT carousel fraud The Swedish tax administration also reported that they not only look at the company but also at its network.

The HMRC (Her Majesty's Revenue and Customs) in the UK applies predictive analytics to identify high risk VAT traders. Data is taken from the VAT population (returns, debt information, trader characteristics and audit visit outcomes). A behaviour model is used to identify behaviour based on logistic regression analysis.[14]

Box 7 provides a good practice example from Belgium of a data analytics tool, namely MiningWatch.

Box 7. MiningWatch data analytics tool, Belgium

MiningWatch is a data mining tool which uses predictive modelling to define fraud risks in three different sectors: construction, cleaning, and the hotel and catering sector. MiningWatch has over 60 predictive automated models that run. Based on the MiningWatch predictive models, search results rank companies according to their risk level: red (high), orange (elevated), green (medium), and blue (low).

MiningWatch can calculate a score for an employer based on data mining and prepares a score card with five variables, listed below in order of importance:

1. Strong personnel turnover
2. Few recent declarations in the DIMONA[15] system
3. Low business turnover
4. Tax variables such as VAT debts
5. Not declaring client listings.

This analytical tool supports inspectors to choose and target their inspections based on the predictive risk modelling of fraud (including undeclared work, abuse of part-time working schemes, and bogus self-employment).

Inspectors have some freedom to choose investigations based on their own initiative from the list of risky businesses produced. Since the beginning of 2015 MiningWatch is available for all inspectors of the FPS Social Security in Belgium. In 2015, 25% of investigations resulted from the use of Mining Watch by inspectors.

[14] 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.

[15] The DIMONA system is an electronic system all employers are required to use to register a new employee with the National Office for Social Security

The models are dynamic, they are monitored regularly and closely, so that adaptations are made as and when necessary. If a model drops below an acceptable level of predictions, then new models are re-designed. The models are complemented by Network analysis. If a company that has many links to companies with confirmed cases of infringements, then the ranking of the company increases, further prioritising it as a target for inspection.

Source: http://ec.europa.eu/social/BlobServ- let?docId=18372&langId=en

The effectiveness of a data mining tool can be assessed by first 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 5 describes these outcomes.

Table 5. Outcome of the prediction

		Prediction	
		Positive	**Negative**
Outcome	**Positive**	**TP**	**FN**
	Negative	**FP**	**TN**

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

The accuracy is measured by (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 to avoid 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.

The efficiency or positive predictive value is 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.

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

Return on investment from the use of data analysis tools

An intervention strategy of this nature (i.e., investing in the use of data analysis tools and data collection) involves high costs, both in terms of the costs of acquiring data, the costs of staff to conduct data analysis and the investment in technology. It is therefore important to assess the "return on investment".

Return on investment (ROI) should be more clearly measured firstly to help enforcement authorities understand the impact of more advanced data usage on the outcome of inspections. Secondly, 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.

Some enforcement authorities have done this in small steps. They have firstly shown how the use of existing databases to select workplace inspections result in a higher rate of detections of undeclared work than conducting "random" inspections, and then calculated for example:

- the increase in tax or social insurance revenue
- the additional revenue generated from penalties, and/or
- the number of employment contracts converted into declared contracts.

This is then compared against the costs of the data analysis to provide a **revenue-to-cost ratio** (e.g., €5 extra revenue generated for every €1 invested).

However, there appears to be little publicly available information on the revenue-to-cost ratio of data mining initiatives. Given the large investments made to develop data analysis tools, this is perhaps surprising.

This is a hindrance to enforcement authorities who wish to conduct a cost-benefit analysis (CBA) of the likely impact of developing databases and data analysis tools. This is used to evaluate the total expected cost of the implementation of a data analysis tool (i.e. in terms of additional human or financial resources) compared with the total expected benefits (direct benefits - results selected by data analysis tool – and indirect benefits - harmonisation of the data, administrative gains, behavioural change) in order to determine whether the proposed implementation is worthwhile. Therefore, the expected return of investment could be estimated. The cost of persons engaged in monitoring and inspections 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. One of the few revenue-to-cost figures in the public domain comes from the UK (see Box 8).

Box 8. Revenue-to-cost ratios of developing data analysis tools: the case of the Connect tool of Her Majesty's Revenue and Customs (HMRC), United Kingdom

Her Majesty's Revenue and Customs (HMRC) in the UK launched Connect, one of its main analytical tools, in 2010. It ingests over three billion data items and looks towards matching them and producing connected entities. In total, it brings together 40 data sets with 22 billion

lines of data and 600 million documents.

There are some 250 data analysts and 4,000 users of the Connect tool, and there have been 13 million searches. 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 gain a comprehensive picture of its taxpayers and the data (HMRC and third-party data) relating to them.

Recently, HMRC has been using the data within Connect to create maps of undeclared work, overlaying the 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.

Evaluating the revenue-to-cost ration, HMRC report that the additional tax identified by *Connect* is far greater than the cost of the data mining tool. HMRC have spent £90 million on developing this system, but until now, it has helped secure an additional £3 billion in tax revenues. More than 7 out of 10 enquiry case selections are generated by *Connect.*

Source: http://ec.europa.eu/social/BlobServ- let?docId=18525&langId=en

The FPS Social Security in Belgium reported that data mining tools result in undeclared work being twice as likely to be discovered compared to random selected audits.

Communication between the back-office and front-office

Data mining does **not replace the need for inspectors**. However, it can help target resources and lead to efficiency gains. Involving inspectors at all stages of the data gathering and mining process is important to: gain their trust; ensure that systems are accessible to them; and enhance the effectiveness of the data mining systems.

Good communications between the back-office (the data analysis experts) and the front-office (inspectors), and vice versa, are essential if data analysis is to be effective. Many inspectors may perceive that data analysis will de-professionalise their role by taking away their ability to make qualitative judgements on what risky businesses to inspect and to use their local experiential knowledge gained over many decades of working as an inspector. For this reason, great care needs to be taken when introducing data-led approaches to targeting inspections.

In some economies, this has been done by asserting that in the first instance, a certain percentage of inspections should be based on data analysis and the remaining percentage left to the judgement of the inspector. Another approach is to provide the inspector with for example 200 risky businesses from the data analysis and ask them to select 100 of these for inspection.

There is also a need for inspectors to be trained to use the databases to select risky businesses for inspection and when in the field to check workplaces against the databases (e.g., employment register). This can often be done using a peer-to-peer approach with a small number trained in each office who then pass on this learning to their colleagues.

There is also a need for a bottom-up approach to be built into the design of the data analysis function with inspectors feeding back up to the data analysts the results of their inspections. There is in addition a need for inspectors to help identify the “red flags” or “alarms” that data

analysts use to identify potentially risky businesses. This allows the predictive model to be fine-tuned. Box 9 describes how this has been achieved in Belgium.

Box 9. Involving inspectors in developing data analysis tools: the case of MiningWatch, Belgium

To improve the effectiveness of the MiningWatch data analysis tool, inspectors have been engaged in developing the variables used. A distinction is made between inspectors who are "power users" who will make good use of the wealth of data available, "casual users" who will use it but not further it, and "anti-users" who are non-believers in data mining. The number of anti-users has been declining over the 15 years since its introduction, largely due to the positive results arising from the use of MiningWatch.

The IT team in MiningWatch is complemented by a network of 50 expert investigators. The characteristics of each predictive model are presented to expert inspectors and their feedback is requested. This process is repeated until the characteristics from a risk profile are accepted by inspectors. The model is then monitored to see if the profile stays predictive.

Indeed, feedback takes three forms. Feedback is collected in the short-term immediately after inspection, in the medium-term such as the confirmation of fraud type X or Y when an inspection case concludes, and in the long-term after combining cases over time or across networks such as verdicts of legal cases, the proportion of the contributions actually collected etc.

For inspectors, MiningWatch is a tool to help them select inspections. Practice has shown that inspectors that did not use MiningWatch had worse results from their inspections so its adoption has become ever more widespread since it helps inspectors achieve their targets in terms of numbers of investigations, infringements or income and thus use it more.

Source: https://ec.europa.eu/social/BlobServlet?docId=18372&langId=en

These databases also need to be easily accessible for inspectors in the field to help them detect and prevent undeclared work. In Estonia, for example, new IT tools have been introduced for the ETCB tax auditors. For example, mobile offices have been introduced, which tax officials can use during on-site inspections to perform and record all operations. Their new mobile working environment means that they can access all tax and employment-related data they need for inspections on-the-spot. A mobile app has also been developed, which enables officials carrying out inspections to use smartphones to gain direct access to the Employment Register, including a photo of the employee, alongside his/her employment data.

5. Conclusions

The key findings on data collection, sharing and analysis are:

Data collection

- **Data collection** is the process of gathering data from internal and external sources.
- These data bases need to:
  - collect the data/fields/variables required to detect undeclared work;
  - have access to these data on a real-time/up-to-date basis;

- be available to all relevant levels of the organisation who need these data, including inspectors.

- The major challenges preventing the development of data bases are that enforcement authorities:
  - need the financial resources available to develop these databases;
  - need the technical skills available to develop these databases;
  - need appropriate legislation on personal data and privacy safeguards that enables access to such data for enforcement authorities, and
  - need the "political support" to develop these databases.
- The **quality of the data** being used is key and as more data becomes available, a key question is how to get accurate data. Either initiatives have to start with good data or resources are needed to ensure that the data being used is clean, accurate and reliable.
- The practical approach to data collection uses the available data that has been collected for other purposes to detect and prevent undeclared work. A **strategic approach to data collection** starts by asking the following question: "What data/information does the enforcement authority need to identify instances of undeclared work?" This question could be usefully addressed by authorities.

Data sharing

- **Data sharing** is the process of making data available to other users.
- **Data protection** and **data security** are key issues that need to be built into any data gathering and sharing system from the very start. **Privacy by design**, or its variation "data protection by design", is a multifaceted concept, involving various technological and organisational components, which implement privacy and data protection principles in systems and services**.**
- A second barrier to the exchange of data relates to **political will** and **trust**. Data sharing requires political will and trust between the different parties involved as well as a clear idea of what data needs to be shared.
- **Interoperability of the data shared by agencies is also crucial**. Clarifying definitions of variables and terminology between agencies is another important step.

Data analysis

- **Data analysis** involves either **data matching** (i.e., the large-scale comparison of records or files collected or held for different purposes, with a view to matching two or more sets of collected data) or **data mining** (i.e., a set of automated techniques used to extract buried or previously unknown pieces of information on potential instances of undeclared work from large databases, records or files collected for other purposes).
- For this to be effective in detecting and preventing undeclared work, there is a need for:
  - **Up-to-date data** to be available.
  - The databases containing the data to be **inter-operable**.
  - The appointment of **specialised staff** who can administer and produce intelligence using data mining and matching.

- A well-functioning **data analysis tool**.
- Resources to be available to fund such analysis, which requires "proof of concept" in terms of the "**returns on investment**".
- Such analyses to be made **easily available to inspectors** to help them in the field detect and prevent undeclared work. Data mining does not replace the need for inspectors, but it can help target resources and lead to efficiency gains. Involving inspectors at all stages of the data gathering and mining process is important to gain their trust.

- There appears to be little information on the **critical assessment of data mining initiatives**. It is important to define useful indicators to measure the performance and effectiveness of tools and to ascertain the return on investment.

Next steps:

- **Mutual learning** is useful for members of the Western Balkan Network Tackling Undeclared Work not only for those who are at an early starting point, but also for those authorities who are more advanced in the process who can learn from each other. There is no need to "reinvent the wheel".
- This could be taken forward, for example, by an authority hosting a **Mutual Assistance Project (MAP)** where others can be invited to the host authority to discuss how data collection, sharing and analysis could be improved in the host authority.
- There might be also limited opportunity for a **staff exchange** to take place where staff members from one enforcement authority can visit another authority perhaps more advanced to learn about how they have developed their data collection, sharing and analysis functions.

References

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.

De Wispelaere, F. and Pacolet, J. (2017a*). Data Mining for More Efficient Enforcement: a learning resource*. European Platform Tackling Undeclared Work, Brussels. https://ec.europa.eu/social/BlobServlet?docId=18746&langId=en

De Wispelaere, F. and Pacolet, J. (2017b*). Data Mining for More Efficient Enforcement: a toolkit*. European Platform Tackling Undeclared Work, Brussels. https://ec.europa.eu/social/BlobServlet?docId=18826&langId=en

European Commisison DG TAXUD (2010). *Compliance Risk Management Guide for tax administrations*, European Commission, Brussels.

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

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

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

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

International Labour Organization (2013). *Labour Inspection and Undeclared Work in the EU*, ILO, Geneva.

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

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

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

OECD (2004), *Compliance Risk Management: Managing and Improving Tax Compliance*, OECD, Paris

Rangra, K. and Bansal, K.L. (2014). "Comparative Study of Data Mining Tools", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 4, No. 6.

Williams, C.C. and Puts, E. (2017). *2017 Platform Survey Report: organisational characteristics of enforcement bodies, measures adopted to tackle undeclared work, and the use of databases and digital tools*, European Commission, Brussels. https://ec.europa.eu/social/BlobServlet?docId=18747&langId=en

List of Abbreviations

CBSS – Crossroads Bank for Social Security

DG TAXUD – European Commission's Directorate-General for Taxation and Customs Union

ENISA – European Union Agency for Network and Information Security

ETCB – Estonian Tax and Customs Board

FN – False negative

FP – False positive

FPS – Federal Public Services

GEIU – Grey Economy Information Unit

HMRC – Her Majesty's Revenue and Customs

ICT – Information and Communication Technology,

ILO – International Labour Organisation

MAP – Mutual Assistance Project

MoU – Memorandum of Understanding

TN – True negative

TP – True positive