NEBOT Paper 6

Linked Beneficial Ownership Data? Challenges and Opportunities

Network of Experts on Beneficial Ownership Transparency, NEBOT

Civil Society Advancing
Beneficial Ownership Transparency

Linked Beneficial Ownership data? Challenges and opportunities

Authors: Mihaly Fazekas, Viktoriia Poltoratskaia

Contributors: Marco Vianello, Christoph Trautvetter

Reviewers: Louise Russell-Prywata

European Commission, Directorate-General for Financial Stability, Financial Services and Capital Markets Union. Fazekas, M. & Poltoratskaia, V. Linked Beneficial Ownership data? Challenges and opportunities: Network of Experts on Beneficial Ownership Transparency Policy Paper 6, Publications Office of the European Union, 2023.

Civil Society Advancing Beneficial Ownership Transparency (CSABOT) is a project that implements the Preparatory Action – Capacity Building Programmatic Development and Communication in the Context of the Fight against Money Laundering and Financial Crimes. This project is performed by Transparency International Secretariat (TI-S), together with Tax Justice Network (TJN), Transcrime – Università Cattolica del Sacro Cuore (Transcrime – UCSC) and the Government Transparency Institute (GTI), under a contract with the European Union represented by the European Commission. The opinions expressed are those of the authors and do not necessarily represent the views of all NEBOT members.

Table of contents

Abstract 4
Part I. Policy goals 5
Part II. Types of linked data and linking challenges 8
Challenges of linking data 9
Part III. Case studies 11
Case No. 1: Politically-connected firms and public procurement data: Case of Bulgaria 11
Case summary 11
Goals and datasets involved 11
Data-linking activities and challenges 12
Uses of linked data: investigations, policy analysis 12
Lessons learned 14
Case No. 2: Beneficial ownership of German real estate 15
Case summary 15
Country background 15
Data-linking activities and challenges 16
Uses of linked data 17
Lessons learned 17
Case No. 3: EBOCS (European Business Ownership and Control Structures) project 18
Case summary 18
Goals and datasets involved 18
Data-linking activities and challenges 19
Uses of linked data: investigations, policy analysis 19
Lessons learned 20
Part IV. Technical recommendations 22
Conclusions 25
References 26

Abstract

One of the most widely-accepted policy goals of beneficial ownership registers is to help tackle money laundering and financial crime. In order to further this goal, this paper first identifies types of data needed to track dark money and assesses how these can be combined with beneficial ownership information. Second, it offers practical examples where data-linking has been done, offering insights into both its benefits and technical challenges. Third, reflecting on the lessons from case studies, we also provide technical recommendations on how data-linking can be done better, and how it can be made easier. The research is based on a combination of literature review and primary data analysis of selected case studies.

Part I. Policy goals

Corruption involving illicit financial flows and money laundering can affect democratic institutions and actors, as well as generally undermine the integrity of the political system. When political institutions are vulnerable to capture, fundamental principles of good governance and political accountability can be compromised. Therefore, fighting corruption and illicit financial flows is not only a goal in itself, but is also desirable for to its larger impact on democratic mechanisms.

The creation of publicly available beneficial ownership (BO) registers is an essential step for fighting corruption, tracing dark money flowing into EU political systems and hence often threats to democracy and good government. To facilitate the adoption of BO registers by national governments, in 2015 the fourth EU anti-money laundering directive (AMLD) was launched. It gave EU member states a two-year time frame to transpose the Directive into national legislation. This Directive was largely built on recommendations of the Financial Action Task Force ("FATF") and it introduced new approaches to risk assessment and data protection standards, as well as precise definitions of politically exposed persons and beneficial ownership. For example, the already specified risk-based approach in the third EU anti-money laundering Directive was further enhanced by limiting exemptions for lower risk entity types. Additionally, each legal entity would be individually assessed through the use of specific risk variables, establishing a system of evidence-based control over money laundering and terrorist finance.

BO registers can provide a wide variety of possibilities for tracing illicit financial flows, as well as increase the effectiveness of investigations. For instance, having direct and open access to BO registers can enable proactive investigations by government agencies. Typically, prior to BO data publication, investigators needed to file complex and lengthy data requests, while publicly-available registers enable them to proceed without additional bureaucratic procedures and avoid the need to have pre-established evidence in order to access the data. Moreover, open registers resolve potential legal and technical obstacles to data-sharing between government departments. When it comes to the private sector, publicly-available BO registers can help companies to better assess the risks of their business relationships such as using a certain supplier or buying a particular company. BO registers offer an independent and trusted data source to conduct know-your-client risk assessments. Additionally, all

companies have the same access to BO data at a low to minimal cost, lowering the costs of doing due diligence considerably, benefiting smaller companies in particular which have fewer resources for such checks.

Improving risk assessments with the help of open BO registers is also helpful for risk prevention. They allow policy makers to implement the necessary mitigation measures and improve the regulatory environment. By revealing the potential risk factors in advance, it is possible to save time and efforts on anti-corruption interventions and prevent wrongdoing in the first place. Moreover, public BO registers provide access to civil society and journalists to conduct their own independent investigations and therefore also assist in improving government agencies' investigations.

However, the data in BO registers needs to be complemented by other datasets to realise the full potential of the above benefits. BO data by itself cannot provide much valuable information - it merely establishes the links between entities (individuals, companies, etc.). Only when BO data is linked to other datasets containing information on potentially corrupt transactions (e.g. government contracts) or the results of corrupt deals (e.g. real estate) can corruption risks be better assessed. Therefore, the goal of this paper is to provide insights to potential benefits and challenges from linking BO registers to other datasets (i.e. asset declaration, public procurement, PEP data, etc.), and the possibilities that data-linking opens up for government agencies, NGOs and civil society.

Currently, there are very few BO registers in open access and in standardised format which can be used by the general public or civil society and academics. Therefore, one of the most common substitutes is comprehensive company ownership data provided by private sector data aggregators such as Bureau van Dijk (BvD). Such companies typically provide access to their datasets for a fee, often very expensive, especially when the user needs to gain access to the full database rather than individual records. BvD offers one of the most comprehensive ownership datasets covering around 400 million companies all over the world, including the information on their ownership and various financial indicators. Unlike BO registers, companies like BvD use the information provided by official company registries, therefore the data is validated differently than in BO registers and the quality of data depends on the country-specific regulations. Thus, while company ownership registry data can be used as a good substitute for BO data, the information is limited by the countries that it covers, as well as the quality of data in these countries.

In the following sections, we present the types of data which can be linked to BO

data, as well as potential challenges associated with data-linking. Then we outline three in-depth case studies presenting the actual practices of data-linking with procurement data, real estate data and business register data. We conclude by offering technical recommendations to support actors in their attempts to use open data for preventing corruption and fraudulent behaviour.

Part II. Types of linked data and linking challenges

First, the value of BO data is greatly enhanced if it is directly linked to other company information. Ideally BO data is published as part of already-available company information such as registry attributes (date of incorporation, location of headquarters, etc.), financial data (turnover, number of employees, etc.), and management information (names of chief officers).

Second, measurement of corruption, money laundering and terrorist finance risks typically requires transactional data which describe the exchanges during which money is moved to the benefit of malevolent actors. For example, BO data linked to government contracts allows for tracking sources of corrupt income for a company (supplier) and hence the BOs behind it. Public procurement data enables tracking if the legal entity (supplier) has signs of fraudulent or corrupt behaviour, such as benefitting from tailored tendering terms, e.g. short time periods for submission such that it cripples competition, or a generally low level of competition (e.g. a single bid submitted on a competitive market). Moreover, if BO data is linked not only to procurement data but also to data on political office holders (see below), it is possible to trace personal ties between buyers or suppliers, hence revealing conflicts of interest.

Third, data on assets such as real estate holdings can be used to further enhance the analytical value of BO data. Where the real ownership of an asset class, say real estate in a particular city, is of interest, BO data offers the crucial link to individuals ultimately owning properties. The value of such data-linking is revealed when certain individuals or groups of individuals are targeted by policy, e.g. by sanctions or taxes. A particular high-value case of such data-linking is when BO data is linked with politicians' and bureaucrats' asset declarations. Asset declaration data is in itself a great tool to trace corruption as it reveals conflicts of interest and points at unjustified assets. Linking such data to BO data can help verify the content of asset declarations submitted by politicians and can also reveal links between individuals, for example business associates of politicians, indicating conflicts of interest.

Fourth, data on the individuals themselves such as official political positions the person holds has the potential for greatly enhancing the usefulness of BO data. For

example, data on politically exposed persons can support the tracing of dark money and corrupt money flows. Presence of politically-exposed persons (PEPs) in the chain of companies' ownership is considered to be a high risk in itself, requiring further investigation. By linking PEP data to BO registers, it is possible to identify such people in the ownership structure. In most cases, PEPs will try to avoid public scrutiny and would rather create a long chain of companies through which it is difficult to identify the full list of owners and beneficiaries. Therefore, matching the two datasets can help to easier establish the network structure.

Challenges of linking data

In all these cases, linked data can help to reveal corruption risks related to a company or group of companies. However, data-linking can pose lots of technical challenges. Differences in units of analysis, time coverage and data accuracy can influence the results. However, most of these issues have potential solutions and require multiple steps of documenting and analysing datasets prior to matching.

The first step is to map all the datasets, listing the full scope of variables and the unit of observation for each. This is a necessary step to resolve two potential issues: duplicated variables (or interconnected ones) and multilevel observations. For example, the most common issue with matching BO registers to other datasets can be that one dataset has individual-level information (e.g. politically-exposed persons), whereas the other has company-level information. An important step here is to identify whether there are any variables in both types of datasets that can serve as unique identifiers and help in matching. For instance, if the individual-level dataset has a variable on the company owned or related to the individual, the rows can be collapsed and aggregated to the company level. Alternatively, if there is a possibility to match individual-level information to company IDs, the dataset previously containing information on the organisational level can be complemented with information on individuals owning the company and thereafter matched to the individual-level data.

The need for unique identifiers is another issue to solve. The need for IDs which are unique and not duplicated in at least one of the datasets is a necessary requirement to avoid thousands of duplicates after merging. In case of multiple repeated IDs in the datasets, each ID will be filled with repeated information from the same ID coming from a different dataset; therefore, if there are three identical IDs in one dataset and two in the other, the final dataset will have six rows with the

same IDs and repeated information. Having at least one master dataset with unique identifiers will solve this problem by matching many to one, and therefore no de-duplication will be needed. There are a few ways to solve this issue and get at least one master dataset for matching. The first solution is to collapse the rows and get aggregated information per each unique ID. If collapsing affects numeric variables, average or median values can be taken. When it comes to categorical values, the analytical solution is more complicated and requires the development of methodology for such cases. The most obvious solution is to leave the most frequent category. Alternatively, the ratio of categories can be calculated (in cases of binary outcomes).

Overlapping variables can cause another issue for dataset size and future analysis (as inclusion of correlated indicators might inflate the significance of predictors and the model in general). Therefore, each dataset should be thoughtfully mapped before linking, as well as analysed with descriptive statistics tools prior to any further matching steps. For instance, in cases when there are two variables with similar meanings yet different operationalisation or coding mechanisms, the one of higher quality should be left. Checking for quality requires both quantitative and qualitative assessment, i.e. what is the percentage of missing values, what is the variation in the values, as well as how this variable was recorded and verified. The threshold for "good enough" quality is another analytical decision to make, as there are no universal standards that can be applicable to all kinds of datasets. Depending on how valuable or accessible certain information is, the threshold might significantly differ.

Part III. Case studies

The third part of this paper provides case studies showing the value of data-linking and how linked datasets are indispensable for tracking down and stopping the flow of dark money into politics. By providing examples of datasets complementary to BO registers, such as real estate data, this section will demonstrate particular schemes that can be revealed through working with linked data as well as how data-linking and data use is best done in practice.

Case No. 1: Politically-connected firms and public procurement data: Case of Bulgaria

Case summary

Institutional and governance challenges are a key constraint reducing Bulgaria's economic potential and private sector productivity. Bulgaria continues to lag behind most EU countries on governance indicators. The gap with the rest of the EU is most pronounced along institutions critical for economic growth such as the rule of law, control of corruption, and government effectiveness. One critical institutional area where governance weaknesses and state capture by private interests are evident is public procurement. Linking public procurement data to BO data to reveal politically-connected firms can help to detect potential signs of corruption and conflicts of interest in multiple ways. While the presence of political connections is not necessarily proof of corruption, by using data from public procurement and BO registers it is possible to verify whether PEPs were using their personal connections for private benefit.

Goals and datasets involved

For public procurement data, we use two sources for the analysis. First, all tenders and contracts were collected from the previous national e-procurement portal, AOP. Second, we also collected all publications from the new national e-procurement portal, EOP. We collected the data by using automated web-scrapers which are adapted to the specificities of the source websites and data repositories.

For BO information, we used data provided by Bureau Van Dijk Orbis, offering company-level information with extensive data on the corporate ownership structure. For collecting data on politically-exposed persons, the list was provided by

the Center for the Study of Democracy, complemented with information from the Panama and Pandora Papers as well as the Magnitsky Act.

Data-linking activities and challenges

The first step prior to analysis of any dataset is data cleaning, especially when it comes to variables which are needed for the matching. For instance, buyer names can be spelled in various ways even within the same dataset, and therefore all redundant characters should be removed as well as the letter case aligned. Next, the missing rate for the variables should be checked as well as how the missing data points are stored (e.g. whether they stored as “99” or “9999” or “NA”), as this can further influence the outcome of the analysis. Ensuring the correct calculations for numeric variables is also important (e.g. confirming the unit of measurement and rounding).

Next, one of the main challenges related to data merging was linking the names from the politically-connected persons list to the BvD shareholder names with corresponding IDs. Getting IDs was important for further analysis and revealing the network of ties between companies and shareholders. We tried to re-construct the algorithm used by BvD for the transliteration of Cyrillic names in the Latin alphabet, and after a number of tests were able to secure an adequate result. However, there were too many duplicated names (e.g. "Georgi Ivanov Georgiev" could refer to 86 different persons in the Orbis data with different IDs). We tried several methods to find a reliable way of merging the data, but the available PEP data did not allow for an unambiguous merge, and therefore some of the companies had to be dropped. The next step was to match bidder and buyer names from the procurement dataset to the company data from Orbis and get BvDIDs. After the transliteration from Cyrillic to Latin and removing all redundant characters, the matching rate of these datasets was around 85%.

Finally, the PEP data was merged with the public procurement data. The number of unique PEP-connected firms in the public Procurement data was 197 (both matched from the buyer and bidder side) out of the 4566, which leads to around 36 500 contracts if only matched on BvDIDs (so time invariant, which is the baseline).

Uses of linked data: investigations, policy analysis

In order to check if there is any significant relationship between politically-connected firms and corruption risks in public procurement, first corruption risks in public procurement were calculated. Following academic literature as well as World Bank publications, we define corruption in public procurement as the allocation and performance of public contracts by distorting principles of open and fair procurement in order to benefit

some connected actors to the detriment of all others. The resulting composite score, called the Corruption Risk Indicator (CRI), which can be considered an objective proxy measuring institutionalised corruption in public procurement, is a risk indicator that identifies situations where corruption tends to happen more often. The CRI allows for consistent comparisons across time, sectors, regions, and organisations, and can be further expanded and build upon using additional corruption proxies. For ease of interpretation, the CRI is calculated in the following way:

Each individual risk indicator is recoded as low (0) or high (1) risk with sometimes an in-between medium (0.5) category added.

The CRI is the arithmetic average of these defined individual risk indicators. It is calculated for each contract.

As a result, the CRI falls between 0 and 1, with 1 representing the highest observed corruption risk and 0 the lowest.

After matching data on politically-connected companies to public procurement data, we created a set of binary variables taking a value of “1” in cases where there is a politically-connected shareholder present and “0” where there are none in both the bidding and buying organisations.

Table 1: Regression results for politically-connected companies (PC) and Corruption Risk Indicator (cri)

Regression Results

	Dependent variable:		
	cri		
	(1)	(2)	(3)
PC company, time invariant	0.013***		
	(0.001)		
PC buyer company, time invariant		0.008***	0.008***
		(0.001)	(0.001)
PC bidder company, time invariant		0.029***	0.030***
		(0.003)	(0.003)
PC buyer and bidder (interaction), time invariant			-0.003
			(0.007)
Constant	0.267	0.266	0.266
	(0.286)	(0.286)	(0.286)
Observations	199,085	199,085	199,085
Log Likelihood	37,134.130	37,165.500	37,165.640
Akaike Inf. Crit.	-73,562.260	-73,623.000	-73,621.280
Note:	*p<0.1; **p<0.05; ***p<0.01		
	Included controls not shown are: Buyer location, Buyer type, Supl. location, Contract type, Year FE, Market FE, Contract value deciles		

The results of the analysis (Table 1) show that there is indeed a significant positive relationship between politically-connected companies and corruption risks, controlling for buyer and contract type, location, and fixed effects for year and market. Both politically-connected buyers and bidders increase the potential corruption risks in public procurement.

Lessons learned

Through matching public procurement data to the list of politically-connected companies, it was possible to establish the positive relationship between politically-connected organisations and corruption risks in the tendering process. Such results would not be possible if these two datasets were analysed separately. The list of politically-exposed persons does not provide valuable information in itself for identifying and preventing corruption. Politically-connected organisations can simply be defined as such in cases when a person who was an active businessman decided to go into politics, or the other way around. The more important question is whether such a person is willing to use their personal ties and connections for private gain. The analysis conducted on the Bulgarian case shows that this assumption has reasonable grounds. The presence of politically-connected companies in tendering procedures

increased the corruption risks by lowering competition, setting unrealistic decision and advertisement periods, or simply by increasing the buyer's dependence on the same supplier.

However, merging this type of data from different sources is a challenging task from a technical point of view. This is particularly relevant for datasets with different alphabets (Cyrillic vs. Latin) as well as in the absence of unified IDs across sources. For such complex cases, there is first a need for an algorithm which can transliterate text from different sources in the same style, which will help to reduce the time spent on matching. Second, there might be information loss to some extent due to the absence of IDs by which organisations can be matched across datasets.

Case No. 2: Beneficial ownership of German real estate

Case summary

With corrupt money from Russia and other places infiltrating financial markets and democratic societies in mind, the G7 communiqué of June 2022 reconfirmed the commitment to BO transparency and its importance for fighting corruption and safeguarding national security and democracy. This adds another until now somewhat-neglected goal to BO transparency, i.e. identifying assets bought with corrupt money, and boosted the debate around global wealth registers. Because real estate makes up more than half of all assets in any developed country, connecting BO information to real estate ownership would be the first and biggest step towards achieving these goals. With this discussion in mind, we tried to combine administrative data on real estate ownership with Orbis and the BO register to identify the BOs behind companies owning German real estate.

Country background

As in many other countries, the question of "who owns German cities" is high on the German agenda both as part of the fight against money-laundering and tracing Russian assets as well as in the context of the policy debate around exploding housing prices and gentrification. A series of studies from the UK to Dubai, France, Norway and finally Germany are currently trying to tackle this question. They face different issues of data availability.

1. While data on legal owners is publicly available as open data in the UK, France and Norway, real estate ownership information is not publicly available in Dubai and Germany. In Germany, real estate ownership is recorded at local registers and exchanged with the sixteen cadastres at the level of and under the jurisdiction of the federal states. The German study used freedom of

information requests to obtain data, which were successful in some places and rejected in others. The study in Dubai profited from a leak.

2. While the real estate data in France contains a unique identifier (company ID) for legal owners, the German data does not systematically provide such information and poses challenges related to partially outdated and incorrectly or differently spelled names.

3. While BO data is available as open data in the UK and France, the German BO register provides public access on a case-by-case basis and at a cost of 1.96 EUR per extract. Additionally, because the German BO register was set up in parallel to the company register with poorly-monitored exemptions from the duty to register, less than 10% of companies were registered by 2020. Despite a major reform in 2021, this number was still at around 50% in mid-2022.

Data-linking activities and challenges

The data obtained from the freedom of information requests, i.e. the name of companies owning real estate in Germany, was linked to Orbis using several algorithms to clean up different spellings of the same name and spelling mistakes prevalent in the data (i.e. separating company name and type, correcting for standard company types and matching based on alphabetically-sorted name-letters; for more details, see Miethe, Trautvetter 2022). While the Orbis data is very comprehensive for German companies, this matching only reached a 69.73% coverage due to the limitations of the source data from the registers and the limitations of the matching algorithms (a manual match for a subset of the data including historic company names increased the matching to nearly 100%). 91% of the companies matched (and most likely about the same number in the original sample) were German companies. Orbis provides information both on all available shareholders as well as global ultimate owners defined as individuals or companies that directly or indirectly own more than 50% of shares. Again, Orbis coverage for shareholders of German companies is very comprehensive, thanks to the German company register providing public information on all shareholders for most company types. In contrast, Orbis does not have information from the German BO register. Through an iterative process, we managed to identify natural persons behind all shares of the companies owning real estate in 79.87% of cases. For 4.4% of all cases, the ownership chains ended in an anonymous company in a secrecy jurisdiction. For a selection of these cases (39 out of 1 297 companies), we obtained information from the German BO register (or BO registers from other countries where applicable) manually. For 23% (9 cases), there was no entry in a BO register available, mainly due to the gaps

in the German register. For another 23% (9 cases), the BO register contained additional information on shareholders. For the remaining 54%, the BO register only contained information on the person controlling the company, usually the German manager. This meant that – due to data quality issues and the definitions used for BOs – the majority of real estate ownership structures that could be identified as suspicious based on the structure visible in company registers and Orbis appeared unsuspicious in the BO register.

Uses of linked data

The analysis shows that linking data from the (German) real estate register to company ownership and BO data can serve two major policy goals. It can help to identify the majority (by value) of assets with unclear and/or suspicious ownership for further analysis by law enforcement. And it can – to some degree – help to answer the question of "who owns the city" by providing information on the degree of concentration of ownership and to identify major owners. A recent example from Berlin helps to illustrate this: Journalists identified four Berlin-registered companies owning Berlin real estate and in turn being owned by three companies from the BVI. While at the time of reporting none of the four companies were registered in the BO register, by July 2022 (following the second deadline to register), only one was registered. While the data analysis cannot identify the BOs of those companies, it can a) identify how many plots are owned by those BVI companies directly or indirectly throughout Germany, and b) for the first time provide an answer as to how often and where such anonymous structures are actually used. The results are consistent with the findings from other countries and encouraging: Only a small share of real estate and a very small share of real estate owners use anonymous corporate structures to hide their ownership, with a strong but not exclusive focus on big cities. While this makes targeted analysis by law enforcement possible, it does not mean that this analysis is expendable, because even a tiny share of national real estate means many billions of Euros of corrupt money hidden from scrutiny.

Lessons learned

Improving the analysis of real estate ownership and the identification of German assets with unclear and/or suspicious ownership would require four major improvements to data availability and data linkage:

1. Make real estate ownership information available for research by clarifying the legal basis for accessing this data.

2. Create a unique identifier for companies owning real estate in the real estate register (i.e. a company ID and/or the BO register ID) as promised in the coalition

agreement of the current German government.

3. Make BO data available for bulk analysis.

4. To obtain information on the value of the assets or the share of apartments owned in a certain city by any of the owners, additional information, i.e. on the purchase price and the number of apartments per cadastral plot, would need to be collected.

Case No. 3: EBOCS (European Business Ownership and Control Structures) project

Case summary

EBOCS (European Business Ownership and Control Structures) is an example of a linked BO and Business Register which resulted in a project covering multiple countries and visualising ownership structures. The project was established by an international consortium led by the European Business Registry Association and consisting of a number of partners coming from the business registry world. It provides simplified and unified access to Beneficial Owner Register data and Business Register data on business ownership and control structures for financial analysis and investigative purposes, thus increasing the level of transparency of legal entities.

Linking BO data to business register data helps to reveal connections and ties between companies and individuals on a national as well as on a cross-border level, and helps actors like Financial Intelligence Units, Law Enforcement Authorities and others to identify (ultimate) owners of European legal entities for anti-money laundering and anti-terrorist financing purposes. This aims to support the disruption of international crime networks through better detection and prevention of financial, economic and other related crimes.

Goals and datasets involved

EBOCS provides real-time information on 22 ML companies and 50 ML officers and owners coming from seven Business Registers (Estonia, Italy, Spain, Ireland, Latvia, Romania and United Kingdom) and three Beneficial Ownership Registers (Latvia, Ireland and Spain). The national registers, Business as well as Beneficial Owners, provide official information. A central visualisation tool was developed to allow end users, usually counter-crime agencies, to intelligently access the EBOCS information services.

Data-linking activities and challenges

The Beneficial Ownership Register is clearly a very important source of information, pointing out the ultimate business owner; however, the whole picture can be broadened quite significantly by adding the information from the Business Register, highlighting every single appointment and ownership (even small shares) of a specific individual. This can be done not only at a national level, which would already be an outstanding achievement, but even at a cross-border level.

One of the main challenges related to data merging was linking individuals on a cross-border level. On a national level, individual IDs help to identify specific businesspersons with certainty, revealing the network of ties with companies. But individual IDs have national relevance only; as soon as we cross the border, we require human assistance to identify and match businessmen. A European unique "person" identifier, which at the moment does not exist (every country has its own national individual identifier), would be a significant step forward in the process of matching individuals in different jurisdictions.

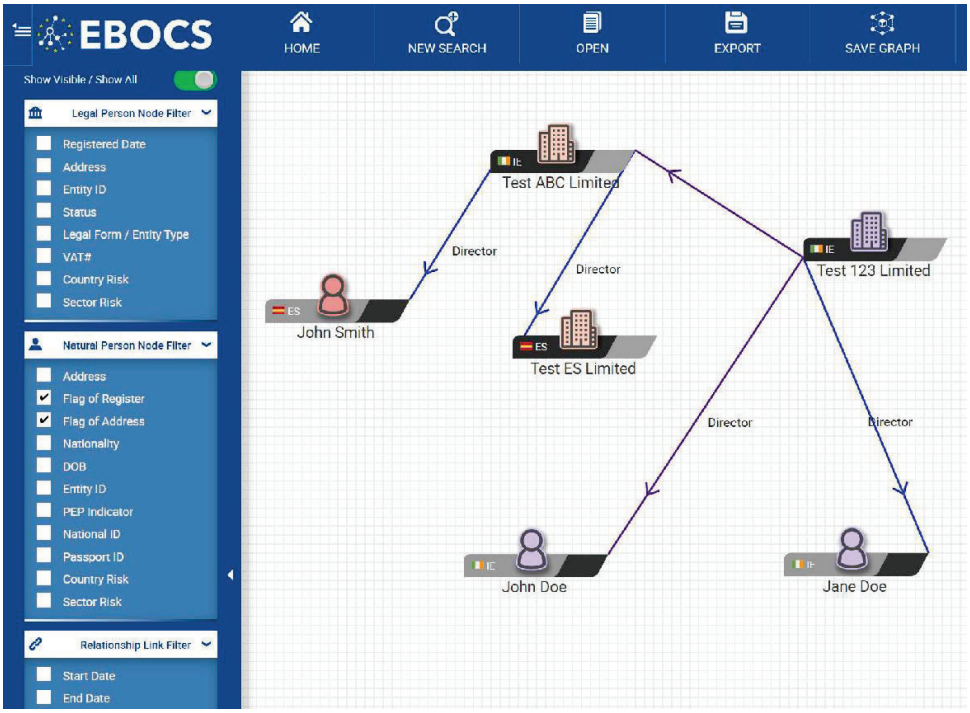
Moreover, EBOCS's services architecture was designed to easily integrate and connect, with a long-term view, many other sources of information, such as the enterprise's bank accounts database, or the land and property register, to make the whole analysis even more powerful.

Uses of linked data: investigations, policy analysis

The EBOCS platform provides the user with a *Visualisation Tool*, an online graphical tool that aggregates ownership and control structure information from primary data sources. The information is retrieved in real time from the official national data repositories and is presented to the end user in a graphical representation. It provides users with a simplified representation of how natural and legal entities are linked to each other at both a national and cross-border level.

The typical investigation starts with a question like: "What companies does Mr. John Smith have connection with?". The search would start in a specific jurisdiction, where the user would identify the companies of which John Smith is the ultimate beneficial owner. Then, thanks to the business register information, the analysis will be enhanced with all John Smith's appointments and ownerships.

At this stage we will have a graphical representation that indicates Mr. Smith is (for example) a beneficial owner of Company A and Company B, board member of Company C, general director of Company D, and owner of 10% of Company E.



The next step will be to search Mr. John Smith in a different jurisdiction. Once identified, the original graph will be expanded with ties (ultimate ownership, appointments, simple ownership) to companies based in the new jurisdiction. The same search will be run again in every relevant jurisdiction.

The final graph will reveal any significant relationship between companies and Mr. Smith: which companies Mr. Smith owns, partially or completely, which companies are under his control (being a board member) and which companies might be under his control (through second-level ownership ties).

The counter-crime agencies will hold a comprehensive and Europe-wide view of the business connection and properties of the individual under investigation, and the overall picture could be enhanced also by adding individual land properties, for example by connecting national land registries data, etc.

Lessons learned

The number of sources of business and economic information has increased dramatically over the last few years, but these data sources do not talk to each other, forcing the end user to ask for many different access permissions, download a significant amount of data, standardise them and eventually draw manual connections in order to get a full and comprehensive picture. The whole process gets even more complicated as local economies are turning into global

economies which are increasingly interconnected on a cross-border level.

It is clearly important to have access to trustable information (possibly certified), but it has become even more important to be able to use tools and services that gather information from those sources, create a network of ties and linkages automatically, highlight connections and dependencies, and make it easy to examine and investigate. Reliable data provision is no longer enough, especially when investigating for anti-money laundering and anti-terrorist financing purposes; we need to assign the proper value to every piece of information and make it easier to interpret linked data.

Part IV. Technical recommendations

In order to make the data-linking process easier, a few technical steps have to be taken by the users. First, one should identify the potential data sources that can be linked to the BO data. After identifying potential data sources, the most promising datasets have to be mapped in detail. This detailed mapping should consider the following metadata features:

Scope: What percentage of the relevant population is covered in the dataset. For example, what is the share of the total public procurement spending in a country which is reported in the tendering and contracts dataset? Scope also encompasses the time period covered by the data, including considerations such as whether the dataset is regularly updated.

Depth: Data depth measures the amount of information available on each observation. This requires listing all the relevant variables available in the dataset and cross-checking this list with the desirable variables for corruption measurement purposes.

Accuracy: The accuracy of data captures the completeness and truthfulness of data compared to the represented actor behaviour. The most basic check of data accuracy is the prevalence of missing values. Moreover, it is also easy to look for apparent data errors such as typos or nonsensical information (a company name typed up instead of a contract value in a public procurement announcement).

Accessibility: Data accessibility implies that the data is machine-readable, easily downloadable and processible. If data access requires complicated and error-prone web-scraping, this may represent considerable barriers to data use for measurement purposes.

Interoperability: Mapping includes assessing how different datasets can be linked in a meaningful way that allows for combining information. For example, if asset declarations data cannot be connected to specific public organisations – i.e. people reporting their assets cannot be connected to the institutions they are affiliated with – then we cannot connect asset declarations to contracting risks of those public organisations. Furthermore, connecting people to organisations on its own is often not enough; information on the time of affiliation is also important. When we know the period of office for the official declaring his/her assets and the

corresponding awarded contracts, we can begin to unpack whether certain procurement processes were corrupted for private gain.

One should differentiate between the steps that a user can take in order to ensure the data accessibility and accuracy and structural problems which can only be overcome by data providers (state agencies, private companies, etc.). There are three main data-related issues noticed by researchers when accessing data:

Absence of identifiers which can be used cross-nationally. Usually the company IDs are country-specific, which makes it very difficult to match a company from country A to the company in country B. Using company names becomes the only possible solution for such a problem, which also requires efforts on the users' side to clean and control for spelling.

Absence of centralised data registers. Some countries do not administer a unified centralised data register, having local registers instead. Due to the country's administrative division (e.g. federal state) this cannot be overcome by introducing centralised registers, but the datasets do have to be standardised (with the same variable names, data coverage, identifiers, etc.).

Restricted or paid access to the datasets. There are many reasons why private companies or state agencies do not provide free access to data, including data protection policies. Yet when it comes to using the data for corruption prevention goals, a certain exception for civil society actors or academics and journalists should be made.

Next, data-linking can be performed. In order to do so, a few criteria should be applied to the datasets:

The unit of observation should be established for all datasets and aligned to the same level. For example, there are a few datasets on state subsidies and grants provided to certain companies. Some of the datasets will contain information on companies, and therefore the level of observation is company. Others provide information on subsidies and grants, and therefore the unit of observation is the subsidy or grant call. In some cases it is quite challenging to merge datasets of different levels of observation, as in the absence of unique IDs, the row will be multiplied many times, ending up with identical observations. The IDs of the two merging datasets should be unique, so that in the merging process it will be clear which row corresponds to which ID.

At least one of the datasets should serve as a "master" dataset and contain unique IDs to which other data can be matched. Otherwise, one might end up with multiplied IDs in the main dataset, which should be avoided for further linking. For example, if there is company-level data with an address as a unit of analysis. The

only ID by which it is possible to merge this dataset to the main one is company ID, but they are multiplied because the same company might have multiple addresses.

The final list of variables should be of a high quality without repetitive and incomplete columns. For instance, in cases when there are two variables with similar meanings yet different operationalisation or coding mechanisms, the one of higher quality should be left. Checking for quality requires both quantitative and qualitative assessment, i.e. what is the percentage of missing values and what is the variation in the values, as well as how this variable was recorded and verified.

Conclusions

As demonstrated in the case studies, linked data can significantly boost the possibility for investigating and tracing illicit financial flows. This can be done through using various datasets, including beneficial ownership data, public procurement, real estate registers, company registers and others. Data-linking helps to identify inconsistencies across databases, as well as reveal otherwise hidden connections between companies or individuals.

However, there are many challenges along the way to getting a good match between data and being able to extract as much information as possible from the linked datasets. The absence of unique identifiers, especially when it comes to working with multiple countries, imposes significant limitations that cannot be overcome simply or easily by advancing the technical skills of the people working with data. Different units of analysis require additional efforts to align the datasets and can frequently result in information loss due to the higher level of observations. Finally, the variables themselves can limit comprehensive analysis due to the low quality of observations, missing values, data errors and other issues.

Putting additional efforts into developing and monitoring the implementation of data standards in governmental agencies as well as making data open to the general public and NGOs and allowing them to use it for independent investigations would significantly boost dark money tracing and increase the efficiency of monitoring.

References

Beneficial ownership data in procurement. Open Ownership, 2021

A new global standard on BO transparency. Transparency International, 2021

Strengthening the future global standard. Transparency International, 2021

Preventing abuse of the financial system for money laundering and terrorism purposes: the 4th directive.

A Beneficial Ownership Implementation Toolkit, OECD, 2019

Enhancing Government Effectiveness and Transparency: The Fight Against Corruption. Chapter 9: Beneficial Ownership Transparency. World Bank, 2021

Anonymes Immobilienvermögen und international Besitzketten. Miethe, Jakob and Trautvetter, Christoph 2022 (in German).

Contact:
csabotproject@transparency.org