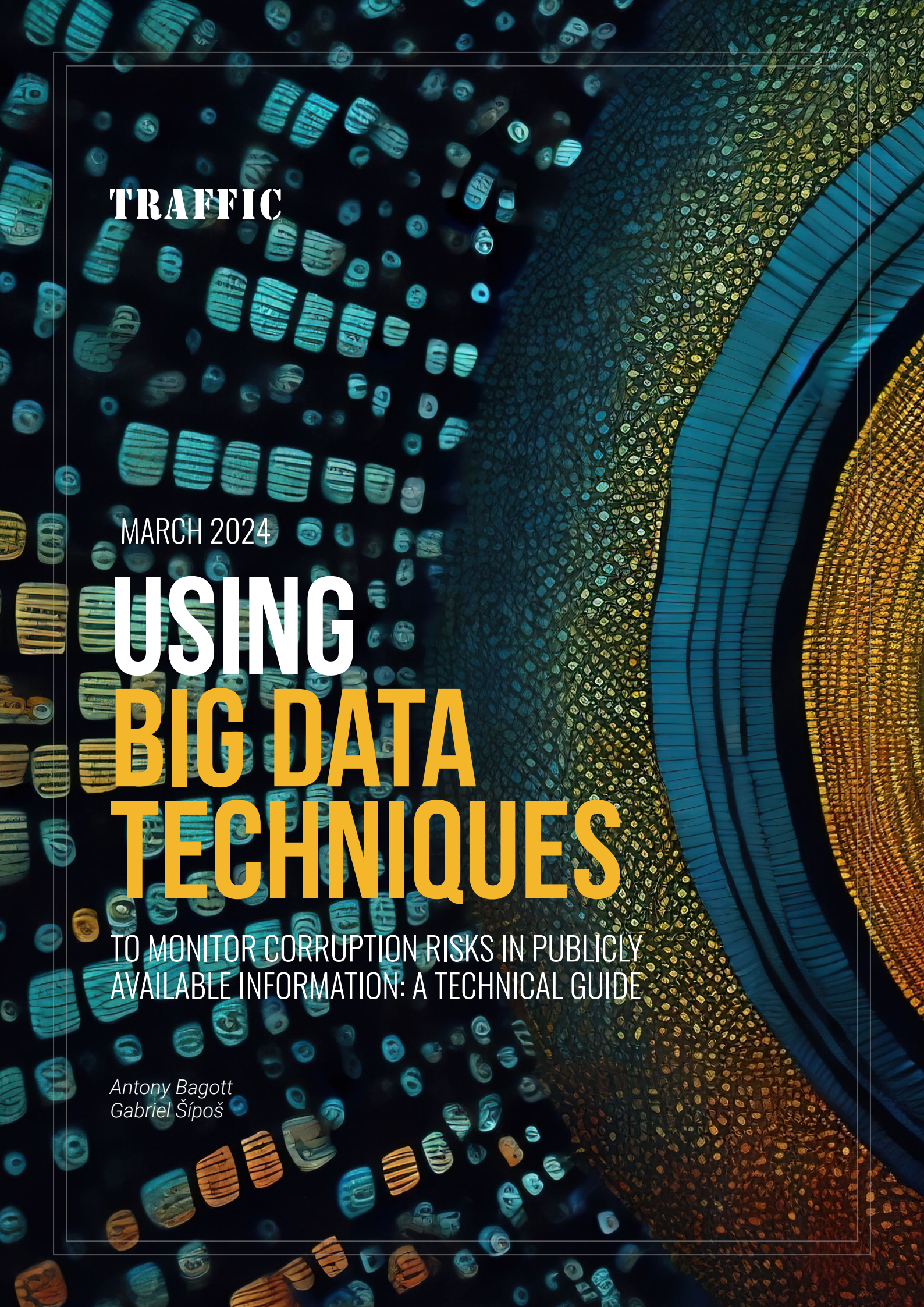


TRAFFIC

MARCH 2024

USING BIG DATA TECHNIQUES

TO MONITOR CORRUPTION RISKS IN PUBLICLY AVAILABLE INFORMATION: A TECHNICAL GUIDE

Antony Bagott
Gabriel Šípoš

TRAFFIC REPORT

ABOUT US

TRAFFIC is a leading non-governmental organisation working globally to ensure that trade in wild species is legal and sustainable, for the benefit of the planet and people.

The designations of geographical entities in this publication, and the presentation of the material, do not imply the expression of any opinion whatsoever on the part of TRAFFIC or its supporting organisations concerning the legal status of any country, territory, or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The research on which this Guide is founded was made possible by the generous support of the American people through the United States Agency for International Development (USAID). The contents are the responsibility of the authors and do not necessarily reflect the views of USAID, the United States Government, or individual Targeting Natural Resource Corruption (TNRC) consortium members.

ABOUT ABOUT TARGETING NATURAL RESOURCE CORRUPTION

The Targeting Natural Resource Corruption (TNRC) project is working to improve biodiversity outcomes by helping practitioners to address the threats posed by corruption to wildlife, fisheries and forests. TNRC harnesses existing knowledge, generates new evidence, and supports innovative policy and practice for more effective anti-corruption programming. Learn more at tnrcproject.org

PUBLISHED BY:

TRAFFIC International, Cambridge, United Kingdom.

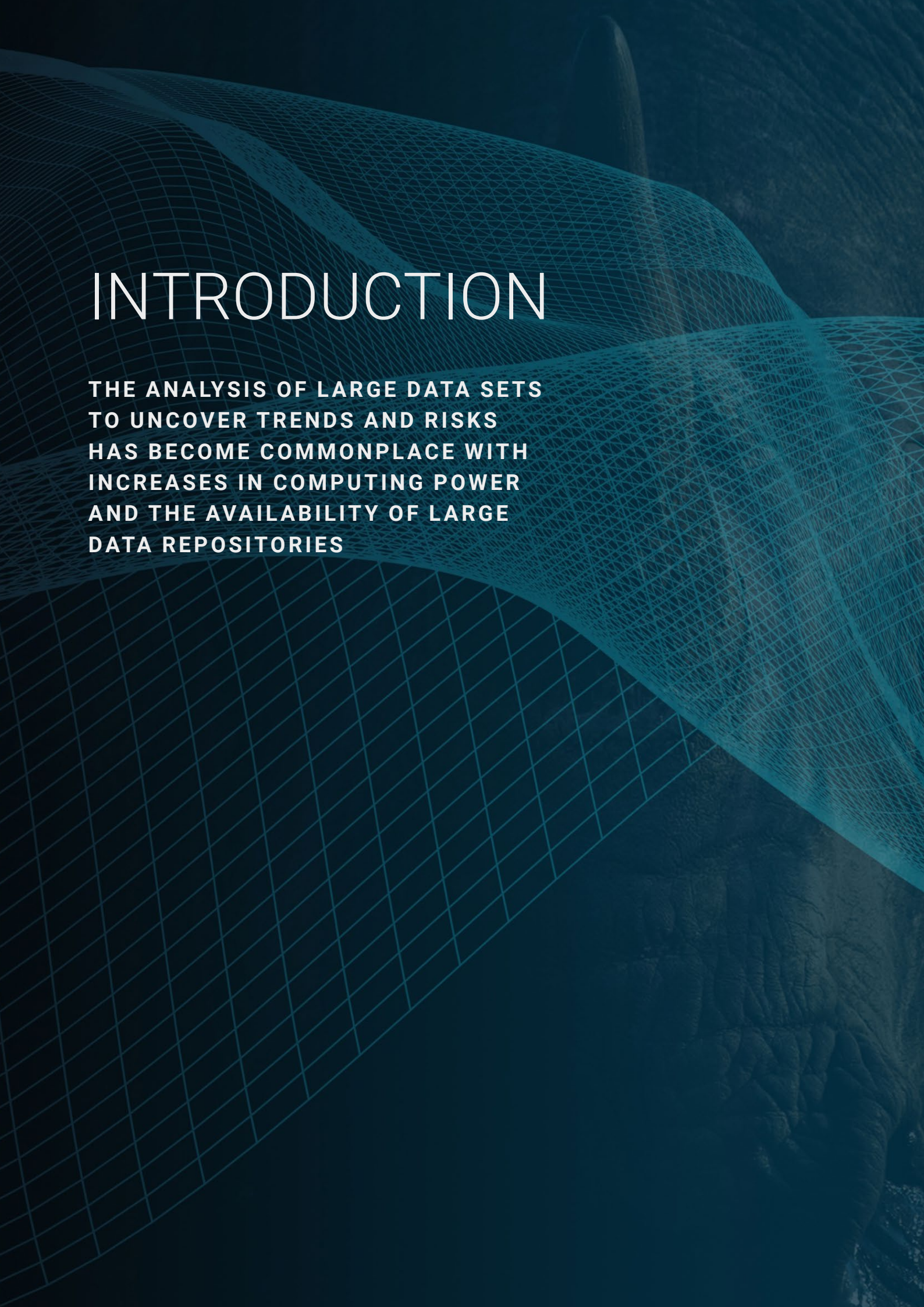
SUGGESTED CITATION

Bagott, A., Šípoš, G. TRAFFIC **(2024).** *Using Big Data Techniques to monitor corruption risks in publicly available information: a technical guide.*

DISCLAIMER

This publication is made possible by the generous support of the American people through the United States Agency for International Development (USAID). The contents are the responsibility of the author(s) and do not necessarily reflect the views of USAID, the United States Government, or individual Targeting Natural Resource Corruption project consortium members.

CONTENTS



INTRODUCTION

THE ANALYSIS OF LARGE DATA SETS TO UNCOVER TRENDS AND RISKS HAS BECOME COMMONPLACE WITH INCREASES IN COMPUTING POWER AND THE AVAILABILITY OF LARGE DATA REPOSITORIES

After having piloted such approaches to explore corruption risks in forestry concessions in 2022, we have sought to apply our learnings to additional regions or countries in which forestry data are widely available.

By using existing technology in an innovative way, TRAFFIC aims to determine whether "big data"[1] approaches to analyse publicly available information can provide credible evidence for monitoring and investigation where there is the potential that corruption in the forestry sector has occurred. But our model can also be implemented as a preventative tool – when managed well, increasing the visibility of conflicts of interest in contract procurement can serve to decrease instances of corruption.

In this report, we offer technical guidance on using tools to analyse large datasets and reveal the potential corruption risk of individuals or companies involved in the forestry sector. We also outline indicators of corruption risk and describe the challenges faced when working with this kind of information to assist those seeking to apply similar techniques in the field of conservation as well as anti-corruption.

OUR APPROACH

DATA PROCESSING MODEL

The data processing model describes the steps taken to collect, store, sort, analyse and enrich the data. It outlines the major actions, decisions, data types and analytical tools required to reach each result.

The model is presented in the flowchart on the next page. This is followed by a full description of each stage in the process.

IDENTIFICATION OF CORE DATA SOURCES:

The first step is to identify the core data sources, i.e., the publicly available sources of information that will provide us with names of entities (such as people and companies) and their involvement in forestry (such as their roles and their harvesting rights).
Key data sources from which to obtain this information include:

- *Lists of politically exposed persons (PEPs)*[2]
  We search for PEPs within the focal country, aiming to identify those who are linked (even indirectly) to forestry or who have allocated logging permits/concessions to companies.
  In some cases, PEP lists will contain information about income and expenditure, assets, liabilities, gifts, real estate, vehicles, and more.
- *Registries of forestry concessions*
  We search for information relating to the allocation of logging rights within the focal country. As a minimum, this should include the names of the companies involved, the responsible person(s) within each company, and the PEP(s) who allocated the rights. Where available, we also include additional information such as the location allocated, the permitted amount of harvestable timber, and the purpose for harvesting.

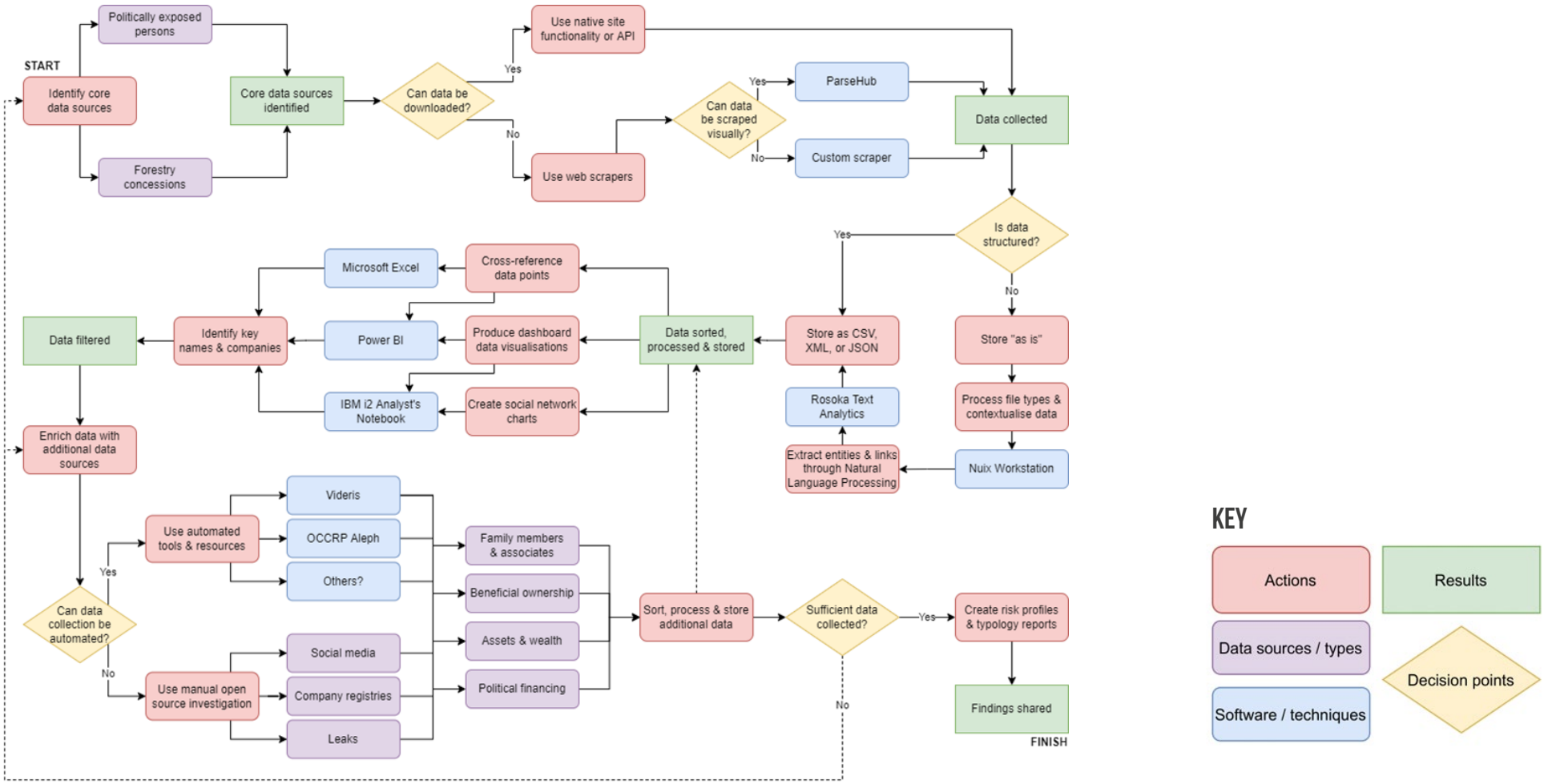
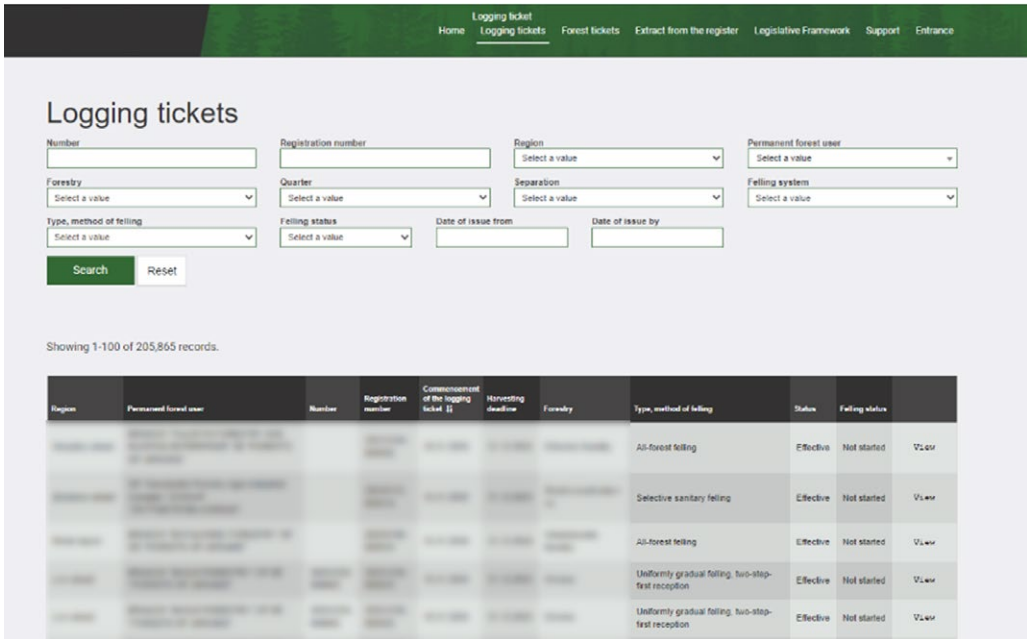


FIGURE 1

A flowchart outlining the data processing model used. This model assumes that all data is obtained from online sources

FIGURE 2

A website listing forestry concessions granted in a country over the last decade. Some sections with potentially sensitive information have been blurred.



DATA COLLECTION

Where large volumes of data are involved, automated data collection processes such as web scraping are typically far more efficient than manual processes. The efficiency of these processes depends on the available functionality on the site, the consistency of the site structure, the complexity of the webpage design, the presence of images and/or text, and the technical ability of the person performing the scraping.

Depending on the permissions and structure of the source website, we can either download the data via the native site functionality, scrape the website using a pre-built visual web scraper, or scrape the website using a custom web scraper.

- *Native site functionality*
  If the site permits users to download the data – for example, through an API (Application Programming Interface) or simply through an "Export" button, we use this method.
- *Pre-built visual web scrapers*
  If the required data cannot be downloaded using native site functionality, we use pre-built visual web scraping software such as ParseHub to allow for efficient data collection. Visual web scrapers work best in situations where the site structure is consistent, the webpages are laid out in a coherent way, and the data consists of text rather than images.
- *Custom web scrapers*
  In cases where a visual web scraper is unable to collect the data, it may be necessary to build a custom web scraper in-house. These would typically be set up for each website. Due to the high cost and time implications of this method, it should be reserved for situations in which no alternative method will work, no other source contains the same information, and only when there is a high chance that it will yield a large quantity of valuable data.

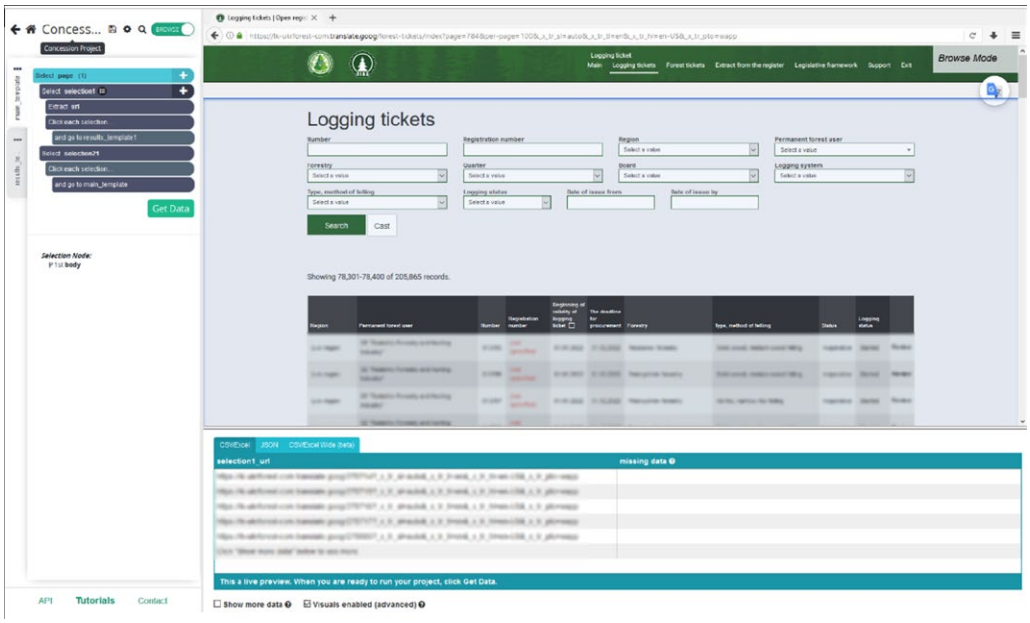
DATA STORAGE AND PROCESSING

The way the data is initially stored depends on its format following its collection. If the collected data is *structured*, in tabular format for example, it can be stored in a format such as CSV, XML or JSON for later analysis and import into other software.

If the data is *unstructured*, in narrative format as a PDF for example, we need to perform additional processing on it before it can be understood by most analytical software. The processing is as follows:

FIGURE 3

ParseHub, a visual web scraping tool that provides a simple "point and click" interface to select and scrape elements of a webpage. Some sections with potentially sensitive information have been blurred.



- ***Contextualisation of data***
  Browsing the contents of a large number of files with different file types can be a time-consuming process within Window's native file explorer. Digital investigation tools such as Nuix Workstation allow users to examine thousands of unstructured data types, pick out relevant information, and identify entities within the text through regular expressions (regex). Nuix Workstation also performs Optical Character Recognition (OCR) on text documents, ensuring that documents are machine-readable.
- ***Natural Language Processing (NLP)***
  NLP tools such as Rosoka Text Analytics can find links between entities in the text. When paired with visual analysis software such as IBM i2 Analyst's Notebook, it is possible to view a visual representation of the relationships between entities (people, objects, locations and events) within the text.

FIGURE 4

Rosoka Text Analytics, which works as a plugin to IBM i2 Analyst's Notebook to visually display connections between people, objects, locations, and events. This image has been blurred to remove potentially sensitive references to people and/or places.

IDENTIFICATION OF ENTITIES OF INTEREST

Once the data is in a format that can be transferred into other analytical software, it can be queried and analysed to identify entities of interest, i.e., key names and companies.

- *Cross-referencing data points*
  A simple way to cross-reference datapoints across two datasets is by using the Fuzzy Lookup Add-In for Excel.[3] This ensures that spelling variations, alterations, and typos of the same entity are accounted for. It also detects matches when forename(s), middle names, and surname(s) are presented in different orders or are omitted altogether – which is particularly useful when the name is arranged differently according to the naming convention of the language used, or when one part of the name is retained in one source but omitted in another. Users can adjust the level of similarity that causes a match.Pre-built visual web scrapers
- *Producing dashboard data visualisations and social network charts*
  Dashboard visualisations of data can be produced using a range of programs, including Power BI and IBM i2 Analyst's Notebook. These visualisations can be queried manually, but the programs also provide a number of automated features that facilitate the process.
  i2 Analyst's Notebook allows users to create a network chart of the entities involved in the datasets and the relationships between them – for example, a professional relationship between an employee and an employer, a financial relationship between a company and a politician, a familial relationship between a father and son, and so on.
  These analytical tools can help users find entities with multiple degrees of separation, or those with a statistically significant number of relationships with other entities.
- *Applying a scoring system*
  In cases where a range of different corruption risk factors from various datasets can be exported on a large scale, it may help to apply a scoring system. We relied on anti-corruption experts inside and outside of TRAFFIC to identify corruption risk factors. More information about the specific risk factors we explored can be found in the Risk Factors section of this report.
  The scoring system combines these various risk factors, applies weighting, scales them accordingly, and produces a final risk score.[4] This score can then be used to assess the relative corruption risk of a large number of different individuals and companies.

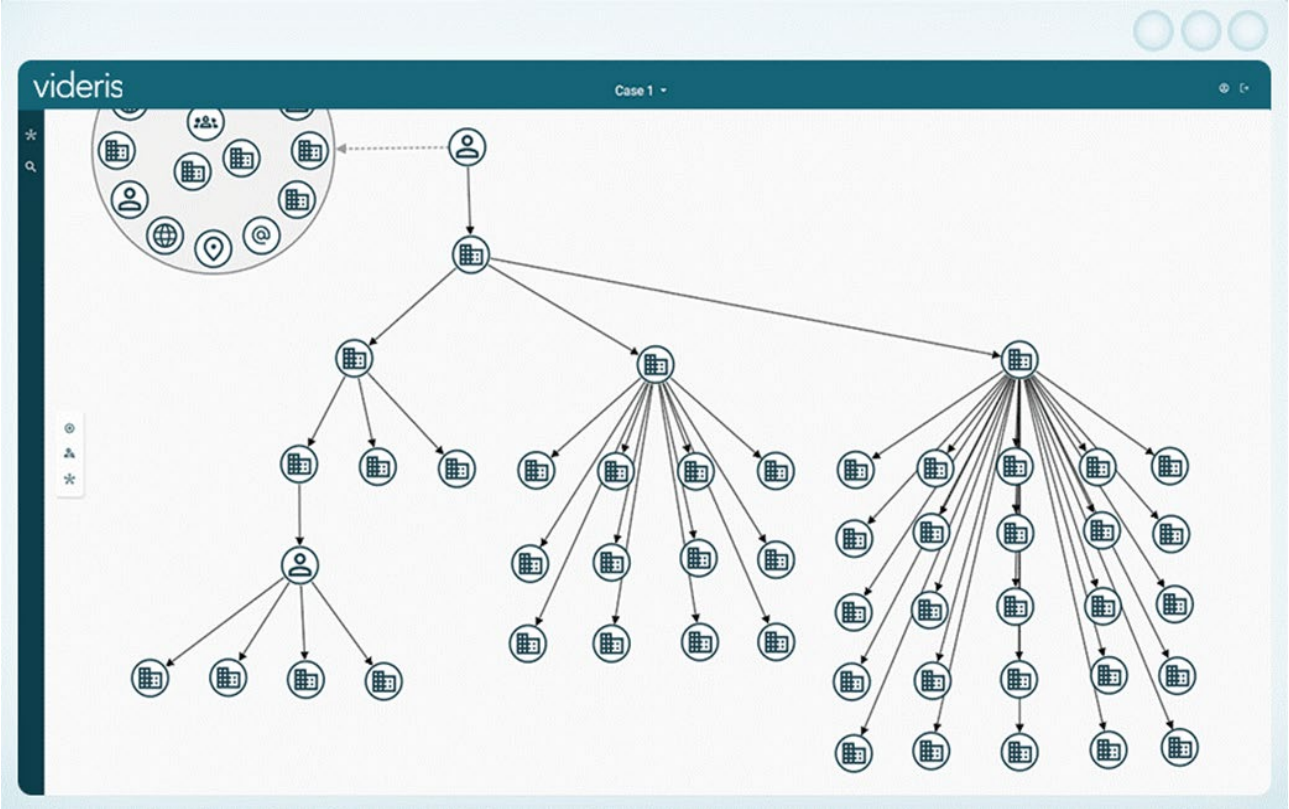
DATA ENRICHMENT

Once entities of interest have been identified, the data we hold can be enriched using social media sites, company registries, and leaked documents. This process usually requires some level of manual open-source investigation, but it can be partially automated through online investigations platforms such as Videris and through data repositories such as OCCRP Aleph.

Additional potential insights include evidence of beneficial ownership, political financing, sanctions, and any held assets or wealth. Lists of company officers and evidence of social and professional associations may also prove useful.

FIGURE 5

Videris, an online investigation platform that allows users to search for up-to-date data about individuals and companies from a range of online sources, visualise the information in a relationship chart, and enrich it further as needed.



RISK FACTORS

Risk factors are used to provide context to the relative corruption risk of each forestry actor. They are particularly important when applying a scoring system. In this case, risk factors should be split into "scores" and "multipliers".

"Scores" are used for factors that indicate an increased risk of corruption. The scores are added to each other to create an initial risk score. "Multipliers" are used for factors that – by themselves – do not necessarily indicate additional corruption risk, but could increase the impact of any existing risk(s). These are used to multiply the initial risk score to create the final score.

RISK FACTOR "SCORES" THAT HAVE BEEN USED EFFECTIVELY INCLUDE:

- *Sanitary logging*
  Concessions listed as "sanitary" logging could be based on documentation with false information, i.e., falsely stating that the trees are infected or dead. According to an expert we consulted, between 3% and 6% sanitary clearance is standard. Therefore, anything over 10% could indicate a level of risk, particularly if a large volume of sanitary wood is permitted to be logged within a certain concession area.
- *Percentage of auctions won*
  A single person or company being consistently or regularly successful in their bids for forestry contracts could indicate a greater risk of corruption, bribery, or nepotism in the bidding process - especially for large contracts (i.e., those with a higher revenue).
- *Unexplained wealth*
  If PEPs have unexplained wealth, such as undeclared property or assets, monetary assets that exceed their salary, luxury cars, and so on, there is a higher potential risk of corruption.
- *Price by volume*
  Contracts that were exceptionally cheap (i.e., with a low price per cubic metre of wood) or exceptionally expensive (with a high price per cubic metre of wood) could indicate corruption, e.g., someone receiving a suspiciously good deal could indicate that nepotism or bribery played a role, or someone paying an excessive amount for a contract could indicate a risk of money laundering, etc.

RISK FACTOR "MULTIPLIERS" THAT HAVE BEEN USED EFFECTIVELY INCLUDE:

- *Number of contracts*
  In cases where forestry actors are corrupt, those with a larger number of contracts are more likely to have a negative impact on forestry areas (and are more likely to be involved in higher-level corruption) than forestry actors with a smaller number of contracts. However, contract number alone is not an accurate indicator of risk.
- *PEP status*
  Forestry actors who are also PEPs - or are related to/affiliated with PEPs - are more at risk of having a conflict of interest when dealing with the allocation of timber harvest rights. However, in countries where forestry areas are often state-owned by nature, PEP status alone is not an accurate indicator of risk.

OTHER FACTORS THAT WE HOPE TO INCLUDE IN FUTURE METHODOLOGIES INCLUDE:

- *Tree cover loss data*
  By comparing forestry concession harvesting limits with satellite data on tree cover loss using the interactive map provided by Global Forest Watch, we could determine whether certain forestry actors appear to have exceeded their permitted harvesting over a set period of time. This risk factor is complicated by the irregularity of tree sizes due to variations in species and age, meaning it could be difficult to determine the volume of wood that equates to a certain area of tree cover loss. Before undertaking this kind of analysis, it is recommended to research local laws regarding the minimum size of trees that are permitted to be harvested. Basing any estimates on that figure would allow for at least a ballpark figure of the volume of extraction taking place.
- *Political cycle dynamics*
  Undertaking a comparative analysis of the political cycle (i.e., changes in government) with the concession data could highlight cases where political influence has had a statistical impact on where and to whom harvest rights are allocated.
- *Natural calamity data*
  Although not a risk factor in itself, natural calamity data can help to distinguish justified sanitary logging from illegal sanitary logging.

CHALLENGES AND LESSONS LEARNED

IDENTIFICATION OF CORE DATA SOURCES

We have encountered challenges with the automated collection of data from some sources due to the nature of the way the data is presented. These challenges have been listed below:

- *Website design*
  It is not always easy to access information on multiple concessions or permits at once (e.g., through an API or a bulk download to Excel). In these cases, using a web scraper is necessary to avoid an inefficient data extraction process.
- *Sources with imagery*
  Sources that present their concession data through an image, such as a map, are useful for exploring individual concessions but prevent the use of a standard visual web scraper due to their unstructured nature. In these cases, a custom web scraper may be required.
- *Poor quality photocopies*
  Some sources provide data as photocopies of text documents in PDF format. In some cases, these photocopies are of poor quality and have not yet undergone OCR to read the text contained. This can prove a problem for OCR software.
  Quality improvements can be performed upon PDFs through image manipulation software. Improvements can include changes to contrast, brightness, sharpening, black/white balance, and more. These improvements are limited by the initial quality of the original source. In all cases, visual improvements constitute a manual task, so should only be performed when there is a high chance that the target document will yield valuable data.
- *Variation in data availability*
  Different websites make available different levels of data. For example, some concessions sources might only focus on a specific region or a certain type of contract. This makes it difficult to make comparisons

between countries and requires the user ensures they fully understand the data that is made available or omitted.
Where there is a high level of variability in the availability of datasets, using risk factors that are relative rather than absolute – i.e., looking at the ratio rather than the actual difference between separate datapoints – can help to soften the impact of any missing information.
For example, focusing on the proportion (rather than the absolute number) of sanitary logging contracts per forestry actor can ensure that even when certain types of contracts are omitted, comparisons can continue to be made. However, this should be performed on a case-by-case basis and care should be taken to avoid misrepresenting the data.

- *Lack of clear metadata and guidance*
  Some Understanding the exact nature of the data being made available is often not straightforward, and often requires the user to undertake significant research to avoid making ill-informed conclusions.
  Where sufficient metadata or guidance does not exist, speaking with or reading reports by other researchers who have used the same data can help to improve a user's understanding of the data and the efficiency of their analysis.
  Having a user with an understanding of forestry is beneficial as they can understand any nuances in the data and have experience of finding third-party resources (such as NGO reports) to help inform them.
- *Language barriers*
  The language skills of the user undertaking the research can affect the speed and accuracy of the analysis.
  Translation tools such as Google Translate and DeepL can help to an extent, but care should be taken to avoid mistranslations.
  For example, when translating proper nouns (e.g., family names) that have an equivalent literal meaning (e.g., "Smith", "Miller", "Hunter", etc.), online translation tools regularly provide the literal translation rather than the family name itself.

DATA COLLECTION: WEB SCRAPING

Visual web scrapers present several technical challenges. These challenges and their solutions have been listed below:

- *Data scraping produces no results*
  Certain data scraping projects produce no results in web scraping platforms. This issue may be caused by a variety of reasons, including:

  Elements fail to be selected on the page
  Visual web scrapers aim to intelligently select the appropriate elements (i.e., the images, text, etc.) on each web page, based on the pattern of elements that have been selected on previous pages. Sometimes, this does not proceed as planned due to inconsistent page design. In these instances, the web scraper needs to be "trained" using additional pages.
  Authentication required
  Some websites require a login before data can be scrapped. In these instances, the action of logging into the website will need to be added as the first step to the project.
  Long load times
  Web scrapers will occasionally run too quickly for slower websites. In these instances, the project can be set to wait for a specified amount of time for elements to appear. This will decrease the speed of the scrape, depending on the amount of time set to wait per element.
  Web scraper is blocked
  Some sites have technology that prevents web scrapers from accessing the data. In these instances, the IP Rotation feature can be turned on to avoid this. However, doing so will significantly decrease the speed of the scrape.

- *Data appears with unreadable characters*
  When exported to CSV and viewed in Excel, datasets that contain non-Latin script (e.g., Cyrillic script) do not appear as expected. This is because the CSV export does not support UTF-8 encoded data.
  To view the script, we need to tell Excel to treat the file as UTF-8 encoded. To do this, we import the file into Excel instead of opening it directly. When doing so, we choose the appropriate encoding (UTF-8) and the data is imported correctly.

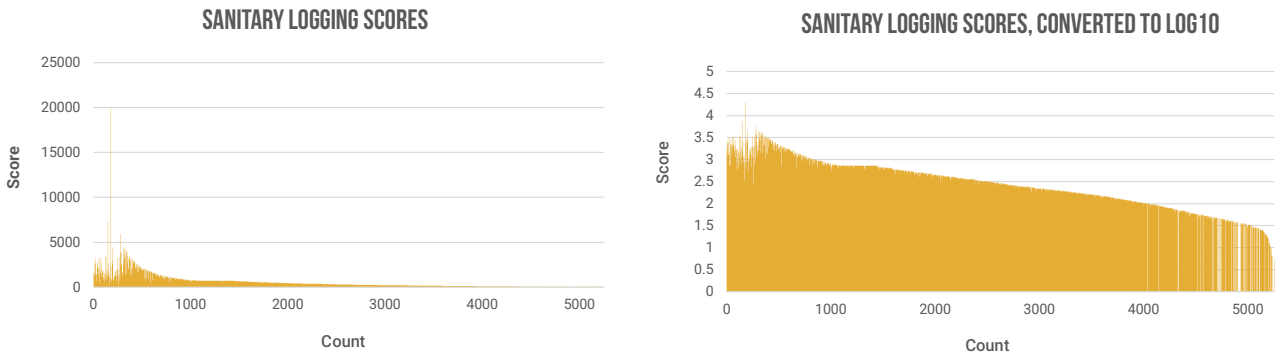
- *Document data is in another language*
  To allow for consistency between datasets and ease of use, it may be necessary to translate datasets from their original language into English. For large datasets, the entire document can be translated through third-party translation tools available online.
- *Datasets do not convert into CSV*
  Certain files are too large to be converted into CSV. Attempting to convert the file into CSV will simply result in the software crashing. In these cases, they should be downloaded as another format such as JSON, and any analysis of the data should be run using the JSON file.

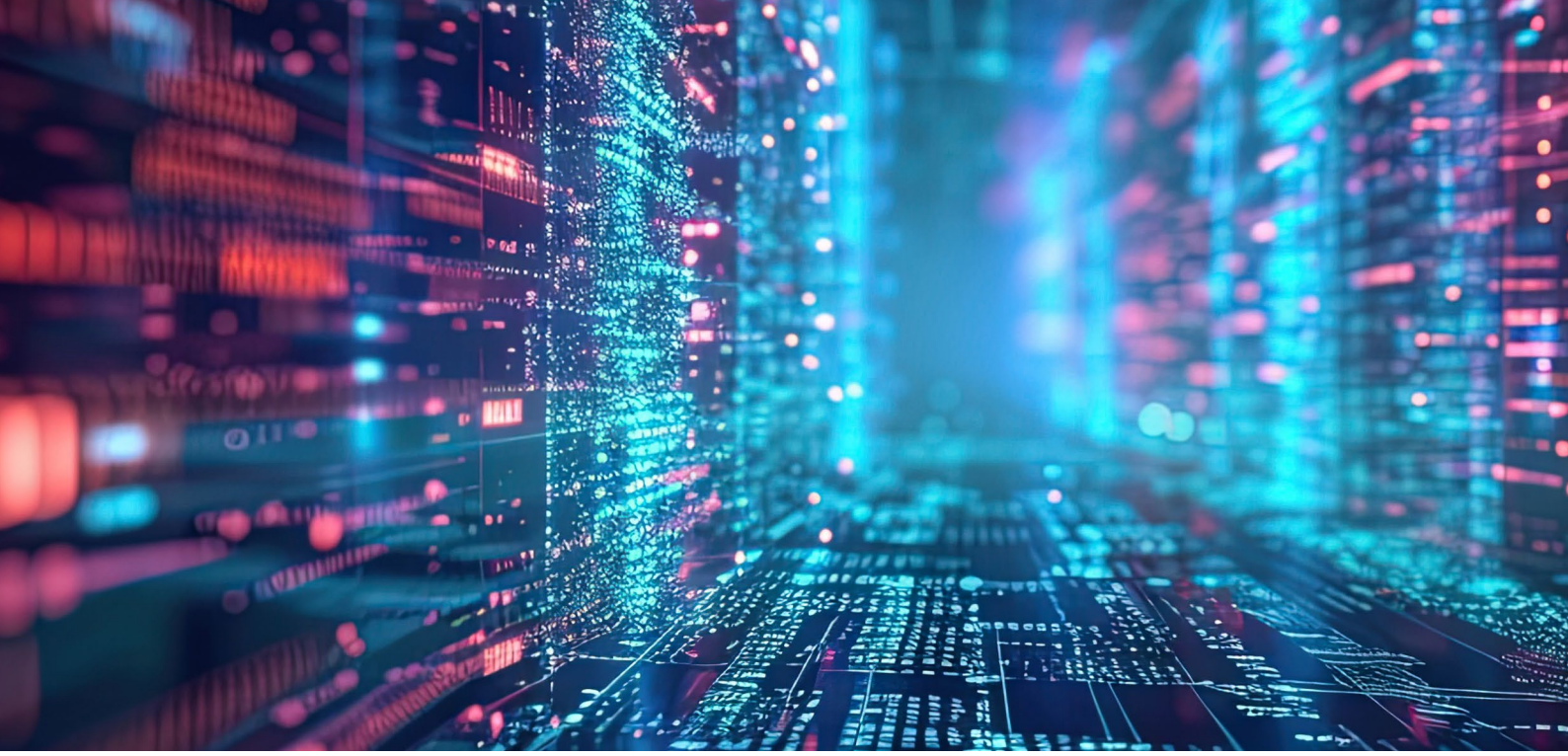
IDENTIFICATION OF ENTITIES OF INTEREST: APPLYING A SCORING SYSTEM

- *Wide diversity of parameters and datapoints results in an unbalanced range of scores*
  When corruption risk factor scores are based on a diverse set of parameters that span a wide range with a handful of anomalous datapoints on one or both of its ends, the scores can be skewed. The handful of data points at either end of the range end up garnering very large or very small scores, while most data points remain in a homogenous group in one section of the range. This reduces the effectiveness of the final risk score.
  Converting scores to a logarithmic form (e.g., base 10) solves this problem. It essentially changes the curve in the "line" of datapoints, improving the spread of scores across the range. This can be performed easily using the inbuilt functionality within a wide range of platforms, e.g., the LOG function in Excel, or the log10() function in R. Scaling these scores down to manageable numbers (10, 5, 1, etc.) improves the readability of the data while also allowing us to add a "weighting" to each score based on our determination of the severity of the risk factor.

FIGURE 6

A comparison of charts of sanitary logging scores (based on the percentage and volume of sanitary logging in the analysed forestry contracts) before and after conversion to logarithmic form, showing the improvement in the spread of scores across the range.





RESULTS

Due to the sensitive nature of the data, we cannot share the specific results of our analysis here. Instead, we can outline the type of results that we produced. The results of our data processing model include the following outputs:

- *Scoring Compilation*
  The Scoring Compilation is a list of all individuals and companies involved, allowing for a large-scale comparison of the relative potential corruption risk between entities. Our Scoring Compilation includes all datasets used, as well as the formulas employed to reach our final scores.

FIGURE 7

An anonymised version of TRAFFIC's Scoring Compilation, showing each of the score categories that contribute to the total risk score for the forestry performers/PEPs/bidders in the list.

Forestry performer/PEP/bidder details			Scoring						
Name from original source	Name in Ukrainian	Category	TOTAL SCORE	Sanitary logging score (Sliding scale from 0 to 10 based on percentage and volume of "sanitary" contracts.)	Auctions won score (Sliding scale from 0 to 5 based on percentage and revenue of auctions won.)	Unexplained wealth score (sliding scale from 0 to 5 based on unexplained wealth of a PEP.)	Revenue/volume score (+1 if low or high auction revenue per volume of timber.)	Number of contracts multiplier (sliding multiplier from x1 to x5 based on number of contracts)	PEP score multiplier (x2 multiplier if the forestry actor is a PEP.)
[Redacted]	[Redacted]	Performer	35.81	7.53	0.00	0	0	4.7527087	
[Redacted]	[Redacted]	Performer	30.42	6.08	0.00	0	0	5.0000000	
[Redacted]	[Redacted]	PEP	28.56	5.85	0.00	5	0	1.3161249	
[Redacted]	[Redacted]	Performer, Bidder	21.59	6.53	4.15	0	0	2.0223072	
[Redacted]	[Redacted]	Performer, Bidder	20.93	7.81	4.19	0	0	1.7444232	
[Redacted]	[Redacted]	Performer, PEP	20.35	7.15	0.00	1.74403183	0	1.1436159	
[Redacted]	[Redacted]	Performer, PEP	19.69	7.99	0.00	0.875331565	0	1.1104738	
[Redacted]	[Redacted]	Performer, PEP	19.34	6.58	0.00	1.850132626	0	1.1461653	
[Redacted]	[Redacted]	Performer, PEP	18.65	7.24	0.00	1.850132626	0	1.0263437	
[Redacted]	[Redacted]	Performer, PEP	18.07	7.10	0.00	0.543766578	0	1.1818568	
[Redacted]	[Redacted]	Performer, PEP	17.90	6.93	0.00	0	0	1.2914808	
[Redacted]	[Redacted]	Performer, PEP	17.71	6.88	0.00	0	0	1.2880814	
[Redacted]	[Redacted]	Performer	17.28	5.86	0.00	0	0	2.9460378	
[Redacted]	[Redacted]	Performer, PEP	15.67	7.24	0.00	0	0	1.0824304	
[Redacted]	[Redacted]	Performer	15.54	6.75	0.00	0	0	2.3010410	
[Redacted]	[Redacted]	Performer, PEP	15.43	5.31	0.00	2.393899204	0	1.0008498	
[Redacted]	[Redacted]	Performer, PEP	15.30	6.45	0.00	1.087533156	0	1.0144466	
[Redacted]	[Redacted]	Performer, PEP	15.05	7.05	0.00	0	0	1.0671341	
[Redacted]	[Redacted]	Performer	14.61	7.01	0.00	0	0	1.8480586	
[Redacted]	[Redacted]	Performer, Bidder	14.54	5.56	0.00	0	0	2.6154663	
[Redacted]	[Redacted]	Performer, PEP	14.42	6.65	0.00	0.543766578	0	1.0025494	
[Redacted]	[Redacted]	Performer	14.19	8.16	0.00	0	0	1.7401742	
[Redacted]	[Redacted]	Performer, PEP	14.18	5.31	0.00	1.631299735	0	1.0212450	
[Redacted]	[Redacted]	Performer, PEP	13.86	6.65	0.00	0	0	1.0424899	
[Redacted]	[Redacted]	Performer	13.62	6.53	0.00	0	0	2.0843425	
[Redacted]	[Redacted]	Performer	13.61	6.69	0.00	0	0	2.0359040	
[Redacted]	[Redacted]	Performer, Bidder	13.52	7.37	4.78	0	0	1.1130232	
[Redacted]	[Redacted]	Performer, PEP	13.45	6.70	0.00	0	0	1.0033992	
[Redacted]	[Redacted]	Performer	13.41	7.34	0.00	0	0	1.8268536	
[Redacted]	[Redacted]	Performer, PEP	13.36	6.65	0.00	0	0	1.0050988	
[Redacted]	[Redacted]	Performer, PEP	13.31	6.58	0.00	0	0	1.0110474	
[Redacted]	[Redacted]	Performer	13.31	6.61	0.00	0	0	2.0138092	
[Redacted]	[Redacted]	Performer, Bidder	13.28	5.97	4.29	0	0	1.2931804	
[Redacted]	[Redacted]	Performer	13.24	6.31	0.00	0	0	2.0962396	
[Redacted]	[Redacted]	Performer, PEP	13.18	5.15	0.00	1.306366048	0	1.0212450	
[Redacted]	[Redacted]	Performer, Bidder	13.10	7.79	4.41	0	0	1.0739124	
[Redacted]	[Redacted]	Performer, Bidder	13.10	5.88	3.44	0	0	1.4062035	
[Redacted]	[Redacted]	Performer, Bidder	12.94	7.02	4.07	0	0	1.1657106	

- *Risk Profiles*

  Based on the information provided in the Scoring Compilation, we can create risk profiles of specific entities to allow for a deeper insight into high-risk individuals or companies.

  The Risk Profiles created by TRAFFIC consist of a personal profile detailing the main information known about the PEP, an overview of the total risk rating and the contributing risk factors, a list of suggested actions, a typology chart, and a catalogue of the sources used.

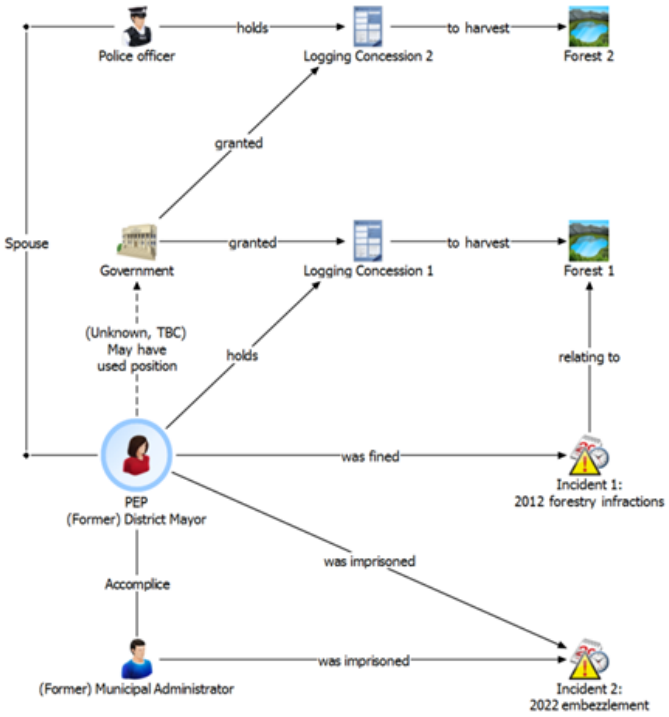


FIGURE 8

An anonymised typology chart detailing the connections between a PEP, their associates, logging concessions, and any corruption-related incidents.

CONCLUSION

Our results will be shared with relevant financial institutions to follow up on the high-risk individuals and companies that we've identified to allow further research into their financial and business relationships and take action where needed. By sharing the methodology and the process that we've undertaken, we also hope to allow others to apply similar techniques to uncover potential corruption in our focal countries and elsewhere.

It should be noted that although the data processing model outlined here can help to identify potential instances of corruption (indeed, in one case we independently identified a PEP as a high corruption risk a few months before they were convicted for embezzling public funds), we – as an NGO – cannot accurately verify the data on our own. It is with the support of financial institutions that we can take our insights further and find verifiable proof of corruption through their existing systems and processes. A suggestion for taking this work further would be to collate this proof and verify if the risk ratings in our Scoring Compilation match up with real-world corruption.

In sectors with a large number of repeated contracts such as forestry, the use of big data techniques can be a fruitful method of monitoring and investigating corruption risks. As long as the data are made publicly available, this control system can be used not only by internal government agencies but also by media or civil society, helping to ensure accountability even when supply chains are suspected of being mired by corruption. Furthermore, if data from timber supply chains – including cross-border trade – were made available in other countries, big data approaches such as these could bring further opportunities for ensuring that trade in natural resources is legal and sustainable.

ENDNOTES

[1] Throughout the project, we processed types of information that are defined as big data (e.g., social media interactions) and employed approaches that are typically used on big data (e.g., automated collection, processing, and analysis). However, it should be noted that not all the datasets we encountered can be defined as "big data": i.e., extremely high-volume, high-velocity, and/or high-variety datasets that require automated analysis to reveal patterns or trends.

[2] In some countries, there exist official PEP lists consisting of information that the PEPs themselves have submitted. This is because public officials in these countries are required to declare information about their assets. However, there is massive variability in these regulations across the globe. Where these regulations do not exist, in certain cases independent transparency organisations have compiled similar lists. In other cases, these lists don't exist at all – in which case, it is more difficult to undertake this kind of work.

[3] The Fuzzy Lookup Add-In uses the Jaccard similarity coefficient to measure the similarity between finite sets of objects. This is defined as the size of the intersect of the two sets divided by the size of the union of the two sets, as per the following (where J is the Jaccard distance, A is set 1, and B is set 2):

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

[4] Converting the initial scores to logarithmic scale can help to even out any excessively large ranges. The following formula can then be used to "scale" the scores, i.e., to convert the measurement r from its original range (from r min to r max) into a weighted target range (from t min to t max):

$$r = \frac{r - r_{\min}}{r_{\max} - r_{\min}} \times (t_{\max} - t_{\min}) + t_{\max}$$

IMAGE CREDITS

Cover	jensenartofficial / Pixabay
2	A. Walmsley / TRAFFIC
4	TheDigitalArtist/ Pixabay edited with Carla McMahon / iStock
6	franganillo / Pixabay
12	TRAFFIC / A. Walmsley
14	geranimo / Unsplash
17	BrianPenny / Pixabay

JANUARY 2024

WORKING TO ENSURE THAT TRADE IN WILD SPECIES IS LEGAL AND SUSTAINABLE, FOR THE BENEFIT OF THE PLANET AND PEOPLE

TRAFFIC

TRAFFIC
+44(0)1223 331 997
traffic@traffic.org