COrruption Risk Indicators in Emergency

Co-funded by the European Union

TECHNICAL GUIDELINES FOR THE PLATFORM

Deliverable WP 4.1

Andrea Nelson Mauro - onData

Alessio Cimarelli - onData

Davide Del Monte - info.nodes

Grant Agreement number: 101038790 — CO.R.E — ISFP-2020-AG-CORRUPT

Index of contents

1. Introduction

The WP4 objective is to define precisely how the tools for data collection and data analysis must work and must be used by the project team and by common users.

The partners aim to:

1. define the characteristics (UX, data formats, front-end, back-end) of the database
2. define the data characteristics of the data to be collected, inserted and analyzed
3. define a common methodology for the analysis of the data collected, accordingly with the general methodology and the definition of indicators as for the WP3.
4. define and implement a data visualization dashboard, based on the data collected, that will be used as the “home page” of the project website, in order to provide information about the indicators of corruption in a simple, accessible and captivating way.

The present document is focused on:

1. define the characteristics (UX, data formats, front-end, back-end) of the database the “Definition of the database”.

2. Definition of the database

Database development must take into account the two main features:

- gathering data from all partners referring to public contracts, such as contract details, suppliers, information about the implementation. This data has been mapped preliminary in order to identify reliable data sources with the help of the project partners.
- allowing basic and in-depth data analysis in order to point out common patterns and possible outliers while comparing data among different administrations, different tenders, different suppliers whether at national or international level.

This definition has been clarified during the implementation of the data collection process (WP 4.2). In fact, Info.nodes coordinated a series of one-to-one meetings with project partners to identify and analyze data sources. During these meetings it raised the opportunity of gathering reliable data from several public data sources in Italy, Ireland, Spain and Portugal

In particular, Italian and Portuguese public agencies have been releasing public contracts data with the so-called Open Contracting Data Standards format. This format entails the possibility to manage advanced comparisons among different data sources.

As for Ireland, the data sources that have been identified are lacking several pieces of information, but they will be enough for at least a preliminary setup of the database.

In order to ensure that the database meets the needs of the project objectives and the operational needs of the partners, info.nodes prepared a first draft of the “Technical Guidelines for the Platform” outlining the main steps for its implementation and shared it during the meeting in Barcelona (September 2022) to collect comments and feedback

3. Setting up of the database

The first steps to setup the database requires that info.nodes, Dataninja and the University of Perugia clearly identify and detail:

- indicators of corruption in emergency that must be calculated;
- variables for each indicator to be extracted from data sources;
- analytical processes to be carried out in order to extract insights from the data.

A preliminary test activity (e.g. the calculation of some indicator with a limited amount of data through "R for Statical Computing") is needed to identify the possible limits of the analysis and in order to scale the analysis up for all the public contracts published in the 4 countries participating in the project.

In particular info.nodes and Dataninja will implement a structure for the database that includes:

1. the download of data from its data sources (sample or dump)
2. the preliminary process of cleaning and normalizing this data in order to be comparable each other
3. the extraction of the variables from this data, that will inform the indicators system

Once finished, this process will be refined and standardized for the final database creation.

Regarding the technologies that Dataninja selected for the project purposes, there are several factors that impacted the decisions:

- Public contracts data is a multidimensional object that includes information about the contract and the tender, but also about participants, suppliers, implementation and beyond. It means that the project cannot simply rely on a SQL/NOSQL choice for our database.
- Italian and Portuguese agencies have been releasing public contracts data in accordance with Open Contracting Data Standards (so called OCDS). Info.nodes and Dataninja have previous experience on OCDS standard as they suggest evaluating the selection of a full text search engine in order to cope with the potential large amount of text within a single contract.
- the database should be scalable in the future, including the possibility to hold more data from the countries included into the project and beyond. In fact, a database of public contracts should be capable of collecting a large amount of data and to analyze within a comfortable time range.

Due to the requirements of scalability and sustainability of the project - new Countries included, new databases connected, new indicators and new variables to be calculated - Dataninja outlined the specific stack of hardware and software to serve those analyses.

For these reasons info.nodes and the partners should:

- Manage preliminary analyses and further ones through on-premises machines;
- Compare hardware resources whether on-premise and in cloud, in order to understand their efficiency in terms of analysis and costs;
- Select a full text search engine as the main instrument for the database creation, to fully analyze the texts included in the public contracts;
- Set up an environment where to manage analyses through R for Statistical Computing (configuring R libraries).

4. Define possible use cases and common queries

For implementing this task, statisticians and analysts from University of Perugia and Dataninja have been managing specific analyses on sample data. The main objective of this task is to identify data for calculating single variables. This sample data comes from Italian national agency for public contracts (ANAC) that has been releasing data with OCDS format.

Analysts have been managing these analyses for testing the availability of data, the feasibility of calculations, and the possibility of replication on large scale and datasets.

The output of this task is:

- the list of data needed for those calculations;
- the calculation formulas for extracting all the variables needed;
- the technical requirements for automating these calculations within a software environment.

To ensure that the database functions meet the needs of researchers and journalists, the platform will be opened at a later time (indicatively starting from February 2023) to a group of selected users (anti-corruption experts and investigative journalists especially), so that they can provide feedback, comments and suggestions. to increase its usability.

5. Design and implement different ways to access data (download, API)

This task refers to accessing data from different sources. The state of public contracts data sources has been mapped by University of Perugia, Info.nodes and Dataninja. We can leverage the public contracts portals of Italy and Portugal where data can be downloaded with Open Contracting Data Standards format. These data were downloaded from those sources for normalization and preliminary analysis. Dataninja decided to manage only downloaded data in a first implementation stage, in order to make the project as efficient as possible.

In a second implementation stage, Dataninja will connect the database with external sources through public APIs.

6. Define a series of insights to be shown as main indicators

Given that the University of Perugia selected 13 indicators of corruption in emergency so far, as the ones that must be used for the project implementation, info.nodes and Dataninja managed to valorize those indicators as the ones to be shown to the final users.

To calculate those indicators, we identified a number of variables to be extracted from data (as in the document “4.2 Guidelines for data collection and data analysis”). Those variables are informed both by raw data extracted from public contracts and third-parties data sources at national levels.

After a preliminary analysis, the partners noted that several variables will be calculated with a low degree of complexity, while other variables refer to data that are closed or accessible only under paywall (e.g., data about Italian companies is accessible only by paying users).

Taking in account all these aspects, the project database for public contracts will serve the calculation of variables and indicators that will be used for creating the project dashboard and the public contracts analysis.

More information about the dashboard will be available in the document “4.4 Concept note – data visualization”.

7. Technical solutions to ensure openness and reusability

The evidence from those preliminary activities showed us that the better instrument for creating the database should be ElasticSearch (ES). In fact, ES will help us and even the final users to analyze data fully leveraging all texts included within each public contract. On the other hand, ES will allow us to implement and automate all calculations of variables and indicators that will serve all possible use cases, such as for instance the project dashboard for indicators of corruption risks, the data reporting, and API service to export indicators elsewhere on the web.

Elasticsearch is a distributed, free and open search and analytics engine for all types of data, including textual, numerical, geospatial, structured, and unstructured. Elasticsearch is built on Apache Lucene and was first released in 2010 by Elasticsearch N.V. (now known as Elastic). Known for its simple REST APIs, distributed nature, speed, and scalability, Elasticsearch is the central component of the Elastic Stack, a set of free and open tools for data ingestion, enrichment, storage, analysis, and visualization. Commonly referred to as the ELK Stack (after Elasticsearch, Logstash, and Kibana), the Elastic Stack now includes a rich collection of lightweight shipping agents known as Beats for sending data to Elasticsearch. This choice makes higher the possibility of reusability of the software created for this project.

All software components will be released with open licenses and accessible also from different repositories (e.g. Github).