IT TOOLS AND METHODS FOR MANAGING MIGRATION FLOWS

# D6.1

## Report on the specifications and architecture of the EMT platform

**Georgios Stavropoulos**
**Nikolaos Grevekis**
**Ilias Iliopoulos**

**Centre for Research and Technology (CERTH)**
**June 2021**

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 882986".*

| Deliverable Factsheet | |
|---|---|
| **Title and number** | *Report on the specifications and architecture of the EMT platform (*D6.1) |
| **Work Package** | WP6 |
| **Submission date** | 14/06/2021 |
| **Authors** | Georgios Stavropoulos, Nikolaos Grevekis & Ilias Iliopoulos (CERTH) |
| **Contributors** | Diana Suleimenova & Alireza Jahani (BUL); Mattia Di Salvo (CEPS); Lenka Dražanová (EUI); Asli Okyay (IAI); Tobias Heidland & Finja Krüger (IfW); Gergana Tzvetkova (CSD); Emma Teodoro, Andrea Guillen, André Groger, Daniel Morente & Colleen Boland (UAB); Haithem Afli (MTU); Georgios Gogolos, Stelios Gkouskos, Konstantinos Kalampokis, Anna Pappa & Natalia Oumnova (TRC) |
| **Reviewers** | Derek Groen (BUL), Mehwish Alam (FIZ) & Cristina Blasi (UAB) |
| **Dissemination level** | PU (Public) |
| **Deliverable type** | R (Report) |

| Version Log | | | |
|---|---|---|---|
| **Issue Date** | **Version** | **Author** | **Change** |
| 17/05/2021 | V0.7 | Georgios Stavropoulos | Initial version sent to reviewers |
| 21/05/2021 | V0.7 | Derek Groen, Diana Suleimenova & Mehwish Alam | Review sent to Author |
| 31/05/2021 | V1.0 | Georgios Stavropoulos | Initial version including changes and suggestions made by reviewers, sent to CO |
| 08/06/2021 | V1.2 | Emma Teodoro, Lilian Mitrou, | Legal/ethical review |
| 14/06/2021 | V1.3 | Georgios Stavropoulos, Nikolaos Gevrekis | Addressed legal/ethical review comments. |
| 14/06/2021 | V1.3 | Cristina Blasi | Coordinator Check |
| 15/06/2022 | V2.0 | Georgios Stavropoulos, Nikolaos Gevrekis | Final version after EC review |

## Executive Summary

This deliverable aims to provide a first view on the design principles of the EUMigraTool that will be developed within the ITFLOWS project.

The EUMigraTool (EMT for short) is a software platform that will integrate all the knowledge created within the ITFLOWS project. It will provide to relevant stakeholders a set of tools to enable them to do simulations and predictions on various migration aspects, ranging from the number of people expected to leave a certain region within selected countries of origin towards EU, to potential challenges when migration populations arrive in EU territories.

The EUMigraTool is being designed under Task T6.1, which concentrates on the basic design definitions of the EMT. T6.1 started in month 3 and ends on project month M9 with the submission of the present deliverable. Over the next period, we will work on the development of EMT and aim to release the first version of the tool on project month M18. We will report the release details in the deliverable D6.2 including any updates to the EMT design.

The design of the EMT consists of 2 major components:

(1) the front-end that is what the user sees, providing a set of intuitive tools to set up the required prediction and simulation use cases (s)he wants to analyse; and

(2) the back-end, which is responsible for collecting, storing and managing all the data, along with performing all the necessary processing to produce the simulations and predictions.

In this deliverable, we present the data sources that will be utilised in the EMT, along with their respective function. Next, we present the user requirements, which are extracted from the 1st end-users' workshop that was held under the actions of WP7. The rest of the document presents the design principles of the EMT, its architecture, functionalities, and some initial screenshots, starting from the underlying concepts, to a high-level description and a detailed analysis of the individual components and the technologies to be used.

**Keywords**: EUMigraTool, simulation, prediction, requirements, architecture.

# Table of Contents

# Abbreviations

**ACLED**: Armed Conflict Location & Event Data Project
**AIDA**: Asylum Information Database
**API**: Application Programming Interface
**BUL**: Brunel University London
**CAMEO**: Conflict and Meditation Event Observations
**CEPS**: Centre for European Policy Studies
**CERTH**: Ethniko Kentro Erevnas kai Technologikis Anaptyxis
**CESSDA:** Catalogue of the Consortium of European Social Science Data Archives
**COO**: Countries of Origin
**CPI**: Consumer Price Index
**CRED**: Centre for Research on the Epidemiology of Disasters
**CRI**: Associazione della Croce Rossa Italiana
**CSA**: Coordination and Support Action
**CSD**: Centre for the Study of Democracy
**EASO**: European Asylum Support Office
**EC**: European Commission
**EM-DAT**: Emergency Events Database
**EMT**: EUMigraTool
**ESS**: European Social Survey
**EU**: European Union
**EUI**: European University Institute
**EVS**: European Values Study
**FAO**: Food and Agriculture Organization
**FIZ**: Fiz Karlsruhe–Leibniz-Institute fur Informationsinfrastruktur GMBH
**FRONTEX**: European Border and Coast Guard Agency
**GA**: Grant Agreement
**GDP**: Gross Domestic Product
**GDPR**: General Data Protection Regulation
**GKG**: Global Knowledge Graph
**GQG**: Global Quotation Graph
**GRG**: Global Relationship Graph
**HDX**: The Humanitarian Data Exchange
**HTTP**: Hypertext Transfer Protocol
**IAI**: Istituto Affari Internazionali
**IfW**: Institut für Weltwirtschaft
**ILO**: International Labour Organization
**IMF**: International Monetary Fund
**IOM**: International Organization for Migration
**IPC**: Integrated Food Security Phase Classification
**IPR:** Intellectual Property Rights
**ISSP**: International Social Survey Programme

**MTU**: Munster Technological University
**MVC**: Model, View and Controller
**NGO**: Non-Governmental Organisation
**OCC**: Open Cultural Center
**OECD**: Organisation for Economic Co-operation and Development
**OIT**: Oxfam Italia Onlus
**OSM**: OpenStreetMap
**REIGN**: Rulers, Elections, and Irregular Governance
**RIA**: Research and Innovation Action
**SOAP**: Simple Object Access Protocol
**SPI**: Standard Precipitation Index
**TB**: Terabyte
**ToR**: Terms of References
**TRC**: Terracom AE
**UAB**: Universitat Autònoma de Barcelona
**UB**: Users Board
**UCDP**: Uppsala conflict Data Program
**UN:** United Nations
**UNDP**: United Nations Development Programme
**UNHCR**: United Nations High Commissioner for Refugees
**UNODC**: United Nations Office on Drugs and Crime
**URL:** Uniform Resource Locator
**WDI**: World Development Indicators
**WP**: Work Package
**XML:** Extensible Markup Language

# 1. Introduction

The scope of this deliverable is to describe the overall architecture of the EMT and to provide a blueprint of the entire system. The primary objective is to provide accurate migration flow predictions, to equip practitioners and policy makers involved in various stages migration management with adequate evidence and to propose solutions for reducing potential conflict/tensions between migrants and EU citizens, by considering a wide range of human factors and using multiple sources of information. Technological partners will identify and define the core functionalities and specifications of the EMT, based on the end-user requirements defined in D7.1 and designated data sources instructed by the partners responsible for WPs related to data collection.

Regarding the first stage, EMT will develop several comprehensive models of mixed migration (labour related and refuge seeking), which will take both the origin and the receiving countries perspective and cover multiple countries at both ends. Countries of origin include Venezuela, Mali, Nigeria, Morocco, Iraq, Afghanistan, Syria and Eritrea. Regarding the destination countries, EMT will focus on Spain, Greece, Italy, Sweden, Netherlands, Poland and Germany. The predictions will be validated in at least three specific EU Member States, namely Spain, Italy and Greece. In relation to datasets, this deliverable states the data sources and datasets that will be used at this early stage of development according, to the Data Management Plan (deliverable D1.1). The data sources are further described and updated in revised deliverable D1.1, as well as are updated on the ITFLOWS website.

The deliverable is structured and organised as follows:

**Chapter 1:** is the introduction of this report, outlining the scope of this document and the relation to other deliverables.

**Chapter 2:** provides an overview of the datasets that will be given as input to the EMT to produce reliable migration predictions. It describes the origin of data and the allocation in WPs.

**Chapter 3:** presents the end-user requirements that will be translated to formal specifications for the EMT and will lead to the design of its architecture.

**Chapter 4:** presents the approach adopted for the design of the architecture of the EMT system and outlines the key principles that have been followed for setting up

the conceptual architecture and delivering the detailed specifications of the software modules that compose the EMT system.

**Chapter 5:** describes a conceptual architecture of the ITFLOWS EMT system, which specifies its theoretical models identified firstly in the ITFLOWS concept.

**Chapter 6:** exhibits the conceptual system architecture analysis conducted here, capturing and analysing the fundamental components of the EUMigraTool. The primary objective is to map the EMT initial concept into a high–level pipeline, which is further broken down into explicit functional components. These functional components comprise the basic software components of the EMT.

**Chapter 7:** studies the static structure of the EMT. It provides a detailed analysis of the EMT software architecture, while it also provides specific details for the feature extraction process implemented.

**Chapter 8:** presents the implementation view of the proposed architecture. In accordance with the literature -relevant to the design of system architecture- the implementation view is decomposed into the development and physical view of the examined system.

**Chapter 9:** provides a few real-life examples of how the EMT will be used.

**Chapter 10:** provides a brief discussion upon the entire document, and conclusions are drawn.

## 1.1. Glossary

| Term | Definition |
|---|---|
| **Source Country / country of origin** | Country from which migrants start their journey. |
| **Destination country** | Country in which migrants end their journey. |
| **Attitude** | General evaluation of a situation or particular entity or person. |
| **Migrant integration** | The process of accepting and including migrants into society. |
| **Migration flows** | The number of migrants entering or leaving a country (source or destination). |
| **Large-scale model** | Model that includes several connected countries, a continent or even a global scenario |
| **Small-scale model** | Model that focuses on a country of conflict and neighbouring countries (primary displacement) |

# 2. Data sources

## 2.1. Migration drivers and intentions: origin, transit, and destination countries

This section outlines the aims of the analyses to be carried out in tasks T3.1-T3.4 and briefly explains which data sources will be used by different tasks/deliverables (see also Table 1). Further details regarding the data sources and variables to be used by D6.1, D6.2, D6.4 and D6.3 can be found in **Appendix I.**

Task T3.1 and task T3.4 constitute the qualitative dimension of the analysis. The two tasks are closely linked and complement one another. Our approach is premised on seeing migration processes and outcomes as a function of the mutually constitutive interaction between structure and agency. [1] Task 3.1 focuses on how macro- and meso-level structural factors in the contexts of origin, destination and transit shape the broader context in which (EU-bound) migration decisions are made, while accounting for the effects of relevant EU policies. Task 3.4, through conducting and analysing a total of 90 interviews with migrants and asylum seekers in Greece, Italy and Spain, seeks to have a better understanding of how individuals from different national contexts—but also with different backgrounds in terms of race, class, ethnicity, sexual orientation and religion— perceive and interact with these structural factors in making decisions about where, when, and how to move. The analysis will also shed light on migrants' and asylum seekers' experience upon arrival to the three EU countries.

In both tasks, we focus on mixed migration from the Middle East and South-Central Asia along the so-called Eastern Mediterranean Route, from West, East and North Africa along the Central and Western Mediterranean Routes, and from South and Central America along the Atlantic Air Route. Within these regions, Task 3.1 will carry out in-depth case studies on selected countries of origin, namely, Afghanistan, Syria and Iraq; Eritrea, Nigeria, Mali, Morocco (and possibly also Tunisia); and

[1] De Haas, H. (2011), "The determinants of international migration. Conceptualising policy, origin and destination effects.", Working Paper no. 32. Oxford: International Migration Institute.
Bakewell, O. (2010) "Some Reflections on Structure and Agency in Migration Theory." *Journal of Ethnic and Migration Studies* 36 (10): 1689–1708.

Venezuela, Colombia and Honduras. Due to the fragmented (Collyer 2010) and non-linear nature of mixed migration flows reaching the external borders of the EU,[1] these case studies will go beyond a mere focus on countries of origin, and focus on the entire migratory complex linking origin countries, contexts of transit (or alternative destinations before the migratory process is resumed with the aim to move towards the EU), and EU destinations. Inquiring into drivers across different contexts in which the migratory process unfolds, the analysis aims to identify which different "configurations" of structural factors[2] are relevant for explaining patterns and trends of mixed migration from specific contexts of origin to specific contexts of (EU) destination, while shedding light on how journeys connecting these two contexts are shaped. In Task 3.4, the bulk of the interviews will be conducted with migrants and asylum seekers (proportionally taking into account women in the research participant selection) from these countries as to complement Task 3.1 analysis focusing on structural factors. A smaller number of interviews with other nationalities with growing relevance in terms of irregular arrivals and asylum applications in the EU will also be conducted.

In terms of data sources, Task 3.1 will use statistics provided by the European Border and Coast Guard Agency (FRONTEX)[3] on irregular crossing by migration route and data provided by the Directorate General of Eurostat – European Statistics on asylum applications (both broken down by nationality) so as to look into the evolution of irregular arrival and asylum application trends over the past ten years (taking into account gender disaggregation, to the extent possible). In addition to identifying visible peaks and drops in in the volume of migrants and asylum seekers arriving to the EU from the origin countries under consideration, this will also allow us to identify shifts in the main routes/entry points used by migrants and asylum seekers from particular nationalities at concrete points of time. As for the transit

[1] Crawley, H., F. Duvell, K. Jones, S. McMahon and N. Sigona. 2017. Unravelling Europe's 'Migration Crisis': Journeys Over Land and Sea. Policy Press Bristol University Press; McMahon, S. and N. Sigona. 2018. Navigating the Central Mediterranean in a Time of 'Crisis': Disentangling Migration Governance and Migrant Journeys. Sociology 52 (3): 497-514.
[2] Van Hear, N., O. Bakewell, and K. Long. 2018. Push-Pull plus: Reconsidering the Drivers of Migration. *Journal of Ethnic and Migration Studies* 44(6): 927–4
[3] European Border and Coast Guard Agency (FRONTEX), "Migratory Map: Detections of illegal border crossings statistics download (updated monthly)"

contexts, secondary data, i.e., reports and situation analyses shedding light on the dynamics and trends along major migration routes towards Europe (e.g., the IOM (International Organization for Migration)'s Displacement Tracking Matrix, Mixed Migration Centre's 4mi data, Frontex Risk Analysis reports, or field-based reports by the Global Initiative Against Transnational Organized Crime and United Nations Office on Drugs and Crime - UNODC) will be used to identify significant shifts in terms of routes, hubs, migrant smuggling dynamics as well as patterns of border crossing along the journeys towards Europe. As for the analysis of how macro- and meso-level structural conditions change and combine across contexts of origin, transit, and destination (and how relevant EU policies impact on these structures) so as to shape migration patterns and trends in the last decade (to be identified in the first step), we will rely on process tracing.[1] Data sources will include source-country and route-specific academic and grey literature (e.g., reports by relevant UN Agencies, European Asylum Support Office - EASO country of origin reports), as well as relevant policy documents (mainly from the EU, also from source/transit countries).

As task T3.1 will mainly rely on qualitative data, i.e., academic literature, reports, situation analyses, and policy documents, those sources are not expected to be directly introduced into the EMT. Although this kind of textual data might also limited use for the modelling purposes of the EMT, it could still provide relevant information: the analysis of how macro- and meso-structures in the contexts of origin, destination and transit acts in tandem to shape the wider context in which (EU-bound) migration decisions are made, and how EU policies and those in relevant transit contexts impact on the opportunity and constraint structure faced by (potential) migrants at different stages of the migratory process. Therefore, qualitative data will provide context-specific and in-depth background information necessary to capture by large-scale quantitative data, which can be ultimately useful for nuancing the model. Task 3.4 will not use secondary data but will produce primary qualitative data through the semi-structured interviews. The interviews

[1] Venesson, P. (2008). Case studies and process tracing: Theories and practices. In D. Della Porta and M. Keating (eds.), Approaches and methodologies in the social sciences: A pluralist perspective. Cambridge: Cambridge University Press.

will be transcribed, anonymised, and qualitatively analysed on the basis of a shared codebook. Due to data protection requirements, access to original transcripts will be kept limited to the partners collecting the data (CRI, OCC, OIT) until the end of the project, while access to anonymised transcripts will be kept limited to the partners conducting the analysis (IAI and UAB).

Task T3.2 focuses on irregular mixed migration flows, looking at flows to the EU and into neighbouring countries. Irregular mixed migration flows to the EU usually entail fragmented journeys across several countries, which conditions could affect movement of people along the route. For instance, violence and political instability, as well as changing weather conditions and mass disasters could affect the practicability of routes. Similarly, increasing instability in countries of origin can trigger sudden displacement of people, both in neighbouring countries and towards the EU. D3.1 will therefore focus on irregular mixed migration flows to the EU to try to disentangle the role played by conditions in transit countries in shaping migration flows. D3.4, instead, will look at movement into neighbouring countries and therefore focus on first-time displacement triggered by growing instability in countries of origin.

Statistics on irregular crossing by migration route from FRONTEX will be used in D3.1, covering all arrivals detected since 2009 with a monthly frequency. The high dimensionality of the FRONTEX dataset (i.e., by citizen and route with monthly frequency) allows to control for the different situations across the main five migration routes to the EU. In addition, being interception-based, these figures capture with certainty the time of arrivals of migrants, therefore making them suitable to try disentangling the impact of changing situations in transit countries on the development of migration flows to the EU.

In D3.4, instead, UNHCR Operational Portal Refugee situations will be used for a selection of relevant countries and regions (e.g. Syria, Horn of Africa, South Sudan, Venezuela). UNHCR Operational Portal Refugee in fact provides granular information on the number of crossing from key origin countries (of international displaced people) into neighbouring countries. Depending on data availability of

each specific refugee situation, figures are available either at a daily or monthly frequency and usually broken down by each neighbouring country. The high frequency of observations will therefore help understand the time-structure of first-displacement into neighbouring countries trigged by growing instability in origin countries.

Both deliverables will therefore take a quantitative approach and rely on longitudinal datasets with a high time frequency to regress migration flows against a set of explanatory variables. The aim is to assess whether an increase in instability (e.g. growing number of conflicts or mass disasters) leads to an increase in irregular mixed migration flows. When looking at flows to the EU, the high dimensionality of the dataset based on FRONTEX data will allow to control not only for the situation in the country of origin at different point in times, but also for the situation along the transit routes.

Event-based data on both conflicts and mass disasters will provide the number of events as well as the number of deaths/casualties in each month: the Georeferenced Event dataset from the Uppsala conflict Data Program (UCDP) and the Emergency Events Database (EM-DAT) from the Centre for Research on the Epidemiology of Disasters (CRED) are the two main data sources used for conflict and disasters respectively. The regression analysis will therefore assess whether an increase in the number of conflict events and mass disasters has a significant effect in affecting migration flows, and whether this effect is positive or negative.

In addition, other contextual variables capturing level of political stability will be sourced from the Rulers, Elections, and Irregular Governance (REIGN) dataset, which provides monthly information on several political dimensions, focusing not only on the leader of each country, but also on the overall political system: e.g. type of government, risk of *coup d'état*, number of months since an irregular election took place. For instance, an increase in the risk of *coup d'état* indicates growing political instability, either in the country of origin or in the transit route, that might lead more people to migrate or might affect migration networks across transit routes. Moreover, the relationship between the risk of *coup d'état* and migration flows

might differ from that with, for instance, mass disasters and conflict events, both in terms of magnitude and timing of their effect on migration. A mass disaster might in fact push people to migrate immediately due to a sudden and significant change in their living conditions, while an increase in the risk of *coup d'état* might have a delayed effect if, for instance, the risk is not immediately "perceived" by the population or if it does not materialise in a concrete political crisis.

The REIGN dataset provides also the Standard Precipitation Index (SPI) for each country and month, for which values of 0 correspond to historically average levels of rainfall. The SPI will then be complemented by the FAO index on Temperature Change to capture worsening weather conditions in origin countries and transit routes and understand their possible effects on migration flows. Months with exceptionally lower or higher levels of precipitation and/or temperature might affect migration flows compared to a "standard" month. Once again, relationships with migration flows might differ depending on whether these changes in weather conditions concern countries of origin or transit, as well as if they are limited to a single month or if they concern longer periods.

Finally, different control variables capturing the socio-economic conditions of origin and transit countries will be used (e.g. GDP, unemployment rate, trade openness, government expenditure on health). The World Bank Development Indicators (WDI) is the main data source used; however, being that WDI reflects only an annual basis, monthly values will be estimated using temporal disaggregation techniques. Other specifications will also test alternative indicators at a monthly frequency sourced from IMF, as the CPI, interest rates (e.g. deposit and saving), exchange rates.

**As ITFLOWS is still in an early development phase, more details on the exact use of these data sources and the relevant methods will be provided in the WP3 specific deliverables, as well as in D6.2.**

Task T3.3 is divided into two parts. The goal of the first one is to follow Böhme et al.'s (2020)[1] approach of using Google Trends search indices for migration-relevant

[1] Böhme, M., A. Gröger, and T. Heidland (2020): "Searching for a Better Life: Predicting International

keywords in origin countries to (i) create a real-time measure of bilateral migration intentions to EU destinations, and (ii) practically implement short-term k-step ahead forecasting of refugee arrivals to the EU using Google Trends as a leading indicator.

The data used in this task can be categorised in terms of the dependent variable (to be predicted), "classical" migration predictors (push- and pull-factors), and the Google Trends Indices. The resulting dataset features monthly frequency at the bilateral migration flow level between origin and destination countries. The dependent variable used is provided by EUROSTAT and captures asylum applicants as well as first-time asylum applicants, respectively.

To proxy for origin country push-factors of different dimensions, we rely on violent conflict indicators from ACLED, political events from REIGN and ELVI, agricultural production from FAO, natural disaster events from EM-DAT, labour market conditions from ILO, and macroeconomic performance from IMF. In terms of EU destination country pull-factors, we include the same data sources as for origins plus a set of detailed economic production indicators from EUROSTAT.

Finally, we use Google Trends for around 200 different topical migration-related keywords in combination with all EU destination country names (e.g. "asylum Germany"). We cover a diverse set of origin countries worldwide using 10 official local languages in which we extract Google Trends (i.e. Arabic, Dari, English, Farsi, French, Hausa, Pashto, Portuguese, Spanish, Turkish).

The econometric approach employed in the forecasting exercise remains to be determined at this stage. We plan to experiment with different linear and non-linear models as well as variable selection procedure motivated by recent advances in machine learning.

Migration with Online Search Keywords", Journal of Development Economics, vol. 142, 102347; see also Golenvaux, Nicolas, et al. "An LSTM approach to Predict Migration based on Google Trends." *arXiv preprint arXiv:2005.09902* (2020).

As mentioned above, this work is still in its early stages, thus no further details can be provided at this time. WP3 deliverables and D6.1. will provide a more detailed view on the data and the methods to be used.

The second part of T3.3 aims to identify *push* factors informing migration from the source countries, in order to assist in predicting the migration flows into EU countries. It will achieve this by using Twitter-based data (see Section 6.4). As ITFLOWS is a Research and Innovation Action (RIA) project, with a targeted TRL 6, the Consortium selected Twitter as the social media of preference to be implemented, due to its API availability and existing experience from the Consortium. This work can be extended to other/more social media platforms, once/if the project develops to a commercial solution.

**Table 1: Open data sources regarding origin, transit and destination country data**

| Source | Link | Geographical coverage - Destination | Geographical coverage - Origin | Related Tasks/Deliverables |
|---|---|---|---|---|
| **Eurostat** | https://ec.europa.eu/eurostat/data/database | Europe | na | T3.1 (D3.2); T3.3 (D3.6) |
| **FRONTEX** | https://frontex.europa.eu/along-eu-borders/migratory-map/ | migration routes (e.g. Western, Central, Eastern Med) | worldwide | T3.1 (D3.2); T3.2 (D3.1) |
| **ACLED** | https://acleddata.com/#/dashboard | worldwide | na | T3.3 (D3.6) |
| **EMDAT** | https://www.emdat.be/ | worldwide | na | T3.2 (D3.1 - D3.4); T3.3 (D3.6) |
| **WDI** | https://databank.worldbank.org/source/world-development-indicators | worldwide | na | T3.2 (D3.1 - D3.4) |

| | | | | |
|---|---|---|---|---|
| **IMF** | https://data.imf.org/?sk=4FFB52B2-3653-409A-B471-D47B46D904B5 | worldwide | worldwide | [Potentially] T3.2 (D3.1); T3.3 (D3.6) |
| **Google Trends** | https://trends.google.com/trends/?geo=US | worldwide | worldwide | T3.3 (D3.6) |
| **Rulers, Elections, and Irregular Governance (REIGN) dataset** | https://oefdatascience.github.io/REIGN.github.io/menu/reign_current.html | worldwide | worldwide | T3.2 (D3.1 - D3.4); T3.3 (D3.6) |
| **ILO** | https://ilostat.ilo.org/data/ | worldwide | worldwide | T3.3 (D3.6) |
| **UCDP - Georeferenced Event Dataset** | https://ucdp.uu.se/ | worldwide | worldwide | T3.2 (D3.1 - D3.4) |
| **FAOSTAT** | http://www.fao.org/faostat/en/#data/ET | worldwide | worldwide | T3.2 (D3.1 - D3.4); T3.3 (D3.6) |
| **UNHCR - Operational Portal** | https://data2.unhcr.org/en/situations#_ga=2.231374239.1479040944.1599118541-1126213969.1593164972 | selected countries for each specific situation | selected countries for each specific situation (e.g. Syria, Somalia, South Sudan) | T3.2 (D3.4) |
| **ELVI** | https://oefdatascience.github.io/REIGN.github.io/menu/elvi_current.html | worldwide | worldwide | T3.3 (D3.6) |

## 2.2. Destination countries

### 2.2.1. Data on Integration of migrants

This section provides the sources for comparative electronic reports (feeding on databases) on the socio-economic situation in the project's focus Member States (Spain, Greece, Italy, Sweden, Netherlands, Poland and Germany). These reports in notebook format, i.e. feeding on live data instead of being a static .pdf document. The reports will consist of visualisations based on the most recent data in the database underlying the EMT and interpretations or explanations of these data for the user.

The necessary data for T4.2 comes from public sources. To ensure that indicators are standardised, available for all EU Member States, and are in line with the data that other EU-funded activities use or are in keeping with the work in different DGs, we use Eurostat data. This data has the advantage of usually being comparable across countries. When available, data series that are compiled by the national statistical offices or through other standardized formats such as the EU Labour Force Surveys, these data are collected centrally and standardized by Eurostat. Their platform thus provides vast amounts of indicators that can be used to compare country differences. Since this is the most important data requirement T4.2 has, Eurostat is our main data source. For individual countries there are potentially additional national data as well. However, lacking comparability across countries, these are not useful for the reports compiled in T4.2. Also, they will be of limited usefulness for prediction models as in WP3 in WP5.

Some subtasks of the work package such as D4.1 will rely on non-Eurostat data. For example, D4.1 will use individual level panel data from Germany, the Demographic and Health Surveys and the proprietary Gallup World Poll). These are neither relevant for EMT nor available for sharing there, as they require individual applications or even payment for data access. In line with data security requirements, those subtasks in WP4 requiring other data shall operate based upon the use of separate, local IT infrastructures and will therefore not provide data to the EMT. They will however share results based on analyses of these data such as reports in traditional and static formats such as .pdf via the EMT platform.

The Eurostat indicators on the integration conditions cover basic demographics,

social and economic aspects as well as migrant-specific measures.

Factors such as the population size, the age structure, education levels, unemployment, and poverty rates are well known in the academic literature to include migration flows (see WP3) or attitudes towards migration (see WP5). These factors may affect both migrants' potential for economic integration, create push and pull effects if these are not already captured by the WDI (WP3), and differences in attitudes towards migrants (WP5) that could affect social integration and which are particularly relevant for the multilevel analyses in WP5 (WP5). Further indicators capture amenities and access to public goods as well as the overall satisfaction with life and institutions.

In addition to these, we include asylum-specific indicators, which matter for the project both as outcomes variables (e.g. inflows of asylum seekers) or as indicators of the context in which asylum seekers or recognized refugees live (e.g. their number relative to the local population). These indicators are also provided via Eurostat and are provided by the different member countries agencies or statistical offices. These include breakdowns of asylum applicant numbers by different demographic characteristics and factors that can directly indicate integration, such as numbers of first-time residence permits.

**Table 2: Data sources for standardised economic and social integration conditions**

| Source | Link | Geographical coverage - Destination | Geographical coverage - Origin | Related Tasks/Deliverables |
|---|---|---|---|---|
| **Eurostat** | https://ec.europa.eu/eurostat/data/database | Europe | n.a. | T4.2 (D4.2); T5.2 (D5.2); T5.3 (D5.3) |

### 2.2.2. Public attitudes to migration data

This sub-section outlines the analyses to be carried out in WP5 (Tasks T5.1, T5.2, T5.3) and briefly explains the data sources that will be used by the different tasks/deliverables (see also Table 3).

Task T5.1 provides a multi-disciplinary meta-analytical overview of the state of knowledge regarding the most prominent micro- and macro-level factors affecting attitudes to immigration from the past 10 years. The task is based on a review of the empirical literature in five social science disciplines (economics, sociology, political science, psychology and migration studies).

The first step of the task is to identify the relevant journals for each discipline. The aim of this meta-analysis is not the selection of all studies, or a representative selection of all studies regarding attitudes to migration, but rather the selection of "best" studies published in the last decade regarding factors affecting attitudes to immigration. Published work in the top-ranked academic journals has gone through the process of rigorous peer review and is therefore supposed to be of high quality and report more reliable results. We have chosen top-ranked journals in each of the five disciplines as a combination of the journals ranking in Google Scholar, Scimago Journal and Country ranking and Clarivate Analytics and rankings.

The sample of articles for the meta-analysis is selected based on a search for all articles fulfilling the established criteria directly in the peer-reviewed journals' own database, within the publication time frame 2009-2019. Admittedly, the ten-year timeframe is not based on a rigorously defined pre-established criterion but has been selected as an ad hoc reflection of recent developments in the empirical research of attitudes toward immigration. The first pre-selection was done by at least two independent coders, who used the search terms "immigrant" and "immigration". In cases where they disagreed, a third coder assessed the inclusion of the article. This procedure created a list of articles published in top thirty journals in each of the five disciplines (150 academic journals in total) that investigate factors affecting attitudes to immigration (this led to a selection of about 800 academic articles in total). Attitudes to immigration can refer to many attitudes and we have identified roughly around 150 different types of attitudes to immigration in the academic literature. These can be roughly grouped into 10 higher-ordered groups of attitudes, which range from attitudes towards allowing immigrants the same rights and opportunities as citizens, policy preferences regarding immigration, feelings towards immigrants, attitudes regarding immigrants´ contribution to

society, and prejudice towards immigrants. To investigate each of these different types of attitudes is beyond the scope of the project. We, therefore, look at two groups of attitudes, which we consider the most relevant for the ITFLOWS project.

First, we look at immigration policy preferences, which are usually questions asking respondents to clarify whether they believe their country should allow more or less (unskilled, labour, Muslim, Jewish etc.) immigrants to come to the country. This concept engages with policy debates about levels of immigration as well as with the criteria for entry, such as debates around the introduction of points systems that privilege potential migrants with higher skills and/or language skills. Qualifications for entry can also be conceptualised as varying according to acquired and ascribed immigration criteria. Acquired immigration criteria consist of those individual competencies and attitudes (such as commitment to the way of life of the destination country) that in principle immigrants could attain if they wish. Ascribed immigration criteria, in turn, are categorical qualities related to inherent, collective characteristics of a social category (such as being of a certain race). This distinction between ascribed and acquired characteristics mirrors the classic distinction made in the literature between ethnic and civic conceptions of the nation. Second, we look at attitudes regarding the "contribution of immigrants to the society". These are mostly attitudes regarding the effect of immigration on society and whether the respondents believe that immigration is beneficial to the community (in terms of the economy, culture etc.). While the first set of attitudes mostly deals with attitudes regarding the future, the second group of attitudes deals with the present effects of immigration. In this sense, the two sets of attitudes complement each other. Moreover, it is plausible that factors affecting attitudes regarding the effects of immigration might affect attitudes to immigration policy differently and thus it is important to analyse more than one set of attitudes.

The criteria for the final selection of the dependent variables (attitudes to immigration) together with a detailed protocol describing the inclusion criteria (e.g. how attitudes toward immigration are defined, the unit of analysis for the dependent variables etc) will be described in detail in the deliverable D5.1.

After selecting the studies with our dependent variables of interest, we report the most commonly included variables in models explaining immigration attitudes. We analyse results from previous empirical studies regarding these factors and summarize the findings via quantitative methods, the so-called meta-analysis.

We therefore create a dataset which summarises the results of previous studies for each of the socio-demographic factors affecting attitudes to immigration. This unique and original dataset is then further analysed and allows us to determine whether a micro- or macro- level factor is indeed found to be significantly affecting attitudes to immigration based on recent empirical research.

The factors we highlight and analyse as the most prominent in influencing attitudes to immigration are the following, as shown in *Table 3* below; at individual level, we concentrate on a series of sociodemographic variables such as age, education, gender, place of residence (rural versus urban) and being a minority (for instance, the respondent is a non-citizen of the country he resides in or is a member of an ethnic minority), on a series of economic factors such as income, being (un)employed, social class, type of occupation (high skilled versus low skilled), economic satisfaction at the individual level (example questions "how satisfied are you with your economic situation" or "are you facing income difficulties?"), economic satisfaction at the national level (example question "how satisfied are you with the current state of national economy"), and other individual characteristics such as the level of religiosity, left-right political self-identification, ideology (liberal versus conservative), contact with minority (respondent has minority friends or contact with minority) and interpersonal trust. We also investigate how factors at the regional as well as country level might affect individual attitudes to immigration. We particularly focus on how regional/country unemployment level, regional/country GDP per capita and regional/country minority share affect individual attitudes to immigration.

**Table 3. Factors affecting attitudes to immigration studies as part of D5.1**

| Individual level | Contextual level |
|---|---|
| Age | Regional unemployment level |
| Education | Regional GDP per capita |
| Place of residence | Regional minority share |
| Minority background | Country unemployment level |
| Income | Country GDP per capita |
| Social class | Country minority share |
| Employment status | |
| Type of occupation | |
| Economic satisfaction (personal) | |
| Economic satisfaction (national) | |
| Religiosity | |
| Left-right political orientation | |
| Ideology | |
| Contact with minority | |
| Interpersonal trust | |

As Task T5.1 (and the relevant deliverable D5.1) relies on academic journal articles as the main data source, those sources are not expected to be directly feeding the EMT,. Nevertheless, based on Task 5.1, which identifies the relevant factors affecting attitudes to immigration *globally* and is based on statistically representative samples from all over the world, Task 5.2 applies the knowledge obtained from T5.1 specifically into the European context. Task 5.2 and deliverable D5.2 provide an original contribution by testing whether factors identified globally as the most important in affecting individual attitudes to immigration also apply specifically to the European context. Therefore, Task T5.2 relies on an original analysis of European citizens' attitudes to immigration and provides a direct answer to the question what the main factors are contributing to differences in attitudes to immigration between individuals within countries as well as between European countries.

In order to investigate European citizens' attitudes to immigration and immigrants

we will possibly make use of several European academic surveys, such as the European Social Survey (ESS) data, Eurobarometer, European Values Study (EVS) and International Social Survey Programme (ISSP) data. The European Social Survey is an open-access high-quality academic survey collected in a number of European countries bi-annually since 2002 providing information of several types of attitudes to immigration. The Eurobarometer data has been collected by Eurostat since 1974 in EU28 countries plus candidate countries, and is also publicly available. The European Values Study is a cross-national and longitudinal survey conducted in several waves. Finally, the International Social Survey Programme (ISSP) is a cross-national collaboration programme conducting annual surveys on diverse topics relevant to social sciences. Especially the European Social Survey is one of the highest quality attitudinal surveys regarding attitudes of European citizens using a rigorous methodology and providing a reliable and unified source with regards to attitudes in many European countries.

These attitudinal surveys will be combined with data sources on unemployment, GDP per capita and minority share at regional and country level for the destination countries. This information is planned to be taken from Eurostat (unemployment, minority share) and OECD data (unemployment, GDP per capita). After running a series of multilevel regression models for a number of European countries, deliverable D5.2 will empirically identify factors affecting individual attitudes to immigration in Europe. For instance, we will be able to determine whether younger or older respondents are more likely to feel negatively towards immigration, whether those individuals with higher levels of education are more pro-immigration compared to those with lower levels etc. Similarly, we will be able to identify whether individuals living in regions and countries with higher share of immigrants are more or less likely to feel positive about immigration compared to those with lower shares etc.

End users been consulted on design and functionality during the proposal stage and also in the first EMT workshop at M6 (January 2021). For the final EMT tool, this aims to allow users to look at various demographic distributions (e.g. age, educational levels, gender, rural/urban divides etc.) of the population in destination

countries and classify regions and countries as more or less likely to view immigration positively. Thus, for the final EMT tool, the selection of data sources regarding demographic distribution are crucial. At the current state of developments, Eurostat and OECD are considered the most promising data sources.

**Table 4: Data sources for standardised economic and social integration conditions**

| Source | Link | Geographical coverage - Destination | Geographical coverage - Origin | Related Tasks/Deliverables |
|---|---|---|---|---|
| **Eurostat (unemployment, minority share, age distribution, educational levels, gender distribution and rural-urban divides of the population)** | https://ec.europa.eu/eurostat/data/database | Europe | n.a. | T5.2 (D5.2); |
| **European Social Survey** | https://www.europeansocialsurvey.org | Europe | n.a. | T5.2 (D5.2); |
| **Eurobarometer** | https://www.gesis.org/en/eurobarometer-data-service/search-data-access/eb-trends-trend-files | Europe | n.a. | T5.2 (D5.2); |
| **European Values Study** | https://dbk.gesis.org/dbksearch/GDESC2.asp?no=0009&DB=E | Europe | n.a. | T5.2 (D5.2); |
| **International Social Survey Programme** | http://www.issp.org/data-download/by-topic/ | Europe | n.a. | T5.2 (D5.2); |
| **OECD data (unemployment, GDP per capita, age distribution, educational levels, gender distribution and rural-urban divides of the population)** | https://data.oecd.org/economy.htm | Europe | n.a. | T5.2 (D5.2); |

*All the above data sources are publicly available.*

## 2.3. Big Data Sources

Big Data is the broad term used for data sets overly complex or large that conventional data processing applications are insufficient. This term has gained popularity in the present and it is used sometimes to define the exponential data development and availability, both structured and unstructured. In this project, we use machine learning algorithms to extract helpful insights and statistical trends from big data sources like GDELT, Google Trends and Twitter. Big data, when used with relevant machine-learning methods, can provide analysis and insights that are otherwise impossible to retrieve. This way, the EMT will gain significant advantage in providing information to its end-users.

### 2.3.1. GDELT Project

The GDELT Project is an initiative constructing catalogues for human societal-scale behaviours and beliefs from countries all around the world. It also includes catalogues and data regarding news sources, events across the world and their context. GDELT includes data from 1979 to the present. Some older datasets are available in a yearly and monthly granularity, while the newest datasets are being updated every 15 minutes. Data files record events by using Conflict and Meditation Event Observations (CAMEO) coding.

The database is one of the highest-resolution inventories of the media systems of the non-Western world and operates in near real time. It is described as a key for developing technology that studies the worlds society.

The GDELT database provides 15 minutes updates, real-time translation of news written in 65 different languages, Real time measurement of over 2,300 emotions and themes, Relevant imagery, videos and social embeds, Quotes and Event discussion progression.

The GDELT Project offers numerous datasets but EMT will be, mainly, using the following, after a discussion with GDELT representatives as well as internal consortium review of GDELT database:

- GDELT 2.0 Event Database (GDELT Master), which by itself includes:
  - The GDELT Global Knowledge Graph (GKG) and

- The GDELT Mentions CSV dataset.
- The GDELT Global Quotation Graph (GQG) and
- The GDELT Global Relationship Graph (GRG).

Data from GDELT will be downloaded into the EMT's data repository (and updated accordingly), and be fed to the relevant algorithms.

The *GDELT Global Knowledge Graph*, includes themes for reporting economic indicators like price grouping and heating oil price for infrastructure topics and social issues like marginalization and burning in effigy. It includes lists of recognized infectious diseases, ethnic groups and terrorism organizations and, in the 2.0 Database, there have been added more than 600 global humanitarian and development aid organizations. .

The *GDELT Global Quotation Graph*, compiles quoted statements from news all around the world. It scans each article monitored by GDELT and compiles a list of all quoted statements within, along with sufficient context to allow users in many cases to establish speaker identity. This dataset covers 152 languages with minor limitations in capturing some quotes. Each quote is supplemented with a fragment of text before and after the quotation. The dataset is updated every minute but is generated every 15 minutes for download.

The *GDELT Global Relationship Graph,* contains the assertions and relationships made in global press every day. The dataset, ultimately, has rea time updated verb-centred ngrams. The articles are using part-of-speech tags, which may cause misclassification is some cases. Each verb is accompanied by up to 6 tokens before and after the verb. The ngrams are only generated around verbs, creating a fixed context around each verb phrase, capturing the statements of action related in the article. Like the GQG dataset, GRG is updated every minute but is generated every 15 minutes for download.

GDELT offers a huge database and the size of the data is too much to download it all. For reference and according to GDELT's web site, the size of just GKG alone is 2.5TB for a single year. That is why there are written Python programs that will download

exactly which data is required based on the name of the dataset and the date. After downloading from GDELT, the data will then be uploaded to the ITFLOWS CKAN repository as further explained in the upcoming sections.

### 2.3.2. Google Trends

The Google Trends platform provides timebound data about relative search intensities for (any combination of) specific keywords, the so-called Google Trend Index. We follow Böhme et al. (2020) in using these indices for migration-related keywords (such as "visa" or "immigration") and destination-related terms (such as "Germany" or "Spain") as a proxy for migration intentions. The resulting time series thus provides relative search intensities for specific migration and destination queries, reflecting migration intentions with a bilateral (origin-destination) dimension.

We work with an extended list of keywords of approximately 200 unique search terms which represent expressions related to migration, link to a list of European destination countries. We extract the data at monthly frequency for the major origin countries of migration to the European Union. We use keywords in local languages of the respective origin country and cover the following languages: Arabic, English, Farsi, Fula, French, Hausa, Pashto, Portuguese, Spanish, Turkish. These trends are then merged with data on bilateral migration flows from Eurostat and a range of different control variables (see appendix).

# 3. User requirements

## 3.1. Methodology

To extract the User Requirements for the EMT application the report of the end user workshop that took place on 20 January 2021 under the scope of WP7 had to be carefully read and analysed. In deliverable D7.1 as part of WP7, the project was presented to the end users that include various NGOs across the destination countries mentioned in previous sections. Through the user's board first workshop, end users were able to give feedback and make some points that are required for the EMT to work properly and as they were expecting it to be. The workshop produced valuable input and feedback that will help EMT emphasise on some specific aspects the end users consider important and essential. The results of the workshop are available in deliverable D7.1.

In order to move from the users' input to the requirements, the results of the workshop were analysed and restructured in the form of the user requirements' tables presented in the next section.

The consortium's aim was to define the user requirements in such a way that the input received during the workshop (D7.1) was fully covered, considering that the workshop participants are the potential users of the EMT and thus their requests need to be implemented as much as possible, while at the same time abiding by legal and ethical guidelines, in the way data is received and processed within the EMT (e.g. making sure that no identifiable data is used within the EMT). EMT will comply with Privacy by Design (PbD) principles which will be analysed on the next chapter.

The devised user requirements will further be analysed and transformed into system/technical requirements over the next period, with the later guiding the developments of the EMT.

## 3.2. User Requirements

At the current stage of the project, all the following requirements are in an "under investigation" state. They will be further analysed from a technical point of view, and their implementation will be based on the priority indicated in the present document, combined with the required effort for their implementation. An updated list of the user requirements will be provided in deliverable D6.2, along with the first release of the EMT.

| **Title: Strengthening Protection in mixed flux of migrants** |
|---|
| **Code:** Req.01 |
| **Description:** The EMT needs to be able to detect and identify individual needs among migrants prior to their arrival in Europe. |
| **Priority:** High<br>**Justification**: This is particularly relevant for end-users in order to further analyse mixed migration flows and distinguish people in need of International Protection (eligible for asylum status) from those with different migration situation who,-in the absence of legal routes, use of illicit means or methods. |
| **Source:** T3.4, D7.1 |
| **Implementation:** This requirement is met as explained in *Section 6.3*. The Small Scale model is able to identify conflict locations, total number of forcibly dislocated people, camp locations and registration numbers and even simulate links between camps and conflict locations. |

| **Title: Entry Quota and Resources Allocation** |
|---|
| **Code:** Req.02 |
| **Description:** EMT needs to be able to account for entry quote when forecasting migratory flows. Having a real number of migrants arriving to a particular country and region, would help NGOs understand the human effort and material resources that need to be allocated in that particular territory before the arrival. For instance, the information provided by the tool could be suitable for creating a tailored response as far as necessary resources such as blankets, clothing, food etc. |
| **Priority:** High<br>**Justification**: EMT needs to enable NGOs and other organisations to plan and organise their |

| |
|---|
| own or external resources and support the management of the resources dedicated to better accommodate migrants once they finally arrive in the EU. |
| **Source:** D7.1 |
| **Implementation:** This requirement is met as explained in *Section 6.2*. The Large Scale model is able to forecast the exact number and a prediction interval of asylum applications of M migrants from country of origin X to a country of destination Y. |

| **Title: Accurate and updated Statistical Data** |
|---|
| **Code:** Req.03 |
| **Description:** EMT needs to include data in a structural manner from the main international organisations and databases dealing with displaced persons and refugees. Many NGOs have only partial data in an unstructured format, which makes it difficult for them to understand the trends and changes in the routes and drivers. The EMT will provide simple and clear diagrams for NGOs on the main trends of migration flows. |
| **Priority:** High<br>**Justification:** Better understanding of the trends in migration flows in the EU. Also such data is crucial for validation and/or calibration purposes, which will allow to train the algorithms of the EMT during and after the lifetime of this projects |
| **Source:** T3.2, T4.2, T5.1, D7.1 |
| **Implementation:** This requirement is met as explained in *Sections 6.2, 6.3 and 6.4*. All of the models of the EMT will provide data in a structural manner and include visualizations and raw data files that are understandable and easy to work with. |

| **Title: Awareness and Transparency** |
|---|
| **Code:** Req.05 |
| **Description:** EMT needs to be able to improve the awareness and transparency over the arrival of migrants, and to change the narrative about migrant communities. |
| **Priority:** High<br>**Justification:** This is necessary for the NGOs and municipalities in order to prepare resources, and shape social policies (in the case of municipalities) according to the real impact of migrants in the territory. |
| **Source:** D7.1 |
| **Implementation:** This requirement is met as explained in *Section 6.4*. The twitter analysis model is aiming on identifying the attitude towards migration and particular migration trends. |

| **Title: Tensions Defusal** |
|---|
| **Code:** Req.06 |
| **Description:** EMT needs to be able to predict potential tensions and highlight the areas of interest. |
| **Priority:** High<br>**Justification**: This requirement is of a high priority because NGOs and municipalities have pointed out an increasing causality between the negative sentiment towards migration in a territory and the lack of integration by migrants in those territories. |
| **Source:** D7.1 |
| **Implementation:** This requirement is met as explained in *Sections 6.4 and 6.5*. Similarly to the previous User Requirement, the Twitter analysis model identify potential tensions and the general attitudes towards migration in a given country. |

| **Title: User Friendly** |
|---|
| **Code:** Req.07 |
| **Description:** The EMT should contain intuitive, consistent and efficient features for end-users, such as:<br>-Clear and fast registration process<br>-Notifications that will inform users and assist them on the decision process<br>-Personalisation elements (name of the user in the welcome screen, storage of preferred settings)<br>-Ability for users to give feedback in regard with EMTs functionality interface so that it can be further improved based on the user's needs<br>-User roles, depending on the identified roles, users will only have access to features and functionalities that are necessary for their assigned tasks.<br>-Clear and simple design to enhance decision making<br>-Prominent CTA (call to action) buttons for the users to easily spot them and make use of them<br>-It will also include an uncluttered and simple User Interface with clear explanation of input and output data, helping not to confuse the end users. |
| **Priority:** High<br>**Justification**: This requirement is of high priority, because it is significant to be clear to the end users what steps they have to follow in order to achieve their goals. At the same time, it allows the EMT to be up to date and adapted to users' needs, increasing user engagement and offering |

| |
|---|
| a more interactive functionality. Last but not least, different layers of user roles permit the use of data to those ones involved with specific responsibilities avoiding data misuse or overlapping user responsibilities. |
| **Source:** D7.1 |
| **Implementation:** This requirement is met as explained in *Section 6.6 and throughout Section 7.1 and its subsections*. The frontend of the EMT is in keeping with all modern practices for a clear, easy to use and interactive User Interface for its users. The visualizations are easy to understand. Moreover, there are both forums and ticketing systems so that users can interact with each other and the developers of the project. |

| **Title: Components Analysis for Emergency Plan** |
|---|
| **Code:** Req.08 |
| **Description:** The EMT needs to map vulnerability, nationality, age range and gender variability, as these are the criteria that most NGOs need to identify in order to determine that type of assistance to be provided. |
| **Priority:** High<br>**Justification**: EMT needs to have the possibility to draft an emergency preparedness plan knowing migrant nationality, characteristics and vulnerability. |
| **Source:** D7.1 |
| **Implementation:** This requirement is met as explained in *Section 6.2*. The Large Scale model is able to forecast asylum applications of specific nationalities, age groups and gender, as long as this information is provided from the data sources the model is using. |

| **Title: Information misuse prevention** |
|---|
| **Code:** Req.09 |
| **Description:** Data misuse is the use of data and information in a different way that it was originally intended for. The misuse of EMT can entail the closing of borders, deterring violence and even political purpose to gain consensus with anti-migration policy. It is important to prevent that by setting conditions on how data can be used through user agreements, data privacy laws and other regulators. This will be achieved by introducing an access mechanism consisting of three aspects: Registration & Data usage, Authentication and Authorization. Registration & Data usage refers to the fact that users who wish to benefit from EMT, must both provide verifiable identity information (e.g. an e-mail and/or mobile number) and also, accept an agreement/license which articulates the terms and conditions pertaining to the usage of the data offered within EMT, in order to create an account. Authentication involves |

| |
|---|
| securing access to a user's account by employing tools such as 2FA, in addition to the basic username & password scheme. Details on Authentications are provided in section 7. Last but not least, Authorisation refers to the mechanism that will be restricting user access, to specific resources of EMT, whether that be data or functionality, via utilizing a hierarchy of user roles. Every user will be assigned a role upon registration, and every role will entail certain scopes that essentially define the privileges of any user assigned to the role. |
| **Priority:** High<br>**Justification**: For NGOs it is highly important that there is enough secure measures in the tool that prevent from being used by other actors for different purposes. |
| **Source:** D7.1 |
| **Implementation**: This requirement is met as explained in *Section 6.6 and throughout Section 7 and its subsections*. Both the backend and frontend will be developed with all the necessary security modules and practices. The backend will only be able to respond to authorised requests, and the frontend will require a valid email in order for an account to be created. Additionally, the project coordinators will review all account registration forms. |

# 4. EMT Design Approach

The overall architecture of the EMT prediction tool - a software intensive structure - is the composition of several subordinate system modules, which comprise software elements, their externally visible attributes and properties of those elements, and the relationships among them.

A hierarchical approach was taken in this deliverable to determine the software architecture of the EMT system. Specifically, this practice starts from the ITFLOWS concept and it preliminary analyses its basic concepts through an abstract conceptual manner. Then, the high-level architecture is defined by elaborating the conceptual modules with the available models of the system as well as the external data stored in the CKAN's repository. The top level of Figure 1 is described in section 5. Moving lower, the high-level architecture (middle part) is described in section 6, while the bottom part (Dynamic structure) is described in section 7. Moving deeper in the hierarchy and by interpreting the Use Cases as well as their explicit end-user functional requirements, the static structure of the EMT system is determined analytically by presenting the core functional modules. The following figure illustrates this approach.

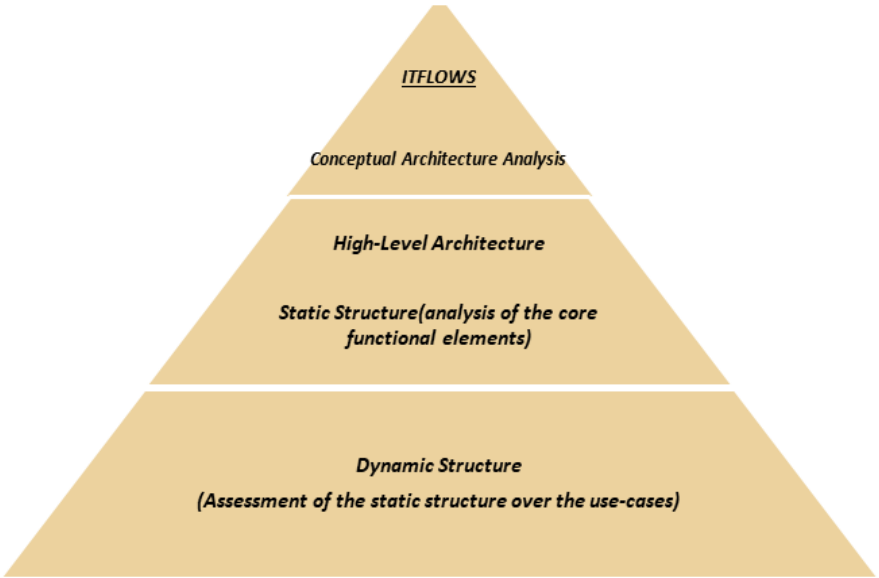


**Figure 1: EMT hierarchical approach pyramid**

## 4.1. Data protection, privacy and security by design

This section provides details on data protection and privacy by design in earlier stages of EMT devolvement, in order to define a standard secure software development lifecycle. We implement Privacy by Design (PbD) by building privacy into the design, operation, and management of the EMT system. Doing so, we are ensuring privacy through every phase of the data lifecycle (e.g. collection, use, retention, storage, disposal or destruction) as this has become crucial to avoiding legal liability and maintaining regulatory compliance.

The amount of personal data should be restricted to the minimal amount possible (data minimization) close to none. EMT will try to avoid and limit the need to collect and process personal data. If needed to collect personal data, the data subject will be adequately informed whenever his/her data is processed (transparency). According to data protection regulations (e.g., GDPR), the data subject has the right to control his/her data, including access, review and/or deleting his/her own data. Moreover, all personal data will be hidden from public view. Anything that should be stored in a database will be encrypted and/or will be anonymised beforehand. To avoid risk of privacy abuse, personal data will be stored in separate databases from the rest of the EMT infrastructure. A privacy policy will be enforced in the EMT system, compatible with legal requirements, and EMT will apply the highest privacy settings by default.

Following is a list of some privacy implementation within EMT:

- EMT shall provide a form of consent before collecting any data.
- EMT shall summarise the content of data and why it is needed before collection in order to avoid collection of sensitive and unwanted data.
- EMT shall implement privacy enhancing techniques that include anti-tracking, encryption of sensitive data (such as emails and passwords for authentication processes) and secure file sharing in order to avoid unwanted exposure of data.
- EMT will store data anonymously whenever possible and applicable (e.g. Tweets will be fully anonymized).
- EMT shall store personal and sensitive data in separate databases from the rest of the EMT infrastructure to limit loss or exposure of these data.

## 4.2. Design Principles

This section describes the design principles of the EMT architectural elements, fully describing their functional specifications and the interactions among them and the environment or the end-user when applicable. Such design principles have been established in line with the Privacy by Design principle (Article 25 GDPR), meeting all its requirements both from a GDPR perspective and from a design and end-user perspective

**Principles explained in D2.4 on the ITFLOWS regulatory model**

These include human rights, well-being, data agency, transparency, accountability and awareness of misuse. Details on this matter are expected in deliverable D2.4.

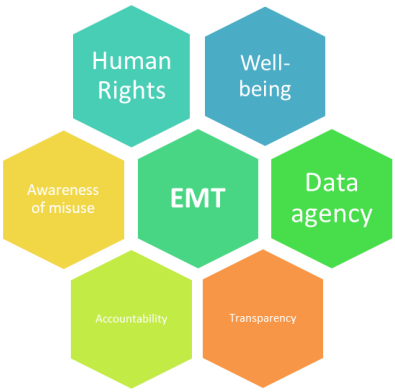


**Figure 2: Basic principles from the ITFLOWS design and regulatory model**

**RESTful API[1]**

We will first analyse the EMT backend, which will be built as an API (Application Programming Interface) and will follow the API-first approach. It means that the server will be build first, allowing the development of frontend servers and other apps and websites on top of the same functions and conditions regarding communication to the backend server. More specifically EMT backend will be a REST API. REST (Representational state transfer) is a set of architectural constraints that any API developer can implement to an application. These sorts of apps send a Representation of the information or resource, as a state, from the frontend and the backend server acts accordingly, ultimately changing this state to the needs of the

[1] Definition of RESTful API can be found in: https://en.wikipedia.org/wiki/Representational_state_transfer

request made by the user. There are several criteria an API must meet in order to be considered RESTful.

- The frontend servers' and clients' requests must be managed through HTTP.
- Client-server communication must be stateless, meaning the requests are different and not connected to each other.
- Information transmitted between various components of the application in a standardised form.
- A system for organizing each type of server requests that are not visible to the client.

REST APIs are faster and more lightweight than other forms of applications and they offer increased scalability over similar approaches, such as SOAP (Simple Object Access Protocol) applications. They have become one of the most common types of applications, mobile and web.

**MVC Pattern**

EMT will follow the JavaScript MVC pattern. MVC comes from Model, View and Controller and, as obvious as it is, it decomposes an application into these three different components. In a simple explanation, a request arrives, and a *Controller* is being called changing some information/data from the *Model* and sends that information to the *View*. In other words, the Model is responsible for managing the data, the View is representing that data in various ways depending on the user and the Controller is responsible for manipulating the data accordingly. As seen in *Figure 3* below, the User can see the *View* and can use *Controllers*. The *Controller* can manipulate the *Model* meaning the *View* will be updated. The User can see the updated *View*, which is the goal.

This pattern is clear and simple. It allows the development of complex and modern applications which requires numerous interactions between the frontend and backend across multiple parts of the application while producing more modular and reusable code.

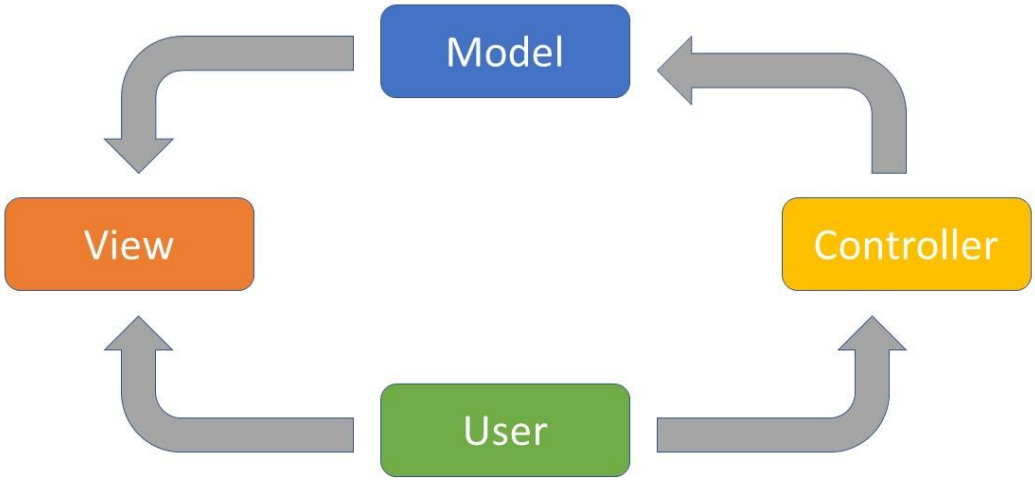


**Figure 3: The MVC pattern**

**Containerisation using Docker**

Scalable architectural solutions are crucial for meeting and sustaining the demand of large, expanding, and elastic device networks. EMT will be used from different end users and from various platforms. To promote rapid onboarding and to be able to run the application on any platform depending on the user needs, EMT will follow a containerised approach that will allow developers to program each service over the most suitable operating system and programming language. Similar to virtual machines, containers allow to package and dissociate applications from the environment in which they will be running. This decoupling will allow container based EMT services to be deployed easily and consistently, regardless of the user environment. However, containers offer several benefits with respect to VMs. Instead of virtualising the hardware stack (as VMs), containers virtualise at the OS level, i.e., multiple containers can be running over the OS kernel directly. Hence, containers are far more lightweight than VMs, as they can be run much faster, and use a fraction of the memory. Some other benefits containerizing an application are a consistent runtime environment, sandboxing the application, which increases security, only requiring a small size on the disk and offer low overhead.

More specifically, EMT will be running on Docker. Docker offers containers for developing applications. These containers are isolated from each other and run their own software and files/code. They are able to communicate with one another

through pre-defined channels. These containers are hosted on the software called Docker Engine. After containerizing the application in multiple packages, the app can run on any OS computer, further meaning that it can run in a variety of locations which enables EMT to be portable and easy to use on any device.

Docker consists of three components, the *Software* component called "dockered", the *Objects* that are units used to create an application in Docker and the *Registries* which is a repository of Docker images. Docker, finally, offers three tools.

- **Docker Compose**, which defines and runs Docker applications that use multiple containers and allows users to run commands on all of them at once.
- **Docker Swarm**, which offers a native functionality for creating clusters of Docker containers which unites a group of Docker engines into one virtual Docker engine.
- **Docker Volume**, which prevents the deletion of files that are copied or created inside a container after stopping that container.

## 4.3. Architecture definition process

The process is described in six different steps. The details of these steps are as follows:

- **STEP 1:** The crucial part of correctly determining the architecture of a system is the definition of the end-user requirements and their respective use cases.
- **STEP 2:** Simultaneously with the use cases, the technical specifications of the EMT system are determined to define the software structure as far as inputs/outputs are concerned. A conceptual decomposition of the system into the fundamental structures in terms of hardware and software requirements is performed. The latter has been achieved by taking into consideration of the conceptual analysis of the software requirements.
- **STEP 3:** The interface between user requirements and technical specifications comprised the model's functional requirements. Specifically, each sub-use case has been interpreted as a sequence of functionalities to fulfil a specific task. This assisted in the determination of the essential functionalities that the system should retain. Additionally, by assessing the resulting performance

requirements -tightly connected to the software requirements- with the functional requirements, the static model of the EMT architecture has been established

- **STEP 4:** The static model of the EMT system constitutes a detailed analysis of the essential software parts to be implemented. It comprises the set of algorithmic routines that will be developed within the lifespan of the ITFLOWS project. In addition, the static model holds a detailed description and definition for each software functional component providing explicit details about their functionality, their in-between dependencies, their operation requirements as well as their indicative performance requirements.
- **STEP 5:** Upon the determination of the static model of the EMT system, its functionality is examined by assessing its behaviour during use in the various scenarios of sub-use cases, including the way each module acts.
- **STEP 6:** It is apparent that some software functional components will be missed during the steps 3 and 4 since their functionality is hard to be foreseen. These functional components are commonly the ones used as interpreters between the core functional elements, while their necessity becomes apparent immediately after an indicative integration such as the one the dynamic structure offers. Therefore, multiple iterations among the steps have been performed to conclude on a valid and globally descriptive static model.

These steps have already been completed and presented throughout this document. Nevertheless, the development of the EMT is active and dynamic, so modifications are expected over the next period, as the work in other related WPs is also on going. We will have the details when the responsible WPs will further progress. All the updates in the EMT design, will be reported in deliverable D6.2, along with the release of EMT initial version.

# 5. Conceptual Architecture

This section describes a conceptual architecture of the ITFLOWS EMT system, which specifies its theoretical modules identified firstly in the ITFLOWS concept.

EMT will get input data from various sources described in the appropriate Work Packages and their respective Tasks as seen in *Figure 2* above. Qualitative and Quantitative data will be gathered from tasks that are part of WP3. EMT will gather both raw data and analysed / processed data from various sources, as explained in previous sections of this deliverable, and software or models created through the work of WP3, WP4 and WP5. There are also data regarding migration patterns, economic and social assessment and policies, and procedures as part of WP4. Meta-analysis of factors and causes of attitude variations are part of WP5. Finally, there are some additional data from UB that will also be included in EMT.

After getting the input data, it is stored in the EMT data repository, which is CKAN. Once stored in CKAN, datasets will be updated from their sources based on each datasets frequency as described in each individual source or dataset details (e.g. some datasets are updated yearly, while others are updated daily or even every 15 minutes). Using such a data repository brings many advantages, such as easier management of the data and the ability to frequently update data. EMT will then be able to get the data from the repository and use it accordingly. There are two main services that are offered and use the data: the prediction, and the simulation. Each service needs different data and needs different analysis.

Once the analysis is done, the results will be visualized and showed in the frontend, ultimately meaning the user. The end user will be able to configure the way these results are shown, change visualization settings and run multiple predictions and simulations.

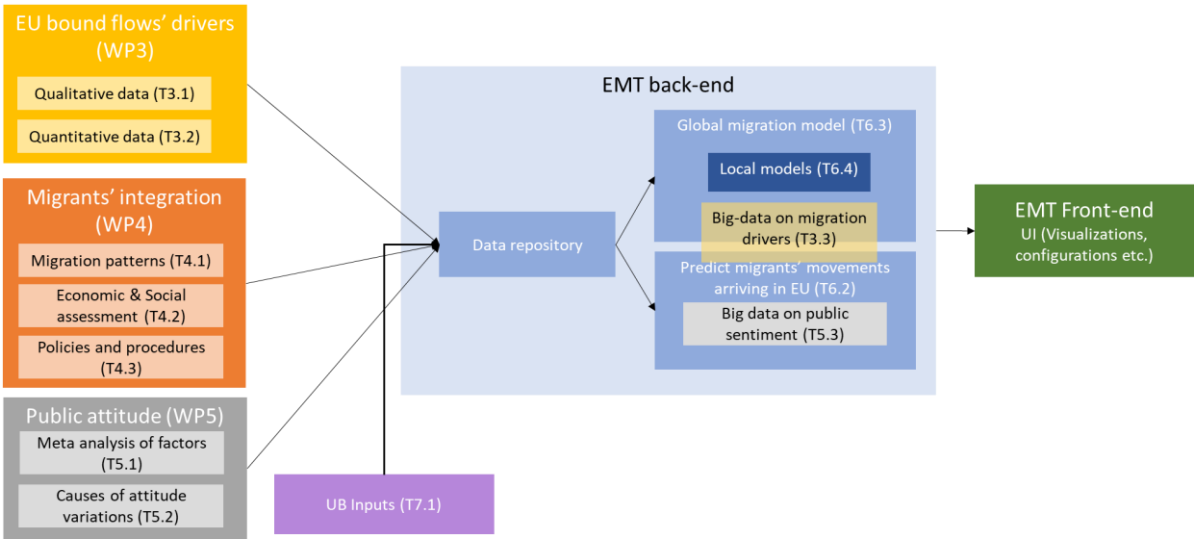


**Figure 4: EMT conceptual architecture**

# 6. High-Level architecture analysis

In this chapter, a high-level architecture analysis of the ITFLOWS system is conducted. It is oriented towards the extension of the conceptual breakdown of the subordinate modules for each task to be fulfilled. Therefore, to build a concrete foundation, where the software architecture will be held, an abstract analysis that outlines the basic functional components for each conceptual module of the ITFLOWS system should be determined.

In this section, the modules from the conceptual study will be analysed into high-level functional modules mitigating from the ITFLOWS concept to an abstract system architecture.

## 6.1. Data repository

CKAN is a powerful Open-Source data portal platform that makes data accessible by supplying tools to streamline publishing, sharing, finding, and using data. It can be of wonderful use to data publishers (national and regional governments, companies, and organisations) wanting to make their data open and available (either privately or publicly).

The reason we choose CKAN out of all the possibilities is its open-source nature and the fact that it comes with no financial cost. This means that you can use it without any license fees, and you keep all rights to the data and metadata you enter. In our case, CKAN will serve as a data repository for the large-scale model, and it will be automatically updating its content using automated parsing through public APIs. In addition, all data added in CKAN will be completely anonymised and non-identifying, in order not to raise any legal/ethical risks.

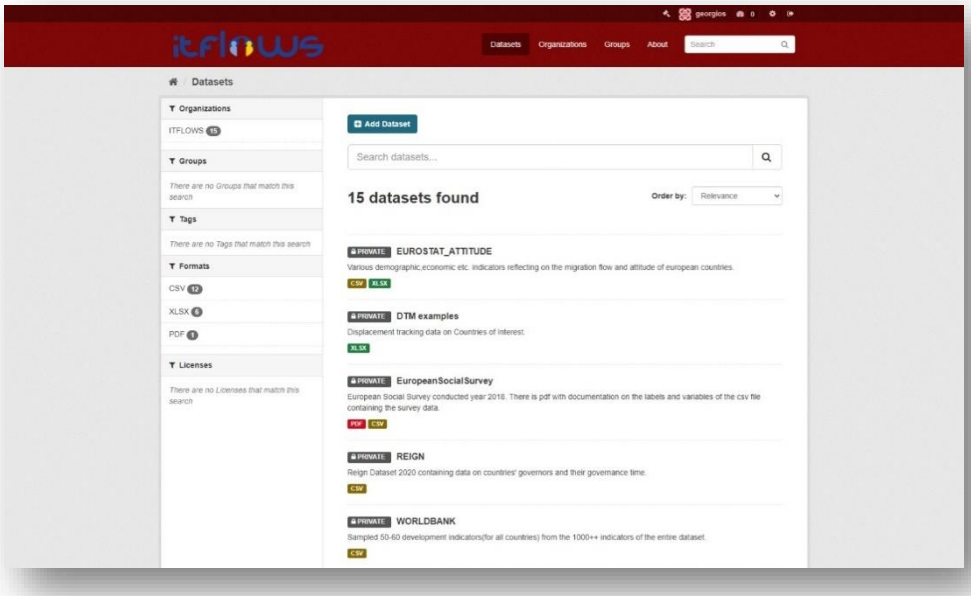


**Figure 5: CKAN repository with some datasets already uploaded**

## 6.2. Large-scale model

A high-level analysis of the large-scale model consists of the following stages:

- ***Data Processing:*** Data coming from the CKAN repository will be cleaned in terms of data formatting and missing values. Both categorical and numerical imputation methods will be applied, and the data of least value will be removed.
- ***Feature Extraction:*** At this stage we extract the most essential features for migration flow and attitude prediction using both traditional machine learning algorithms (like support vector machine and linear regression) as well as state of the art deep learning architectures (e.g., VGG16). These features include several indicators like violence, economic growth, public health care reach, climate anomalies, national sentiment towards migrants, political situation etc.
- ***Artificial Intelligence Module:*** Having acquired the best features possible, as far as quality and relevance is concerned, we use them as input to several machine learning classification algorithms (including both traditional and state of the art approaches) to get a realistic estimate of the probability of a nation's migration inflow/outflow and attitude towards migrants. AI outputs will be accompanied by comments, as to how the conclusions were drawn. It will also examine the best way to avoid biased datasets and ensure a realistic estimate.

➢ ***End-User Requirements Satisfaction:*** This is the last stage of the model where we format the output of the Artificial Intelligence Module to the desired information specified by the end-user functional requirements. For example, one of the outputs will be an estimation of the amount M of migrants coming from a X country of origin to a Y country of destination. Confidence intervals will be provided of course for these estimations. This relates to Req02, as it provides a real number of forecasting asylum applications. It also meets Req03, as it provides a structural manner of displaying the data. Finally, it complies with Req08 in that the model is able to predict nationality, age and gender (these are defined in the datasets that are fed into the model). These User Requirements are described in detail in *Section 3.2, User Requirements*.

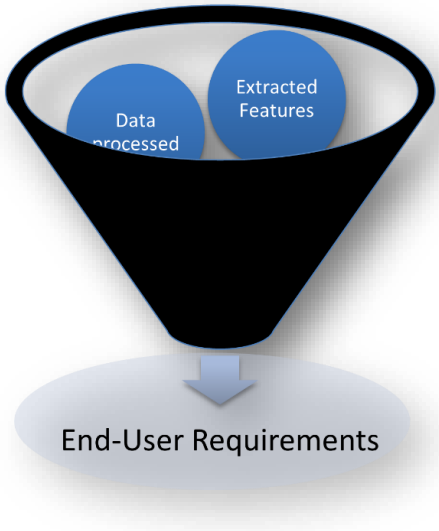


**Figure 6: Depiction of the A.I. Module**

## 6.3. Small-scale model

The construction and execution of small-scale models require a generalised and automated simulation development approach (SDA), which has six main phases: situation selection, data collection (through CKAN), model construction, model refinement, simulation execution and analysis (Suleimenova et al., 2017).

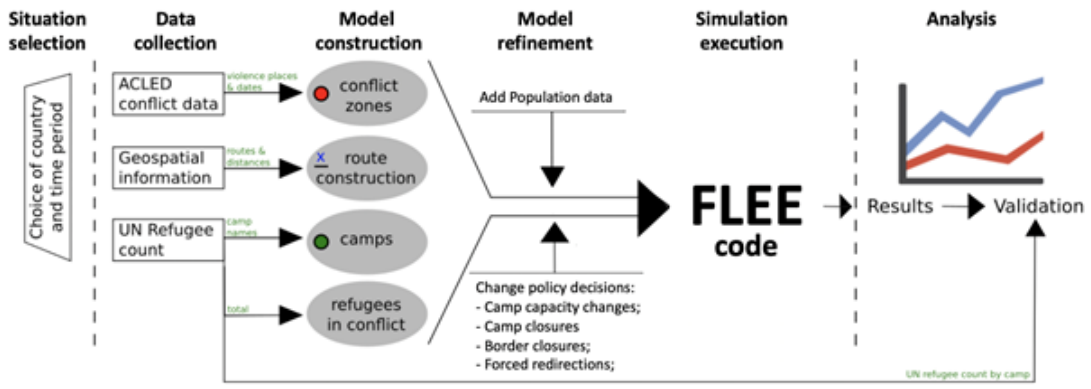


**Figure 7: Simulation development approach to predict the distribution of incoming forced population across destination camps**

First, we select an origin country and period for simulation and prediction, such as

- **Mali**: 1 March 2016 - 1 January 2020,
- **Nigeria**: 1 March 2016 - 1 January 2020,
- **Syria-Iraq**: 1 January 2016 - 1 January 2020,
- **Venezuela:** 1 March 2016 - 1 December 2019.

The methodology underpinning this selection and timing will be detailed in Deliverable 6.3.

Second, we obtain relevant data to the conflict from three main data sources that are:

- The United Nations High Commissioner for Refugees (UNHCR, https://data2.unhcr.org) presents data for the total number of forcibly displaced people in the conflict situation, the camp locations in neighbouring countries and their population capacities.
- The Armed Conflict Location and Event Data Project (ACLED, https://acleddata.com) provides the locations and dates of battles that have taken place in the conflict.
- The OpenStreetMap (OSM, https://www.openstreetmap.org) Platform identifies locations of major settlements and route information (i.e., links and distances) between the various camps, conflict zones and other settlements.

(all data is anonymous and unidentifiable)

In turn, these data sources provide input data for the small-scale origin and receiving countries:

1. Conflict locations automatically obtained from the ACLED database.
2. The total number of forcibly displaced people obtained from UNHCR.
3. Camp location names in neighbouring countries.
4. Camp registration number counts.
5. Links and distances between conflict locations and camps.

Third, we construct our initial model and create, among other things, a network-based agent-based model. We then refine the initial model with population data to improve simulation and incorporate the actual population counts of the origin locations using the City Population database (https://www.citypopulation.de) or other population sources. To execute the constructed model, we run simulations using the Flee agent-based code, which is optimised for simplicity and flexibility and provides a range of scripts to handle and convert forced population data from the UNHCR database. Once the simulations have completed, we analyse and validate the results against the full UNHCR numbers (data is stored into CKAN for process and visualised in EMT).

The small-scale model provides data in order to satisfy the User Requirements as they are described in Section 3.2. Req01 is satisfied by analysing the origin countries and their neighbouring countries. The data is provided and presented in a structural manner in order to meet Req03, and there will be various diagrams to illustrate the main trends in migration flows from the origin countries.

## 6.4. Big data analytics components for migration drivers

*Big Data* sources will be used in the second sub-task of Task 3.3 (WP3). The goal of this task is to use *Big Data* sources to identify the *push* factors of immigration from the source countries. These identified push factors will assist in predicting the migration flows into EU countries. To achieve this goal, the following steps will be performed: (i) the first step is to extract tweets from Twitter based on keywords, (ii) consider low resource languages from source countries, (iii) automated analysis of these tweets for extracting *push* factors. This will help the EUMigraTool in validating the prediction of the migration flows coming from different sources of data.

**We should mention at this point that all data regarding tweets are anonymised and not identifiable. Moreover, no one will have access to source material, but only to processed information.**

During the phase of data extraction, multilingual keywords were used based on the languages of the source countries, e.g., English, Spanish, French, Persian, Pashto, and Urdu, etc. for extracting textual information and metadata from Twitter. This metadata already contains the geographical information, e.g., Nigeria, Venezuela, Mali, and Afghanistan, etc. which is captured while crawling the tweets. The collection of tweets related to the countries of origin will be based mainly on the language (and dialect) and an estimated location. If we take the example of Syrian users, ITFLOWS will be focusing on collecting public data of users of Levantine Arabic (spoken in Lebanon, Jordan, Syria, Palestine, and Israel) language who are located (based on the Twitter API information) at least in the following locations: https://data2.unhcr.org/en/situations/syria . Since the location is only approximated, there will be no discrimination based on the nationality in this task.

In order to broaden the ranges of languages and the content, we will deploy an iterative methodology -- expand the multilingual keywords, extract data, analyse and extract keywords such as hashtags. In addition to analysing the opinion of possible migrants in countries of origins, we plan to augment our models with the data from transition countries and official sources of information. Time-based constraints will also be provided for initial filtering when the migration was at its peak.

Potential ethical challenges that could arise from using Tweets, will be solved [*by*] applying the following points:

- **Data collection**: Users are typically not approached directly to solicit their informed consent to take part in research. Instead consent is given by the user's acceptance of Twitter's Terms of Service.[1] These state: "By submitting, posting or displaying Content on or through the Services, you grant us a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy,

reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods (now known or later developed). This license authorizes us to make your Content available to the rest of the world and to let others do the same".

- **De-identification methods (Authorship Obfustication) for natural language processing tasks:** multiple steps need to be addressed. ITFLOWS technological partners (MTU and FIZ) will extract identifiers from text, and they will anonymise [*pseudonymise*] the data set used for NLP tasks. For example, all addresses, names, and so on by using named entity recogniser will be removed [*replaced with randomly generate identifier*].

Furthermore, we will perform Named Entity Recognition related to a specific hashtag, to identify particular migration intentions. Such information can improve predictions of migration trends, and thus, will be useful for WP6. For achieving this task, we will exploit frequent pattern mining algorithms adapted to Twitter data analysis.

This approach falls in line with User Requirements Req03, by providing data in a structural manner. It also fulfils Req05 and Req06, by providing useful insight into attitudes towards migration in destination countries.

## 6.5. Big data analytics components for public sentiment

This task will comprise an analysis of social media environments (Twitter only for the scope of ITFLOWS and within the confines of the acceptable use policy) in the EU Member States which are the destination countries. The goal of this task is to identify the public attitudes in countries of destination, in order to manage the potential tension resulting from the migration flows. The task would also measure the correlation between public attitudes and other factors such as GDP per capita and unemployment rate. The details of the analysis is under investigation/development, and details will be provided in the relevant deliverables (WP5)

The analysis of the public attitudes would be from the destination countries where the data will be extracted from Twitter using migrations-related keywords. Keyword selection will consider the gender dimensions of biases such selection presents, to the extent possible. The tweets will also be filtered based on trending hashtags related to the current migration scenarios including references to geographical entities (i.e., mention of a location in the tweet or the location of origin of the tweet). Thereby georeferenced opinions of the masses in the country where the migrants are currently residing will also be mined, i.e., is there an increasing hatred towards the migrants?

With the help of Hate Speech Detection, the tweets will be classified into hate speech or non-hate speech. Further analysis will help in identifying if the hatred is expressed towards a community or nation or individuals. Machine/Deep Learning-based methods will be exploited for analysis and classification purposes. As such, a multi-lingual approach is an important aspect to be addressed. In addition to providing descriptive statistics and mapping of sentiments, we will conduct econometric analyses to improve the understanding of the determinants of hate speech and related phenomena. For this, we will rely on geographically disaggregated data at the destination country level.

Similar to Section 6.5, this analysis contributes to the satisfaction of Req03, Req05 and Req06 by providing structured data in an understandable way, as well as an understanding of the attitudes towards migration and potential risk of tension in the destination countries.

## 6.6. EMT backend infrastructure

The core technology for developing the backend of the EMT web application is JavaScript and the web app will be running on Docker. As already mentioned in section 4.1, the backend development will follow the JavaScript MVC which consists of three components. The *Model*, which manages the data, the *View*, which manages the representation of information, and the *Controller*, which is the main functionality of the API.

The end user interacts with the frontend server to make requests to the NodeJS backend server. The server can communicate with CKAN and other Python programs and scripts responsible for getting and pushing data to CKAN. The main database of the application is MongoDB, which is being managed by Mongoose through NodeJS.

The NodeJS backend server will, also, be able to communicate, through the Controllers, with other Python based programs that are responsible for analysing the data sets and provide the end user with useful information based on the service being called.

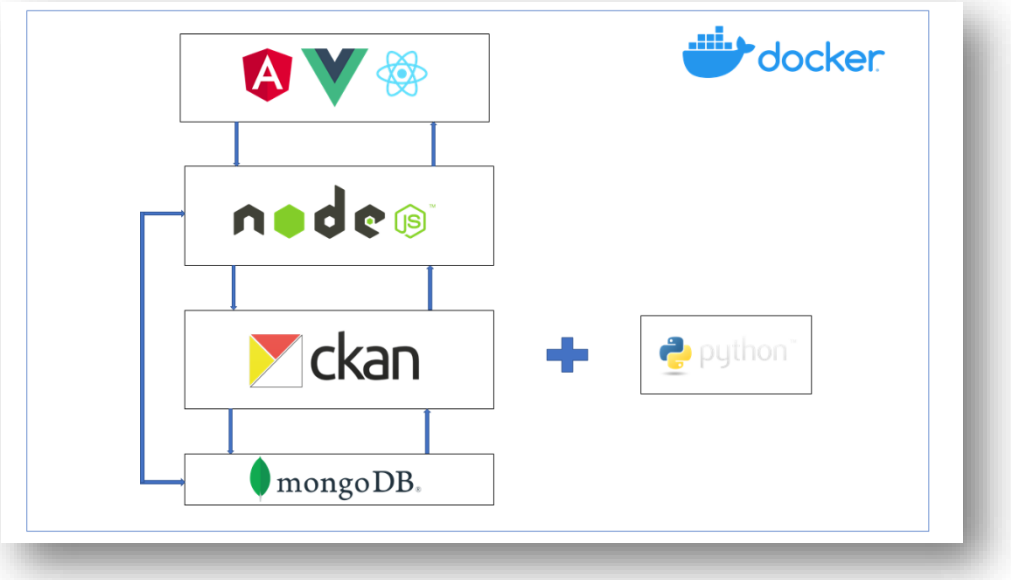


**Figure 8: EMT backend ecosystem**

These technologies allow the development of fast and responsive applications even if it is using and analysing Big Data. As a result, the User Interface is user friendly and interactive.

The EMT application can extract reports in various types depending on the needs of the user.

The backend server of the EMT has been designed to ensure an easy and secure

experience for every user. This is achieved with the help of the frontend server as well, by providing all the necessary authentication and authorisation processes all modern web applications use. This falls in line with and satisfies Req07 and Req09 from the User Requirements described in *Section 3.2*.

# 7. Detailed analysis of the EMT

To accurately describe the static structure of the EMT we first need to define the back-end support of the model. All data that is lawfully made available in the public domain. It will be stored in the CKAN repository where it will be periodically updated using Python scripts.

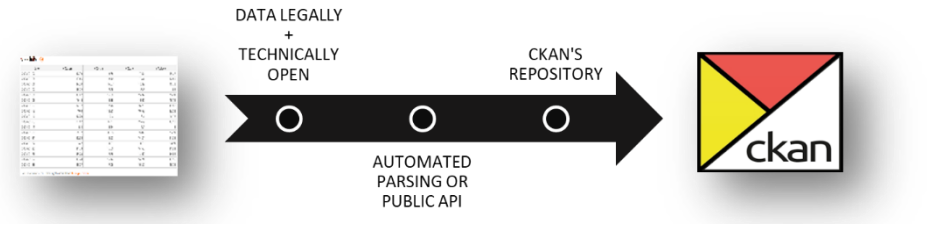


**Figure 9: CKAN's back-end data fetching pipeline**

The most crucial part of a machine learning prediction attempt (after ensuring the data quality) is that of feature extraction. The main functions of the EMT tool will be the quantitative prediction of a European country's inflow of migrants in a certain time, as well as the estimation of a nation's attitude towards migrants that are already residing in a particular territory within the EU. To achieve a significant statistical confidence in our predictions we need to select our model's input accordingly.

Both functions' pipelines look identical except for the inputs (features) and outputs (end-user requirements) of the model.

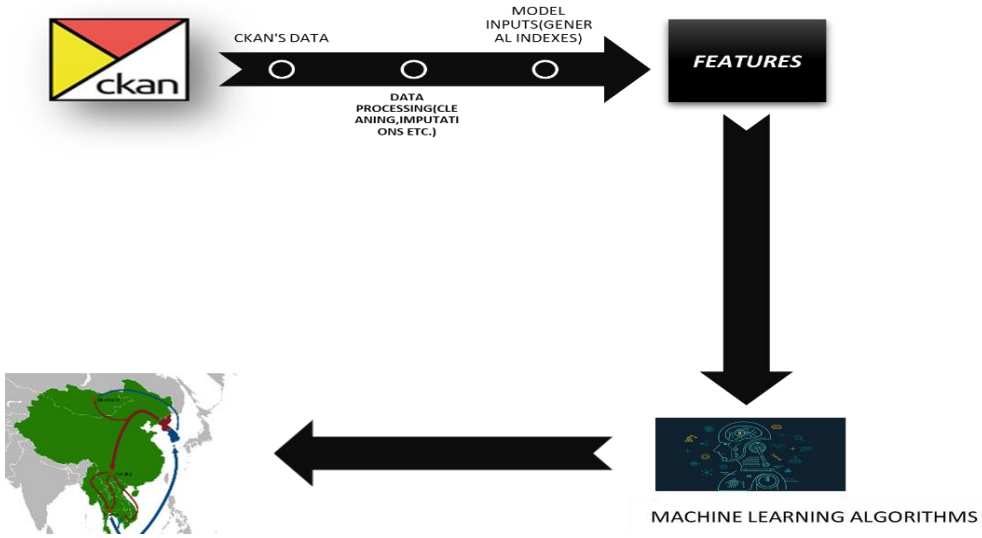


**Figure 10: First main pillar of the prediction model regarding the Migration Flow**

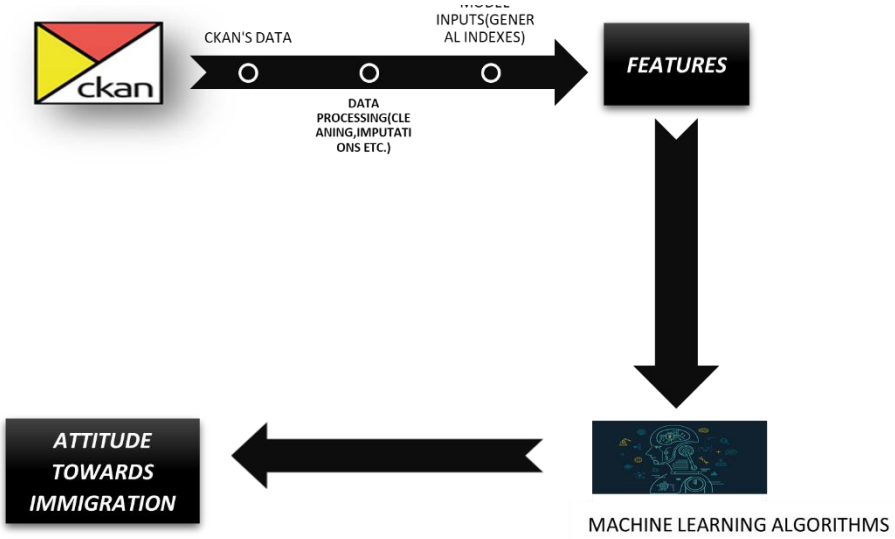


**Figure 11: Second main pillar of the prediction model (developed in WP3 & WP5) regarding the Anti-migration Attitude**

## 7.1. EMT Portal

A portal is a type of website / web app which combines webapps / websites / information from various sources, that are usually accessible individually, into one point of accessibility. All this information is in a central location within the EMT system. EMT is able to combine datasets from various sources (as mentioned in previous sections), providing novel analysis models which are not publicly accessible. Due to the many advantages that come with it and the fit it entails for the purposes and the technologies used for this project, a portal will be developed for ITFLOWS. The end-result will allow the registration and logging in to the ITFLOWS platform and the access into data and information through many dashboards and filtering options, in that ensuring adaptability to the specifics of each use-case and need.

The development aims to create a user-friendly system through which the user can be authenticated and in turn interact with the EMT system in an easy, fast, and self-explanatory way. The whole design of the environment and user interaction will be based on the feedback gained from the users' board through various iterations and the relevant deliverables such as 7.1. Report on Users Board Participatory Feedback

(D7.1). As it is designed ITFLOWS Portal, will consist of the following two components:

- The **User Authentication system** through which the user can register, login to the system by being authenticated and be authorised to use certain system's features and capabilities. Details will follow in section *7.1*.
- The **User Interface** through which the user can interact with the system.

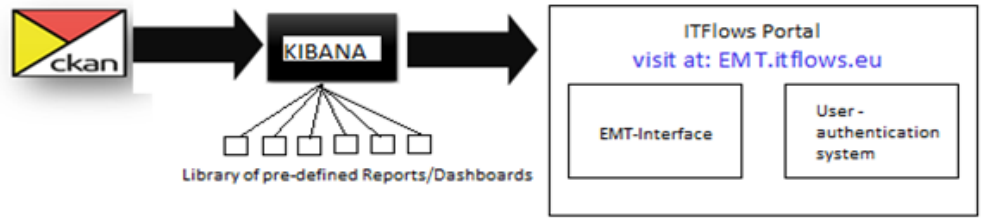


**Figure 12: Depiction of the ITFLOWS Portal conceptual architecture**

The figure above depicts the ITFLOWS Portal with the User authentication system and the EMT interface (User interface). it also highlights the flow of data from CKAN (open-source software) to Kibana as well as the proposed URL for it to be accessible by the end-users EMT.itflows.eu. Kibana is a visualization tool that implements data from Elasticsearch. It provides various chart visualization (pies, lines, maps) and can analyse a large amount of data. The two components of the ITFLOWS Portal, along with their functionalities will be described elaborately in the next sections.

This approach satisfies Req07 by providing a very clear and easy User Interface, personalisation elements and an easy, accessible feedback system and forum support. The user authentication system and registration that are both described in the following sections also assist the backend server in meeting Req09, by providing authentication and authorization processes to prevent data misuse.

### 7.1.1. User Authentication System

User authentication is the process that allows a service to verify the identity of a user who wishes to make a connection to a network resource and assess if shall be granted access or not.

For the needs of the ITFLOWS project, it is necessary to deploy a subsystem that will handle providing authentication and authorisation mechanisms. These services will be offered by the User Authentication subsystem, as part of the ITFLOWS Portal. This subsystem will both regulate the entry of users to the ITFLOWS Portal (authentication) and decide their access rights within the Portal (authorization). Access to specific reports and functionalities based on the type of user will be supported, in case it is considered necessary for various use cases such as the role within the organization of the user accessing the EMT or the access by different entities, such as NGOs and municipalities.

The Authentication process will include an e-mail form that will be sent to each user wishing to create an account. The form will include a link that must be followed in order to validate the address and indicate that the user has access to that email account. After registration is completed, EMT will provide 2FA (two Factor Authentication) for the user to be able to login to his/her account and use the app. The 2FA will either be completed through an email service which will contain a link that must be followed or a code that must be inserted in the login form, or through a mobile phone app that will provide the appropriate authentication code for the login form.

All sensitive data, such as email address and password, will be encrypted upon the registration process, so not even admins and/or developers will have access to these kinds of data.

### 7.1.2. Register - Login to the ITFLOWS Portal

For potential user, to enter the ITFLOWS Portal, they shall create an account through a registration process and define a personal password and a valid e-mail address (email address should be linked to an authorised party). Once the user has been successfully registered, a corresponding account is created and stored in the users' database allowing user to login to the ITFLOWS Portal.

User's login is achieved by entering credentials in the corresponding login form on the ITFLOWS Portal website. Credentials are being sent to the User authentication

system and checked for their validity against the stored information of the database. If successful, user is then redirected to the home page of the ITFLOWS Portal.

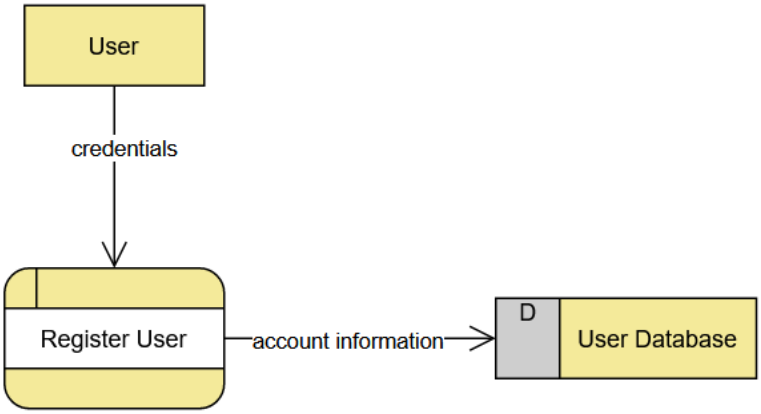


**Figure 13: Registration process**

### 7.1.3. User Authentication system and user capabilities within the ITFLOWS Portal

As mentioned above, the User authentication system allows different access rights, according to the principle of privacy by design. This is made possible by having a hierarchy of user types. Each user, upon registration, is assigned to one of the available types of users. For each type of user, a few operations are additionally defined. For example, municipalities and NGOs' will have access to distinct levels of information than the ITFLOWS project managers. Also, NGOs' access both the prediction of arrivals and sentiments functionalities, whereas municipalities access only the latter. In general, the system will be flexible, it is up to the consortium to decide who will have access and in what functionalities.

The operations of a user type describe the capabilities of any user characterised by this type, within the Portal. This means that a user has access only to those operations and data that are defined by their type. The type of features a user can have lies to the fact that the end users might have different types of responsibilities.

### 7.1.4. EMT Interface

The User Interface is the way through which the user can interact with the system. For instance, all the pages that someone can see in the browser of a web application,

on YouTube, on google, etc. constitute a user interface. At the same time, it consists of all the graphic elements, images, buttons, navigation bar, etc and the logic behind accessing data and functionality.

As shown in *Figure 9* above, the work on the back end and the creation of pre-structured reports and dashboards happens with Kibana, before it becomes available to the end-user with the implementation of iframes. Data from CKAN will be transformed to the proper form (e.g., csv) transferred to Kibana and converted to reports and the data visualisations by **using the logstash env**. Kibana is part of the EMT system and it is not an external service, meaning all visualization services are being processed internally within EMT.

One of the reasons that Kibana was selected, is that it offers an intuitive way to create visualisations, through the Lens functionality and that the representation of data is more suitable to the ITFLOWS case being vivid, flexible, and easy to process.

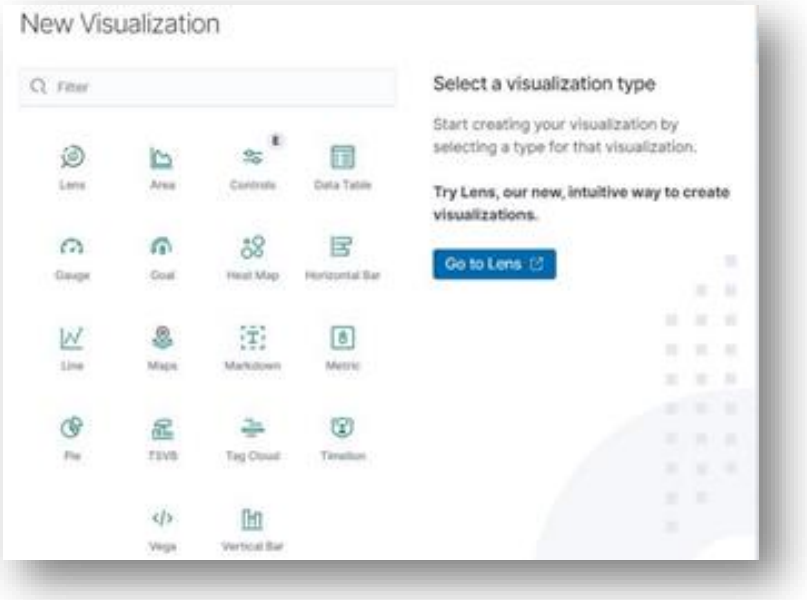


**Figure 14: Kibana Visualisation Select Screen**

Various reports will be combined to create the desired dashboards that will correspond to the use-cases described by the users board. Lens functionality allows the creation of reports as well their share through iFrame code (which will include all our reports along with the Filter option) so that it can be viewed by the end-users.

As the final system will be used by a matrix of users with different background and skillset, the design and development of a user-friendly and intuitive interface is crucial. The proposed solution will have a structure as shown (for demonstration purposes we've used the dataset from Greece - Data on forcibly displaced populations and stateless persons from data.humandata.org) on the image below:

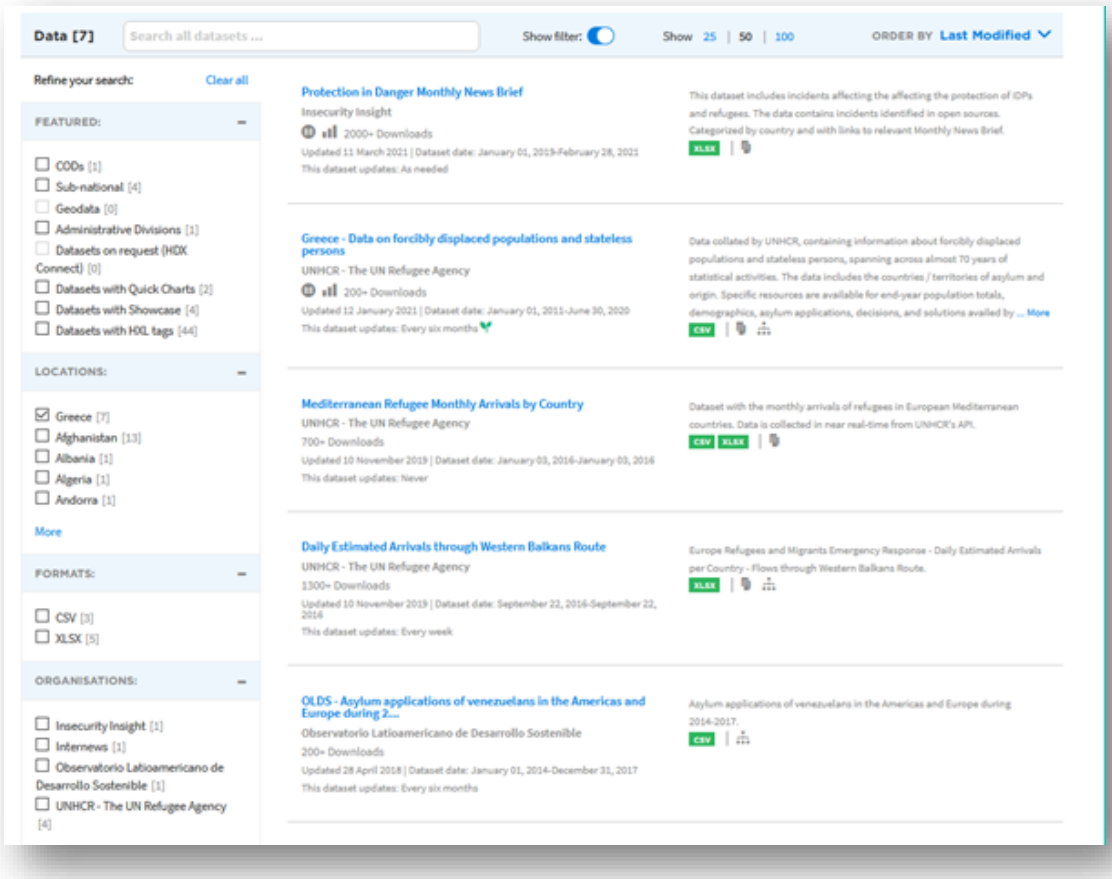


**Figure 15: HDX (Humanitarian Data Exchange) Main Page**

A user will be in position to define a number of variables and filters (country of origin, asylum submissions tec.) depending the information that would like to obtain and then access various reports that their format has been defined according to the identified needs of the different groups of users.

A selected report will highlight the available information, through pie charts, bar charts and maps for the specified use-cases, such as showing the migration flows from the countries of origin to the countries of destination to help user obtain valuable insights such as the image below:

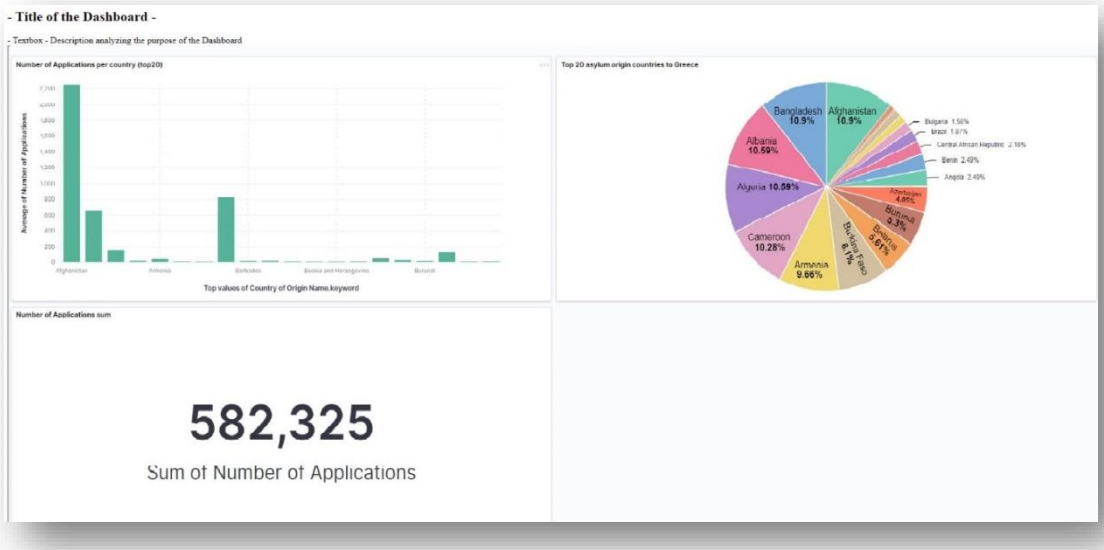


**Figure 16: Sample Kibana Dashboard**

Kibana and Elsticsearch features fall in line with user requirements Req03 and Req07 as described in *Section 3.2*. Req03 is met by providing data in a structural manner and in a very easy and understandable way to the users. The simple yet sophisticated visualizations provide a clean and easy-to use User Interface, which increases user engagement and offers more interactive functionality.

## 7.2. Features Extraction from Dataset Pipeline

All the features above will be used as inputs to the machine learning algorithms used for the predictions.

### 7.2.1. Flow-Related Features

An extremely crucial factor of human migration throughout the history has been the presence of extreme violence in the form either of war or political upheaval.

The benefits of violence detection in videos/images are significant for the ITFLOWS project because violent conditions and major upheavals make up the main reasons of migration. Violence metrics can be developed upon detection which can give us a probability of an imminent creation of migration flows.

Their development will be based on various methods such as image feature extraction using both Computer Vision and Deep Learning methods, audio feature extraction using audio to text transcriptions, text feature and embeddings extraction using NLP etc.

At this point, we should mention that no personal data will be stored in the databases as Tweets are stored anonymously and no identifiable video or audio will be stored within the EMT database.

The main pillars of the migration flow prediction are the following:

**1)** ***Violence Detection on data from GDELT Project.*** Using data from the GDELT Project repository, we apply both computer vision techniques to the images and natural language processing to the audio transcripts of the videos to extract meaningful features for violence detection. We trained a Deep Neural Network using transfer learning and the HRUN Dataset of Kaliatakis et al. to perform detection on the occasions of violence involving refugees

***2)*** ***Socioeconomic indicators*** *(Unemployment, GDP, Health Reach, Extreme Poverty, etc.).* Migration is a multidimensional phenomenon as far as the factors driving it are concerned. Undoubtedly, various socioeconomic indexes reflecting on a country's well-being and prosperity can be extremely useful for the task of migration flow prediction. For example, indicators such as unemployment, GDP, access to healthcare, and the extent of extreme poverty can be of a great use to the model.

***3)*** ***Climate and weather anomalies*** *(in the form of indicators like water level, temperature, floods, droughts as well as any other severe phenomenon linked to climate change).* It is important to here the difference between climate change and other pre-existing factors (i.e. droughts have been a cyclical push factor in sub-Saharan Africa for many decades, independent of the increased challenges due to climate change). With the climate change consequences becoming alarmingly more evident, it is likely that a significant amount of people will face harsh climate change impacts resulting in devastating droughts and/or floods, extreme weather, destruction of natural resources etc. Currently, forecasts vary from 25 million to 1 billion environmental migrants by 2050, moving either within their countries or

across borders, on a permanent or temporary basis, with 200 million being the most widely cited estimate, according to a 2015 study carried out by the Institute for Environment and Human Security of the United Nations University (https://ehs.unu.edu/blog/5-facts/5-facts-on-climate-migrants.html). As a result, it indeed makes sense to use climate and weather anomalies indicators, such as water level, floods, droughts, and tropical cyclones, as a strong migration predictor.

### 7.2.2. Attitude-Related Features

The main pillars of the anti-migration attitude estimation are the following:

- *Hate Speech and Sentiment Analysis of Big Data (tweets).*
- *Demographic and Socioeconomic Indicators.*
- *Data from European Surveys.*

Hate Speech and Sentiment Analysis of geo-tagged tweets are two of the most prominent features that will be used as input to our model. Both are numeric values reflecting on a nation's anti-migration sentiment.

When it comes to attitude towards migration, there are several indicators showing significant statistical relation with anti-migration attitudes and include both economic and demographic ones. These indicators will be represented numerically and will also be included in the models input features.

As far as the European Survey data are concerned, an attempt on using text summarisation techniques (supervised/unsupervised machine learning and natural language processing) and text embeddings will be included in the model's input.

For this functionality, it is important to mention that the EMT will build upon the best practices of other H2020 that have utilised similar data (e.g. VOX-POL, DANTE, ASGARD) and will take into account SELP issues regarding the design.

## 7.3. Backend Analysis

The core technology for developing the backend of the EMT web application is JavaScript. More specifically, they are NodeJS and ExpressJs. Nodejs is a runtime

environment for JavaScript code that also offers a package manager for downloading the required libraries. ExpressJS is a modern web application framework for building APIs.

Moreover, the web app will be running on Docker. Docker containers allows the application to easily run on any system that can run the Docker runtime environment.

### 7.3.1. Getting the Data

Python programmes are pulling data sets from various Data Sources. Some of these sources include GDELT and Eurostat, as specified in Chapter 2.3. After filtering the data to our needs, the data files are pushed and stored to CKAN.

The EMT web app API will, then, be able to access CKAN and get the data files that are needed. After the data files are successfully pulled from CKAN, EMT can analyse the data and produce various results.

The backend service responsible for communicating with CKAN will make a proper request to the CKAN URL connected to our account. After the proper filtering in the data sets is completed, the service handles storing the datasets to our CKAN account.

### 7.3.2. Services

Even though there are quite few services integrated into the backend, the two core services that are provided to the user are the Prediction and Simulation services.

The *Prediction* controller will be called after a specific route is accessed from the end user through the frontend. This controller will then run an external program responsible for analysing the data. This program will get the required inputs and after analysing, it will return some information on the backend controller. The controller is then responsible for sending the results back to the frontend to be displayed to the end user.

The *Simulation* controller is working very similar to the Prediction one. The controller is being called after a specific route has been accessed from the user and

after analysing the appropriate data through an external programme, it will send the results to the frontend server to be displayed to the user.
As already mentioned, there is a service for accessing the CKAN API from the backend server. This service is responsible for getting and storing data sets to our CKAN account through the *get* and *store* actions of the CKAN API.

Moreover, the application will support user accounts. The backend will provide authentication and authorisation services to the users. Only Authorised users will have access to specific routes. Users will, finally, be able to search and select specific data sets that they want results from.

# 8. Implementation View

In this chapter, the implementation view of the proposed architecture is presented. In accordance with the literature -relevant to the design of system architecture- the implementation view is decomposed into the development and physical view of the examined system.

An extremely difficult but equally important functionality of the EMT will be the simulation of the migrants' trajectory route (maximum likelihood). This will enable end users to properly prepare and inform the involved countries avoiding unpleasant situations for both migrants and local population.

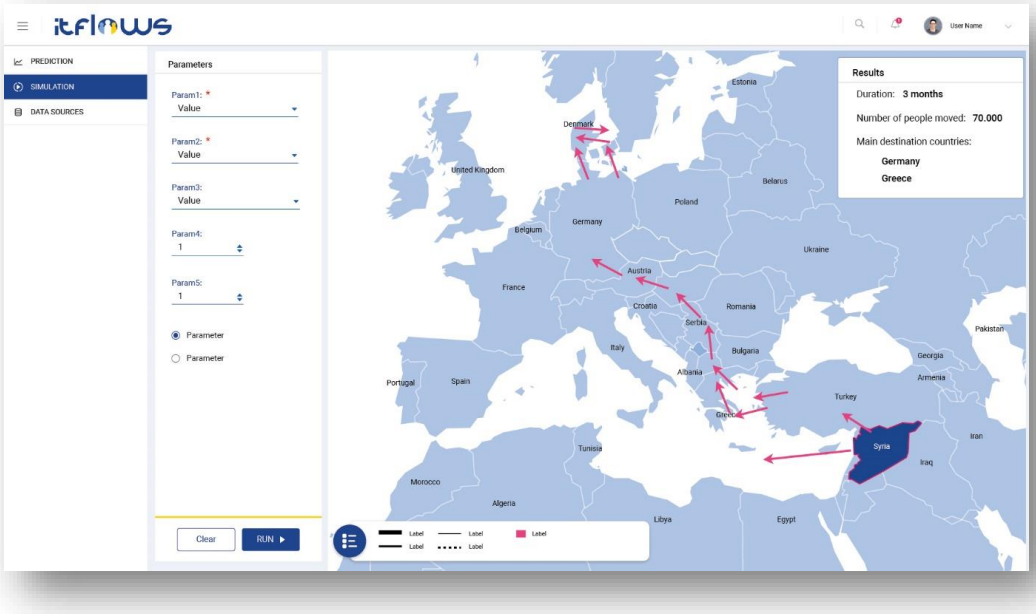


**Figure 17: Simulation of migrants' trajectory route**

A detailed overview of the various possible routes will also be provided by the EMT giving end users and policy makers a better strategic understanding of the imminent flows.

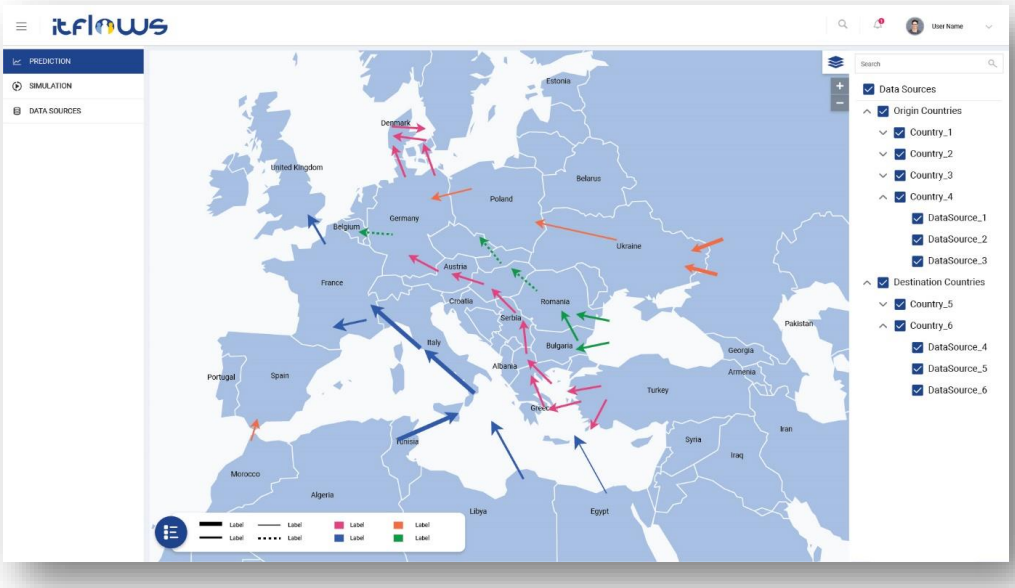


**Figure 18: Prediction of migrants' trajectory routes**

Prediction of migration outflow probability in origin countries is an information of significant value for the end-users of the EMT. Therefore, a detailed prediction is provided including a dynamic range of prediction time frames, choice of countries of origin and historical data used for the indicators.

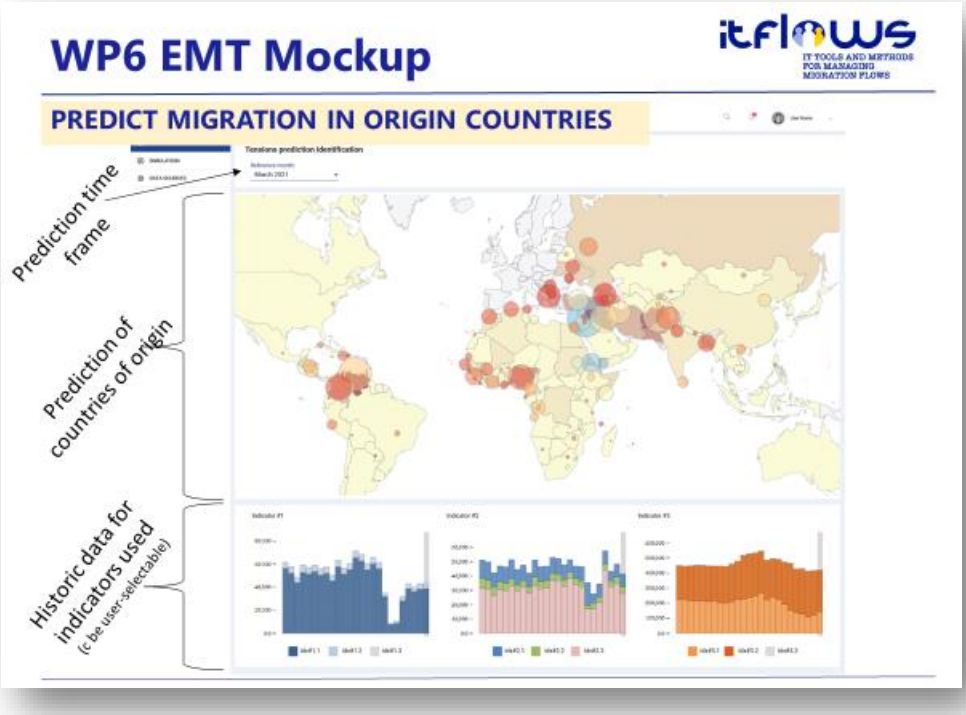


**Figure 19: Predict migration in origin countries**

Regarding migrants' smoothest settlement in the destination countries, identifying risks of tension towards them in the EU is extremely important. For that reason, a detailed map of the EU countries is provided with colour graded tension information. Tension predictions will be based countries of origin and countries of destination as a two-way algorithm. Each destination country will have the ability to show which source country creates more or less conflicts. Moreover, each source country can show which destination country has the most or less tension incidents for migrants originating from this source country.

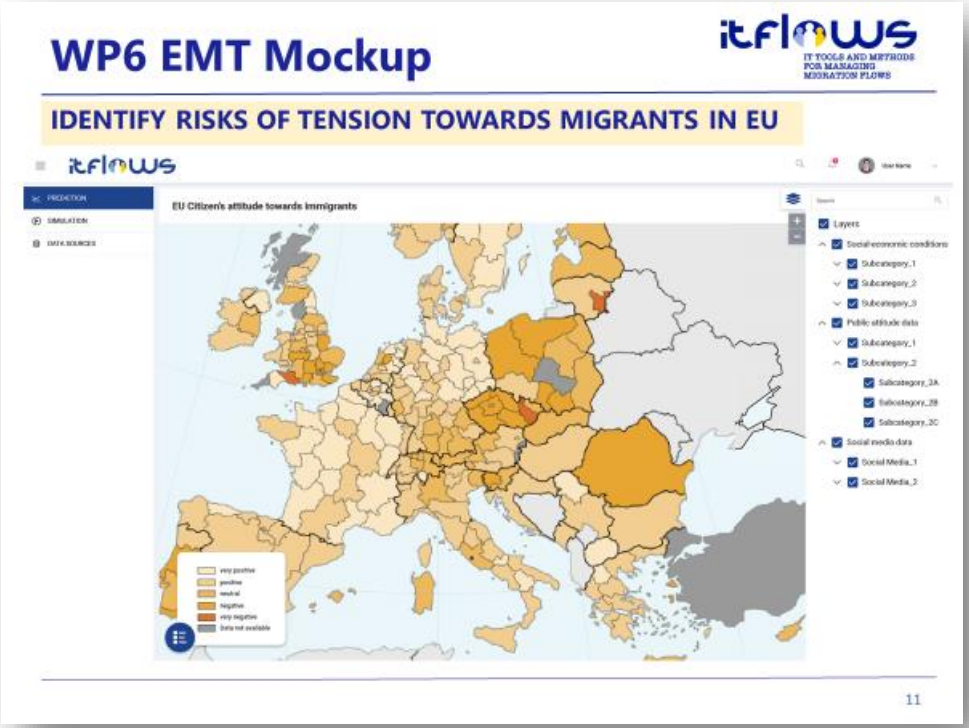


**Figure 20: Prediction of risks of tension towards migrants in EU**

Detailed analytics describing the anti-migration attitude will be provided for every country of choice.

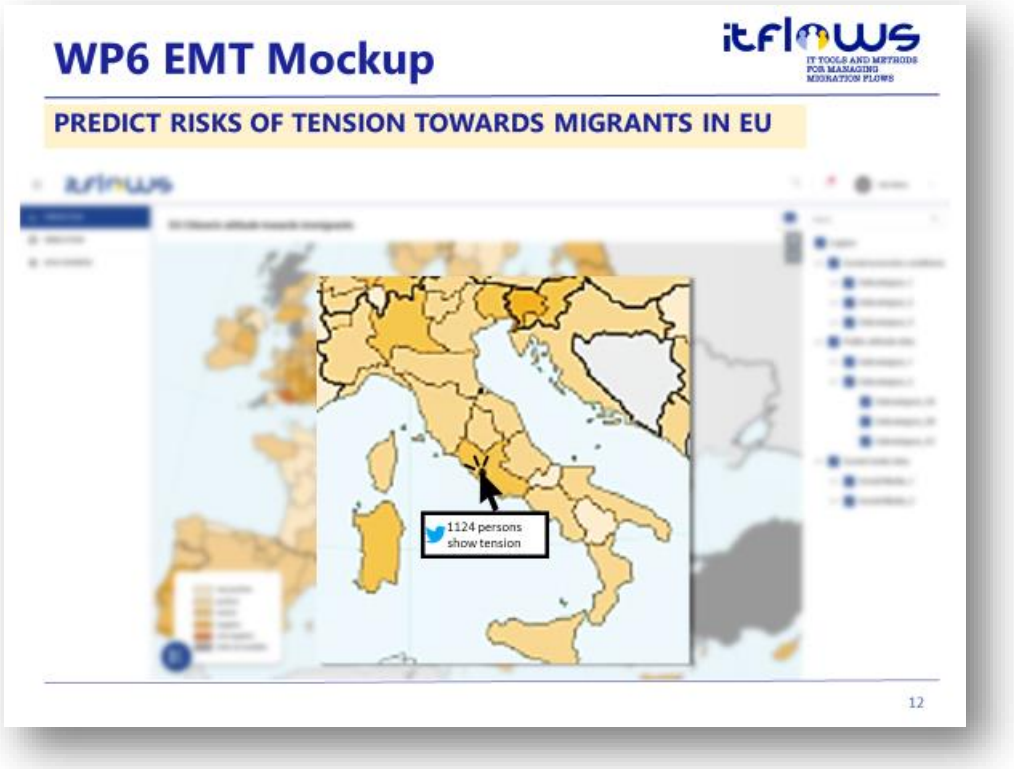


**Figure 21: Detailed analytics on risks of tension**

# 9. Conclusions and Future Work

This deliverable presented the data sources that the EMT application will utilise and analyse in order to provide the appropriate models regarding migrants' movements predictions and simulations, the two basic services the EMT application will provide to the end users. Moreover, it presented the user requirements as gathered from the work of WP7 (D7.1) and further elaborated within the consortium.

In addition, D6.1 summarises the methodology that will be used for developing the EMT application. This includes all the steps needed in order to create an application with friendly UI, capable of providing the required information tailored to the needs of the end users. Alongside the methodology, we discussed some of the basic components of EMT, including the core services of *prediction* and *simulation*, and the concepts and mock-ups of the User Interface that are focused in creating a friendly UI that every end user can easily operate.

The next step is the development of EMT and its services/tools. It is required to keep an open communication line with the end users to get continuous feedback that will help the development in terms of functionality and user experience. The development will take place in the coming months and the goal is to publish the first version of EMT with the D6.2 deliverable, which is scheduled for project month M18.

# 10. Appendix I

| Time Frequency | Geographical coverage - Destination | Geographical coverage - Origin | Related Tasks/Deliverables | Variables to be used by D3.2 (T3.1) | Variables to be used by D3.1 (T3.2) | Variables to be used by D3.4 (T3.2) | Variables to be used by D3.6 (T3.3) |
|---|---|---|---|---|---|---|---|
| | Europe | n.a. | **T3.1 (D3.2); T3.3 (D3.6)** | Asylum and first time asylum applicants by citizenship, age and sex (2008-2020), with specific attention to countries of origin: Afghanistan, Syria, Iraq, Eritrea, Nigeria, Mali, Morocco and Tunisia, Venezuela, Honduras, Colombia - so to identify asylum application trends from these countries of origin. | | | teiis010 Domestic producer prices - total industry (excluding construction)<br>teiis011 Import prices - total industry<br>teiis012 Import prices - manufacturing<br>teiis013 Import prices - intermediate goods<br>teiis014 Import prices - capital goods<br>teiis015 Import prices - consumer durables<br>teiis016 Import prices - consumer non-durables<br>teiis020 Domestic producer prices - manufacturing<br>teiis030 Domestic producer prices - energy<br>teiis040 Domestic producer prices - intermediate goods<br>teiis050 Domestic producer prices - capital goods<br>teiis060 Domestic producer prices - consumer durables<br>teiis070 Domestic producer prices - consumer non-durables<br>teiis080 Production in industry - total (excluding construction)<br>teiis090 Production in industry - manufacturing<br>teiis100 Production in industry - energy<br>teiis110 Production in industry - intermediate goods<br>teiis120 Production in industry - capital goods<br>teiis130 Production in industry - consumer durables<br>teiis140 Production in industry - consumer non-durables<br>teiis500 Production in construction<br>teiis550 Building permits - monthly data<br>asylum_applicants<br>first_time_asylum_applicants |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| monthly | migration routes (e.g. Western, Central, Eastern Med) | worldwide | **T3.1 (D3.2); T3.2 (D3.1)** | Detection of irregular crossings (2009-2020) at the Eastern, Central, Western Mediterranean Routes and the Western African Route by nationality, with specific attention to countries of origin: Afghanistan, Syria, Iraq, Eritrea, Nigeria, Mali, Morocco and Tunisia - so to identify trends of irregular arrival from these countries. | Detection of irregular crossings (2009-2020) at the Eastern, Central, Western Mediterranean Routes, the Western African Route, and the Western Balkan Route. | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| event based | worldwide | na | **T3.3 (D3.6)** | | | | COUNT_Battles No. of "Battles" events<br>COUNT_Explosions/Remote_violence "No. of ""Explosions/Remote violence"" events"<br>COUNT_Protests No. of "Violence against civilians" events<br>COUNT_Riots No. of "Protests" events<br>COUNT_Strategic_developments No. of "Riots" events<br>COUNT_Violence_against_civilians No. of “Strategic developments" events<br>FATAL_Battles Reported fatalities from "Battles" events<br>FATAL_Explosions/Remote_violence "Reported fatalities from ""Explosions/Remote violence"" events"<br>FATAL_Protests Reported fatalities from "Violence against civilians" events<br>FATAL_Riots Reported fatalities from "Protests" events<br>FATAL_Strategic_developments Reported fatalities from "Riots" events<br>FATAL_Violence_against_civilians Reported fatalities from “Strategic developments" events |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| event based | worldwide | na | **T3.2 (D3.1 - D3.4); T3.3 (D3.6)** | | Count of events (possibly also by category) and people killed and affected | Count of events (possibly also by category) and people killed and affected | COUNT_Complex_Disasters No. of Complex Disaster Events<br>COUNT_Natural No. of Natural Disaster Events<br>COUNT_Technological No. of Technological Disaster Events<br>DAMAGE_Complex_Disasters The amount of estimated damage caused by Complex Disaster Events, given in US$ ('000) at the moment of the disaster event.<br>DAMAGE_Natural The amount of estimated damage caused by Natural Disaster Events, given in US$ ('000) at the moment of the disaster event.<br>DAMAGE_Technological The amount of estimated damage caused by Technological Disaster Events, given in US$ ('000) at the moment of the disaster event.<br>FATAL_Complex_Disasters No. of deaths caused by Complex Disaster Events<br>FATAL_Natural No. of deaths caused by Natural Disaster Events<br>FATAL_Technological No. of deaths caused by Technological Disaster Events |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| yearly | worldwide | na | **T3.2 (D3.1 - D3.4)** | | To be explored/tested: sp_pop_totl - Population, total; sp_pop_dpnd_yg - Age dependency ratio, young (% of working-age population); ny_gdp_pcap_pp_kd - GDP per capita, PPP (constant 2017 international $); ny_gdp_mktp_pp_kd - GDP, PPP (constant 2017 international $); sl_uem_totl_zs - Unemployment, total (% of total labor force) (modeled ILO estimate); sl_uem_1524_zs - Unemployment, youth total (% of total labor force ages 15-24) (modeled ILO estimate); nv_agr_totl_zs - Agriculture, forestry, and fishing, value added (% of GDP); ne_trd_gnfs_zs - Trade (% of GDP); ne_gdi_ftot_zs - Gross fixed capital formation (% of GDP); sh_xpd_ghed_pp_cd - Domestic general government health expenditure per capita, PPP (current internat); sh_sta_bass_zs - People using at least basic sanitation services (% of population); | potentially a selection of those used in D3.1 | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | it_net_user_zs - Individuals using the Internet (% of population) | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| monthly, quarterly | worldwide | worldwide | **[Potentially] T3.2 (D3.1); T3.3 (D3.6)** | | Indicators to be possibly used: Consumer Price Index; value of Exports and Imports; interest rates on saving and deposit | | m_gdp Commodity Import Price Index, Individual Commodites Weighted by Ratio of Imports to GDP x_gdp Commodity Export Price Index, Individual Commodites Weighted by Ratio of Exports to GDP xm_gdp Commodity Net Export Price Index, Individual Commodities Weighted by Ratio of Net Exports to GDP |
| daily, monthly, yearly | worldwide | worldwide | **T3.3 (D3.6)** | | | | Selected keywords: T3.3 is currently working with a list of approximately 200 topical keywords in 10 diffferent origin country languages (i.e. Arabic, Dari, English, Farsi, French, Hausa, Pashto, Portuguese, Spanish, Turkish). in combination with the destination country name. Concrete/finalised list will be provided at a later stage. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| monthly | worldwide | worldwide | **T3.2 (D3.1 - D3.4); T3.3 (D3.6)** | | risk of coup d'etat, the wide type of government (democracy, interim, non-democracy), number of months since an irregular election took place, standardized index on precipatation | risk of coup d'etat, the wide type of government (democracy, interim, non-democracy), number of months since an irregular election took place, standardized index on precipitation | age "an approximation of the leader's age calculated by subtracting the leader's birth year from the current year" anticipation "a dummy variable that equals 1 if there is an election for the de facto leadership position coming within the next six-months" change_recent "a dummy variable that equals 1 if the de facto leader changed due to an election in the previous six months" defeat_recent "a dummy variable that equals 1 if an incumbent political party/leader won an election in the previous six months" delayed "a dummy variable that equals 1 if a previously scheduled/expected election is cancelled by choice or through exogenous factors (e.g. regime change)" direct_recent "a dummy variable that equals 1 if a direct (popular) election took place in the previous six months" elected "whether the de facto leader had previously been elected (1) or not (0) to their respective office" election_now "a dummy variable that equals 1 if there is an election for the de facto leadership position taking place in that country-month" election_recent "a dummy variable that equals 1 if there is an election for the de facto leadership position that took place in the previous six months" exec_ant "a dummy variable that equals 1 if there is an executive election to determine the de facto leader coming within the next six-months" exec_recent "a dummy variable that equals 1 if there is an executive election took |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | place in the previous six months"<br>government the regime type of the country during the observed country-month<br>indirect_recent "a dummy variable that equals 1 if an indirect (elite) election took<br>place in the previous six months"<br>irreg_lead_ant "a dummy variable that equals 1 if an irregular election to determine<br>the de facto leader is expected within the next six months"<br>irregular irregular (no description found in explanatory doc)<br>lastelection "an inverted decay function that measures the time since the last election<br>for the de facto leadership position within the country"<br>lead_recent "a dummy variable that equals 1 if any electoral opportunity (nonreferendum)<br>to change leadership took place in the previous six months"<br>leg_ant "a dummy variable that equals 1 if there is a legislative election to determine the<br>de facto leader coming within the next six-months"<br>leg_recent "a dummy variable that equals 1 if there is a legislative election took place in<br>the previous six months"<br>loss loss (no description found in explanatory doc)<br>male the sex of the de facto leader<br>militarycareer "equal to 1 if the leader's primary career and/or source of authority<br>comes from their career in the military, police force or defense ministry"<br>nochange_recent "a dummy variable that equals 1 if the de facto leader did not<br>change following an election in the previous six months"<br>prev_conflict "equal to the number of on-going violent civil and inter-state conflicts<br>that the country was involved in |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | during the previous month"<br>pt_suc a dummy variable that equals 1 if a successful coup event took place in that month<br>ref_ant "a dummy variable that equals 1 if there is a constitutional referendum coming within the next six-months"<br>ref_recent "a dummy variable that equals 1 if there is a constitutional referendum took place in the previous six months"<br>tenure_months "the number of months that a leader has been in power during their current tenure period"<br>victory_recent "a dummy variable that equals 1 if an incumbent political party/leader won an election in the previous six months" |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| monthly, quarterly, yearly | worldwide | worldwide | **T3.3 (D3.6)** | | | | CPI_ACPI_COI_RT National consumer price index (CPI), annual rate of change - discontinued (%)<br>CPI_MCPI_COI_RT National consumer price index (CPI), monthly rate of change - discontinued (%)<br>CPI_NCPD_COI_RT National consumer price index (CPI) by COICOP, percentage change from previous period (%)<br>CPI_NCYR_COI_RT National consumer price index (CPI) by COICOP, percentage change from previous year (%)<br>EAP_DWA1_SEX_AGE_RT Labour force participation rate by sex and age, seasonally adjusted series (%)<br>EAP_DWAP_SEX_AGE_EDU_RT Labour force participation rate by sex, age and education (%)<br>EAP_DWAP_SEX_AGE_GEO_RT Labour force participation rate by sex, age and rural / urban areas (%)<br>EAP_DWAP_SEX_AGE_RT Labour force participation rate by sex and age (%)<br>EAP_TEA1_SEX_AGE_NB Labour force by sex and age, seasonally adjusted series (thousands)<br>EAP_TEAP_SEX_AGE_EDU_NB Labour force by sex, age and education (thousands)<br>EAP_TEAP_SEX_AGE_GEO_NB Labour force by sex, age and rural / urban areas (thousands)<br>EAP_TEAP_SEX_AGE_NB Labour force by sex and age (thousands)<br>EAR_XEES_SEX_ECO_NB Mean nominal monthly earnings of employees by sex and economic activity (local currency)<br>EES_TEES_SEX_ECO_NB Employees by sex and economic activity (thousands)<br>EES_TEES_SEX_OCU_NB Employees by sex and occupation (thousands)<br>EIP_DWAP_SEX_AGE_EDU_RT Inactivity rate by sex, age and education (%)<br>EIP_DWAP_SEX_AGE_GEO_RT Inactivity rate by sex, age and rural / urban areas (%) |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | EIP_DWAP_SEX_AGE_RT Inactivity rate by sex and age (%)<br>EIP_TEIP_SEX_AGE_EDU_NB Persons outside the labour force by sex, age and education (thousands)<br>EIP_TEIP_SEX_AGE_GEO_NB Persons outside the labour force by sex, age and rural / urban areas (thousands)<br>EIP_TEIP_SEX_AGE_NB Persons outside the labour force by sex and age (thousands)<br>EIP_WDIS_SEX_AGE_NB Discouraged job-seekers by sex and age (thousands)<br>EIP_WPLF_SEX_AGE_NB Potential labour force by sex and age (thousands)<br>EMP_DWA1_SEX_AGE_RT Employment-to-population ratio by sex and age, seasonally adjusted series (%)<br>EMP_DWAP_SEX_AGE_EDU_RT Employment-to-population ratio by sex, age and education (%)<br>EMP_DWAP_SEX_AGE_GEO_RT Employment-to-population ratio by sex, age and rural / urban areas (%)<br>EMP_DWAP_SEX_AGE_RT Employment-to-population ratio by sex and age (%)<br>EMP_TEM1_SEX_AGE_NB Employment by sex and age, seasonally adjusted series (thousands)<br>EMP_TEM1_SEX_ECO_NB Employment by sex and economic activity, seasonally adjusted series (thousands)<br>EMP_TEMP_SEX_AGE_EDU_NB Employment by sex, age and education (thousands)<br>EMP_TEMP_SEX_AGE_GEO_NB Employment by sex, age and rural / urban areas (thousands)<br>EMP_TEMP_SEX_AGE_NB Employment by sex and age (thousands)<br>EMP_TEMP_SEX_ECO_NB Employment by sex and economic activity (thousands)<br>EMP_TEMP_SEX_OCU_NB Employment by sex and occupation |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | (thousands)<br>EMP_TEMP_SEX_STE_NB Employment by sex and status in employment (thousands)<br>HOW_TEMP_SEX_ECO_NB Mean weekly hours actually worked per employed person by sex and economic activity<br>HOW_TEMP_SEX_OCU_NB Mean weekly hours actually worked per employed person by sex and occupation<br>HOW_XEES_SEX_ECO_NB Mean weekly hours actually worked per employee by sex and economic activity<br>HOW_XEES_SEX_OCU_NB Mean weekly hours actually worked per employee by sex and occupation<br>LUU_XLU3_SEX_AGE_RT Combined rate of unemployment and potential labour force (LU3) by sex and age (%)<br>POP_XWAP_SEX_AGE_EDU_NB Working-age population by sex, age and education (thousands)<br>POP_XWAP_SEX_AGE_GEO_NB Working-age population by sex, age and rural / urban areas (thousands)<br>POP_XWAP_SEX_AGE_NB Working-age population by sex and age (thousands)<br>UNE_DEA1_SEX_AGE_RT Unemployment rate by sex and age, seasonally adjusted series (%)<br>UNE_DEAP_SEX_AGE_EDU_RT Unemployment rate by sex, age and education (%)<br>UNE_DEAP_SEX_AGE_GEO_RT Unemployment rate by sex, age and rural / urban areas (%)<br>UNE_DEAP_SEX_AGE_RT Unemployment rate by sex and age (%)<br>UNE_TUN1_SEX_AGE_NB Unemployment by sex and age, seasonally adjusted series (thousands)<br>UNE_TUNE_SEX_AGE_DUR_NB Unemployment by sex, age and duration (thousands)<br>UNE_TUNE_SEX_AGE_EDU_NB |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | Unemployment by sex, age and education (thousands)<br>UNE_TUNE_SEX_AGE_GEO_NB<br>Unemployment by sex, age and rural / urban areas (thousands)<br>UNE_TUNE_SEX_AGE_NB<br>Unemployment by sex and age (thousands) |
| event based | worldwide | worldwide | **T3.2 (D3.1 - D3.4)** | | Count of events and people killed and affected | Count of events and people killed and affected | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| monthly/season/yearly | worldwide | worldwide | **T3.2 (D3.1 - D3.4); T3.3 (D3.6)** | | - Temperature change on baseline 1951 - 1980 | - Temperature change on baseline 1951 - 1980 | cpifao Consumer Prices, General Indices (2015 = 100)<br>cpiFood Consumer Prices, Food Indices (2015 = 100)<br>foodInflation Food price inflation<br>temperature_d Temperature change<br>temperature_sd Standard Deviation |
| daily (in most cases) | selected countries for each specific situation | selected countries for each specific situation (e.g. Syria, Somalia, South Sudan) | **T3.2 (D3.4)** | | | Number of crossing in neighbouring countries | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| monthly | worldwide | worldwide | **T3.3 (D3.6)** | | | | del_regdem "delta of regav_dem, the monthly change in regav_dem value month to month" del_reggdp "delta of regionally averaged GDP, the annual change in the average regional GDP since the previous year" del_reggrow "delta of regionally averaged economic growth, the annual change in the average regional economic growth since the previous year" del_regten "ddelta of average regional government duration, the monthly change in regionally averaged governance duration from the previous month" dem_duration "a numeric variable that measures the logged number of months that a country has had a democratic government" elecViolence1 "whether there was any form of election related violence in the week leading up to an election, on election day or the week after an election" elecViolence2 "whether there was any form of government conducted election related violence in the week leading up to an election, on election day or the week after an election" gdpdiff "the relative difference between a country's GDP per Capita and the region it belongs to" gov_democracy "a dummy variable that is equal to 1 if a country is characterized as a presidential democracy or a parliamentary democracy" gov_interim "a dummy variable that is equal to 1 if a country is characterized as a civilian or military-led provisional government" growth the level of annual economic |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | growth based on GDP per Capita<br>growthdiff "the relative difference between a country's annual economic growth<br>and the region it belongs to"<br>l.elecViolence1 "a dummy indicator for whether the previous national election<br>recorded for a country was violent"<br>l.elecViolence2 "a dummy indicator for whether the previous national election<br>recorded for a country was experienced government conducted violence"<br>lnpop2 the estimated logged population of a country for that year<br>logIMR the logged annual infant mortality rate<br>logpredict "an estimated probability of the risk of a military coup attempt taking place<br>in the country-month"<br>logtenure "the logged number of months that a country's government type has<br>been in place"<br>nvio1 "number of violent events 1, the number of cumulative elections experiencing<br>any form of election related violence before the observed event"<br>nvio2 "number of violent events 2, the number of cumulative elections experiencing<br>government conducted election related violence before the observed event"<br>pcgdp the estimated annual median GDP per Capita for a country<br>political_violence "a numeric variable that measures the relative level (z-score)<br>of political violence experienced within the borders of a country for that year"<br>regav_dem the regional average of democratic duration (logged)<br>regav_growth regionally averaged economic growth<br>regav_IMR the regionally averaged logged annual infant mortality rate |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | regav_pcgdp regionally averaged annual median GDP per Capita<br>SPI the Standardized Precipitation Index (SPI) for each country month<br>tae1 "time after event 1, the number of peaceful election events that have taken place<br>since the last election experiencing any form of election related violence"<br>tae2 "time after event 2, the number of peaceful election events that have taken place<br>since the last election experiencing government conducted election related violence" |