IDIAP COMMUNICATION REPORT

THE EMN COUNTRY FACTSHEETS STRUCTURED DATASET

David Alonso del Barrio Daniel Gatica-Perez

Idiap-Com-01-2026

FEBRUARY 2026

Centre du Parc, Centre du Parc, Rue Marconi 19, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

The EMN Country Factsheets Structured Dataset

David Alonso del Barrio[1], **Daniel Gatica-Perez**[1, 2]

[1]Idiap Research Institute
[2]EPFL
ddbarrio@idap.ch, gatica@idiap.ch

Abstract

Each year, the European Migration Network (EMN) country factsheets deliver an overview of key migration and international protection developments within all EMN Member States and observer countries. The factsheets include both a textual component and a visual component. In this paper, we introduce a curated dataset of the textual component of these reports over 35 countries and 13 years (2012-2024.) The dataset was created to facilitate European-level research on migration, and promote the use of reliable sources about migration in data science and media research, particularly at a time when the spread of online misinformation about migration constitutes a serious issue. The dataset transforms the original document texts into a tabular format, with columns corresponding to country, year, section, subsection, content, and harmonized title section. The dataset is publicly accessible through this link: https://www.idiap.ch/dataset/emn-factsheets

1 Introduction

Migration is a topic of great interest in Europe, with significant media coverage (Eberl and Galyga 2021). It is an issue that generates polarization and opposing narratives, which can lead to mistrust and doubts in society about what truthful information is (Seiger et al. 2025). As it relates to data science and media research, we identified the need to promote the use of data considered reliable, in the sense that it comes from government offices that follow well-established protocols for data collection, and is not based on opinions. In the European Union, one of the main actors in the generation of reliable data about migration is the European Migration Network (EMN).

EMN is a network of National Contact Points (NCPs) composed of national network stakeholders with expertise in migration. All the EMN NCPs are designated by their respective national governments, and they are housed in a variety of institutions, including Ministries of Interior and Justice, specialized government agencies focused on migration, research institutes, non-governmental organizations, and national offices of international bodies. In addition to the EU member countries, EMN has non-EU member countries with observer status, including Norway, Georgia, Moldova, Ukraine, Montenegro, Armenia, Serbia, and Albania. These are countries that cooperate with the EMN by sharing information on migration and asylum, even though they are not EU members.

One of the main roles of the EMN NCPs is to generate reports on migration in their respective countries. These reports come in different formats, such as large studies involving several years of analysis; ad-hoc queries, where one country asks others about specific measures or issues and how other countries address them; annual reports from each country on migration-related issues; or factsheets, which summarize the main points related to migration in a shorter format than the annual reports.

In our case, we decided to create a dataset composed of the textual content of the factsheets. These documents are shorter versions of the extensive annual reports. The text in the factsheets is useful for describing and contextualizing policy change. The curated dataset is ideal for identifying major policy shifts and providing high-level context for Cross-National Comparative Research (CNCR). Aggregated data is essential to understand the macro-level policy context when analyzing migration outcomes. Different stakeholders like non-profit organizations or think tanks can use this type of authoritative comparative data to provide quantitative grounding for advocacy efforts and to benchmark performance across different Member States. Instead of having to open and manually process multiple PDF files to make comparisons, this dataset would facilitate data consultation by filtering by country, year, or section.

In the field of migration, there are various projects that aim to evaluate migration policies through indices, where experts from different countries first evaluate different policies through questionnaires and then provide scores based on different indicators. For example, the project Immigration Policies in Comparison (IMPIC) (Helbling et al. 2017; Berger et al. 2024) measures admission policies (border control, entry criteria, etc) of 33 OECD countries during the period 1980-2018.

Another example is the Migrant Integration Policy Index (MIPEX) (Yavçan and Gorgerino 2025), which assesses integration policies (such as labor market access, education, and citizenship paths) against a standard of equal rights across EU countries, relying on national experts to evaluate policy indicators.

A different approach is the one by Determinants of International Migration (DEMIG) and Quantifying Migration

Scenarios for Better Policy (QuantMig) projects (De Haas, Natter, and Vezzoli 2014; Schreier, Skrabal, and Czaika 2023), since they track specific policy events or changes over time, rather than a static index score. This approach is designed for research into the long-term evolution and effectiveness of policy changes, allowing researchers to study cause-and-effect relationships over decades. Some of the entries in QuantMig dataset came from the EMN factsheets as a source.

Our curated dataset complements these indices and datasets, as it does not focus on evaluation, but rather presents reliable information reported by the countries evaluated in these indices in a structural way as all factsheets, for different countries and years, share a common section structure, which facilitates temporal and cross-country analysis. This facilitates research on how the governments of these countries report on migration issues, facilitating understanding of why some countries perform better than others in these indices and whether there is a correlation with the data provided in the factsheets. It can also support data science and media research as provides structured, reliable data that could be used facts in disinformation research.

2 Data description

EMN factsheets are annual summaries of key migration and international protection developments, trends, and statistics for each EMN member country. They are divided in two big parts, one textual and one visual. In the textual part, each country explains the main points related to topics like asylum, legal migration, integration, or irregular migration, while in the visual part there is a statistical annex that complements the textual part with plots and statistics. Our dataset is focused on the textual part. This part is divided into sections.

The sections are defined by titles like ”LEGAL MIGRATION” ”BORDERS, SCHENGEN AND VISA” or ”TRAFFICKING IN HUMAN BEINGS”, and all these titles are part of column *”section”* in our dataset. Over the years, the titles of these sections in the textual part have slightly changed. For example, the titles presented above appear with some variation in certain years: ”LEGAL MIGRATION AND MOBILITY”, ”BORDERS AND VISAS”, or ”ACTIONS AGAINST TRAFFICKING IN HUMAN BEINGS”. These small variations in the titles motivated the creation of an additional column called *”grouped_title_section”*, such that if needed, it is possible to compare all the sections with the same content in terms of topic over the years. Inside each section, there are cases with subsections (e.g., section: LEGAL MIGRATION AND MOBILITY, subsection: FAMILY REUNIFICATION). All these subsections are part of column *”subsection”* in our dataset. In order to keep the structure of the dataset, in the cases of sections without subsection, we used the term ”Main” as the name of the ”subsection”. Then, column *”content”*, contains the specific content to that section and subsection, and columns *”year”* and *”country”* specify the corresponding year and country of the report.

Figure 1 shows the countries and the number of years in which they have published EMN factsheets. It is important to mention that not all the countries have data in all the years, e.g., the United Kingdom: while reports exists for the period 2012-2018, the country stopped reporting to EMN due to their departure from the EU. There are other countries like Serbia (2023, 2024), Montenegro, Armenia (2022, 2023, 2024), or Ukraine (2024) that began to be reported in recent years, or the cases of Denmark (2012, 2013) and Romania (2012, 2013 and 2014), which were only reported in specific years.

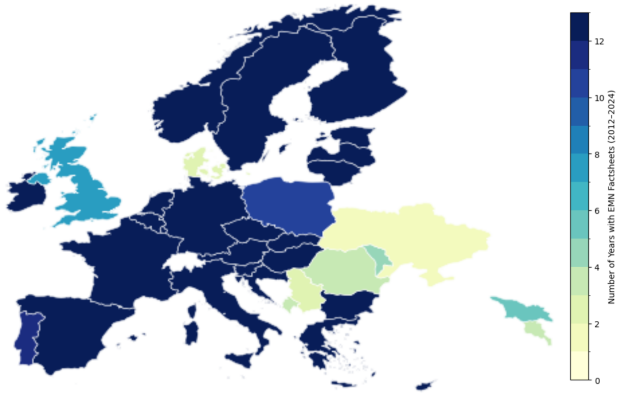

Figure 1: Geographic coverage of the EMN factsheets textual dataset across Europe for the period 2012–2024. Each country is shaded according to the number of years in which an EMN factsheet was available, with darker tones indicating more frequent participation.

Table 1 shows an example of one sample and the total number of countries (35), years (13), section (37), subsections (99), samples (4801) and grouped titles’ section (15) in the dataset.

Table 1: Structure and example content of the EMN factsheets Text Dataset.

Columns	**Example**	**Total**
country	France	35
year	2021	13
section	BORDERS, SCHENGEN AND VISAS	37
subsection	BORDER MANAGEMENT	99
content	A Strategic Border Committee (under the General-Directorate for Foreign Nationals in France, DGEF) was established, along with an Operational Committee for the Border Guard (under the Central Directorate for Border Police, DCPAF), to align and better coordinate the administrations involved in border control or surveillance at the strategic level.	4801
grouped_title_section	BORDERS AND VISAS	15

3 Data Collection and Curation Process

In this section, we explain the process of data collection and curation. The EMN website[1] has a specific section about the EMN factsheets, where anyone can download PDF-format factsheets for various countries and years. We downloaded all the available documents from 2019 to 2024. In addition, we asked the European Commission if there were factsheets from previous years that were not available online. The Commission provided us with the factsheets in PDF from 2012 until 2018, for a total of 325 PDFs. We used the Pymupdf python library to extract the textual part of the factsheets[2]. For PDFs from 2017 to 2020 ,for some reason the factsheets saved the text as an image, so we used the Pytesseract python library [3], which is an optical character recognition (OCR) library that allows to extract the text embedded in images.

For each year, we first extracted the text part and created a .csv file for each year with three columns: country, year and text. It was necessary to manually correct typos generated mainly by the extraction of text from images (e.g., incorrect symbols.) In addition, when extracting the text, the superscripts marking the footnotes and the text of the footnotes were also extracted, so we carried out a thorough cleanup to remove all that text.We acknowledge that some typos may persist, but the content is fully understandable, and such errors are expected to occur in very few cases.

Once the content was cleaned, we merged all the CSV files into one single CSV file with all the content of all years and countries. To pass from 3 columns (country, year and text) to the final version of our dataset with 6 columns (country, year, section, subsection, content and grouped_title_section), we splitted the Text column into section, subsection and content, as described in Section 2.

4 Legal aspects

The European Commission copyright notice [4] indicates that: *"The Commission's reuse policy is implemented by the Commission Decision of 12 December 2011 on the reuse of Commission documents. Unless otherwise indicated (e.g. in individual copyright notices), content owned by the EU on this website is licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence. This means that reuse is allowed, provided appropriate credit is given and changes are indicated."*

The content of our dataset complies with the license, as we are reusing documents provided by the European Commission, to whom we give full credit for the content generated. Likewise, in Section 3, we explained the minor changes we have made, such as removing footnotes and creating a column that in a way standardizes the section titles in order to be able to compare them over the years, recognizing that our curated dataset is not endorsed by the European Commission.

[1]https://home-affairs.ec.europa.eu/networks/european-migration-network-emn/emn-publications/country-factsheets_en

[2]https://github.com/pymupdf/PyMuPDF

[3]https://pypi.org/project/pytesseract/

[4]https://commission.europa.eu/legal-notice_en

5 Potential applications and uses

This dataset allows the content of the EMN factsheets from 2012 to 2024 to be contained in a single file, enabling different uses. Below, we provide a few examples:

- The dataset could be visualized though a dashboard where experts on migration can easily compare between countries and years the different policies applied, instead of opening several PDFs to make that kind of comparison or analysis. Though a filter where users can select year, country, section and subsection, they could compare reports in one single page. In Figure 2, we show an example of an interactive prototype we have built. This could facilitate analysis between countries, or analysis and temporal evolution of a specific country, thereby facilitating the work of stakeholders on migration issues.
- The dataset could be integrated into a RAG system where a citizen or expert can ask about different policies in Europe regarding migration, avoiding hallucinations by constraining the answer to the content of the dataset. This would enable interested individuals (citizens or researchers) to be better informed, facilitating access to resources considered reliable sources that can be consulted when questions arise while reading social media or newspapers, or conducting research on data science or media, thus serving as an additional source of information.
- As the data came from European governments and it is considered a reliable resource, it could be used to extract claims to conduct fact checking by journalists, as a way to combat misinformation that may exist around this topic.
- The dataset could allow a Cross-National Lexical and Terminology Variation Study (Computational Linguistics) to automatically identify country-specific or year-specific jargon for common migration concepts (e.g., different national terms used for "trafficking victim" or "irregular migrant".
- The factsheets represent the official, fact-based documentation of European governments, contrasting sharply with the often polarized discourse on social media. A possible use could be to compare the sentiment in the official EMN reports to the sentiment observed in social media discussions about specific migration trends.

6 FAIR principles

The dataset has been curated and documented in accordance with the FAIR principles (Findable, Accessible, Interoperable, and Reusable), ensuring its long-term usability by the research community, policymakers, and other stakeholders.

- **Findable.** The dataset is deposited in a public data repository and assigned a Digital Object Identifier (DOI).
- **Accessible.** The dataset is openly accessible through this link: https://www.idiap.ch/dataset/emn-factsheets
- **Interoperable.** The dataset is distributed in CSV format, which can be easily processed in Python, R, or other data analysis environments.

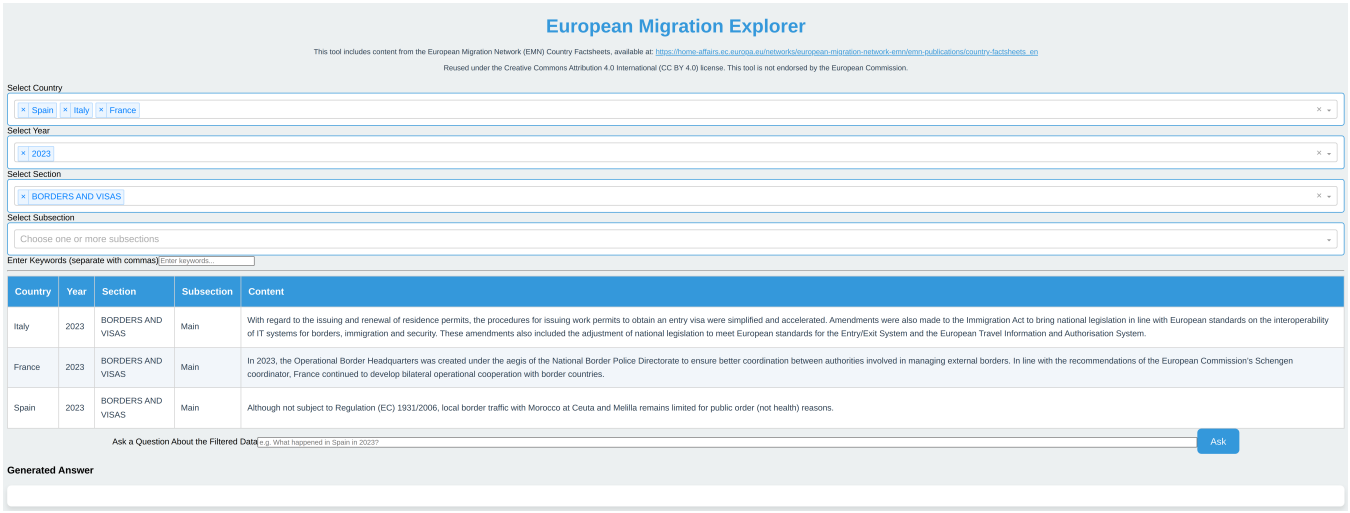

Figure 2: The European Migration Explorer dashboard interface. The top section contains multiselect filters for country, year, section, and subsection, along with a keyword search functionality. The central panel displays a dynamic results table containing the filtered textual data, organized by metadata columns and the extracted content. At the bottom, an interactive query interface allows users to ask natural language questions about the filtered subset, utilizing a Retrieval-Augmented Generation (RAG) approach to provide summarized answers.

- **Reusable.** Comprehensive documentation is provided, including data structure, extraction and cleaning methodology, variable descriptions, and known limitations. The open license (CC BY 4.0) allows reuse, redistribution, and adaptation for both academic and policy-oriented purposes. The data curation ensures internal consistency across years and countries, supporting longitudinal and comparative analyses.

7 Datasheet for dataset

- **Motivation.** This dataset consolidates the textual sections of the European Migration Network (EMN) factsheets (2012–2024) to enable comparative analysis of migration and asylum policies across EMN member countries.
- **Composition.** It contains 4,767 entries covering 35 countries and 13 years. Each entry includes country, year, section title, subsection, grouped title (standardized thematic label), and the corresponding text.
- **Collection and Processing.** Texts were extracted from publicly available EMN factsheet PDFs using semi-automated parsing and manual validation. Titles were harmonized across years, and only the textual sections (excluding statistical annexes) were included.
- **Format and Accessibility.** The dataset is available in CSV format under a Creative Commons Attribution (CC BY 4.0) license. It will be hosted in a public repository.
- **Intended Use.** The dataset supports policy research, text analysis, and NLP applications such trend detection, and cross-country comparisons in migration governance.
- **Maintenance.** Future releases will incorporate new annual EMN factsheets, with each version documented and versioned in the public repository.

8 Conclusions

In this paper, we present a dataset containing the structured textual content of EMN factsheets, which provide annual overviews on the main developments and policies implemented in the area of migration across 35 countries over a 13 year period (2012-2024). This data is designed for diverse applications, serving both migration experts and researchers focused on computational social science. Furthermore, by providing a structured and reliable primary source, this dataset facilitates the reporting and verification of news stories, offering a way to counter the spread of online misinformation regarding migration.

9 Acknowledgments

This work was supported by the ELIAS project, funded by the European Commission (Grant 101120237). We would like to thank the European Migration Network (EMN) for providing us with factsheets from previous years that are no longer publicly available.

References

Berger, V.; Bjerre, L.; Breyer, M.; Helbling, M.; Römer, F.; and Zobel, M. 2024. The immigration policies in comparison (impic) dataset: Technical report v2.

De Haas, H.; Natter, K.; and Vezzoli, S. 2014. Compiling and coding migration policies: Insights from the DEMIG POLICY database.

Eberl, J.-M.; and Galyga, S. 2021. Mapping media coverage of migration within and into Europe. In *Media and Public Attitudes Toward Migration in Europe*, 105–122. Routledge.

Helbling, M.; Bjerre, L.; Römer, F.; and Zobel, M. 2017. Measuring immigration policies: The IMPIC database. *European Political Science*, 16(1): 79–98.

Schreier, S.; Skrabal, L.; and Czaika, M. 2023. DEMIG-QuantMig Migration Policy Database. *QuantMig Project Deliverable D*, 5.

Seiger, F.; Kajander, N.; Neidhardt, A.-H.; Scharfbilling, M.; Drazanová, L.; Deuster, C.; Krawczyk, M.; Blasco, A.; Icardi, R.; Tzvetkova, M.; et al. 2025. Navigating migration narratives: Research insights and strategies for effective communication.

Yavçan, B.; and Gorgerino, M. 2025. MIPEX 2025 – A Roadmap for Inclusive Policy in the EU.