CELSA/23/003

A Data Management Plan created using DMPonline.be

Creators: Sercan Kiyak, Leen d'Haenens

Affiliation: KU Leuven (KUL)

Template: FWO DMP (Flemish Standard DMP)

Project Administrator: Sercan Kiyak

Grant number / URL: CELSA/23/003

ID: 202424

Start date: 01-10-2023

End date: 30-09-2025

Project abstract:

This research project explores how migration is employed as a political tool, leading to observable political polarization expressed through language use. It investigates two key events: the 2015–2016 European migration crisis and the migration resulting from Russia's invasion of Ukraine in 2022. The study also compares Belgium and Hungary, EU Member States with contrasting media systems, migration attitudes, and experiences with these crises, serving as a pilot study for potential broader EU research. The research employs a comprehensive approach, analyzing various layers of migration-related political discourse, including official channels (e.g., politicians' social media discourse, parliamentary speeches), mainstream press, and user-generated content (e.g., tweets in Belgium and online forums in Hungary). This analysis encompasses both professional and non-professional communication. Beyond content analysis, the study explores interactions and diffusion processes and incorporates polling data to understand the interplay between political discourse, media, and public opinion. In short, this research project will investigate: 1) how migration content enters the political sphere; 2) characteristics of migration discourses; 3) audience responses on Twitter (Belgium) and news site forums (Hungary); and 4) language polarization in diffusion processes. This analysis will examine the connection between political discourse and latent opinion climates in Belgium and Hungary, exploring both similarities and discrepancies.

Last modified: 03-01-2024

CELSA/23/003
Application DMP

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects …) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

The research project encompasses a diverse range of data types sourced from academic databases. To ensure consistency and facilitate comparative analysis, we will maintain diversity in political ideology, levels of engagement (lay and public), and official levels across these datasets. The primary storage format will be .csv, with a specific focus on textual content and accompanying metadata. We currently do not plan to generate new datasets but reuse existing sources and curate datasets (from APIs and databases) for our topic.

In accordance with current standards, the size of the dataset is expected to be manageable. The datasets will be drawn from various sources, and below, we provide a tentative list:

1) Hungarian Datasets:

1A) Media: This dataset encompasses the last two decades of the Hungarian online press corpus, including major news sites and discussions within Disqus forums. This will enable the monitoring of press-generated opinions related to migration.

1B) Public: Representing the lay political public, this dataset covers contemporary (last three years) Hungarian politics-related internet texts. It includes content from blogs, social media, and similar platforms.

1C) Political Actors: This dataset comprises texts representing official political actors, incorporating parliamentary speeches spanning 30 years and prime minister speeches covering the last 15 years.

2) Belgian Datasets:

2A) Media: This dataset includes migration crises-related news items from various political tendencies and languages from Belgian news sources, to be collected through iCandid, a digital research infrastructure for researchers at KU Leuven.

2B)Public: Social media messages from Twitter will be collected via iCandid to represent lay-public opinion and the online discussions about migration in the Belgian society.

2C) Survey: Additionally, we plan to reuse Belgian and Hungarian survey data from two research projects: H2020 HumMingBird and OPPORTUNITIES surveys (Approval ID: G-2020-2590) from the SMEC commission.

2D) Political Actors: Texts of speeches by political actors in the Belgian parliament will be scraped from the Federal Parliament API to represent the official level.

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

Data protection officers: Sercan Kiyak (IMS, KU Leuven)

Storage Capacity/Repository:

During the Research: Given that the primary data source is textual, the datasets are anticipated to be of manageable size. They will be stored on the KU Leuven's Teams-site.

After the Research: The majority of our dataset contains media items belonging to media institutions or individual messages on social media, imposing significant restrictions on online sharing. However, we envisage sharing public speech data of politicians dataset via KU Leuven RDR (Research Data Repository).

We are committed to sharing the public segments of our data to enhance accessibility and facilitate reproducibility, including our code. Beyond a five-year timeframe, we do not intend to retain the data.

What’s the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

We do not wish to deviate from the preservation of data principle and plan to keep it stored for 5 years.

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

Our dataset primarily comprises public messages and speeches. For data relating to individual users (e.g., commenters on news items or social media messages), we intend to anonymize it and conduct aggregate-level analysis to study broader public opinion and communication patterns

regarding issues, rather than focusing on the individual level. Consequently, we do not anticipate any ethical and privacy concerns regarding our data usage.

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

We intend to employ textual datasets for topic detection, aiming to identify emerging issues and perspectives in news, public discourse, and official speeches. This computational research approach may necessitate the use of cloud computing services, such as the Google Cloud API. We will exercise caution to ensure that the datasets under analysis are not shared unlawfully.

CELSA/23/003
FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
Hungarian Media Data	Last two decades of Hungarian online press corpus and Disqus forums.	Reuse existing data	Digital	Compiled/aggregated data	.csv	<100GB	
Hungarian Public Discussion Data	Last three years of Hungarian politics-related internet texts from blogs and social media.	Reuse existing data	Digital	Compiled/aggregated data	.csv	<100GB	
Hungarian Politics Data	Texts from parliamentary speeches (30 years) and prime minister speeches (last 15 years).	Reuse existing data	Digital	Compiled/aggregated data	.csv	<1GB	
Belgian Media Data	Migration-related news items from Belgian sources in various languages (via iCandid)	Reuse existing data	Digital	Compiled/aggregated data	.csv	<100GB	
Belgian Social Media Data	Twitter messages representing lay-public opinion on migration.	Generate new data	Digital	Compiled/aggregated data	.csv	<100GB	
Belgian Politics Data	Speeches by political actors in the Belgian parliament.	Generate new data	Digital	Compiled/aggregated data	.csv	<1GB	
Survey Data	Reused Belgian and Hungarian survey data from H2020 HumMingBird and OPPORTUNITIES projects.	Reuse existing data		Survey results	.csv	<1GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

Belgian and Hungarian survey data in the context of our H2020 HumMingBird and OPPORTUNITIES surveys for which we obtained already approval of the SMEC commission. Approval ID: G-2020-2590

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

The datasets primarily consist of public messages and speeches. For data related to individual users, such as commenters on news items or social media messages, we will anonymize the information and conduct aggregate-level analysis. No ethical issues concerning the creation and/or use of the data have been identified.

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes

We will use social media data that is published by users on the public web spaces and social media platforms. We plan to use them at aggregate level and do not intend to infringe any privacy laws. Personal data will be de-identified and analyzed in aggregate form. No personal data will be used in a manner that allows for the re-identification of individuals.

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Our research and subsequent analysis are oriented towards making a substantive social and political contribution to the ongoing discourse surrounding migration. The overarching objective is to provide tangible support to Non-Governmental Organizations (NGOs) and policy-making entities. It is important to underscore that our endeavors do not envision a technological or commercial application. Rather, our primary aim lies in advancing the societal dialogue and informing policy decisions pertinent to this critical issue.

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- Yes

Data procured from social media platforms through APIs is subject to restrictive agreements. Furthermore, sharing collected news item data is precluded to respect the proprietary rights of the companies involved. While these limitations do not impede our investigation of the chosen topics, they do constrain our ability to disseminate the data publicly. Nevertheless, in adherence to best practices, we are committed to retaining our research data for a period of five years, ensuring its availability for quality assessment and further academic scrutiny.

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

Textual data for our research will be collected from established databases such as iCandid and social media APIs and we plan to adhere to their data structures. Moreover, whenever possible example anonymized data and our scripts will be provided in forms of research code notebooks with comments and explanations. Finally, extensive annotations and metadata information will be stored alongside our datasets.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- Yes

We aim to adhere to DDI (Data Documentation Initiative). We will ensure that appropriate metadata is created to facilitate internal

organization and future reference.

3. Data storage & back-up during the research project

Where will the data be stored?

The data will be stored in KU Leuven shared One-Drive service, using Teams-site. It will be shared with project partners under the supervision and anonymized.

How will the data be backed up?

We will use standard data back up services of KU Leuven.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

We will use cloud data storage services of KU Leuven, namely the Teams-site cloud storage.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The project datasets will be shared only among the project partners. In case, datasets are shared outside of the research project (for reproducibility) they would be anonymized. No access or editing rights will be given to third parties.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

Storage option offered by KU Leuven is free and big enough for the project data.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

All data will be preserved for 5 years.

Where will these data be archived (stored and curated for the long-term)?

The data will be stored on cloud servers at KU Leuven Teams-site. The data will be kept in the cloud storage under supervision for 5 years.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

There are no expected costs for data storage and backup.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- No (closed access)

The data will be shared only between the project partners and under supervision.

If access is restricted, please specify who will be able to access the data and under what conditions.

The research data will be accessible to project partners at Belgian and Hungarian academic partners.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- Yes, Privacy aspects
- Yes, Intellectual Property Rights
- Yes, Ethical aspects

Where will the data be made available? If already known, please provide a repository per dataset or data type.

The data will not be made available due to restrctions of databases and APIs. If data from some work packages (such as politician speeches dataset) becomes possible to be shared they would be put onto KU Leuven RDR (Research Data Repository).

When will the data be made available?

Most of the data is not possible to share publicly. Moreover, due to changes in APIs and data limits is not possible to recollected by third parties. However, after the publication of results any work package data which can be shared publicly will be distributed online with the academic community.

Which data usage licenses are you going to provide? If none, please explain why.

We will use the Data Transfer Agreement. We will further look into the specific data usage licences when we evolve in the research project.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

What are the expected costs for data sharing? How will these costs be covered?

There are no expected costs for data sharing.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Sercan Kiyak (IMS, KU Leuven)

Who will manage data storage and backup during the research project?

Sercan Kiyak (IMS, KU Leuven)

Who will manage data preservation and sharing?

Sercan Kiyak (IMS, KU Leuven)

Who will update and implement this DMP?

Sercan Kiyak (IMS, KU Leuven)