TWIN4DEM: Strengthening democratic resilience through Digital Twins

D4.1

Survey, elections, institutions, and socio-economic data procurement document

Grant Agreement No	101178061	**Acronym**	TWIN4DEM
Full Title	Strengthening Democratic Resilience Through Digital Twins		
Start Date	1 January 2025	**Duration**	36 months
Project URL	https://twin4dem.eu/		
Deliverable	Survey, elections, institutions, and socio-economic data procurement document		
Work Package	4		
Deliverable Type	Document, report	**Dissemination Level**	Public
Due Date of Deliverable	31/12/2025	**Actual Submission Date**	19 December 2025
Deliverable Identifier	D4.1	**Deliverable Version**	1.0
Lead Beneficiary	GESIS		
Authors	Fabienne Straßegger (GESIS), Dennis Abel (GESIS)		
Co-authors	Sara Tonelli (FBK), Nina Hänel (GESIS), Sebastian Ziaja (GESIS), Alexia Katsanidou (GESIS)		
Reviewers	Sara Tonelli (FBK), Raul Magni-Berton (ICL), Julien Navarro (ICL)		
Status	☒ Draft	☒ Peer Reviewed	☒ Coordinator Accepted

Document history

Version	Date	Partner	Remarks
0.1	16.11.2025	GESIS	ToC shared with contributors
0.2	30.11.2025	GESIS	Version shared for internal review
0.3	15.12.2025	GESIS	Version shared after internal review
1.0	19.12.2025	GESIS	Submitted version

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Executive Summary

This deliverable documents the data integration strategy for survey, elections, institutions, and socio-economic data for the TWIN4DEM Digital Twin (DT), a pioneering tool aimed at advancing democracy research through computational social science (CSS). The DT replicates core institutional dynamics of democratic systems to simulate scenarios of executive aggrandizement, a process by which elected leaders legally expand executive powers at the expense of parliaments and courts. By modelling these interactions, TWIN4DEM addresses a critical gap in understanding the multidimensional pathways that lead to democratic backsliding.

The objective of WP4 is to provide four integrated datasets for each case study that will feed the Digital Twin developed in WP5. Data sources include socio-demographic and socio-economic data from surveys, elections, censuses, as well as textual data processed and compiled in WP3. The main objectives of WP4 are (a) to identify the data sources that are best fitted for usage in the Digital Twin (T4.1 and T4.2), (b) to clean, harmonize, aggregate, and link them in accordance with legal and ethical provisions, so that they are ready to be used in the models of WP5 (T4.3).

This work in WP4 connects to other WPs by providing the bottleneck of data integration between WP3, WP4, and WP5. This work in project year 1 directly feeds into the subsequent steps in the project years 2 and 3. In particular Task 4.3, "Integrating and updating the data" starts in M13 and builds on the previous steps in WP3 and WP4. This includes aggregating the population data to suitable temporal and spatial units. The goal of this task is to integrate these properties appropriately into the simulation models. In sum, these measures will result in (synthetic) datasets that comprehensively represent the societies examined by TWIN4DEM. The TWIN4DEM integrated dataset (D4.2.) will be released as an outcome of this task in M32.

Table of Contents

List of Figures

List of Tables

ABBREVIATIONS

Abbreviation	Description
CC0	Creative Commons Zero license
CHES	Chapel Hill Expert Survey
CLEA	Constituency-Level Elections Archive
CSS	Computational Social Sciences
CSV	Comma-Separated Values
DMP	Data Management Plan
EEAB	External Ethical Advisory Board
EFA	Exploratory Factor Analysis
FAIR	Findable, Accessible, Interoperable, Reusable
GDPR	General Data Protection Regulation
ISO	International Organization for Standardization (e.g., ISO country codes)
NER	Named Entity Recognition
NLP	Natural Language Processing
OSF	Open Science Framework
V-Dem	Varieties of Democracy dataset
DOI	Digital Object Identifier

1. Introduction

1.1 Purpose of the document

The objective of WP4 is to provide four integrated datasets for each case study that will feed the Digital Twin developed in WP5. Data sources include socio-demographic and socio-economic data from surveys, elections, censuses, as well as textual data processed and compiled in WP3. The main objectives of WP4 are (a) to identify the data sources that are best fitted for usage in the Digital Twin (T4.1 and T4.2), (b) to clean, harmonize, aggregate, and link them in accordance with legal and ethical provisions, so that they are ready to be used in the models of WP5 (T4.3).

Deliverable 4.1 (D4.1) on "Survey, elections, institutions, and socio-economic data procurement document" is the direct output of tasks 4.1 and 4.2 and documents the process to collect and aggregate the data. Task 4.1, "Procuring survey data" (Lead: GESIS, plus: EUR, ICL, CSS, CUNI) gathers surveys from the four case study countries that contain items representing concepts of interest for the population model as defined in WP2. Items will cover demographics, core values, political attitudes, voting intentions and behavior, as well as related issues. The data covers the years 2010-2024. Task 4.2, "Procuring elections, institutional, and official statistics data" (Lead: GESIS, plus: EUR, ICL, CSS, CUNI) further gathers other data types providing information on elections, institutions, and socio-economic traits, covering the years 2000-2024, conditional on the data demand in WP5.

1.2 Structure of the document

This deliverable is structured as follows to provide a comprehensive overview of the data procurement. Following the introduction, Section 2 outlines the legal and ethical provisions guiding data collection, management, and sharing, with particular reference to the project's Data Management Plan. Section 3 describes the identification, collection, and curation of data, including an overview of the main data sources, coordination across work packages, and the operationalization of key model dimensions. This section details survey, election, roll call, institutional, and socio-economic data, and explains how these sources feed into the core components of the Digital Twin (WP5). Section 4 addresses the technical implementation of data harmonization, integration, and linkage across heterogeneous sources, as well as data storage and sharing infrastructures. Finally, Section 5 provides a summary of the main outcomes of Deliverable 4.1 and situates these results within the broader project timeline, highlighting the next steps towards the integrated dataset to be produced in Task 4.3.

2. Legal and ethical provisions – Data Management Plan (D1.2)

TWIN4DEM's Data Management Plan (D1.2) (de Mooij et al., 2025) and the EEAB monitoring activity provide a rigorous framework to ensure that the collection and handling of all data meets Open Science, FAIR (Findable, Accessible, Interoperable, Reusable) and GDPR standards, and that all partners are made aware of their responsibilities. Findability of Twin4Dem software and code will be ensured through publishing on GitHub (https://github.com/TWIN4DEM-project) and integration with OSF to allow for assigning the persistent identifier. Links to both communities will be published and actively promoted through the Open-Source Community on the TWIN4DEM website (https://twin4dem.eu). Accessibility will be provided with open, restricted, and closed access levels, depending on a risk assessment of legal, ethical or contractual obligations. Metadata and code will always be made public under CC0 license. Taking into account the interdisciplinarity of the consortium and extensive and heterogeneous data sources used in different work packages, interoperability and reuse is crucial. The DMP defines core guiding actions:

- Using standard formats: open and non-proprietary file formats will be used for all datasets where possible. Software and code will be written in widely supported languages (i.e., R or Python) and follow language-specific conventions.
- Applying standard metadata schemes: standard metadata schemes will be used for datasets, software and code.
- Language: description of datasets and metadata are in English. Domain-specific vocabularies and ontologies will be used to ensure consistency across datasets and systems.
- Models and algorithms: Python and R languages are used as primary open-source programming languages.
- Libraries and tools: to promote replicability, there is emphasis on those supporting a wide range of computational tasks.

Quality practices on licensing, documentation, naming and versioning will be followed to increase secondary use of the produced code and datasets.

3. Data procurement and integration

3.1 Data identification, procurement, and curation process

The project is based on a complex data structure derived from case study research, social media data, legal documents, official statistics, and survey data. Data and software generation is divided among the work packages as reported in table 1 from the DMP.

Table 1: Data and software summary

Methodology	Data Type	Formats	Data Source	Size	Research Output
T2.1 Definition of decision-making rules (ICL)					
Literature review	Textual data, tabular data	.pdf .docx, .txt, .csv, .xlsx,	Existing datasets, policies, openly available reports, theoretical models, literature	<10 GB	Models, codebook
T2.2 Define and operationalise preferences and behaviours (ICL)					
Literature review	Tabular data	.csv, .xlsx	Literature	<10GB	Toolkit, codebook
T2.3 Define policy scenarios (ICL)					
Focus group	Textual data, transcripts	.,p3, .mp4, .docx, .txt, .jpg	Participants, existing literature	<10GB	Other
T3.1 Creation of a corpus of multilingual textual data (CUNI)					
Web-scraping	Textual data	.pdf .docx, .txt, .csv, .xlsx, .r	Social media, existing datasets, database, openly available reports and policies, archives	<10GB	Dataset: textual corpus, code, software
Content analysis	Textual data	.pdf .docx, .txt, .csv, .xlsx, .r	Openly available reports and policies	<10GB	Dataset
T3.2 Implementation of NLP tools to analyse political discourse (FBK)					
Experiment	Textual data	.py	Existing data, proprietary data, archives	<10GB	Software, textual corpus, NLP models
T3.3 Implementation of NLP tools to classify executive aggrandisement in legal data (FBK)					
Experiment	Textual data	.py	Existing datasets	<10GB	Software, textual corpus, NLP models
T3.4 Development of an interlinked database on executive aggrandizement (CSS)					
NA	Textual data	.r, .py	Processed data from partners	<10GB	Textual corpus, codebook, software
T4.1 Procuring survey data (GESIS) **T4.2 Procuring elections, institutional, and official statistics data (GESIS)** **T4.3 Integrating and updating the data (GESIS)**					
Survey	Tabular data	.rds, .json, .csv, multiple data formats	Existing datasets, openly available reports and policies	<10GB	Code, software, dataset (restricted access)
T5.1 Conceptual design of the DT (LNU)					
NA	Code	.py, .jl	Processed data from partners	<10GB	Models
T5.2 Implementation (UBB) **T5.3 Module integration (UBB)**					
Experiment	Code	.py, .jl	Existing (open) datasets	<10GB	Software
T7.1 Stakeholder definition and engagement (DI)					

Focus group	Audio recordings, transcription	.mp4, .txt, .docx	Participants	100GB	toolkit

The consortium partners have identified, collected, and documented major data sources in the first project year. For this identification and selection process, we have conducted two additional measures: First, we have organized an initial operationalization workshop in September 2025 among the consortium partners to sort, identify and reduce the identified latent constructs, variables, and indicators. The aim was to clarify concepts and ensure alignment between theoretical considerations in WP2, data demand in WP5 and availability in WP4. Follow-up meeting will be held in 2026. The interactive exercise was implemented in a Miro board. Given the extensive amount of identified survey items, we have furthermore initiated a rating exercise among the consortium members in December 2025. We will assess the outcome of this exercise in early 2026 and select the final core item set afterwards.

Year 1 concludes the data identification process as described in the following sections. Project years 2 and 3 will increasingly focus on the integration and preprocessing of these data sources for input into WP5 models.

3.2 Model dimensions and operationalization

TWIN4DEM draws upon existing survey and statistical data sources to capture information about (a) The socio-economic conditions shaping the evolution of democratic systems and (b) The attitudes of citizens appointing political agents as well as (c) Institutional data on decision-making rules in each case study. To do so, existing datasets will be reused following FAIR principles. These include:
(a) Census and official statistics (e.g. Annual Regional database of the European Commission) providing information on population, income, and unemployment at the (sub-)national level;
(b) Electoral data as well as national surveys from the four case study countries containing demographics, core values, political attitudes, voting intentions and behaviour, as well as related issues;
(c) Institutional data on latent and observable properties of the institutions to be modelled will be mostly derived from the Varieties of Democracy (V-Dem) project and complemented with expertise from the national partners.

The latest version of the Twin4Dem model (version 0.2 as of October 30, 2025) (Diosan et al. 2025) includes three core modules simulating the executive, legislative, and judiciary branches. From project year 2 onwards, it will be consecutively fed and developed with real-world data from the case study countries.

In *v0.2*, all agents have the same utility function, with agent-specific weights depending on their position and empirical estimations. Weights determine the relative importance of the six components affecting utility:

1. **Personal political opinion,**
2. **Increasing the power of office,**
3. **Probability of staying in office,**
4. **Probability of being nominated for a prestigious office,**
5. **Satisfying supporting groups,**
6. **Reputation among peers.**

The weights of these six components are not endogenously updated but subject to external input. WP3 and WP4 support this input with empirical data on these six subdimensions. The following subsections outline potential data sources and integration strategies for these six subdimensions.

3.2.1 Personal political opinion

Each agent has an intrinsic preferred opinion about aggrandizement, which varies on the basis of an external input. The current version *v0.2* assumes a binary preference (for or against) but future iterations might be extended to a continuous scale of aggrandizement preferences.
Suitable input information to define the agents' political opinion are extracted from different sources. The primary one is parliamentary speeches, which are being collected within WP3 for each of the four countries of interest (Czech Republic, Hungary, France and the Netherlands). For each country, data are retrieved from existing open platforms or data dumps, and processed to have the same types of metadata such as speech date, speaker's information and topic. Specific NLP tools will then be developed to assign, to each MP (i.e. agent), a stance label, i.e. for or against what is discussed in the speech (task 3.2). Another source of information to define the agents' personal political opinion will be social media accounts or any other public profile made available by the MPs. Given current restrictions concerning data collection from social networks such as Twitter/X and Facebook, other online sources enabling the profiling and stance extraction of agents/MPs are being explored such as Wikipedia/Wikidata.

3.2.2 Increasing the power of office

The power of office is a group-level, time-invariant variable. Executive aggrandizement, if implemented, increases the power of the government but decreases the one of both the parliament and courts. The current version does not require external data input for this subdimension, and it is not expected to be required in future versions.

3.2.3 Probability of staying in office

The utility derived from keeping the office reflects the probability of staying in office in the next period. It is modelled as a logistic function of alignment between the agent action and the preferences of the "supporting" actor (e.g., electorate, supporters or supervisors).

One possible data input for modelling the probability of an agent staying in office is election data, as they provide observable indicators of the extent to which incumbent actions align with the preferences of voters. As executive aggrandizement rarely occurs in a single moment but unfolds incrementally, a process that Khaitan (2019) describes as “killing a democracy with a thousand cuts”, democratically elected leaders must remain in office long enough to implement the institutional changes that gradually weaken horizontal and vertical accountability mechanisms (Khaitan, 2019; Laebens, 2023). The incentive to stay in office is therefore closely linked to electoral dynamics. Such agents may exploit their democratic mandate to evade oversight, neutralize or capture independent checking institutions, and erode electoral, institutional, and discursive accountability. This expands their unilateral authority and limits societal and institutional constraints (Bermeo, 2016; Khaitan, 2019; Laebens, 2023). Voters themselves may at times tolerate greater executive power when they expect policy benefits from leaders they support, even if they otherwise value democratic norms (Singer, 2018). Since vertical (electoral) accountability requires executives to periodically secure electoral support, changes in electoral performance can indicate whether incumbents consolidate political leverage or face growing risks of removal (Khaitan, 2019). Election results thus offer a plausible empirical proxy for estimating how executive behavior shapes political survival.

For this project, election results can serve as one data source for capturing these dynamics over the 2000–2025 period (see table 2). They include systematic information on vote shares, changes in party support, turnout, and seat distributions. A particularly suitable dataset for this type of analysis is the Constituency-Level Elections Archive (CLEA) (https://electiondataarchive.org), which offers harmonized lower-chamber election results across countries and years. CLEA standardizes key variables, like party identifiers, constituency names and vote totals, across diverse electoral systems, enabling consistent national aggregation and cross-country comparison. For Czechia, France, Hungary and the Netherlands, CLEA includes constituency-level vote counts and party-level metadata for most parliamentary elections within the relevant time period. These data make it possible to derive national vote shares and trace electoral trajectories over time.

To achieve full temporal coverage up to 2025, CLEA can be supplemented with official national sources where gaps exist. For Czechia, missing results for the 2025 parliamentary election can be obtained from the Czech Statistical Office (ČSÚ). The Dutch election authority (Kiesraad) can provide national results for the Netherlands for the 2025 parliamentary elections. The parliamentary elections of both Hungary and France are already comprehensively covered by CLEA for the relevant years. These official datasets can then be harmonized to align with CLEA’s variable structure, to ensure compatibility within a unified cross-national dataset. In addition to the parliamentary elections, official national sources further provide presidential election results for France and Czechia, enabling the dataset to capture executive-level electoral dynamics in both semi-presidential and parliamentary systems. Including these elections helps reflect moments where executive mandates are renewed or challenged outside the legislative environment.

Table 2: Election result data sources

Country	Source		Election type	Years
Czechia	Constituency-Level Elections Archive (CLEA)	ČSÚ - Czech Statistical Office	parliamentary	2002, 2006, 2010, 2013, 2017, 2021,(*2025)*
			presidential	2013, 2018, 2023
France		DataGouv	parliamentary	2002, 2007, 2012, 2017, 2022, 2024
			presidential	2002, 2007, 2012, 2017, 2022
Hungary		Constituency-Level Elections Archive (CLEA)	parliamentary	2002, 2006, 2010, 2014, 2018, 2022
Netherlands		Kiesraad (Overheid)	parliamentary	2002, 2003, 2006, 2010, 2012, 2017, 2021, 2023, (*2025)*

Although national data are authoritative and often released more quickly, their structures differ across countries and elections. CLEA mitigates these comparability challenges through its standardized format, which is why it could serve as the primary foundation for this project, with national sources functioning as targeted supplements where coverage gaps exist.

In addition to election results, past legislative and government turnovers can serve as a proxy for the probability of staying in office. For the four countries considered, turnovers can generally be inferred from election results: changes in party control of the legislature or executive indicate a shift in incumbency, while continuity signals political stability. Parliamentary turnovers are captured by shifts in the controlling party or coalition, and presidential turnovers by changes in the executive following elections. These turnover indicators provide a complementary perspective on incumbents' political survival and executive stability and can be integrated into models estimating the probability of remaining in office.

Overall, election data, complemented by information on legislative and government turnovers, provide a rich empirical basis for estimating the likelihood of incumbents remaining in office. In the context of executive aggrandizement such data make it possible to identify periods of electoral strength or vulnerability that may shape the opportunities for, and constraints on, executive power consolidation.

3.2.4 Probability of being nominated for a prestigious office

The probability of being nominated to a prestigious office works analogously to the one of staying in office. The main difference is that, instead of comparing the prospect decision with the preference of external supporting actors, it is compared with the political opinion of selected "powerful" agents. The current version v0.2 operationalizes the preference of the prime

minister, of the president of the court, and of the head of each party in the parliament, depending on whether the agent is a minister, an MP, or a judge, respectively. Given the small n of agents for this dimension, the utility is primarily informed by qualitative assessment based on the case study findings. The current version does not require data input from WP4 and it is not expected to do so in future iterations of the Digital Twin.

3.2.5 Satisfying supporting groups

This dimension depends on the opinion of (non-parliamentary) supporting groups about executive aggrandizement (pro or against aggrandizement). Version v0.2 has not yet established which supporting groups are included or not, which in principle can be citizens and the electorate, institutionalized actors such as business associations and NGOs, or donors.

A major data source for the understanding of preferences of citizens are national and European survey programs. To investigate how executives maintain support among key constituencies and how this helps them expand their power, survey data provide a valuable source, as they capture how citizens evaluate democratic institutions, executive authority, and the legitimacy of institutional constraints. A central finding in recent research is that public opinion does not always function as an effective brake on executive overreach, even though most citizens express support for democracy in principle (Gidengil et al., 2022). Attitudes may shift in ways that create opportunities for stronger executive authority; for instance, when economic prosperity or perceived representation increases support for strong leaders among incumbent-aligned groups (Schafer, 2021). Identification with socially or politically dominant groups can similarly heighten acceptance of norm-violating behavior when leaders are viewed as advancing in-group interests (Schafer, 2021). Further survey-based studies show that political discourse, particularly populist and anti-elitist rhetoric, shapes how citizens interpret executive power. By framing the executive as the authentic voice of “the people” and institutional checks as obstacles to the people's will, such discourse can increase public tolerance for weakened accountability mechanisms, especially among government supporters (Bessen, 2024). Survey data therefore provide systematic insight into how different population groups evaluate democratic constraints, how support for institutional checks varies across them, and how these attitudes change over time. Because vertical accountability ultimately depends on citizens’ willingness to endorse or reject incumbents at the ballot box, such attitudinal indicators clarify whether societal preferences reinforce or weaken the electoral constraints that normally limit executive authority. Consequently, survey measures constitute an essential complement to electoral data for assessing the societal foundations that facilitate or hinder executive aggrandizement.

T4.1 was primarily tasked with screening survey programs and accessible data. Appendix 1 (*D4_1_appendix_1.xlsx*) documents this task. Major survey programs which offer insights into public perceptions on executive aggrandizement include:

- **European Values Study (EVS)** (https://europeanvaluesstudy.eu/)
- **European Social Survey (ESS)** (https://www.europeansocialsurvey.org/)

- **International Social Survey Programme (ISSP)** (https://issp.org/)
- **(Flash) Eurobarometer** (https://europa.eu/eurobarometer/screen/home)
- **European Election Studies (EES)** (https://www.gesis.org/en/services/finding-and-accessing-data/international-survey-programs/european-election-studies)
- **Comparative Study of Electoral Systems (CSES)** (https://cses.org/)

Relevant waves of these programs have been classified according to (1) the field period, (2) number of respondents per case study country, (3) survey methodology, and (4) spatial information. In a first step, extensive research of all potential items from these datasets has been conducted and classified according to several aspects and concepts related to executive aggrandizement. These capture:

- **Understanding of Democracy**: Examines views on core democratic rights, the role of institutions, and the importance of political participation and free elections.
- **Assessment of Country**: This section captures how citizens evaluate the state of democracy in their own country. It covers perceptions of political freedom, electoral integrity, media independence, government transparency, fairness, corruption, and trust in institutions.
- **Governmental Role**: Explores opinions on government power, surveillance, national security, and the balance between civil liberties and leadership.
- **Trust in Institutions**: Assesses public trust in national and international institutions, including government transparency, corruption, and political influence.
- **Satisfaction with Politics**: Covers approval of the government's performance, views on the EU's actions, and satisfaction with how democracy and the political system work.
- **Attitudes & Values**: Explores how important democratic governance is to individuals, views on authority and civil liberties, acceptance of protest, political systems preferences, law obedience, civic participation, and moral values such as trust, justice, and gender roles.
- **Political Action**: Covers individuals' participation in political activities and their confidence in influencing political decisions. Explores how people view their ability to have an impact on politics and whether they feel their voices are heard.
- **National Identity**: This section explores individuals' emotional attachment and pride in their country, as well as their views on patriotism and national values.
- **EU opinion**: Explores individuals' attitudes toward the European Union, including their opinions on its impact on national policies and social benefits, as well as the level of integration they support.
- **Miscellaneous**: Includes questions about personal political positioning on the left-right spectrum, voting habits, and the importance of voting as well as Macro-Level items. It also explores attitudes toward the media, including trust in news sources.

In a second step, a selection of these items will be classified based on a rating exercise among all consortium partners (see section 3.1). This selection will be considered as core set for further classification.

In order to extend the available data points for each case study and capture all relevant subdimensions of perceived executive aggrandizement, we aim to harmonize data from selected survey programs in a next step. Data harmonization allows to make items from different survey programs comparable. This requires (1.) an assessment, whether the survey question texts measure the same constructs and latent variables, and (2.) a system to harmonize potentially varying response scales. Scholars have produced different approaches for this in the last years (Kolen & Brennan, 2014; Roth & Singh, 2024; Tomescu-Dubrow et al., 2024). Taking into account advantages and disadvantages of these approaches and following recommendations from a recent systematic comparison of these approaches (Heizmann, 2025), we propose a harmonization approach based on a selection of the survey programs listed in table 3. We will rely on established procedure for harmonization of survey instruments at GESIS (for further information, see https://www.gesis.org/en/consulting/survey-methods-consulting/harmonization). A starting point for an explanatory baseline and trial will be the linear stretch approach (de Jonge et al., 2014). The Linear Stretch Method is a simple, conventional way to transform survey questions with numerical or ordinal response scales to a common metric and thereby create a first harmonization baseline. For example, 5- or 7-point scales can be linearly rescaled to a 0–10 range by mapping the lowest response category to 0, the highest to 10, and assigning equally spaced values to the intermediate categories (e.g. a 5-point scale becomes [0.0; 2.5; 5.0; 7.5; 10.0]). Based on these transformed values, the sample mean and standard deviation are then computed using the Frequency Approach. Depending on how well the distributions align after linear stretching, more advanced harmonization methods will be considered, such as nonlinear transformations, subgroup-specific adjustments, or model-based approaches (e.g. latent variable models) to better account for differences in measurement properties across survey programs.

After harmonizing and linking the survey datasets, we use exploratory factor analysis (EFA) to uncover the main underlying dimensions of respondents' perceptions of executive aggrandizement and to construct empirical indicators for these latent concepts (Costello & Osborne, 2005). Rather than focusing on individual questions in isolation, EFA looks at how items co-vary and groups them into broader factors that capture shared patterns in the responses. In this way, it helps us translate a large number of observed items into a smaller set of interpretable dimensions that reflect core aspects of the phenomenon we are interested in. For each retained factor, we create composite scores (factor scores or mean indices of strongly loading items), which serve as quantitative indicators of the underlying constructs for subsequent usage in the WP5 Digital Twin.

Using EFA serves three main purposes. First, it allows us to check whether the theoretical dimensions we assume are actually visible in the data, or whether the items suggest a different structure. Second, it provides a systematic basis for refining the item set, for example by flagging questions that do not clearly relate to any underlying factor or that behave inconsistently across contexts. Third, it yields coherent, empirically grounded scales for each dimension of executive aggrandizement, which can then be used for comparison across survey programs, countries, and time.

Table 3: Alignment of major survey programs and elections (x = parliamentary, o = presidential)

Year	Survey Programme	Countries				Party link		National elections			
		CZ	FR	HU	NL	Vote	Intent	CZ	FR	HU	NL
2000											
2001											
2002								x	x/o	x	x
2003											x
2004											
2005											
2006								x		x	x
2007									x/o		
2008											
2009											
2010	ESS - Round 5	x	x	x	x	x		x		x	x
2011											
2012	ESS - Round 6	x	x	x	x	x			x/o		x
2013	ISSP- National Identity III	x	x	x	x	x		x/o			
2014	ESS - Round 7	x	x	x	x	x				x	
	ISSP- Citizenship II	x	x	x	x	x					
	EES 2014	x	x	x	x	x	x				
2015											
2016	ESS - Round 8	x	x	x	x	x					
	ISSP - Role of Government V	x	x	x	x	x					
2017	Joint EVS/WVS	x			x		x	x	x/o		x
2018	ESS - Round 9	x	x	x	x	x		o		x	
	Joint EVS/WVS		x	x			x				
2019	EES 2019	x	x	x	x	x					
2020	ESS - Round 10	x	x	x	x	x					
2021								x			x
2022	Joint EVS/WVS	x			x		x		x/o	x	
2023	ESS - Round 11	x	x	x	x	x		o			x
2024	EES 2024	x	x	x	x	x			x		
2025								x			x

3.2.6 Reputation among peers

The utility from reputation for agents depends on the fraction of the relevant peers (ministers, MPs, judges) that have expressed the same vote on the previous time step (t − 1). The current version v0.2 of the digital twin models reputation among peers as an endogenous process – there is no need for external data.

3.2.7 Other: aggrandizement units

Data on executive aggrandizement will be derived from decisions in the executive, legislative, and judiciary branches. The data collection and processing steps for parliamentary roll call data is described below.

In addition to survey, election and socio-economic data, a systematic collection and harmonization of roll call voting data from the national parliaments across the four case study countries is generated. The dataset forms a harmonized cross-national collection of parliamentary voting behavior, covering the period 2000–2025 and multiple legislative terms. Roll call votes are official records of how individual Members of Parliament (MPs) voted on legislative proposals and offer valuable insights into legislative behavior of the MPs, party cohesion and policy alignment. This data can help to study executive aggrandizement by measuring the extent to which parliamentary votes align with government proposals, revealing how much legislative oversight persists in different political systems.

All roll call data are gathered through transparent and replicable automated scraping procedures using open-source tools such as *rvest* (Wickham et al., 2025) and *RSelenium* (Harrison & Kim, 2022). Each national parliament's website provides public access to voting records, however the structure, format, and accessibility of these records differ considerably across countries. Therefore, country-specific scripts have been developed to extract comparable data elements: metadata on each vote (e.g., voting number, bill title, date, result), summaries of votes by parliamentary party, and the individual-level records of MPs' votes.

For Czechia, votes are collected from the Chamber of Deputies' official site (*psp.cz*) for the 3rd through 9th terms (2000–2025). The scraper systematically retrieves vote summary information and extracts structured data for each recorded vote, including party-level and MP-level information. As the Czechia site is a static website, interactive navigation was not required, and the data could be extracted directly from the page's HTML. For the remaining three countries, the scraping process is automated through a purpose-built *RSelenium*-based R script that interacts directly with the parliamentary search interface to retrieve structured vote records within specified time intervals. The script initiates a local Selenium driver (typically using Mozilla Firefox) to navigate the interactive interface, set date parameters, and extract all recorded roll-call votes by systematically iterating through the individual result pages. This approach is necessary because, in contrast to the Czechia site, the target websites are dynamic and require sequential interaction with page elements to access complete voting

records. Subsequently, the scraper automatically parses through paginated results, collecting all links to individual vote detail pages.

For France, roll call data are obtained from the datan.fr platform, which provides detailed vote tables for legislative terms 14th (2012–2016) through 17th (since 2024). The Selenium-based scraper automates browser navigation and pagination to collect metadata, party-level and individual-level results. Data for earlier legislative terms does exist and could potentially be incorporated to extend temporal coverage, spanning from the 11th (1997–2002) to the 13th (2007–2012) term (Godbout & Foucault, 2013). For the Netherlands, roll call data are retrieved from *openkamer.org*, a civic open data portal that archives parliamentary proceedings, covering terms from 1998–2002 up to 2023–present. This scraper collects bill-level metadata, the final vote outcomes, and information on which parties supported or opposed each bill. For Hungary, roll call data are gathered from the Hungarian National Assembly's official website (www.parlament.hu) for multiple terms spanning 2000–present. The Selenium-based scraper automates browser navigation and collects metadata, party-level and individual-level results.

Table 4: Roll call data indicators

Roll Call Data Overview	**CZ**	**FR**	**HU**	**NL**
Info				
Meeting Number	X	O	O	O
Vote Number	X	X	O	O
Bill Number	O	O	X	X
Date	X	X	X	X
Description	X	X	X	X
Result	X	X	X	X
MP List				
Name	X	X	X	O
Party	X	X	X	O
Vote choice	X	X	X	O
Loyalty (to Party)	x	X	x	x
Vote Summary				
Party	X	X	X	X
Yes	X	X	X	X
No	X	X	X	X
Abstained	X	X	X	X
Not logged in/ Did not vote	X	x	X	O
Excused	X	O	X	O
Cohesion Rate	x	X	X	x

X = Available

x = Can be provided

O = Not available

For each vote, it extracts key components:

- Motion Information: Metadata about the legislative proposal (document/meeting number, title, submitters and vote date);
- Voting Summary: Overall outcome statistics (yes/no/abstain totals, acceptance status);
- Voting Results by Group: Aggregated results by parliamentary faction or party;
- Vote by Name: Individual-level voting records identifying each MP, their affiliated party, and their specific vote (e.g., *Yes*, *No*, *Abstained*).

Table 4 reports the exact details of obtained data.

All tables are parsed into structured data frames, cleaned, and standardized. The cleaning routines translate variable names into English (e.g., "Igen-ek száma" → “Yes votes”), unify numerical formats, and map categorical values to consistent English-language vote codes (“Igen” → “Yes”, “Nem” → “No”, etc.). The scraper also extracts the exact vote date from the page header to ensure correct temporal alignment with other datasets. To ensure cross-country comparability, each table type is processed through uniform transformation functions, resulting in a consistent schema across all cases. The final outputs are stored as structured R lists or exported to CSV for integration into broader comparative datasets.

Because roll call data often include information about individual MPs, data protection and ethical considerations are carefully observed. Only publicly available information is collected, and all identifiers remain consistent with official parliamentary records.

3.2.8 Other: Institutional and socio-economic data

As an agent-focused model, v0.2 so far does not explicitly call for external data input about institutional structures and performance or other macro-level statistics. If this changes during later stages of the project, the consortium is well equipped to respond to this demand. Institutional data on latent and observable properties of the institutions to be modelled can be derived from the Varieties of Democracy (V-Dem) project (https://www.v-dem.net/) and complemented with expertise from the national partners. Taking into account recent research on national-level, institutional aspects of executive aggrandizement (Michel, 2024), we propose core institutional variables in table 5. Aggregate counts of laws and executive acts per year, and identification whether these were initiated by governments or parliaments, might be another important indicator for the model – this data will be derived from the case study work. Harmonized socio-economic and socio-demographic data will be derived from Eurostat (https://ec.europa.eu/eurostat/web/main/data/database), if requested by WP5. A full list of all researched macro-level indicators is attached in appendix 2 (*D4_1_appendix_2.xlsx*).

Table 5: Core institutional and socio-economic indicators

Name	Source	Code	Description
Horizontal accountability	V-Dem (2023, p. 291)	v2x_horacc	Index of horizontal accountability, capturing the extent to which courts and the legislature can and do constrain the executive through review, oversight, and investigations.
Legislative constraints	V-Dem (2023, p. 50)	v2xlg_legcon	Index of legislative constraints on the executive, measuring how far the legislature (including opposition parties) questions, oversees, and investigates executive actions.
Judicial constraints	V-Dem (2023, p. 50)	v2x_jucon	Index of judicial constraints on the executive, indicating the degree to which the executive respects the constitution and complies with decisions of independent courts.
CSO repression	V-Dem (2023, p. 196)	v2csreprss	Expert-coded measure of repression of civil society organizations, indicating how much governments harass, hinder, or shut down CSOs.
Media censorship	V-Dem (2023, p. 201)	v2mecenefm	Indicator of government efforts to censor print and broadcast media, including direct bans and indirect pressures; higher values generally indicate less censorship.
Political liberties	V-Dem (2023, p. 297)	v2x_clpol	Index of political civil liberties, summarizing the extent to which people enjoy freedoms of association and expression in political life.
Freedom of expression	V-Dem (2023, p. 311)	v2x_freexp	Index of freedom of expression, capturing how freely people can discuss politics, the independence and pluralism of the media, and the level of academic and cultural freedom.
Executive respects constitution	V-Dem (2023, p. 115)	v2exrescon	Indicator of whether and to what extent the national executive respects constitutional limits on its power, including compliance with formal rules and court rulings.
Liberal democracy	V-Dem (2023, p. 45)	v2x_libdem	Liberal democracy index combining electoral democracy with rule of law, individual and minority rights, civil liberties, and checks on the executive.
Mobilization for democracy	V-Dem (2023, p. 229)	v2cademmob	Measure of mass mobilization for democracy, capturing the frequency and scale of protest events explicitly demanding more democratic institutions and practices.
Electoral Democracy	V-Dem (2023, p. 44)	v2x_polyarchy	Electoral democracy index summarizing the realization of electoral democracy: elected officials, free and fair elections, freedom of association, freedom of expression/alternative information, and inclusive suffrage.

4. Implementation

4.1 Data integration procedure

Most data sources generated in WP4 will be linkable on the party-level. Of major relevance for our linking procedure is the PartyFacts project (https://partyfacts.herokuapp.com/). PartyFacts is a collaborative database that assigns unique, harmonized identifiers to political parties worldwide and links them across many major datasets (e.g., Manifesto Project, ParlGov, CHES, CLEA, ESS). Each party gets a stable PartyFacts ID plus standardized metadata (name, country, founding/termination, etc.), which allows researchers to treat that ID as a "bridge" variable. We use PartyFacts to integrate several datasets, including the harmonized

survey data by first matching each dataset's party codes or names to the corresponding PartyFacts IDs, and then merging the datasets on those IDs. This way, survey data, election results, and agent-level data that refer to the same party can be combined into a unified dataset.

In a first step, we will build a crosswalk between the harmonized survey data party codes and labels to the PartyFacts ID. For some survey programs, we can rely on established procedures and tools, like the Political Parties Crosswalk (Hill, 2020/2025; Kołczyńska & Powałko, 2022). Remaining gaps in the linking procedure will be approached with deterministic matching and additional fuzzy matching or NLP. As a first step, this will require a reference table of parties of our four case study countries derived from PartyFacts. Next to the PartyFacts ID, it will contain the ISO country code, party names in local languages and English, abbreviations, historical names and changes to their names, and active years and contested during our study period. Next, we will draw the party labels from the survey programs – most of our studies rely on closed-ended party variables. Remaining open-ended text responses need to be processed accordingly to be translated into party labels. Further preprocessing of both reference tables (Unicode normalize, lowercase, removal of whitespace) will simplify linking. For cases where deterministic linking will fail, we will further proceed with fuzzy string matching or named entity recognition (NER), for example based on the stringdist package in R (Loo et al., 2025) or the spacyr package (Benoit et al., 2023) to access Python spaCy models. The entire party crosswalk (reference tables and linking code) will be stored in the Twin4Dem Github repository.

4.2 Sharing and storage

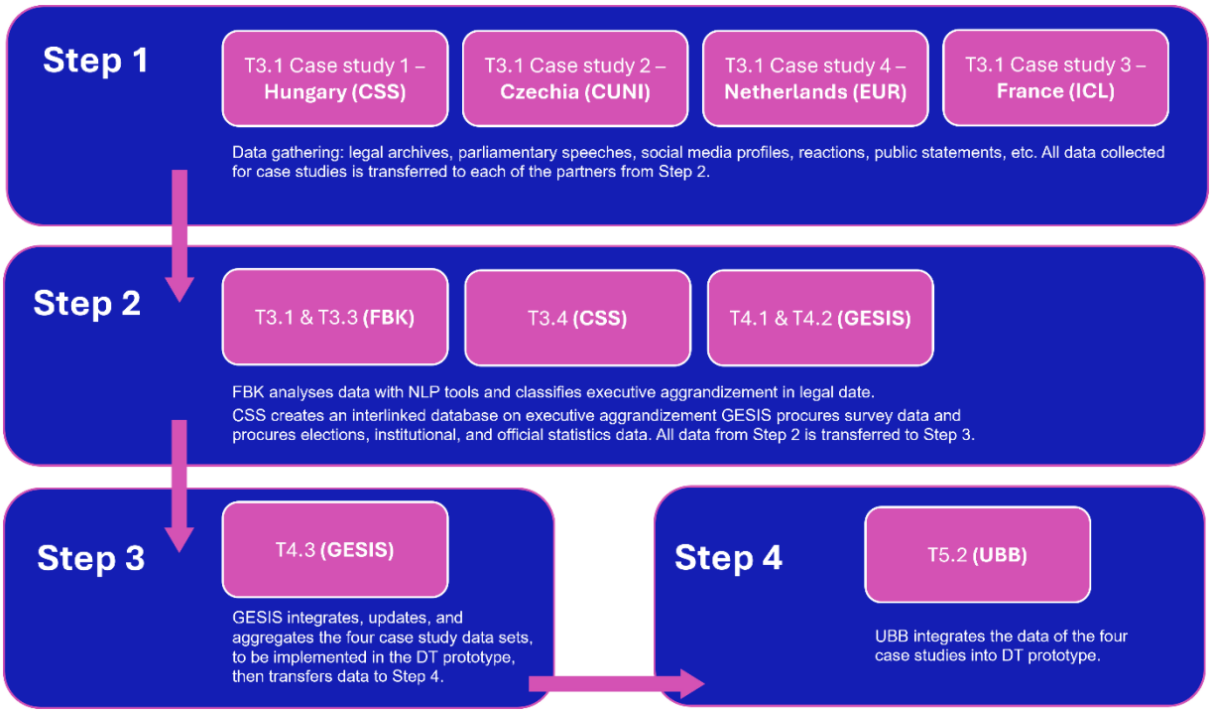
TWIN4DEM's online platform on Github (https://github.com/TWIN4DEM-project) serves as a FAIR data point where our data, algorithms and statistical models, publications, online policy tools and outputs will be available to the public. All datasets will be released with a detailed codebook, specifying the project's own ontology and classification which builds on the most used metadata schemes in comparative politics. The project website will use the standard metadata files of the field, Dublin Core, DDI Lifecycle 3.0 as well as the other community-based metadata schemes used to store reusable datasets. Software and scripts under development will be made available on GitHub with the necessary documentation. In releasing and reusing sensitive personal data, TWIN4DEM will comply with the GDPR's requirements and adopt best practices in the collection and release of survey data developed by GESIS.

The final datasets supporting our publications will be archived in DataverseNL and in the European Open Science Cloud (EOSC), with a permanent identifier (DOI) assigned to each dataset. This allows that after the project's completion, TWIN4DEM's data will be accessible for replication and for secondary use by scientific community. TWIN4DEM online platform will be available for 10 years, while data storage will be organised for unlimited period.

5. Conclusion and outlook

Deliverable 4.1 (D4.1) on "Survey, elections, institutions, and socio-economic data procurement document" is the direct output of tasks 4.1 and 4.2 and documents the process to collect and aggregate the data. This work in project year 1 directly feeds into the subsequent steps in the project years 2 and 3. In particular Task 4.3, "Integrating and updating the data" (Lead: GESIS, plus: EUR, ICL, CSS, CUNI, GESIS, UBB, ETICAS) starts in M13 and builds on the previous steps in WP3 and WP4. This task will integrate the datasets generated in tasks 3.1, 3.3, 3.4, 4.1, and 4.2 (see figure 2 created for the DMP by de Mooij et al. (2025)). This includes aggregating the population data to suitable temporal and spatial units, as described for example in section 3.2.5. This might also require imputing missing information across time and units of all datasets, anonymizing data where required for data privacy, and linking different types of data. This presents significant challenges since data have different formats (text, numeric, etc.) and frequencies, with temporal scales possibly varying from hours to years. The goal of this task is to integrate these properties appropriately into the simulation models. In sum, these measures will result in (synthetic) datasets that comprehensively represent the societies examined by TWIN4DEM. The TWIN4DEM integrated dataset (D4.2.) will be released as an outcome of this task in M32.

Figure 1: Twin4Dem Data Flow Chart



6. References

Benoit, K., Matsuo, A., Gruber, J., & Council (ERC-2011-StG 283794-QUANTESS), E. R. (2023). *spacyr: Wrapper to the "spaCy" "NLP" Library* (Version 1.3.0) [Computer software]. https://cran.r-project.org/web/packages/spacyr/index.html

Bermeo, N. (2016). On Democratic Backsliding. *Journal of Democracy*, *27*(1), 5–19.

Bessen, B. R. (2024). Populist Discourse and Public Support for Executive Aggrandizement in Latin America. *Comparative Political Studies*, *57*(13), 2118–2151. https://doi.org/10.1177/00104140231223738

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, *10*(7). https://doi.org/10.7275/JYJ1-4868

de Jonge, T., Veenhoven, R., & Arends, L. (2014). Homogenizing Responses to Different Survey Questions on the Same Topic: Proposal of a Scale Homogenization Method Using a Reference Distribution. *Social Indicators Research*, *117*(1), 275–300. https://doi.org/10.1007/s11205-013-0335-6

de Mooij, E., Mania, J., & Carvajal, M. (2025). *D1.2—Data Management Plan* (Deliverable for TWIN4DEM: Strengthening Democratic Resilience Through Digital Twins - Grant Agreement No. 101178061).

Diosan, L., Bravo, G., Kopacheva, E., Olar, A., & Mursa, B. (2025). *D5.1—Initial TWIN4DEM DT model* (Deliverable for TWIN4DEM: Strengthening Democratic Resilience Through Digital Twins - Grant Agreement No. 101178061).

Harrison, J., & Kim, J. Y. (2022). *RSelenium: R Bindings for "Selenium WebDriver"* (Version 1.7.9) [Computer software]. https://cran.r-project.org/web/packages/RSelenium/index.html

Heizmann, B. (2025). Good enough? A comparison of different harmonization procedures and their substantive consequences using the example of life satisfaction. *Quality & Quantity*, *59*(2), 1369–1392. https://doi.org/10.1007/s11135-025-02060-7

Hill, S. E. (2025). *ESS PartyFacts Crosswalk* [Computer software]. https://github.com/sophieehill/ess-partyfacts-crosswalk (Original work published 2020)

Gidengil, E., Stolle, D., & Bergeron-Boutin, O. (2022). The partisan nature of support for democratic backsliding: A comparative perspective. European Journal of Political Research, 61(4), 901–929. https://doi.org/10.1111/1475-6765.12502

Khaitan, T. (2019). Executive aggrandizement in established democracies: A crisis of liberal democratic constitutionalism. *International Journal of Constitutional Law*, *17*(1), 342–356. https://doi.org/10.1093/icon/moz018

Kołczyńska, M., & Powałko, P. (2022). The Political Parties Crosswalk for mapping party codes in cross-national surveys to Party Facts IDs. *Political Research Exchange*, *4*(1), 2048957. https://doi.org/10.1080/2474736X.2022.2048957

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices*. Springer New York. https://doi.org/10.1007/978-1-4939-0317-7

Laebens, M. G. (2023). Beyond democratic backsliding: Executive aggrandizement and its outcomes. *The Varieties of Democracy Institute*. https://www.v-dem.net/media/publications/UWP_54.pdf

Loo, M. van der, Laan, J. van der, R Core Team, Logan, N., Muir, C., Gruber, J., & Ripley, B. (2025). *stringdist: Approximate String Matching, Fuzzy Text Search, and String Distance Functions* (Version 0.9.15) [Computer software]. https://cran.r-project.org/web/packages/stringdist/index.html

Michel, J. (2024). *The Subnational Roots of Democratic Stability* [UCLA]. https://escholarship.org/uc/item/1vm85974

Roth, M., & Singh, R. K. (2024). *questionlink: Harmonizing single item survey questions on the same construct.* (Version 0.0.0.9) [Computer software]. https://github.com/MatRoth/questionlink

Schafer, D. (2021). A Popular Mandate for Strongmen: What Public Opinion Data Reveals About Support for Executive Aggrandizement in Turkey, 1996-2018. *South European Society and Politics*, *26*(3), 355–382. https://doi.org/10.1080/13608746.2022.2034689

Singer, M. (2018). Delegating Away Democracy: How Good Representation and Policy Successes Can Undermine Democratic Legitimacy. *Comparative Political Studies*, *51*(13), 1754–1788. https://doi.org/10.1177/0010414018784054

Tomescu-Dubrow, I., Wolf, C., Slomczynski, K. M., & Jenkins, J. C. (Eds.). (2024). *Survey Data Harmonization in the Social Sciences* (1st ed.). Wiley. https://doi.org/10.1002/9781119712206

Wickham, H., Software, P., PBC [cph, & fnd. (2025). *rvest: Easily Harvest (Scrape) Web Pages* (Version 1.0.5) [Computer software]. https://cran.r-project.org/web/packages/rvest/index.html

Annex 1

Overview of survey programs and items. See *D4_1_appendix_1.xlsx* in the Twin4Dem repository.

Annex 2

Overview of core macro level indicators. See *D4_1_appendix_2.xlsx* in the Twin4Dem repository.