OSCARS
Open Science Clusters' Action
for Research & Society

Funded Project

ONTOLISST - Thematic ontologies in social science research data

Presenter: Timea Venczel, ELTE Centre for Social Sciences, Research Documentation Centre
https://orcid.org/0009-0008-6174-1722

Implemented by

Tampere University

FINNISH SOCIAL SCIENCE
DATA ARCHIVE

ELTE CENTRE FOR SOCIAL SCIENCES
RESEARCH DOCUMENTATION CENTRE

ELTE CENTRE FOR SOCIAL SCIENCES
POLITICAL AND LEGAL TEXT MINING AND ARTIFICIAL INTELLIGENCE LABORATORY (POLTEXTLAB)

SciencesPo
CENTRE DE DONNÉES
SOCIO-POLITIQUES

Funded by
the European Union

THE PROCESS

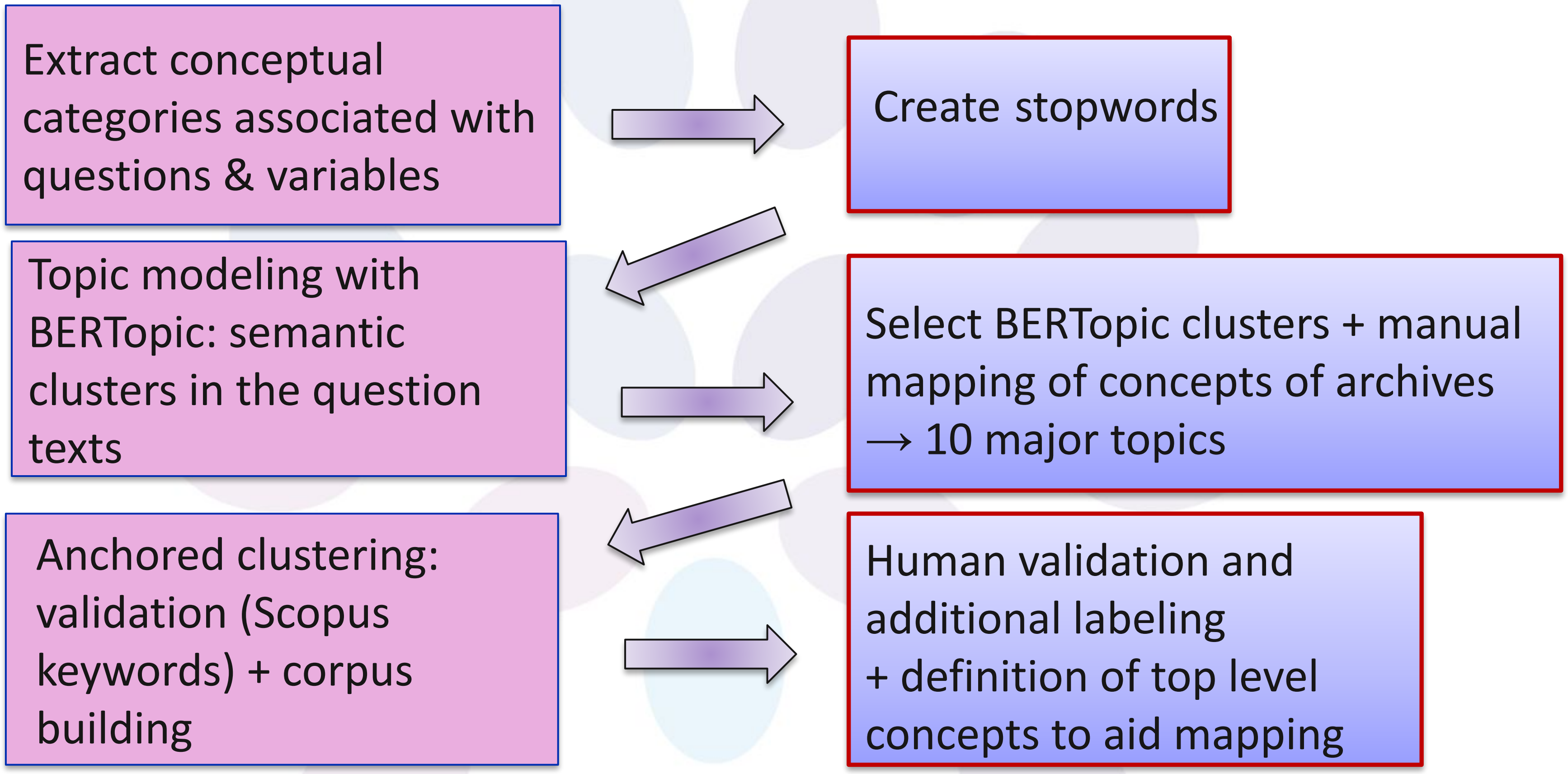
Diversity of metadata ontologies used in social sciences impedes data discoverability, accessibility and interoperability.
→ research-based light social science thesaurus (LiSST) and semi-automated tools for variable-level survey data annotation

Data

Social science research archives → survey metadata + semi-structured interviews (ESS, SOEP, GESIS, GGP, CLOSER, SNDS, ICPSR, FSD, CDSP)

Data scientists

Social scientists



Extract conceptual categories associated with questions & variables

Create stopwords

Topic modeling with BERTopic: semantic clusters in the question texts

Select BERTopic clusters + manual mapping of concepts of archives → 10 major topics

Anchored clustering: validation (Scopus keywords) + corpus building

Human validation and additional labeling + definition of top level concepts to aid mapping

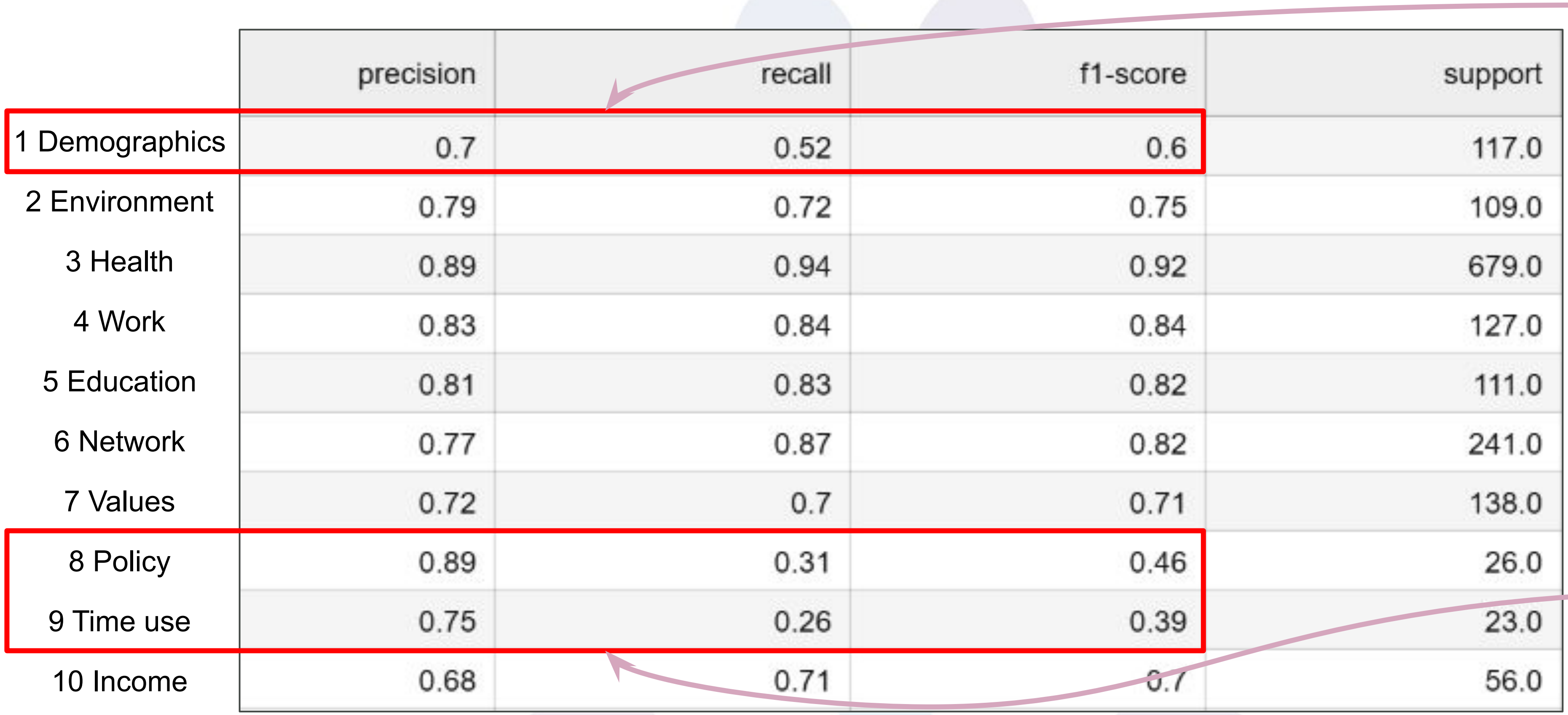
RESULTS: CROSSWALKS

LiSST	ESS	GESIS	SOEP	CLOSER	SNDS	GGP
DEMOGRAPHY, LIFE EVENTS & IDENTITY	Personal and household characteristics Timing of life Gender age and household composition	Demography and population Integration Migration Transnationalization Family planning (SPLIT) Socio-demographics and interview characteristics	Demographics Family planning (SPLIT)	Demographics Life events Pregnancy	Social class	Household members Identification
HEALTH & CARE	Social inequalities in health (SPLIT) Health and care	Family planning (SPLIT) Health and care	Health and care	Physical&mental health; Health care& behaviour; Child development	Diet and exercise Smoking Drinking behaviour Weight	Health Personal care Covid

RESULTS: TOPIC ASSIGNMENT MODEL

Trained a supervised XLM-RoBERTa model on a manually validated subset of archival data and evaluated classification performance across the 10 LiSST categories.

Code level model metrics:



	precision	recall	f1-score	support
1 Demographics	0.7	0.52	0.6	117.0
2 Environment	0.79	0.72	0.75	109.0
3 Health	0.89	0.94	0.92	679.0
4 Work	0.83	0.84	0.84	127.0
5 Education	0.81	0.83	0.82	111.0
6 Network	0.77	0.87	0.82	241.0
7 Values	0.72	0.7	0.71	138.0
8 Policy	0.89	0.31	0.46	26.0
9 Time use	0.75	0.26	0.39	23.0
10 Income	0.68	0.71	0.7	56.0

maybe too general

small sample size

PROBLEMS

OSCARS

Need for balance

Solutions

Diverse data origins & interpretive traditions

- use materials from archives/RIs with different thematic, institutional, and national traditions
- study how conceptual and data practices are shaped by research, archiving and cultural contexts

→ **identify trends & common interests**

Representation of research domains

- determine concepts based on topic clustering
- analyse keywords of highly cited social science papers (Scopus) to include both major and smaller research areas

→ **comprehensive & proportionate coverage of subjects**

Data producers' and re-users' needs

- needs assessment (interviews with archivists)
- validation of results by expert partners /& Scopus
- extend the iterative process through user feedback

→ **continuous refinement and rebalancing**

CONTROLLED VOCABULARY VALIDATION BY
5,000 social science papers ~60,000 unique keywords

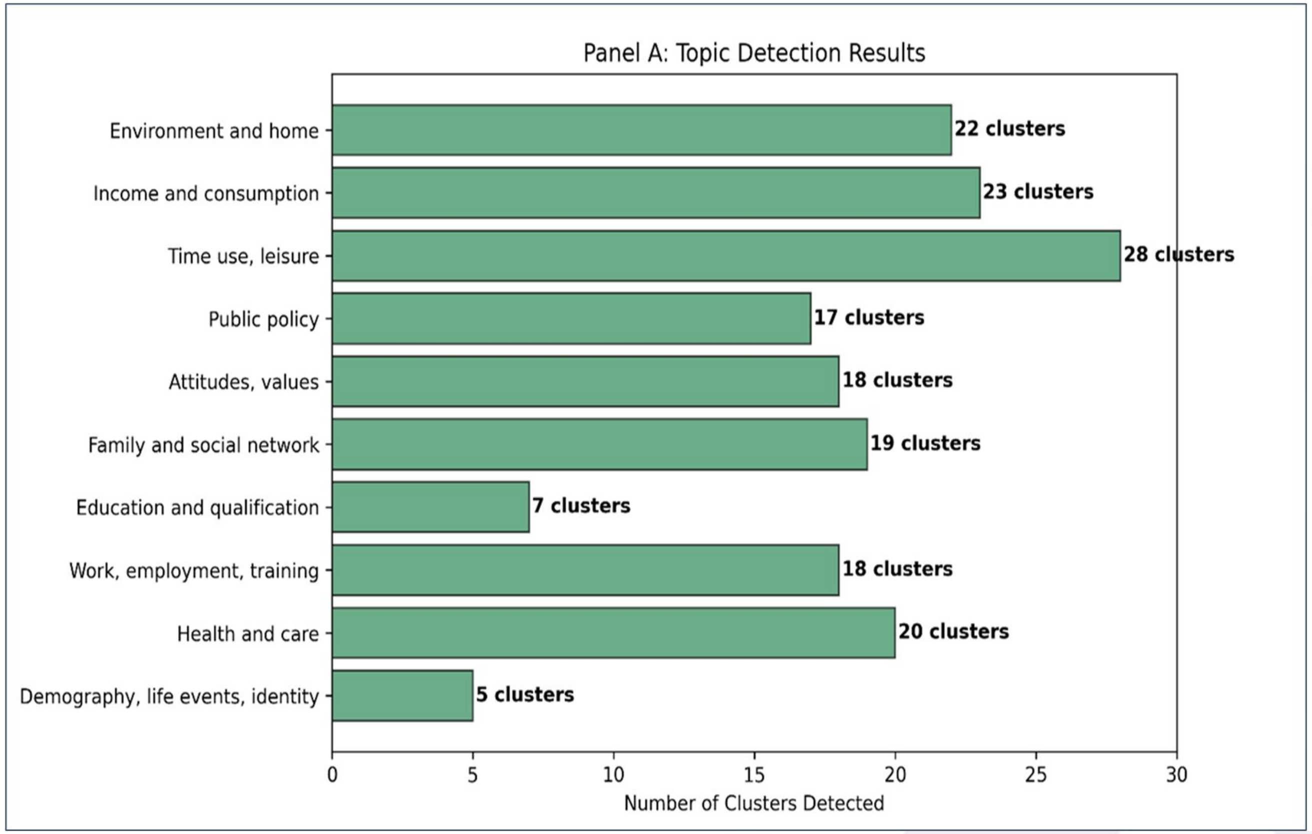
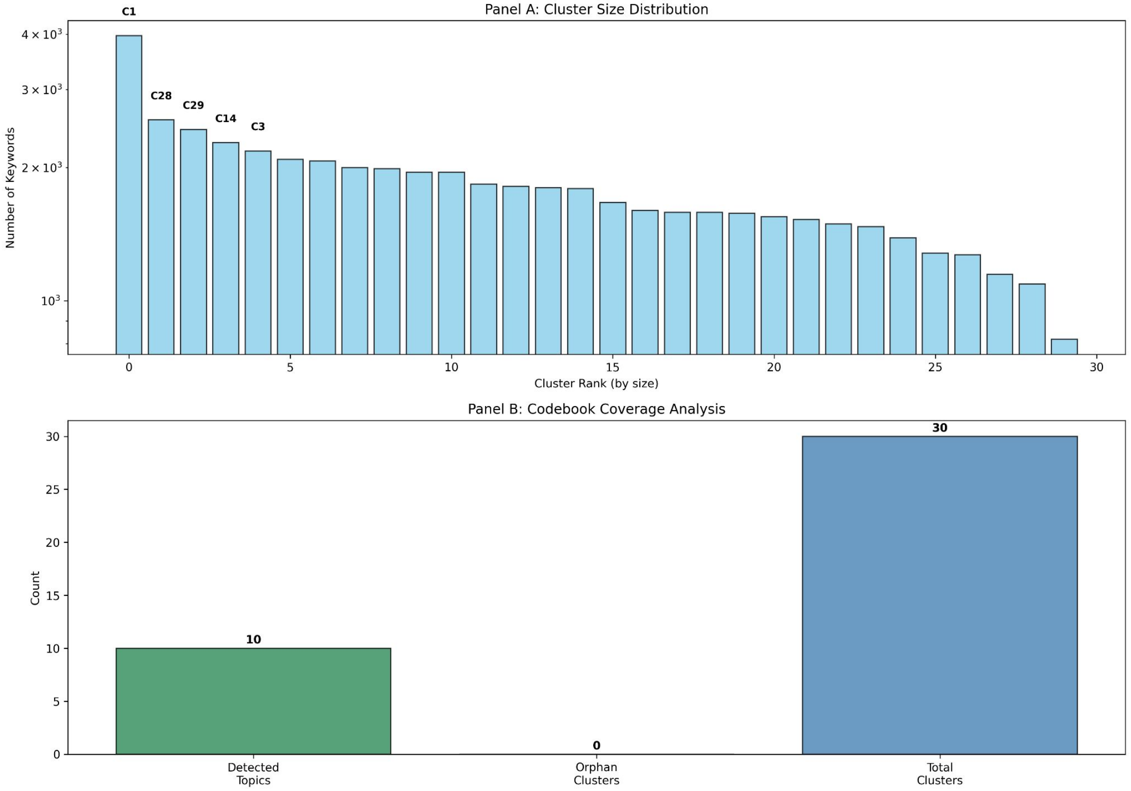

Panel A: Topic Detection Results

Figure 2: Cluster Analysis and Coverage Assessment

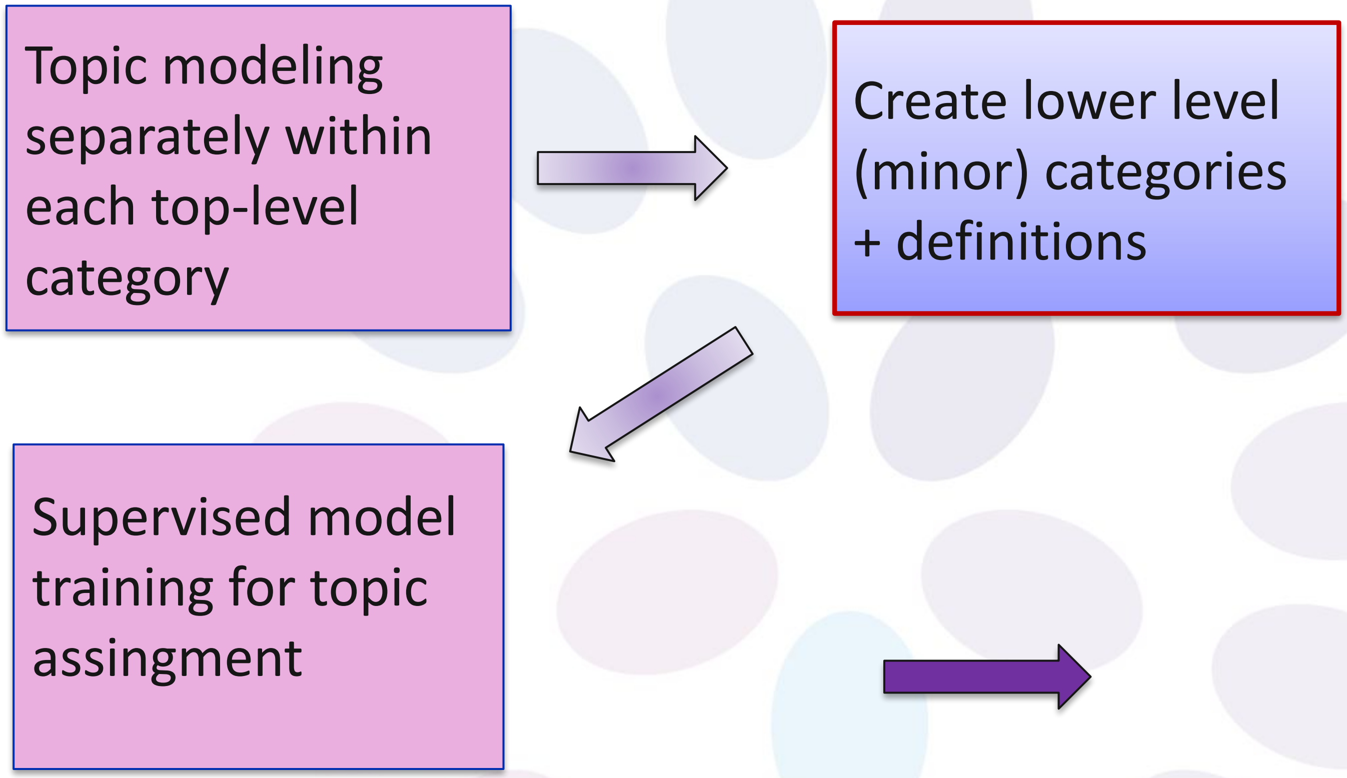

Panel A: Cluster Size Distribution

Panel B: Codebook Coverage Analysis

All 10 major LiSST topics are present in the keyword dataset, most appearing in multiple large clusters.

No “orphan” clusters emerged, indicating that the CV adequately covers the major topics represented in the keyword data.

NEXT STEPS



Data scientists

Social scientists

Topic modeling separately within each top-level category

Create lower level (minor) categories + definitions

Supervised model training for topic assignment

→ LiSST

→ Gold standard corpus

→ Annotation tool

TEAM

Alina DANCIU	Sciences Po, France	https://orcid.org/0000-0002-5126-0078
Jieun JEONG	Sciences Po, France	https://orcid.org/0009-0001-1137-6754
Lucie MARIE	Sciences Po, France	https://orcid.org/0000-0002-7045-4433
Chloé Hertrich	Sciences Po, France	https://orcid.org/0009-0008-4593-097X
Joffrey Bécart	Sciences Po, France	https://orcid.org/0009-0006-2818-5037
Mari KLEEMOLA	FSD, Tampere Univ., Finland	https://orcid.org/0000-0001-8855-5075
Katja MOILANEN	FSD, Tampere Univ., Finland	https://orcid.org/0000-0002-7668-5427
Anne MAUNU	FSD, Tampere Univ., Finland	https://orcid.org/0009-0000-7473-3918
Judit GÁRDOS	RDC, CSS, Hungary	https://orcid.org/0000-0003-1612-2216
Róza VAJDA	RDC, CSS, Hungary	https://orcid.org/0000-0002-5325-4536
Timea VENCZEL	RDC, CSS, Hungary	https://orcid.org/0009-0008-6174-1722
Éva KOVÁCS	RDC, CSS, Hungary	https://orcid.org/0000-0002-4280-9794
Enikő MEISZTERICS	RDC, CSS, Hungary	https://orcid.org/0000-0002-7078-8820
Júlia EGYED-GERGELY	RDC, CSS, Hungary	https://orcid.org/0000-0003-2259-3823
Anna HORVÁTH	RDC, CSS, Hungary	https://orcid.org/0000-0003-3061-9982
Barbara BABOLCSAY	PolTextLab, CSS, Hungary	https://orcid.org/0009-0001-1748-1352
Miklós SEBŐK	PolTextLab, CSS, Hungary	https://orcid.org/0000-0003-0595-2951