SPEAKER PROFILE

Termene AI

200,000	6,000	92%	20+
verified data sources	active paying clients	annual retention	years of experience

Business Intelligence Network for the Romanian economy: 2.7M companies, verified data, predictive models and human expertise, for trade credit, risk, compliance and due diligence. 12 years in the market.

PhD in Economics — "Costin C. Kirițescu" National Institute of Economic Research, Romanian Academy.

THE ROLE, IN SHORT

- ✓ Leads Termene.ro's **research and development**
- ✓ Turns data into **decision-making tools**
- ✓ Builds **automation and applied AI**
- ✓ Partnerships with **employers' associations, chambers of commerce**

Adrian Bănărescu

Director of Research & Development
Termene Just Srl

> *"An official source tells you where the data came from.*
> *Validation tells you whether it is clean.*
> *The gap between the two is exactly where data laundering operates."*

adrian.banarescu@termene.ro

ABOUT THE CONTRIBUTION

Panel: "Language data production, management and market development: overcoming obstacles"

2.7M companies and official Romanian-language documents processed daily: balance sheets, court rulings, official gazettes. That is where the panel's three obstacles come from, concretely - **how you build coverage, how you validate what the source declares, and how you share without losing control.**

Synthetic data – easy to produce, hard to make useful

The real bottleneck is design: **diversity and coverage** – registers, domains, speakers, not volume.
It stays under GDPR if re-identification is possible.

Real data – sharing, not raw transfer

Data sovereignty – the very design principle LDS is built on: controlled access (API, audit log, processing agreement). **Licence the use, do not ship the text.**

"Official source" ≠ "validated data"

A press investigation (Recorder, 2025) showed fictitious balance sheets, self-filed and published as submitted.
The source is official, the data is not — the same holds for any institutional archive.

LDS Country Workshop in Romania · PRECIS Center, UNSTPB

LEGAL FRAME

Termene AI

GDPR: where the chain breaks between public data and training data

"PUBLIC" DOES NOT MEAN "FREE OF GDPR"

official source → verifiable provenance → validation performed → good training data

The full chain has four links, not two. The shortcut "official data, therefore good training data" skips validation — precisely the missing link. Data that is public by law remains personal data: reusing it to train models can be **processing incompatible with the original purpose**.

Anonymisation must be irreversible

Synthetic data does not automatically leave GDPR if generation starts from real personal data. The EDPB standard is irreversible, not "hard to reverse".

Impact assessment before, not after

Any pipeline starting from millions of natural persons — administrators, shareholders — needs a **data protection impact assessment** (DPIA) before the first run, not after.

Scraping registries is a grey zone

Automated collection from public sources and training on it is not neutral: every company has administrators, so the data contains personal data.

Pseudonymisation plus a contract

The clean route: sensitive columns removed or pseudonymised, aggregated sets, and a processing agreement with an explicit basis and purpose.

Exposure does not stop at GDPR: passing contaminated data downstream also opens civil liability.

LDS Country Workshop in Romania · PRECIS Center, UNSTPB

Termene AI

Synthetic data is a method, not a legal status

WHAT GENERATION DOES NOT CHANGE

synthetic generation → ~~automatic exemption~~ → still the legal test

The fact that data was generated does not change its status. The question stays the same: **does this data relate to an identifiable natural person?** The production method does not answer it.

1 The right question

Not “is this synthetic?”, but **does it relate to an identifiable natural person?** Status follows from content, not from how it was generated.

2 The real legal test

If any **reasonable means** could reidentify someone, the data remains personal data, however synthetic it looks.

3 The hard question

How much evidence do you need to confidently call a set **anonymous data**?
That is where deep technical work and legal judgement have to come together.

Status is not permanent — processing can lose it

Even if you begin with anonymous data, you can end up with personal data again: **combining two sets that together re-identify** people, or a synthetic model that unintentionally produces linkable records. The deeper GDPR question is whether identifiability arises **in the specific context**, not whether the set was anonymous on the way in.

Legal status is not obtained by generating — it is demonstrated.

LDS Country Workshop in Romania · PRECIS Center, UNSTPB

DEEP DIVE

Termene AI

Language data production, management and market development

THE PATTERN BEHIND ALL THREE OBSTACLES

Synthetic data · Real data sharing · Security → formally valid process ≠ factually true content

A synthetic generator runs correctly, but that does not guarantee it reflects unobserved reality. A registry publishes correctly, but that does not make its content true. An authority accepts a filing correctly, but that does not verify it. **That confusion is data laundering** — the mechanism behind every obstacle on this panel's agenda.

Termene's own language-technology pipeline

Official Gazette IV › NER transformer › Entities › Cleaning

EU-funded, in production. The hard part is not extraction, **it is cleaning**: the same shareholder appearing under different diacritics created duplicates; addresses needed normalisation and cross-validation rules.

Built-in defence against data laundering – where we are heading

Record taken from the official source — VALIDATION FAILED

The target: records that fail cross-checking across sources – financial statements, the Trade Registry, public tenders, the movables security registry, flagged explicitly, not hidden and not trusted blindly. Not just extracting what one source says, but **checking it against the others and against historical behaviour**. First components are in beta.

THE FINGERPRINT PARADOX

"A real shell company has a fingerprint — a pattern across administrator, address, shareholders, and behaviour over time. A synthetic generator only reproduces the patterns it was told about. Whoever generates synthetic fraud data must already know the fingerprint — and if they know it, they do not need the model."

Data is a warning signal or a confirmation, not a verdict.

LDS Country Workshop in Romania · PRECIS Center, UNSTPB