Using AI for legal form detection

Open Source Tool LENU – Legal Entity Name Understanding

28th Meeting of the Wiesbaden Group on Business Registers – The Hague

5 October 2023

Dominik Jany – Data Scientist (GLEIF)
Zornitsa Manolova – Head of Data Quality Management and Data Science (GLEIF)

Agenda

1. GLEIF intro
2. Entity Legal Forms (ELF) code list
3. Machine learning approach
4. Transformer models

Who is Global Legal Entity Identifier Foundation (GLEIF)?

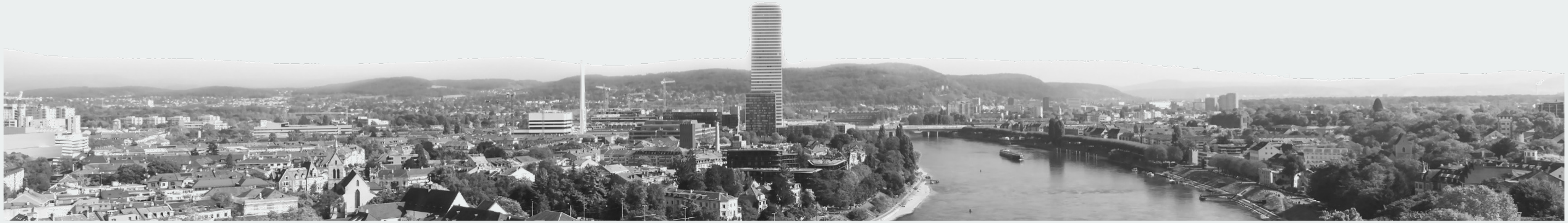
- GLEIF is a **not-for-profit Swiss foundation**, founded by the Financial Stability Board (FSB).
- GLEIF is overseen by **65 regulators and 19 observers** in the Regulatory Oversight Committee (ROC) from more than 50 countries.
- GLEIF Board has **19 independent directors**.
- GLEIF makes available the **LEI data free of charge**.

Partners for LEI issuing (LOUs)

37
and growing

Issued LEIs to date

> 2.4 millions



Legal Entity Identifier (LEI) and Key Reference Data

Who is who? Who owns whom?

"Level 1"

LEI Code 894500451FZY32C0B274

(Primary) Legal Name	Liberty Holdco Ltd.
Registered At	Registry of Companies (General Registry) Registry of Companies (General Registry) Cayman Islands RA000086
Registered As	CO-364933
Jurisdiction Of Formation	KY
General Category	GENERAL
Entity Legal Form	limited liability company (en) MPUG
Entity Status	ACTIVE
Entity created at	2000-03-31T15:00:00Z

- EntityID & Authoritative source
- Entity creation date
- Address information (…)

"Level 2": Direct & ultimate parents & fund relationships

Parents — Hide

Parents	Published Relationship	Lapsed Relationship	Reporting Exception
楽天グループ株式会社 (Direct Parent)	楽天グループ株式会社 (Ultimate Parent)		

LEI Data is available free of charge in various formats:

- GLEIF API: https://api.gleif.org/docs
- LEI Search: https://search.gleif.org
- Golden Copy: https://www.gleif.org/en/lei-data/gleif-golden-copy

 | Author: GLEIF | 28th Meeting of the Wiesbaden Group on Business Registers – The Hague | Public

05-10-2023

Identifying Legal Forms

Easy …right?

- Language proficiency necessary
- Domain knowledge necessary

Netherlands – Same Legal Form?
1. **Vereniging van Eigenaars Rijperduin**
2. **VVE Poldertocht 30-64**
"Vereniging van Eigenaars" = "VVE" → ELF Code: GNXT

Vereniging van eigenaars = Homeowner association

→ ELF Code to the rescue!

Entity Legal Forms (ELF) Code List – ISO standard 20275

A list of all legal forms – in all countries

- ELF Codes identify the distinct entity legal forms in a given jurisdiction
  - Introduced in November 2017
  - Currently 3,250 legal forms in 112 countries (more than 175 jurisdictions)
  - GLEIF acts as maintenance agency
  - Leveraged local expertise of 37 LEI Issuers

ELF Code	Country of Formation	Jurisidiction of Formation	Entity Legal Form Local Name	Abbreviations Local Language
LNBY	Canada	British Columbia	Limited Liability Partnership	LLP;SRL;SENCRL
JDX6	Cayman Islands	Cayman Islands	Special economic zone company	SEZC
B5UZ	China	China	事业单位	
QRZJ	Italy	Italy	Società Cooperativa	S.C.;soc. coop.
M886	United States of America	Alaska	Limited Liability Partnership	L.L.P.;LLP

Standardization of the legal and organizational construct per jurisdiction provides greater understanding of exposure to risk and access to capital

Descriptive Data Analysis of ELF Code List

Issued LEIs
to date

> 2.4 millions

Entity Legal Form
(ELF) Codes

> 3,250

Jurisdictions

175

Jurisdiction	#LEIs	Unique ELF Codes
GB	172.369	34
DE	171.746	29
IT	154.382	51
ES	135.190	43
NL	132.243	44
FR	110.179	197
US-DE	99.712	9
IN	87.968	33
DK	82.312	20
…	…	…

How far can we get with a generic approach?
How much do we need to optimize per Jurisdiction?

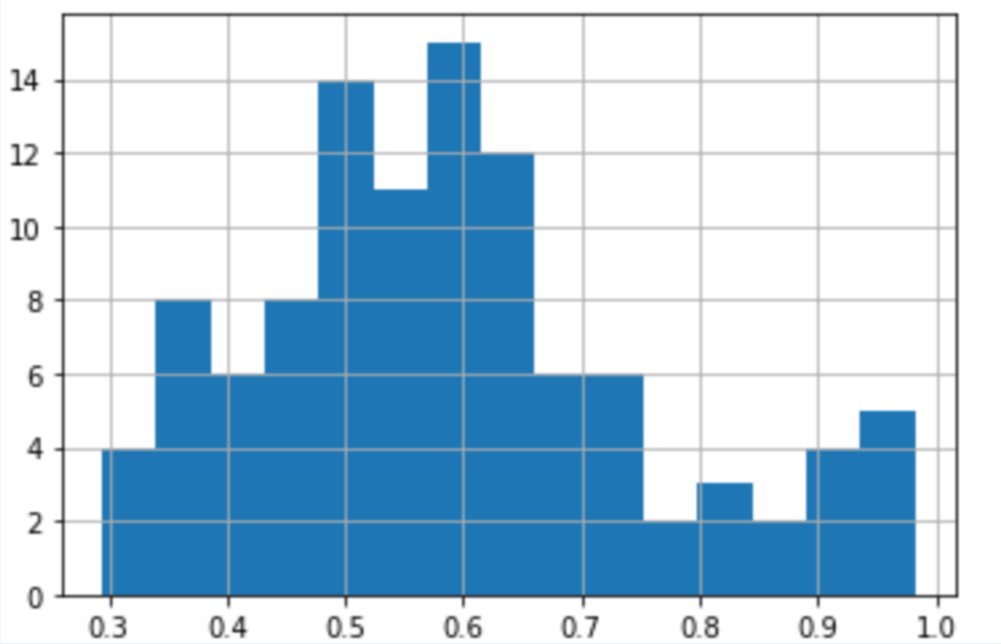
Baseline

“For all LEIs, select ELF Code that appears most frequently in that Jurisdiction”

In most Jurisdictions, that is some form of Limited, e.g. Ltd, GmbH, S.a.R.L., Aktiebolag, Aksjeselskap, ...

⇒ Mean Accuracy: 59%

Accuracy



Abbreviation Matching

- Abbreviations are maintained in ELF Code list
- However, sometimes there is none (about 37% coverage)

Example ELF Code:

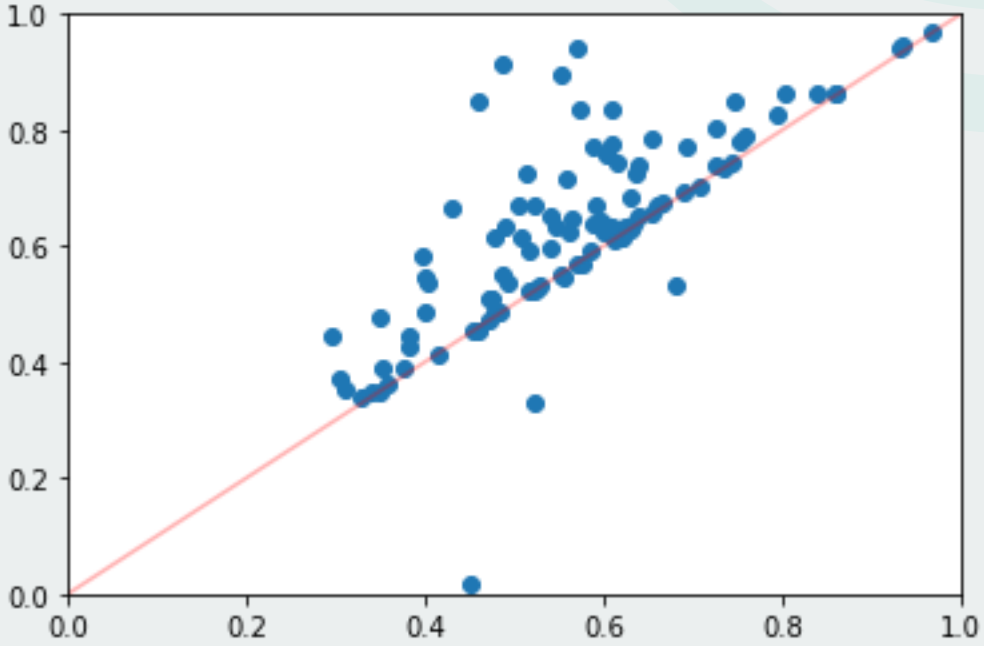
M64D - “Limited Liability Partnership” (US-CO, Colorado)

➔ `Limited; Ltd.; L.L.P.; LLP; RLLP; R.L.L.P.`

⇒ Mean Accuracy: 63%

⇒ Abbreviations Matching performs better than Baseline

Accuracy



Baseline vs. Abbreviations

Machine Learning comes in

Pipeline consisting of

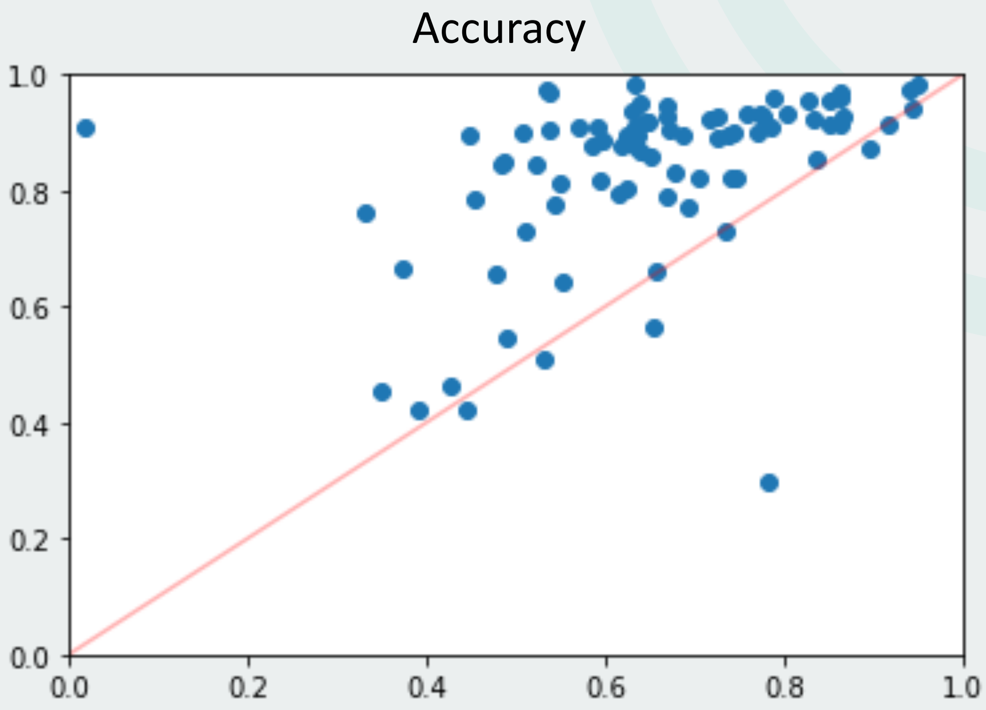
- Abbreviations
- Harmonization + Tokenization
- Naïve Bayes Classifier

⇒ Mean Accuracy: 84%

⇒ Extended ML Pipeline outperforms Abbreviations

Example:

“Katholische Kirchengemeinde Maria Königin Lingen”
➔ SQKS (“Körperschaft des öffentlichen Rechts”)



Abbreviations vs. extended ML Pipeline

⇒ Catches Legal Forms that don’t have abbreviations

Traditional Machine Learning Approach vs Deep Learning Transformers

Scikit-Learn – Naïve Bayes

- Explicit preprocessing necessary
  - Name harmonization
  - Tokenizing
  - Special character handling

→Good results, but limitations regarding:
  - Balanced accuracy
  - Legal names with non-Latin characters

BERT – Transformer Models

- Leverage pre-trained language-specific models
  - From the start, the model ”understands” context and tokens within legal names
  - No language proficiency necessary
  - Multi-lingual models available
- Pre-trained models additionally trained on LEI data

→Improvement of balanced accuracy

→Better performance for non-Latin characters

Jurisdiction	F1 Score sklearn	Balanced Acc. sklearn	F1 Score Transformer	Balanced Acc. Transformer	Comments
ES	0.9429	0.4671	0.9505	0.5044	Uncased BERT works best, because >90% of ES entities are uppercase
JP	0.2746	0.1828	0.9826	0.4293	Massive improvement when using transformer models
GB	0.9424	0.2716	0.9686	0.4089	Transformer improves balanced accuracy
US-DE	0.9593	0.421	0.9878	0.5024	Transformer improves balanced accuracy

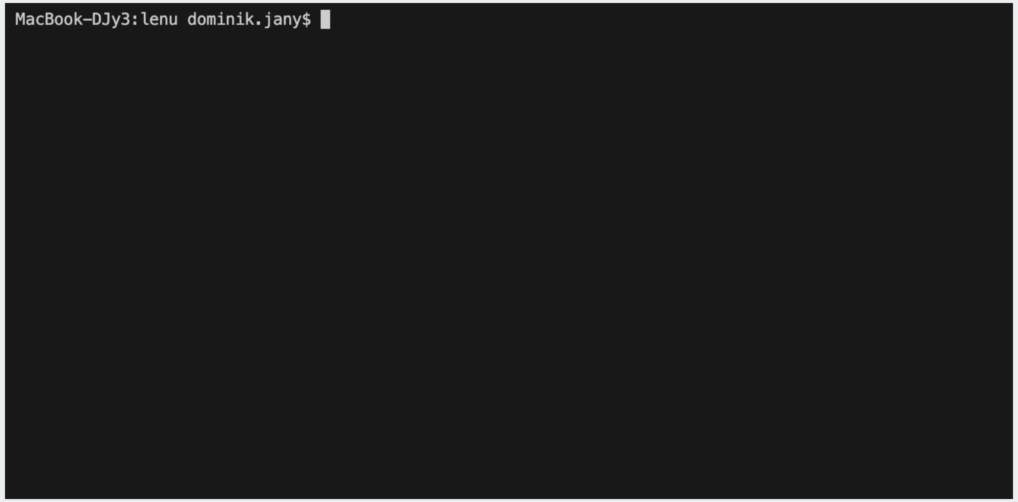
Traditional Machine Learning Approach vs Deep Learning Transformers

Jurisdiction	Model	dataset statistics #samples	classes	Transformer F1-score	macro F1-score	Traditional F1-score	macro avg f1-score
DE	bert-base-german-cased	135079	31	**0.9578**	**0.5812**	0.9433	0.5582
IT	bert-base-italian-cased	104968	50	**0.8752**	**0.2608**	0.8695	0.2270
NL	bert-base-dutch-cased	89748	20	**0.9834**	**0.7582**	0.963	0.6367
ES	bert-base-multilingual-uncased	84231	41	**0.9505**	**0.5191**	0.9429	0.4883
GB	bert-base-cased	74847	29	**0.9543**	**0.347**	0.9424	0.2687
FR	bert-base-multilingual-cased	59973	165	**0.571**	0.1107	0.4408	**0.1545**
DK	danish-bert-botxo	56226	22	**0.9444**	**0.5941**	0.9068	0.4334
US-DE	bert-base-cased	54156	12	0.958	**0.4865**	**0.9593**	0.4505
SE	bert-base-swedish-cased	48083	18	**0.9848**	**0.5711**	0.9721	0.4858
FI	bert-base-finnish-cased-v1	35587	52	**0.9851**	**0.5983**	0.9797	0.5031
LU	bert-base-multilingual-cased	33683	28	**0.8546**	0.3306	0.7455	**0.3703**
NO	bert-base-multilingual-cased	32996	27	**0.9847**	0.4931	0.9853	**0.6048**
AT	bert-base-german-cased	24433	21	**0.9559**	**0.5817**	0.9269	0.5223
BE	bert-base-multilingual-cased	23969	41	**0.5097**	**0.1275**	0.372	0.1172
KY	bert-base-multilingual-cased	20541	13	0.6707	0.3168	**0.6708**	**0.3805**
PL	bert-base-multilingual-uncased	20173	36	0.9709	0.4417	**0.9716**	**0.465**
AU	finbert-pretrain	15350	13	**0.8818**	**0.314**	0.8227	0.2854
IE	finbert-pretrain	15294	19	**0.9249**	**0.4863**	0.8648	0.4300
VG	finbert-pretrain	15086	9	0.833	**0.1768**	**0.8521**	0.1743
CZ	bert-base-multilingual-uncased	14477	52	**0.9908**	0.3824	0.9829	**0.4307**
EE	bert-base-multilingual-uncased	13824	13	**0.9965**	**0.6329**	0.9954	0.6191
CH	bert-base-german-cased	13742	28	**0.9272**	**0.3639**	0.8967	0.3178
HU	bert-base-multilingual-uncased	10041	33	**0.9265**	0.4511	0.917	**0.4897**
JP	bert-base-japanese	9690	12	**0.9828**	**0.44**	0.2746	0.1832
LI	bert-base-multilingual-uncased	9458	13	**0.9525**	0.6616	0.952	**0.7676**
US-CA	finbert-pretrain	6176	14	**0.938**	**0.3897**	0.9275	0.3572
US-NY	finbert-pretrain	4836	10	**0.9541**	**0.5166**	0.9344	0.4771
MX	bert-base-multilingual-uncased	3184	58	**0.875**	0.246	0.8427	**0.2854**
PA	bert-base-multilingual-uncased	2925	7	0.8697	0.3684	**0.8786**	0.3583
BG	bert-base-multilingual-cased	2335	19	**0.5632**	0.1596	0.4385	**0.2205**

LENU – Legal Entity Name Understanding

Open-Source Machine Learning Tool

Enables Organizations Everywhere to Automatically Detect and Standardize Legal Forms



Trained on the 2 million LEI records in the Global LEI Index

It will allow banks, investment firms, corporations, governments, and other large organizations to proactively analyze their master data, extract the legal form from the unstructured text of the legal name, and uniformly apply an ELF code, according to the ISO 20275 standard.

Developed by GLEIF and Sociovestix Labs

https://github.com/Sociovestix/lenu

Some examples where the Transformer shines…

Word attribution calculated by transformers-interpret

Unsere Kinder, unsere Zukunft –
Stiftung der Volksbank Odenwald eG

[CLS]	0.00
unsere	0.11
kinder	0.12
,	0.05
unsere	0.10
zukunft	0.17
-	0.08
stiftung	0.82
der	0.43
volksbank	0.01
oden	0.06
##wald	0.18
eg	0.18
[SEP]	0.00

Counter example:

Volksbank Odenwald eG

[CLS]	0.00
volksbank	0.09
oden	0.06
##wald	0.17
eg	0.98
[SEP]	0.00

These show how the Transformer is taking the sequential statistics into account.

`Models available at: https://huggingface.co/Sociovestix`

LENU Benefits

ELF CODE BENEFITS

Presents the legal form of an entity in a machine-readable format which can be used by AI tools and in other digitized business processes and applications.

ELF CODE BENEFITS

Overcomes problems with legal form data classification that stem from language variations and abbreviation inconsistencies.

ELF CODE BENEFITS

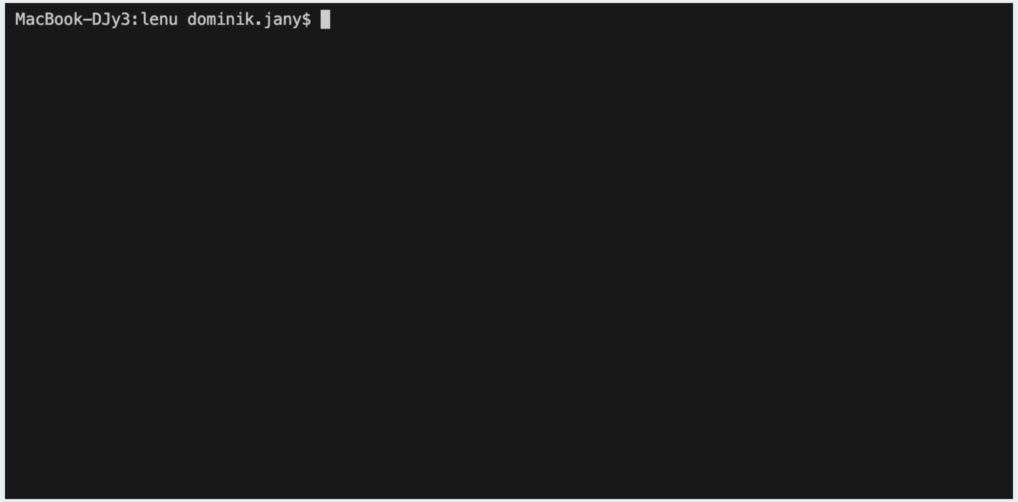
Bypasses the risks and limitations associated with manual engagement with data, including time, inefficiency, human error, and high administrative costs.

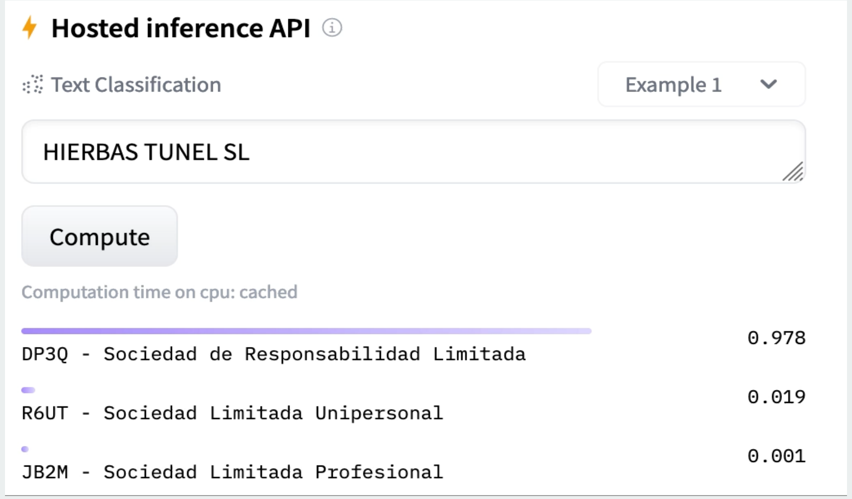
ELF CODE BENEFITS

Automates the standardization of unstructured data (legal forms as part of the organization's name), fostering greater data quality.

Thank you for your interest!
Questions?

- GLEIF API: https://api.gleif.org/docs
- LEI Search: https://search.gleif.org
- Golden Copy: https://www.gleif.org/en/lei-data/gleif-golden-copy
- Open Source Tool: https://github.com/Sociovestix/lenu
- Hugging Face data: https://huggingface.co/Sociovestix





Limitations

- This presentation contains confidential and proprietary information and/or trade secrets of the Global Legal Entity Identifier Foundation (GLEIF) and/or its affiliates, and is not to be published, reproduced, copied, or disclosed without the express written consent of Global Legal Entity Identifier Foundation.
- Global Legal Entity Identifier Foundation, the Global Legal Entity Identifier Foundation logo are service marks of Global Legal Entity Identifier Foundation.