Anonymizing language data within the CEF Data Marketplace

10 March 2021
5th ELRC Conference

Co-financed by the European Union
Connecting Europe Facility

Data Marketplace

The Data Marketplace is **easy-to-use**, **easy-to-explore**, **easy-to-trade**, and **easy-to-trust**, with features that add value for data sellers and buyers equally.

SELLING ON DATA MARKETPLACE

1. **UPLOAD DATA**
   Put your data up for sale in a few easy steps
2. **CLEAN & ANONYMIZE**
   Prepare your data with our cleaning tools and optional anonymization (coming soon)
3. **SET THE PRICE**
   Decide on the right price based on our smart price recommendations
4. **PUBLISH**
   Make the data available on the marketplace and searchable by the buyers

BUYING FROM DATA MARKETPLACE

1. **EXPLORE**
   Choose the language pair to get a glimpse of the available data
2. **FILTER RESULTS**
   Filter by domain, content type and price or find datasets from specific sellers
3. **PURCHASE & DOWNLOAD**
   Found the dataset you like? Complete the payment and download instantly
4. **RATE & REVIEW**
   Validate the quality of purchased data by giving it a score and a review

Data Marketplace - Partners

The Data Marketplace is **easy-to-use**, **easy-to-explore**, **easy-to-trade**, and **easy-to-trust**, with features that add value for data sellers and buyers equally.

Co-financed by the European Union
Connecting Europe Facility

TAUS

FONDAZIONE BRUNO KESSLER

translated.

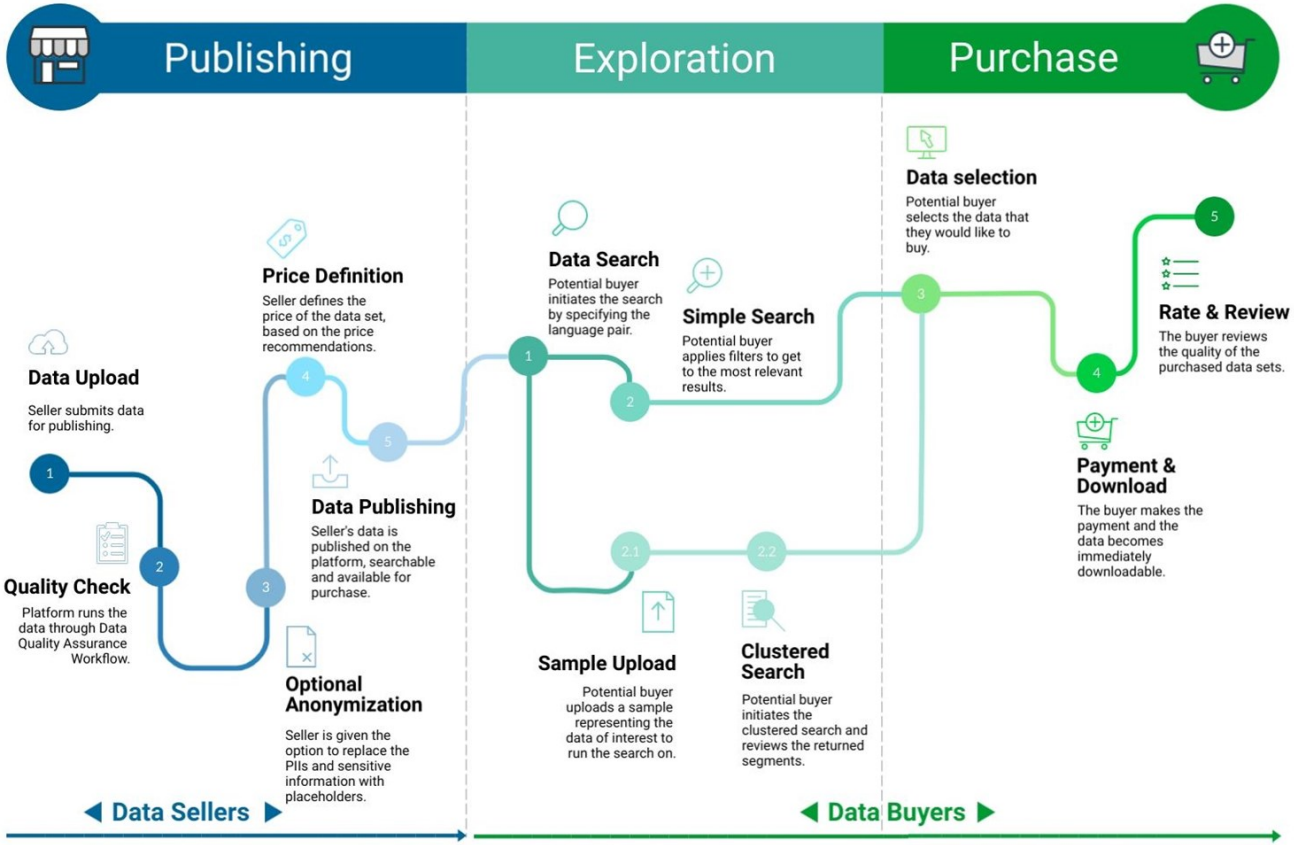
Scope and Objectives of CEF Data Marketplace

CEF Data Marketplace aims to develop a broader market serving the data needs of all stakeholders in the global translation space, across various desired language combinations and domains. It focuses on the development of a universal, secure language data trading platform that will open up a continuous and long-term supply of language data for machine translation and other machine learning applications.

The project is linked to the European Language Resource Coordination ELRC-share index for language tools, and the marketplace software will be shared under open-source licenses.

Co-financed by the European Union
Connecting Europe Facility

Data Marketplace - Workflow



Co-financed by the European Union
Connecting Europe Facility

Anonymization

- The need to anonymize Personally Identifiable Information (PIIs) depends on the context of the text

Publicly shared information (non-sensitive PII)
Come visit our restaurant **“La cucina del Signor Rossi”** in **Via Endrici, 3, Trento**

Private Information (sensitive PII)
The patient **Mr. Rossi** lives in **Trento, Via Endrici, 3.**

- Currently available tools recognize PII but are not able to distinguish between sensitive and non-sensitive PII.
- A high number of false positives (i.e. PII not to be anonymized) will be identified.

Co-financed by the European Union
Connecting Europe Facility

Anonymization Tools

- BERT
- Polyglot
- Translated Pipeline (proprietary)

Co-financed by the European Union
Connecting Europe Facility

Anonymization Tools

BERT NER

Pre-trained BERT, fine-tuned on NER data

Training Data:

- OntoNotes corpus
- 19 types in the markup schema

Technology: Deep Bidirectional Transformers

Co-financed by the European Union
Connecting Europe Facility

Anonymization Tools

Polyglot

Language-agnostic NER approach

Cast as semi-supervised term classification problem based on:

1. Word embedding
2. Wikipedia link structure and Freebase attributes
3. Exact surface form matching.
4. Oversampling of the NER terms

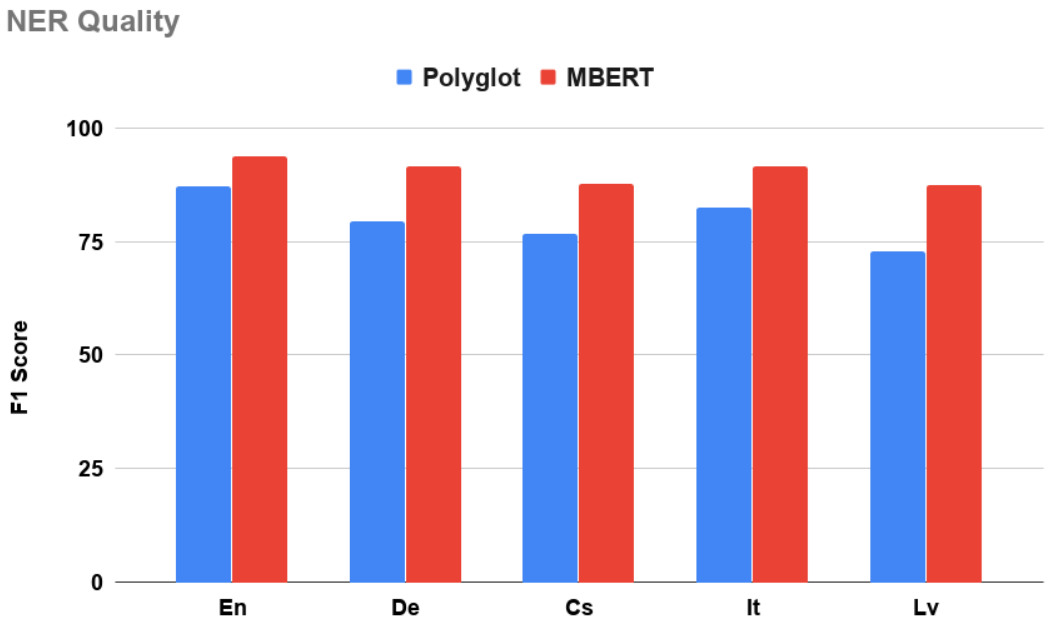
Co-financed by the European Union
Connecting Europe Facility

NER Evaluation

- Test sets manually annotated with person names available for all languages
- Gold standards used for implicit evaluation:
  - **en:** Conll 2003
  - **it:** I-CAB Evalita 2009
  - **de:** Conll 2003
  - **cs:** Cnec 2.0
  - **lv:** TildeNER corpus

Co-financed by the European Union
Connecting Europe Facility

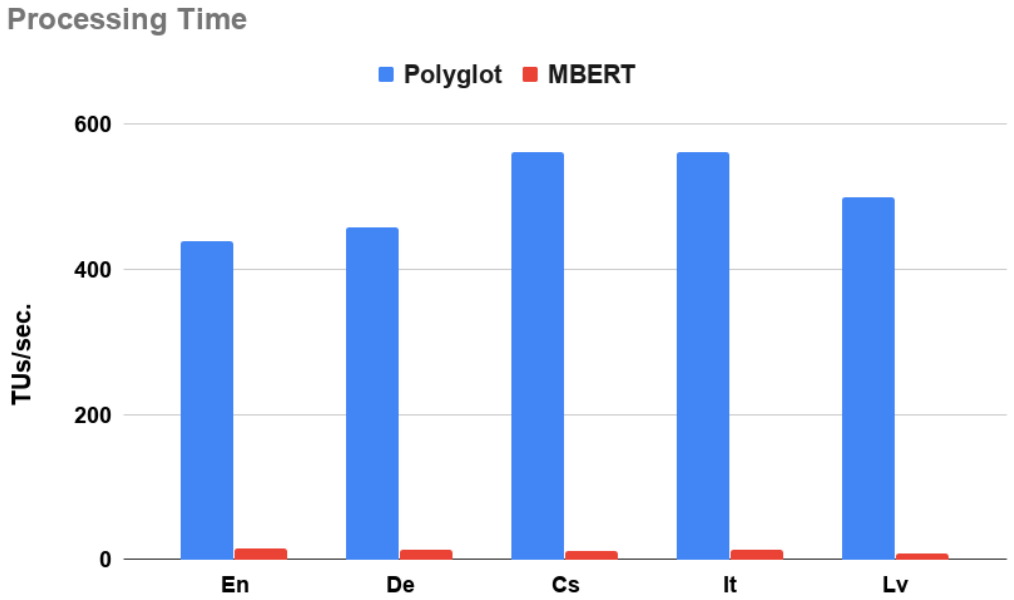
NER Evaluation



- MBERT outperforms Polyglot in all languages
- MBERT finds more entities than Polyglot

Co-financed by the European Union
Connecting Europe Facility

NER Evaluation



- Polyglot is much faster than MBERT
- MBERT processes on average 12 TUs per second

Co-financed by the European Union
Connecting Europe Facility

Translated Pipeline

- Emails
- URLs
- Addresses
- Long integers: phone numbers, credit card numbers
- Alphanumeric codes: driver’s license numbers, identity card numbers, passport numbers, social security numbers, license plate numbers

Co-financed by the European Union
Connecting Europe Facility

API

```
docker run --rm -it --net=host anonymization_service
```

```
curl -X POST -F units=
'id1|en|credit cards 1234-XXXX-YYYY of mr. John Watson and
of Jochen Mass|it|bla bla bla|id2|en|We recommend the
sites bbc.co.uk and cnn.com|it|Paolo Rossi and Giuseppina
Verdi propongono i siti agriturismo.it dolomiti.it
solocane.net'
http://localhost:8080/anonymize_service.php
```

https://github.com/hlt-mt/TM-Anonymizer

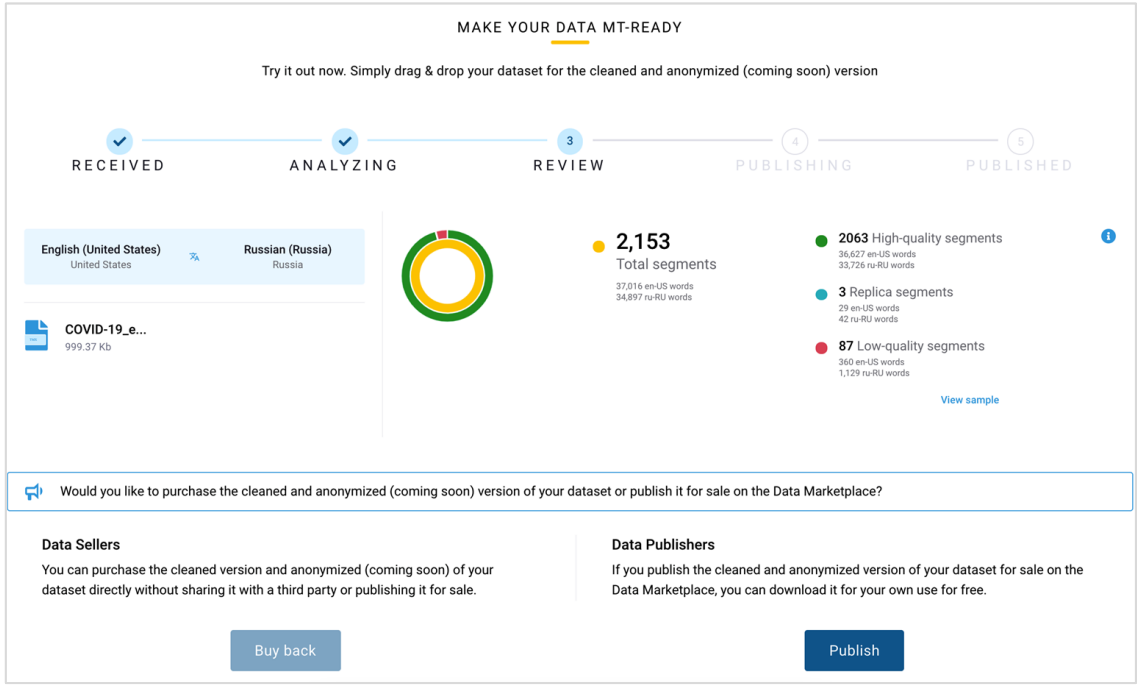

```
{"status": 0,
 "payload":
   [{"id": "id1", "side": 0,
     "annotations":
       [{"type": "CREDITCARD",
         "values":
           ["1234-XXXX-YYYY", "6666-XXXX-YYYY", "7890-XXXX-YYYY"]
        },
        {"type": "PER",
         "values":
           ["John Watson", "Jochen Mass"]
        }
       ]
    },
    {"id": "id2", "side": 0,
     "annotations":
       [{"type": "URL",
         "values":
           [ "bbc.co.uk", "cnn.com"]
        }]
    },
    {"id": "id2", "side": 1,
     "annotations":
       [{"type": "URL",
         "values":
           [ "agriturismo.it", "dolomiti.it", "solocane.net"]
        },
        {"type": "PER",
         "values":
           [ "Paolo Rossi", "Giuseppina Verdi"]
        }
       ]
    }]
}
```

Co-financed by the European Union
Connecting Europe Facility

Data Marketplace - Integration

Uploaders will be presented with recognized PIIs in different categories with examples sentences and they will be able to easily decide if they want to allow or remove certain segments.

In progress ...



The final integration and release is coming June 2021

Q&A

THANK YOU