# CEF Data Marketplace: Powering a Long-term Supply of Language Data

**Amir Kamran**[†], **Dace Dzeguze**[†], **Jaap van der Meer**[†], **Milica Panic**[†],
**Alessandro Cattelan**[‡], **Daniele Patrioli**[‡],
**Luisa Bentivogli**[⋆], **and Marco Turchi**[⋆]
[†] TAUS - Language Data Network, Netherlands {amir,dace,jaap,milica}@taus.net
[‡] Translated, Italy {alessandro,daniele.patrioli}@translated.com
[⋆] FBK, Italy {bentivo,turchi}@fbk.eu

## Abstract

We describe the CEF Data Marketplace project, which focuses on the development of a trading platform of translation data for language professionals: translators, machine translation (MT) developers, language service providers (LSPs), translation buyers and government bodies. The CEF Data Marketplace platform will be designed and built to manage and trade data for all languages and domains. This project will open a continuous and long-term supply of language data for MT and other machine learning applications.

## 1 Introduction

The CEF Data Marketplace project is an initiative co-funded by the European Union under the Connecting Europe Facility programme, under Grant Agreement INEA/CEF/ICT/A2018/1816453. The project has a duration of 24 months and started in November 2019.

With over 350[1] million new Internet users in 2019 and the annual digital growth of 9%, there is insufficient content available in the local languages. The automated translation platforms support merely about a hundred of the 4,000 languages with an established writing system. The CEF Data Marketplace will be the first platform that facilitates the buying and selling of language data to help businesses and communities reach scale with their language technologies while offering a way for the language data creators to monetize their work.

## 2 Platform Description

The platform focuses on the integration and maintenance of the already available technologies for managing and trading translation data. Specifically, the following features will be added to an existing underlying translation data repository:

- An easy-to-use mechanism to upload and annotate data-sets for data sellers, as well as options to upload updates to the data-sets;
- an easy-to-explore mechanism to find the right data for specific languages and domains for data buyers;
- an easy-to-trade transaction system for data sellers to earn monetary rewards by trading their data with data buyers;
- an easy-to-trust reputation system to improve the confidence of data buyers towards the marketplace and to ensure quality of data.

## 3 State-of-the-art Processing Tools

Advanced data processing services will be integrated to enable and facilitate data exchange through the marketplace and to encourage data sellers and buyers to join the platform. These services consist of software for cleaning, anonymizing and clustering the data to ensure that the data-sets available in the Marketplace are of high quality. These services will be provided through APIs and will be available free of charge to data providers or against a fee for users not publishing their data through marketplace. The software will



[1] https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates

also be released open source for the industrial and research communities.

## 4 Data Acquisition Strategy

To acquire as much data as possible with added value for the CEF Automated Translation (CEF-AT) Core Platform our strategy is to create a vibrant, broader market serving the needs of translation providers and translation buyers across various desired language combinations and domains. The legal framework of TAUS Data is updated to build trust[2] of the data owners to participate in the Marketplace. Clear guidelines are provided about data ownership to safeguard the copyrights and to support the royalty-based model.

## 5 Acknowledgement


*The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein.*

[2] http://hdl.handle.net/11346/TAUS-PZNM