Data Collection and Curation

Version 1.1

Service Offering Description
Document Status:

<table>
<thead>
<tr>
<th>Status</th>
<th>Final</th>
</tr>
</thead>
</table>

Document Approver(s):

<table>
<thead>
<tr>
<th>Name</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Document Reviewer(s):

<table>
<thead>
<tr>
<th>Name</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEF PAO</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Summary of Changes:

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Created by</th>
<th>Short Description of Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1.0</td>
<td>December 2016</td>
<td>Anabela Neves</td>
<td>1st draft</td>
</tr>
<tr>
<td>V1.1</td>
<td>August 2017</td>
<td>CEF PAO</td>
<td>Final (reviewed and formatted)</td>
</tr>
</tbody>
</table>
# Table of Contents

**APPROACH AND PURPOSE OF THE DOCUMENT** ................................................................. 4

1. **INTRODUCTION** ........................................................................................................... 5
   1.1. Purpose of the service ............................................................................................... 5
   1.2. Users ......................................................................................................................... 5
   1.3. Scope ......................................................................................................................... 5
   1.4. Benefits ..................................................................................................................... 6

2. **ROLES AND RESPONSIBILITIES** ............................................................................. 7
   2.1. Data Providers .......................................................................................................... 7
   2.2. European Language Resource Coordination (ELRC) Consortium ....................... 7
   2.3. National Anchor Points (NAPs) for data collection ............................................... 7
   2.4. MT@EC/eTranslation Team (European Commission) ............................................ 8

3. **HOW TO USE THE SERVICE** .................................................................................... 9

4. **TERMS AND CONDITIONS** ..................................................................................... 10

5. **ANNEX** ..................................................................................................................... 11

6. **CONTACT INFORMATION** ....................................................................................... 12
**APPROACH AND PURPOSE OF THE DOCUMENT**

The present document is the Service Offering Description (SOD) of the Data Collection and Curation service. Key content includes an explanation of the roles and responsibilities and the steps to use the service.

This document describes the European Language Resource Coordination (ELRC) effort launched by the European Commission to identify, gather, document, store, browse and access language and translation data relevant to national public services, administrations and governmental institutions across all 30 European countries participating in the CEF programme with a view to developing high-quality automated translation systems for EU languages for CEF Automated Translation.

**Glossary**

The key terms used in this Service Offering Description are defined in the CEF Definitions section on the CEF Digital Single Web Portal:

https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/CEF+Definitions

The key acronyms used in this Service Offering Description are defined in the CEF Glossary on the CEF Digital Single Web Portal:

1. INTRODUCTION

Multilingualism is a core value for Europe, home to 24 official languages. Though linguistic diversity is a cherished feature of European identity, multilingualism can also lead to barriers in the digital world. Fortunately, language technologies such as automated translation can help to crack language barriers between people and nations, enabling a truly Multilingual Europe.

Significant amounts of valuable linguistic data are generated every day in all Member States, and in the CEF-affiliated countries, non-governmental and private organisations. A large part of this data can be very valuable as language resources for developing and training automated translation systems.

1.1. Purpose of the service

The overall goal of the European Language Resource Coordination (ELRC) is to collect language resources from and for public service administrations in all EU Member States and CEF-affiliated countries in order to improve the quality, coverage and performance of CEF Automated Translation services in the context of current and future CEF digital online services (CEF DSIs) such as Online Dispute Resolution (ODR), Electronic Exchange of Social Security Information (EESSI), eJustice, Safer Internet etc. As such, all data collected by the ELRC will be used by the European Commission (EC) to support the development of CEF eTranslation service and its adaptation to the relevant CEF digital services.

ELRC promotes the collection and curation of large general domain corpora, whether monolingual (e.g. official corpora of national languages) or multilingual parallel corpora (e.g. aligned corpora from translation memories) as well as domain-specific corpora and terminology resources (e.g. lexica and dictionaries) in the fields of consumer rights, culture, legal domain, social security, health, public procurement, etc.

All data resources gathered in the ELRC initiative will be used to develop high-quality automated translation systems for EU languages for the eTranslation service.

1.2. Users

The Data Collection and Curation service is intended for the following type of users:

- **Data providers**: providers of language and translation data relevant to national public services, administrations and governmental institutions across all 30 European countries participating in the CEF programme in the form of large general domain corpora, whether monolingual or multilingual parallel corpora as well as domain-specific corpora and terminology resources (e.g. lexica and dictionaries) in the fields of consumer rights, culture, legal domain, social security, health, public procurement, etc.

1.3. Scope

The data collection and curation services include:
• Indicating possible data sources through ELRC-SHARE (these can be either general or domain specific data sets and can be monolingual or multilingual);

• Uploading language resources through ELRC-SHARE;

• Technical processing of language resource data sets; ELRC provides a range of services to public administrations, helping them to prepare language resources for submission to CEF Automated Translation.

• Clarifying legal aspects of submitted data and their re-use conditions; and

• Onsite assistance for technical and legal issues.

1.4. Benefits

The Data Collection and Curation service is an intermediate step aimed at further developing the CEF eTranslation service and therefore generates as such no immediate benefits for its users (Data Providers) until the latter are end users of the eTranslation service. The long-term benefits are such that the language resources can be used to build, improve and/or evaluate natural language systems such as machine translation engines thus producing high quality machine translations.

The Data Collection and Curation service has been designed to generate a list of benefits to the end users of the eTranslation service:

• **Increased language coverage**: collection of resources to cover the languages from the 30 European countries participating in the CEF programme. Additional coverage for languages with currently less resources.

• **Increased domain coverage**: collection of domain-specific corpora and terminology resources (e.g. lexica and dictionaries) in the fields of consumer rights, culture, legal domain, social security, health, public procurement, etc. allowing the training of domain specific engines for the eTranslation service.

• The technical processing services guarantee that the provided language resources will lead to higher quality automated translation systems.
2. Roles and Responsibilities

This section describes the main roles of the Data Collection and Curation service.

2.1. Data Providers

Role: Data providers can be any public administration or organisation which holds data potentially useful for improving CEF Automated Translation.

Responsibilities:

- Identify data that can be shared on the ELRC-Share Repository
- Submit the data resources via the ELRC website

2.2. European Language Resource Coordination (ELRC) Consortium

Role: Coordinators of the Data Collection and Curation effort

Responsibilities:

- Coordinate the Data Collection and Curation effort
- Promote ELRC in national events and workshops
- Identify, gather, document, store, browse and access language resources and translation data
- Provide a helpdesk for language resource/data providers
- Manage the ELRC-Share Repository
- Liaise with the European Commission

2.3. National Anchor Points (NAPs) for data collection

The National Anchor Points are individuals who are appointed in the framework of the European Language Resources Coordination (ELRC) action to support the data collection process in each of the 30 participating countries. There are usually two NAPs in each of the countries; one is a technological representative (Technology National Anchor Point) and the other representative from the public services administration (Public Services National Anchor Point).

The Technology NAP is a highly regarded language or language technology expert. He or she often has a distinguished academic or research background, or represents a national language institution.

1 http://www.lr-coordination.eu/resources
The Public Services NAP is a representative of national public services, public administration or a ministry. He or she acts as a liaison contact person to the national, regional and local administrations, and is able to effectively mobilize and spread the word about the importance of language resources and the ELRC effort among the public authorities/ministries in each country. Together the NAPs form the Language Resources Board, the governance body of the European Language Resource Coordination effort.\(^2\)

**Role:** Support the data collection process in each of the 30 participating countries

**Responsibilities:**
- Act as liaison/contact persons between ELRC and the national, regional and local administrations
- Effectively mobilize and spread the word about the importance of language resources and the ELRC effort among the public authorities/ministries in each country

### 2.4. MT@EC/eTranslation Team (European Commission)

**Role:** Technical support and guidance

**Responsibilities:**
- Liaise with the ELRC consortium and the NAPs on any technical and policy matters related to the data collection and curation

\(^2\) [http://www.lr-coordination.eu/anchor-points](http://www.lr-coordination.eu/anchor-points)
3. HOW TO USE THE SERVICE

Data providers may submit Language Resources to ELRC at any exploitation stage, by:

- Indicating potential sources for relevant data (data sources are identified websites URLs that could be exploited, through a crawling process, for the preparation of Language Resources, in particular parallel corpora to be built up from multilingual websites);

- Directly submitting resources owned by the data providers.

Submissions can be done via the ELRC website at [http://www.lr-coordination.eu/resources](http://www.lr-coordination.eu/resources)

From a user perspective, ELRC-SHARE offers\(^3\):

- **For data owners and contributors**: basic functionalities for registering/login and contributing language resources through a simple web form. A detailed walkthrough for data owners and contributors is available [here](http://www.lr-coordination.eu/resources).

- **For metadata editors** (ELRC members): a user-friendly metadata editing environment for the description of resources. A detailed walkthrough for metadata editors is available [here](http://www.lr-coordination.eu/resources).

- **For the general public**: browsing, simple and faceted search of the resources inventory.

---

\(^3\) Status in April 2017.
4. TERMS AND CONDITIONS

The general terms and conditions of CEF Building Blocks can be consulted in the Master Service Arrangement, available on the CEF Digital Single Web Portal:

https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/Master+Service+Arrangement
5. Annex

6. **CONTACT INFORMATION**

A technical and legal service desk is available through the [European Language Resources Coordination (ELRC)](https://www.elrc.europa.eu) initiative.

A team of technical and legal experts is happy to answer any questions on the technical or legal aspects related to the use, production, collection, processing, and sharing of language resources.

The service desk can be contacted through one of the following channels:

<table>
<thead>
<tr>
<th>Channel</th>
<th>Contact Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telephone*</td>
<td>+33 970 440 522</td>
</tr>
<tr>
<td>Secretariat Support</td>
<td>+49 681-8575 5285</td>
</tr>
<tr>
<td>Skype</td>
<td>CEF-AT-Helpdesk</td>
</tr>
<tr>
<td>E-mail</td>
<td><a href="mailto:help@cef-at-helpdesk.org">help@cef-at-helpdesk.org</a></td>
</tr>
</tbody>
</table>

6.1.1.1.1.1.1. *This number is associated to a web-conferencing desktop (Skype). All requests will be handled in English.

In addition, a [Webforum](https://www.elrc.europa.eu) is also at your disposal to submit any questions you may have. Upon receipt, your question will be processed and answered within 24 hours (simple query) to 5 days (complex query).

All questions and answers relevant for other users are compiled in a [FAQ](https://www.elrc.europa.eu).