I Want to Donate Data. What Now?

Donating data is easy. Depending on the size of the dataset, you do one of the following:

- Send your data to ELRC via email to **data@lr-coordination.eu**, as a zip file.
- Upload data onto the ELRC Repository, as a zip file, at **http://www.lr-coordination.eu/resources**.

Each dataset must have a short description and a title in English.

„Managing linguistic and cultural diversity in Europe is complex. I believe language technologies, such as Machine Translation, can play a significant role in supporting the setup of the Digital Single Market in Europe.“

Víctor Calvo-Sotelo,
State Secretary of Telecommunications and for the Information Society (Spain)

Where Can I Find More Information?

The ELRC has a website with an abundance of information to get you started. Please visit our website at **http://lr-coordination.eu/resources**.

If you prefer to speak to someone directly about your data or any questions you may have, feel free to contact our ELRC Helpdesk. The contact details can be found online at **http://lr-coordination.eu/helpdesk**.

The ELRC Consortium

For further information, please contact:
Prof. Josef van Genabith (ELRC Coordinator)
Dr. Andrea Lösch (ELRC Project Manager)

Deutsches Forschungszentrum
für Künstliche Intelligenz
Multilingual Technologies
Stuhlsatzenhausweg 3
Campus D3.2
D-66123 Saarbrücken

Phone: +49 681 857 75-5285
Email: info@lr-coordination.eu

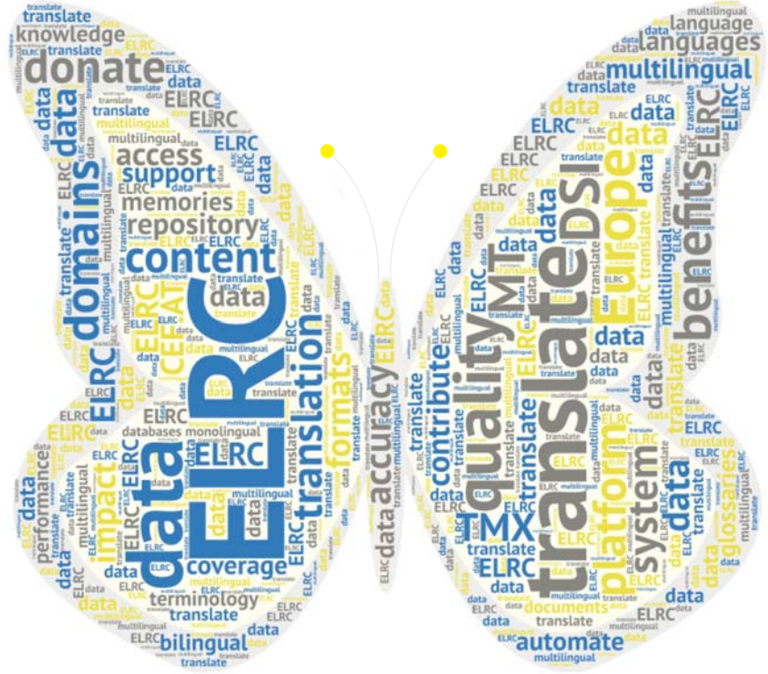
Further information about ELRC is available online at **www.lr-coordination.eu**

European Commission

European Language Resource Coordination
Connecting Europe Facility

Why Your Language Data Matters:

The Butterfly Effect



Why Your Language Data Matters: The Butterfly Effect

Your data matters. Not only to your institution. Your language data can make a difference to hundreds of public service administrations like yours across Europe.

Your language data will be included in a repository full of data, whether monolingual or multilingual, from other public administrations in Europe. This repository provides the basis for the CEF Automated Translation (CEF.AT) platform whose purpose is to enable multilingual Digital Service Infrastructures (DSIs) to put knowledge into the hands of Europe's citizens in any language, anytime.

Donating Data: What's in it for you?

When you donate language data, you are increasing the coverage for the content and languages that pertain to your administration. The most direct, immediate, and significant benefit is that content processed with the CEF.AT platform will be more quickly and accurately translated. The more you give, the better your translated content.

You will also benefit from the data donated by other public administrations because they, too, will contribute content that is relevant to your domains.

State-of-the-art automated translation systems such as CEF.AT rely on huge amounts of language data to optimize their performance. The ELRC Initiative collects this language data in a variety of formats and language combinations and sends it to the CEF.AT platform to enhance system performance and improve the quality of the texts it translates.

What Data Can I Donate?

Bilingual or multilingual content in digital editable formats provides the best results for the CEF.AT platform when it comes to improving coverage and performance. Monolingual data also serves a useful purpose when training automated translation systems.

The objective is to use your data in conjunction with that of others to help you translate your content efficiently. You will have access to a profound amount of multilingual content covering your topics, issues, knowledge bases because translated data is being collected from sources just like yours all across Europe.

Ideally, donated data should include translation memories such as TMX files, or documents in their original format including the source and at least one translated document for alignment.

The type of content that can have a positive impact include:

- internal reports and other documents
- publications and other materials for external use
- web content and brochures
- terminology databases and glossaries

Three Good Reasons to Donate Data to the ELRC

- You can significantly improve the translations you produce for your domains and your languages.
- You have 24/7 access to the system, allowing you to translate urgent content or larger amounts of data anytime.
- You have peace of mind when it comes to confidential content. The CEF.AT platform is completely secure, unlike most other open MT systems.

Quality is Key, But Quantity Helps

The good news is that the bilingual and multilingual data that is being collected from public administrations like yours has primarily been translated by professional translators, so you know the quality of the content going into the repository is as good as yours. This is key when creating parallel data for use on the CEF.AT platform.

While quality plays a critical role in how useful the content will be, the amount of data provided by you is no less important.

Quantity helps achieve a level of consistency in many aspects that are crucial to the translation. The more often specific language is used, the more accurate the generated data will be, significantly reducing the effort required to edit the content for publishing.

> *„As many of the language technology solutions are data-driven, they rely on the wide availability of language data, in other words, language resources. The European Commission supports making the use of language resources in many different ways [...] The Connecting Europe Facility promotes the collection and sharing of language resources in 30 languages to make European public services multilingual."*
>
> *Andrus Ansip,*
> *Vice-President of the European Commission*

What about Confidential Content?

Prior to donating your data, you should:

- ensure that the material is in the public domain or that the necessary licenses have been obtained;
- follow the national Public Sector Initiative (PSI) transposition rules.

Confidential and personal information must be excluded from the dataset you are sharing. The ELRC can support any necessary anonymization tasks that must be performed on your content.