ELRI
A Decentralised Network of National Relay Stations to Collect, Prepare and Share Language Resources

Thierry Etchegoyhen,[1] **Borja Anza Porras,**[2] **Andoni Azpeitia,**[1] **Eva Martínez Garcia,**[3]
José Luis Fonseca,[4] **Patricia Fonseca**,[4] **Paulo Vale**,[4] **Jane Dunne**,[5] **Federico Gaspari**,[5]
Teresa Lynn,[5] **Helen McHugh**,[5] **Andy Way**,[5] **Victoria Arranz**,[6] **Khalid Choukri**,[6]
Hervé Pusset,[6] **Alexandre Sicard**,[6] **Rui Neto**,[7] **Maite Melero**,[8]
David Perez,[8] **António Branco**,[9] **Ruben Branco**,[9] **Luís Gomes**[9]

[1] Vicomtech, Spain - {tetchegoyhen, aazpeitia}@vicomtech.org
[2] Bexen Medical, Spain (Work done while at Vicomtech) - borja.anza@gmail.com
[3] CEIEC, Spain (Work done while at Vicomtech) - eva.martinez@ceiec.es
[4] AMA, Portugal - {jose.fonseca, patricia.fonseca, paulo.vale}@ama.pt
[5] DCU, Ireland - {jane.dunne, federico.gaspari, teresa.lynn, helen.mchugh, andy.way}@adaptcentre.ie
[6] ELDA, France - {arranz, choukri, herve, alexandre}@elda.org
[7] Linkare, Portugal - rneto@linkare.com
[8] SEAD, Spain - maite.melero@upf.edu, dperezf@minetad.es
[9] University of Lisboa, Portugal - {antonio.branco, ruben.branco, luis.gomes}@di.fc.ul.pt

Abstract

We describe the European Language Resource Infrastructure (ELRI), a decentralised network to help collect, prepare and share language resources. The infrastructure was developed within a project co-funded by the Connecting Europe Facility Programme of the European Union, and has been deployed in the four Member States participating in the project, namely France, Ireland, Portugal and Spain. ELRI provides sustainable and flexible means to collect and share language resources via National Relay Stations, to which members of public institutions can freely subscribe. The infrastructure includes fully automated data processing engines to facilitate the preparation, sharing and wider reuse of useful language resources that can help optimise human and automated translation services in the European Union.

Keywords: ELRI, Language Resources, European Infrastructure, Connecting Europe Facility

1. Introduction

The European Language Resource Infrastructure project[1] (ELRI) is an initiative funded within the Connecting Europe Facility (CEF) Programme[2], which started in October 2017 and ended in September 2019.[3] Its main goal has been the development of an infrastructure to help collect, process and share language resources (LR) in the European Union. Seven partners were involved in the project, representing four Member States (MS), namely France, Ireland, Portugal and Spain.

Quality multilingual language resources are of paramount importance to improve translation services, both human and automated, and thus support language equality in the European Union. The development of European Digital Service Infrastructures (DSI), in particular, is tied to the development of transversal services such as eTranslation[4], the automated translation service provided by the Directorate-General for Translation (DGT) to Public Administrations of the European Union. Such services can greatly benefit from language resources produced by public institutions on a daily basis across the European Union.

The ELRI initiative sought to support the collection of quality language resources, by mitigating obstacles identified during the data collection efforts of companion initiatives such as the European Language Resource Coordination project[5] (ELRC). Among the main identified difficulties were the reluctance of data holders to make their data available due to perceived concerns related to Member State regulations and IPR issues, the lack of internal expertise or dedicated staff to take the steps needed to provide appropriately prepared language resources, and the lack of clear short-term incentives to share their resources.

ELRI has addressed some of these issues by providing a sustainable solution deployable at the Member State level, where data checking and processing take place prior to sharing the resources, at the Member State level or beyond, and users can benefit in the short term from fully prepared language resources that can improve their own translation processes, human or automated.

A key contribution of the ELRI project has been the development and deployment of National Relay Stations (NRS), which are web applications that facilitate the collection, preparation and sharing of language resources. Each NRS is available to members of public institutions in the corresponding Member State and its user interface is provided in the language(s) of the Member State, thus providing an environment for LR sharing that is in line with the linguistic specificities of the relevant Member State. National Relay Stations integrate fully automated processing of multilingual resources to reduce the time and effort required for the

[1] www.elri-project.eu

[2] https://ec.europa.eu/inea/en/connecting-europe-facility

[3] See (Etchegoyhen et al., 2019) for more details on the project.

[4] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation

[5] http://lr-coordination.eu/

Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020), pages 35–43
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

manual reviewing and processing of file collections, whilst also providing stakeholders with fully prepared resources in the short term. This integrated processing notably allows the creation of translation memories from raw user data in the form of document collections in multiple languages and the automated cleanup of existing translation memories. ELRI also features a group-based sharing policy where users can select the group(s) with which they intend to share their resources, thus providing the means to share language resources according to the different sets of constraints that may be tied to specific resources.

A major outcome of this initiative was the provision of a sustainable infrastructure that will be maintained after the completion of the project itself, with a detailed governance plan to support the extension of the network to new Member States and EEA countries.

The remainder of this paper is organised as follows. In Section 2. we describe the core objectives and approach of the ELRI initiative. Section 3. presents the components of the infrastructure and Section 4. describes the LR validation process. In Section 5., we describe the activity of the network at the end of the project, including the community of stakeholders that was built and the initial resources that were collected during the 2-year project. Section 6. outlines the sustainability of the solution and the governance plan for countries willing to join the network after the conclusion of the ELRI project. Finally, Section 7. draws conclusions from the project.

2. Objectives and Benefits

The core objectives of ELRI can be summarised as follows:

- Build and deploy an infrastructure to help collect, prepare and share language resources that can in turn improve translation services in the European Union, both human and automated.
- Automate the creation of translation memories and other resources from raw data provided by public institutions and translation centres.
- Provide flexible means to share language resources at the national, European and Open Data levels.
- Prioritise resources that are relevant to Digital Service Infrastructures.
- Contribute to improve the EU automated translation services that are freely available to public institutions.
- Deploy ELRI in France, Ireland, Portugal and Spain, with a future extension to additional member states as a key objective beyond the current action.
- Provide a robust and sustainable infrastructure.

These objectives were aligned with the identified challenges regarding the collection of quality language resources, and aimed to provide the following benefits:

- The provision of flexible means of sharing resources establishes a clear process where compliance with the relevant restrictions can be established at every step.
- Raw language resources are converted automatically into a format useful for translation experts as well as machine translation infrastructures.
- Data sharing with ELRI provides broad compliance verification covering intellectual property rights and the Public Sector Information Directive.
- Language resources can be shared as deemed appropriate by stakeholders, with return benefits for providers as well as users of translation services.
- Data holders can benefit from the automatically prepared resources in the short term to help optimise their own translation processes.
- By sharing their resources, stakeholders can benefit from improved European translation services such as eTranslation and promote language equality for the languages of their Member States.

This set of benefits was at the core of the ELRI project and the infrastructure was designed to achieve these objectives.

3. ELRI Infrastructure

In this section, we provide a summary of the infrastructure developed within the project.

3.1. Architecture

ELRI is a decentralised network composed of National Relay Stations, i.e. the web applications designed to collect, prepare and share language resources. Figure 1 illustrates the currently deployed infrastructure, where Each Member State deploys an instance of a National Relay Station, localised into the language(s) of the Member State and comprising a Web application, data processing engines and a database of language resources.

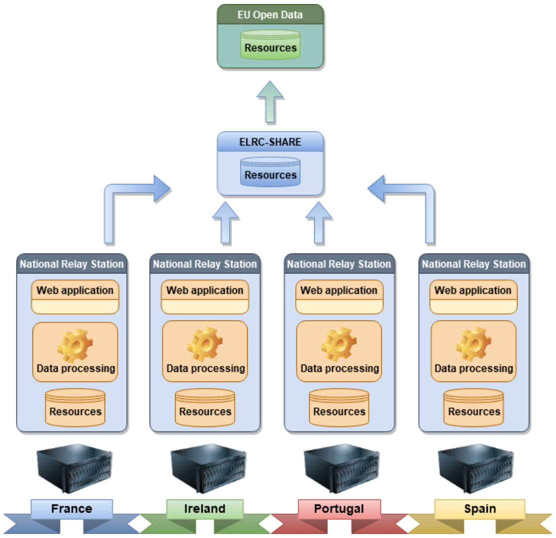


Figure 1: Overview of the ELRI network

The Web application serves as an interface where users of public institutions of the Member State can register and contribute their resources. The data uploaded by users of the NRS are processed by integrated engines, which perform sequences of processing steps to produce structured and clean language resources from raw data. These data processing pipelines, called toolchains, can notably create translation memories from raw document collections in multiple languages or clean existing translation memories. The processed resources are then available for review and validation, a task performed by designated personnel in each Member State.

Prepared resources that are deemed valid are then published in the NRS of the Member State, thus becoming directly available to the users who contributed them, as well as to the other users of the groups with which the data contributors are willing to share the resources. Resources that are shared with the European Commission are then transferred to the ELRC-SHARE repository[6], via API or manual transfer.[7] Additionally, resources that have been shared as Open Data are deposited to the EU Open Data Portal[8], via links to ELRC-SHARE. The communication between the principal components of a National Relay Station is illustrated in Figure 2.

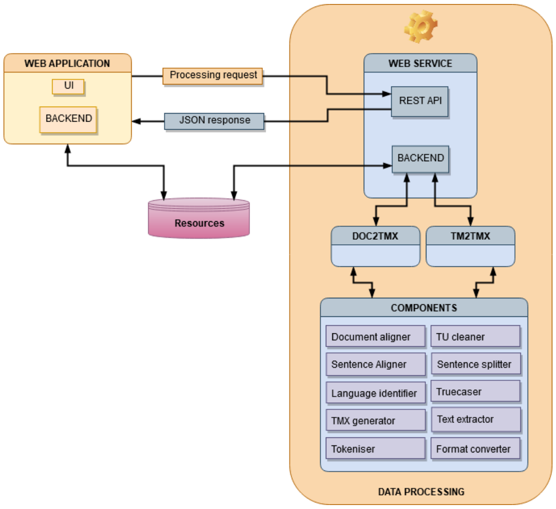


Figure 2: Communication between NRS components

The web application communicates with the data processing component via a web service, with requests sent via a REST API and responses provided as JSON objects. The initial user data as well as those generated by the data processing engines are stored in a shared repository, accessible to both the Web application and the data processing engines. The main components of the NRS software are provided as Docker containers, assembled via docker-compose, and comprise the web application itself, the data processing pipelines, an nginx web server, a solr search server, and a postgres relational database.[9]

The decentralised nature of the network provides robustness for a sustainable service, as eventual discontinuing of one of the NRS nodes would not impact the persistence of the service in the other Member States where it is deployed.

3.2. National Relay Stations

The Web application provides the necessary functionality for users to register, browse the catalogue of resources, download resources available to them and contribute their own resources. The application also handles all actions related to storage and retrieval of language resources, and interfaces with the automated data processing engines.

The application is a fork of the ELRC-SHARE software[10], itself based on the META-SHARE software[11]. The core functionality of the web application includes Web page navigation, user registration and access, data upload, user-provided information, interface with automated data processing functionality, metadata editing, data sharing under group-based policy, data download and email communication with users of the service. Even though modifications have been made to the look-and-feel of the ELRC-SHARE codebase, as well as fixes and adaptations of the user interface to match the requirements established for ELRI, the underlying infrastructure was preserved for the most part, and the metadata established for the resources stored by the system have notably been maintained as is. This ensures compatibility with the requirements of the Automated Translation services of the DGT. There are however three main differences between the original codebase and the ELRI Web application.

First, the application was localised into the language(s) of the four Member States that were represented in the project. The original English content was thus translated into French, Irish, Portuguese and Spanish. The main goal of the localisation process was to provide an environment suited for the users of the NRS in each Member State, also in line with the efforts towards language equality in the European Union. For Ireland, this requirement led to adding a language switch to the user interface, allowing NRS users of that Member State to switch at will between the Irish and English environments.

The second main difference is the integration of automated data processing, described in more detail in the next Section. To be able to process different types of data, the Web application was extended with a functionality to branch files to the appropriate data processing engine, according to file types, and to retrieve the results of data processing. The integration of automated data processing functionalities is one of the key features of the Web application in ELRI, one which allows to accelerate the preparation of language resources and their delivery to the users.

Finally, the third major difference is the inclusion of a group-sharing policy which provides flexible means to share data, acknowledging that sharing restrictions may need to vary for specific resources. Sharing via an NRS is

[6] https://elrc-share.eu/

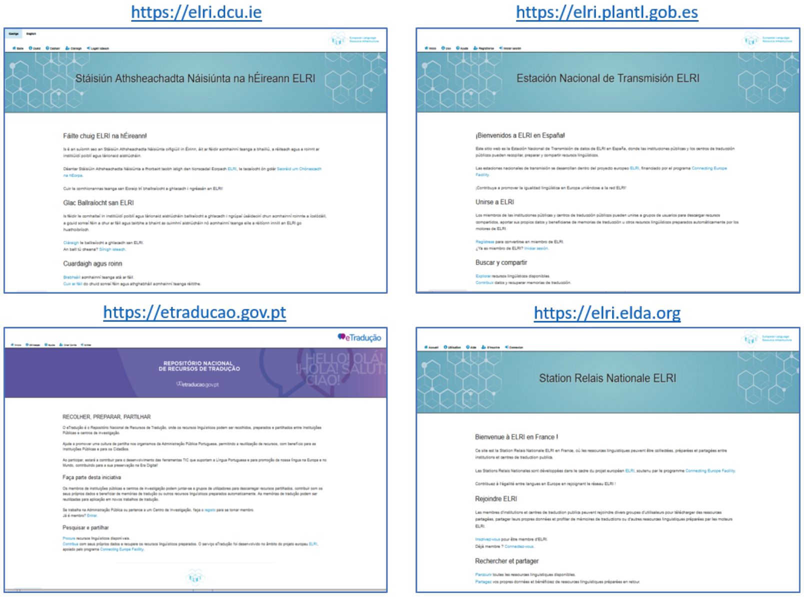
[7] At the end of the project, manual transfer was still necessary, in part because information required for LR publication on ELRC-SHARE, such as LR documentation, could not be transferred at the time via its API.

[8] https://data.europa.eu/euodp/en/home

[9] Further documentation is available at: https://github.com/ELDAELRA/ELRI/tree/master/docker

[10] https://github.com/MiltosD/ELRC2

[11] https://github.com/metashare/META-SHARE



https://elri.dcu.ie

Stáisiún Athsheachadta Náisiúnta na hÉireann ELRI

Fáilte chuig ELRI na hÉireann!

Is é an suíomh seo an Stáisiún Athsheachadta Náisiúnta oifigiúil in Éirinn, áit ar féidir acmhainní teanga a bhailiú, a réiteach agus a roinnt ar institiúidí poiblí agus lárionaid aistriúcháin.

Déantar Stáisiúin Athsheachadta Náisiúnta a fhorbairt taobh istigh den tionscadal Eorpach ELRI, le tacaíocht ón gclár Saoráid um Chónascadh na hEorpa.

Cuir le comhionannas teanga san Eoraip trí bhallraíocht a ghlacadh i ngréasán an ELRI!

Glac Ballraíocht san ELRI

Is féidir le comhaltaí in institiúidí poiblí agus lárionaid aistriúcháin ballraíocht a ghlacadh i ngrúpaí úsáideoirí chun acmhainní roinnte a íoslódáil, a gcuid sonraí féin a chur ar fáil agus tairbhe a bhaint as cuimhní aistriúcháin nó acmhainní teanga eile a réitíonn innill an ELRI go huathoibríoch.

Cláraigh le ballraíocht a ghlacadh san ELRI.
An ball tú cheana? Sínigh isteach.

Cuardaigh agus roinn

Brabhsáil acmhainní teanga atá ar fáil.
Cuir ar fáil do chuid sonraí féin agus athghabháil acmhainní teanga réitithe.

https://elri.plantl.gob.es

Estación Nacional de Transmisión ELRI

¡Bienvenidos a ELRI en España!

Este sitio web es la Estación Nacional de Transmisión de datos de ELRI en España, donde las instituciones públicas y los centros de traducción públicos pueden recopilar, preparar y compartir recursos lingüísticos.

Las estaciones nacionales de transmisión se desarrollan dentro del proyecto europeo ELRI, financiado por el programa Connecting Europe Facility.

¡Contribuya a promover la igualdad lingüística en Europa uniéndose a la red ELRI!

Unirse a ELRI

Los miembros de las instituciones públicas y centros de traducción públicos pueden unirse a grupos de usuarios para descargar recursos compartidos, aportar sus propios datos y beneficiarse de memorias de traducción u otros recursos lingüísticos preparados automáticamente por los motores de ELRI.

Regístrese para convertirse en miembro de ELRI.
¿Ya es miembro de ELRI? Iniciar sesión.

Buscar y compartir

Explorar recursos lingüísticos disponibles.
Contribuir datos y recuperar memorias de traducción.

https://etraducao.gov.pt

REPOSITÓRIO NACIONAL
DE RECURSOS DE TRADUÇÃO

RECOLHER, PREPARAR, PARTILHAR

O eTradução é o Repositório Nacional de Recursos de Tradução, onde os recursos linguísticos podem ser recolhidos, preparados e partilhados entre Instituições Públicas e centros de investigação.

Ajude a promover uma cultura de partilha nos organismos da Administração Pública Portuguesa, permitindo a reutilização de recursos, com benefício para as Instituições Públicas e para os Cidadãos.

Ao participar, estará a contribuir para o desenvolvimento das ferramentas TIC que suportam a Língua Portuguesa e para promoção da nossa língua na Europa e no Mundo, contribuindo para a sua preservação na Era Digital!

Faça parte desta iniciativa

Os membros de instituições públicas e centros de investigação podem juntar-se a grupos de utilizadores para descarregar recursos partilhados, contribuir com os seus próprios dados e beneficiar de memórias de tradução ou outros recursos linguísticos preparados automaticamente. As memórias de tradução podem ser reutilizadas para aplicação em novos trabalhos de tradução.

Se trabalha na Administração Pública ou pertence a um Centro de Investigação, faça o registo para se tornar membro.
Já é membro? Entrar.

Pesquisar e partilhar

Procure recursos linguísticos disponíveis.
Contribua com seus próprios dados e recupere os recursos linguísticos preparados. O serviço eTradução foi desenvolvido no âmbito do projeto europeu ELRI, apoiado pelo programa Connecting Europe Facility.

https://elri.elda.org

Station Relais Nationale ELRI

Bienvenue à ELRI en France !

Ce site est la Station Relais Nationale ELRI en France, où les ressources linguistiques peuvent être collectées, préparées et partagées entre institutions et centres de traduction publics.

Les Stations Relais Nationales sont développées dans le cadre du projet européen ELRI, soutenu par le programme Connecting Europe Facility.

Contribuez à l'égalité entre langues en Europe en rejoignant le réseau ELRI !

Rejoindre ELRI

Les membres d'institutions et centres de traduction publics peuvent rejoindre divers groupes d'utilisateurs pour télécharger des ressources partagées, partager leurs propres données et profiter de mémoires de traductions ou d'autres ressources linguistiques préparées par les moteurs ELRI.

Inscrivez-vous pour être membre d'ELRI.
Déjà membre ? Connectez-vous.

Rechercher et partager

Parcourir toutes les ressources linguistiques disponibles.
Partagez vos propres données et bénéficiez de ressources linguistiques préparées en retour.

Figure 3: National Relay Stations in Ireland, Spain, France and Portugal (clockwise from the top left)

done on the basis of well-defined groups, where users can browse and download only those resources that are shared with a group that they belong to. There are three different groups to which users of an NRS belong by default:

- *NationalOrganisations*: This group includes all registered users of the NRS from a specific country and resources shared with this group are accessible to all registered users of the NRS based in that Member State.
- *NationalOrganisations+EuropeanCommission*: This group includes all registered users of the NRS and the European Commission, via the ELRC-SHARE repository, who may then utilise the shared resources to improve the eTranslation services.
- *OpenData*: This group includes all registered users of the NRS and all users of the free Open Data portal of the European Union.

These default groups are always available to data contributors and aim to cover the most frequent cases of resource sharing. If different sharing needs arise for specific resources, users may request the ad hoc creation of specific groups by contacting the designated staff running the NRS in the relevant Member State. The four localised National Relay Stations are shown in Figure 3.

3.3. Data Processing

As previously indicated, each National Relay Station includes data processing engines which can handle different types of content and file formats, including doc(x), odt, rtf, pdf, tmx, sdltm and plain text.[12] Figure 4 describes the main processing steps for the four major types of data handled by the engines.

The leftmost case in the figure describes the operations needed to handle documents containing translations in two or more languages. This is the most complex scenario and its main steps are summarised below.[13]

The contents of the input files in different formats are first extracted, followed by automated language identification which allows the different text files to be grouped by language.[14] Within each file, the text is then split into separate sentences, to allow further processes to apply. Each sentence is then pre-processed, which mainly includes tokenisation and truecasing; these operations are performed with scripts that are part of the Moses toolkit[15] (Koehn et al., 2007). All document pairs with content in different languages are then automatically aligned with the DOCAL document aligner (Etchegoyhen and Azpeitia, 2016). For all document pairs whose alignment score indicates that the documents are a translation of each other, sentence alignment is then performed on the content, retrieving translations at the sentence level.[16] From the aligned sentences a translation memory in TMX format 1.4b is then gener-

[12] Although processing (collections of) PDF files is possible, the recommendation is to process the editable source files when these are available, as some challenges remain with extracting text from PDF files, potentially resulting in smaller language resources generated from the original data.

[13] Unless otherwise specified, all components are Java components developed by Vicomtech and licensed to the Innovation and Networks Executive Agency (https://ec.europa.eu/inea/en) under conditions.

[14] Text extraction is performed with Apache Tika™ (https://tika.apache.org/).
Language identification is performed with the Cybozu language identification library (https://github.com/shuyo/language-detection/tree/master/src/com/cybozu/labs/langdetect).

[15] https://github.com/moses-smt/mosesdecoder

[16] Sentence alignment is performed with HunAlign (Varga et al., 2005): https://github.com/danielvarga/hunalign

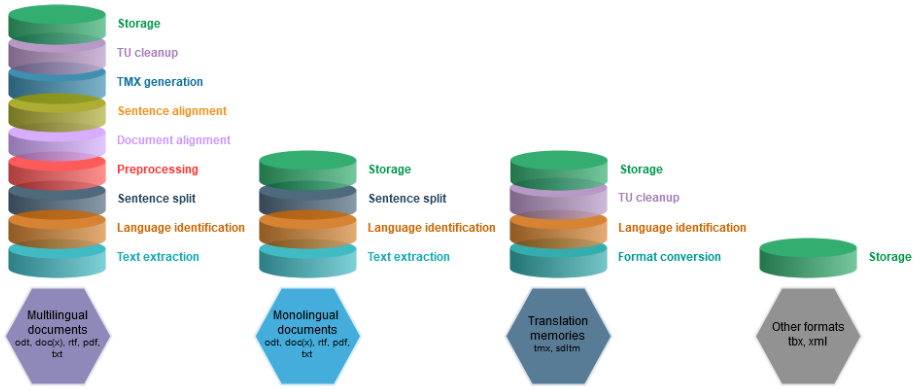


Figure 4: Main data processing scenarios

ated, with the identified sentence translations encapsulated in paired translation units. The entire translation memory is then cleaned up, removing the errors generated by erroneous alignments, filtering translation units that feature content mismatches indicated by marked length differences, unexpected languages or character sequences, for instance; duplicate translation units are also removed automatically. Finally, the clean translation memory is stored and indexed by the system.

The second case from the left is comparatively simpler, as it involves files with content in a single language. In this case, only a subset of the previously described processes apply, namely text extraction, language identification, sentence split and storage. Collections of monolingual files are thus transformed into a single file with one sentence per line. Although not as useful to human translators or automated translation as translation memories, domain-specific monolingual data can be helpful to train machine translation systems via several techniques and the ELRI processing engines are prepared to provide structured resources from strictly monolingual data.

The third case involves existing translation memories as input. In this case, the first step is format conversion, since the system handles translation memories in SDLTM format in addition to the TMX standard. Once converted to TMX, language identification is performed on the translation units as a second step. The translation memory then undergoes the previously described clean-up operations, generating a clean version of the initial translation memory.

Finally, a fourth case was added to the system, as terminology files in TBX format and resources in XML format can be stored and shared in a National Relay Station. In this case, no particular processing is performed, as terminological units cannot be filtered similarly to sentential translations and resources in unpredictable XML format cannot be processed without additional knowledge on the format.

The automated processing component of the NRS software is a Java application which integrates and connects the different components responsible for each processing step. Two major toolchains were designed and implemented: TM2TMX, which handles all processing related to existing translation memories, and DOC2TMX, which manages multilingual as well as monolingual input files.[17]

The overall process is performed with quality components, supporting an optimal creation of structured resources from raw data. For instance, the document alignment step, which is an essential part in multilingual scenarios, is performed with DOCAL, one of the top-performing tools for the task in terms of quality of the alignments and processing efficiency (Azpeitia and Etchegoyhen, 2019). The ability of the NRS software to ingest raw data in multiple file formats and generate structured resources in an automated manner is one of the main features of the ELRI infrastructure.

4. Quality Control and Validation

Language resources uploaded to a given NRS undergo a systematic validation process, summarised below.

The first step involves the contribution of a resource by registered users of the NRS, who upload their data and specify the desired level of sharing for each resource. Once uploaded, the data are then automatically processed via the integrated language processing engines, a process called Ingestion which results in prepared language resources. An important next step in the process is resource validation, which is performed by dedicated personnel on the basis of strict guidelines for quality control. If at any step an issue is detected, the process is put on hold until issues are eventually resolved with the user who contributed the data. An initial review is first performed to detect possible issues with the original data uploaded by the user. This might be the case, for instance, if the files significantly mix content in more than one language, or if the content underwent digital corruption at some point. Resources that pass initial review then undergo quality reviewing, which involves manual examination of samples of the processed data, to determine for instance the quality of the translation units in the case of translation memories generated by the automated language processing engines. Poor alignment quality, which may happen for instance with some input files in PDF format, would result in the resource not being validated and the user being notified of the issue.

[17] The second toolchain shares the initial processing steps in multilingual and monolingual scenarios, as can be seen in Figure 4; despite its name, the output of this toolchain for monolingual input data is a text file, not a TMX.

The third main step in the validation process involves the review of potential personal, confidential or sensitive information. Although users are required to warrant that the data they contribute does not infringe on any legislation, such as the GDPR[18], the ELRI validation process involves a specific step to help determine if the contributed data may nonetheless include such data. For this purpose, a specific tool was developed within the project to process the data under validation and generate a report on detected patterns of sensitive data and named entities.[19] Patterns include national identification numbers, passport numbers, words and phrases, in the relevant national language(s), indicating confidential material or typical formulations related to personal information, among others,that can be easily customised to country-specific needs and circumstances. The tool is meant only as an aid and no guarantee is given that it would fully or adequately capture sensitive or personal information in the data. However, it may help detect the presence of this type of data, in which case the validation process would be placed on hold until the matters are resolved, or eventually abandoned if no resolution is reached.

The final step in the validation process involves reviewing the legal aspects associated with the resource. This includes a review of the licensing scheme selected by the user. By default, the user can select among the main types of licenses typically associated with the sharing of language resources, such as Creative Commons licenses[20]. Reviewers evaluate the selected license and check that the relevant information is available, such as attribution text and IPR holder information, as needed. Additionally, users may provide their own licenses for a given resource, in which case the legal validation will involve a specific examination of the user-provided licensing scheme prior to any further validation. The selected sharing group is also reviewed to set the appropriate metadata, for instance ensuring that resources shared as Open Data allow uses besides the DGT.

Finally, if no issues are detected during the validation process, the reviewer will sign off for publication of the resource, which will then be available for download for the data holder and all members of the selected sharing groups.

5. Network Activity

An important part of the ELRI project was dedicated to building communities of stakeholders across the four Member States involved in the initiative and beyond. In this section, we describe the main dissemination activities, the initial resource collection efforts which took place during the project, as well as key features that support the maintenance of the network and its eventual extension to new countries.

5.1. Stakeholders Communities

As a nationally deployable infrastructure, ELRI was meant to facilitate contacts and interactions with stakeholders, notably via the four institutions in charge of hosting an NRS in their respective Member States: the Administrative Modernisation Agency (AMA) in Portugal, Dublin City University in Ireland, the Evaluation and Language Resources Distribution Agency (ELDA) in France, and the Secretary of State for Digital Advancement (SEAD) in Spain. This key feature of the network proved to be an important factor in building strong communities of stakeholders across the board, to support the continuous collection and sharing of language resources.

Several events were organised during the project to disseminate the goals and benefits of the ELRI infrastructure, resulting in growing communities of users who viewed the approach based on localised National Relay Stations as an important component to handle their respective resources. A series of workshops was notably organised in all four Member States in spring 2019, to provide an open and practical forum on the use of the ELRI services for public institutions. These events drew large attendances overall, demonstrating the interest generated by the ELRI approach and opening the doors to public entities in the different Member States involved in the initiative. In addition to these dissemination events, a large number of direct contacts and interactions with stakeholders took place at the national level during the project, which helped raise awareness on the importance of language resources, digital advancement and optimised translation processes at the national and European levels.

As a result of these community building activities, the National Relay Stations have registered growing numbers of active users from different institutions of the Member States where they are deployed. Figure 5 shows the number of institutions and authorised users by the end of the ELRI project, in September 2019. With 71 participating institutions and 101 authorised users at the time, the National Relay Stations can be considered to have attracted the interest of public institutions in the Member States participating in the Action.[21]

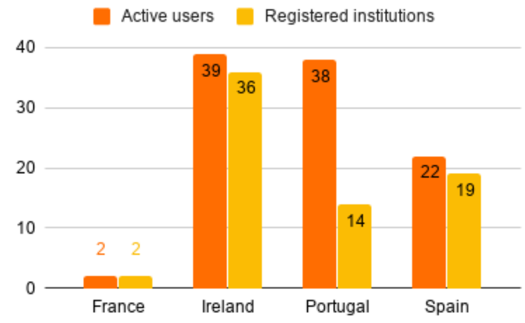


Figure 5: Registered institutions and active users (2019/09)

[18] https://gdpr-info.eu/

[19] Named entity recognition and classification are performed with the SpaCy toolkit: https://spacy.io/

[20] https://creativecommons.org/

[21] As a tentative basis of comparison to evaluate the significance of these numbers, (Lösch et al., 2018) indicates that “more than 58 public sector organisations across Europe had shared their language data with ELRC”, at the end of the 2016-2017 project, which targeted all Member States of the Union and EEA countries.

5.2. Language Resource Collection

Registered users of the different National Relay Stations have contributed a number of initial resources, several of which have been fully validated and published on the corresponding NRS. Figure 6 shows the published resources in the four National Relay Stations as of September 2019.

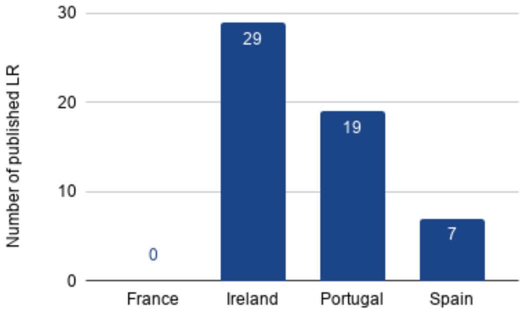


Figure 6: Number of published resources (2019/09)

The collection of resources had thus been initiated by the end of the project, with first sets of resources in all but one Member State. Regarding the French NRS, it should be pointed out that the national events addressing stakeholders took place at a later stage compared with other countries, and that several discussions are currently ongoing with institutions willing to participate and share data. Meetings have taken place for that purpose after completion of the project and initial resources are starting to be uploaded and processed via the French National Relay Station. The sustained National Relay Stations allow resource collection efforts to be adapted to the specific dynamics of the Member States and, in the case of France, ELRI will be available to support an increased sharing of resources over time.

Although the number of published resources is indicative of the initial activity for each NRS, resources vary in terms of content, with users uploading data of varying sizes. Figure 7 illustrates the number of translation units for published resources.

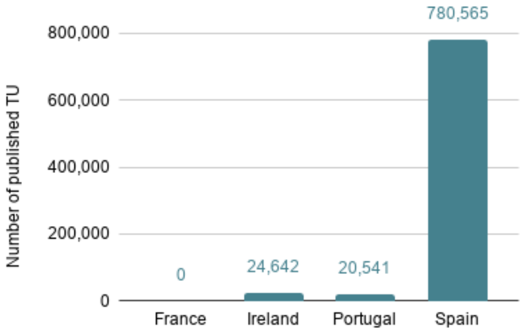


Figure 7: Number of published translation units (2019/09)

As shown in this figure, although the Spanish NRS has published a comparatively smaller number of resources than its Irish and Portuguese counterparts, several of the published resources in that Member State contain large amounts of content, with close to 800 thousand translation units. Although an important factor, the size of the resources, be it the number of translation units or the number of sentences for monolingual data, is only one indicator of the usefulness of a resource, as smaller resources may contain domain-specific information that is of equal importance for both human translators and for the training of accurate machine translation systems.

As previously described, the ELRI infrastructure provides the means to share resources beyond the national level. Figure 8 indicates the percentage of resources shared with the European Commission or as Open Data. Two main conclusions can be drawn from these figures. First, the fact that some resources remain at the Member State level indicates that there is some need for country-based repositories. Secondly, the fact that most LRs have been transmitted beyond the national level shows that ELRI stations can act as a relay in the global data collection effort. It is worth noting that resources that remain at the national level for the time being may be shared further in the future if the relevant data holders consider that the conditions are met for extended sharing of specific resources.

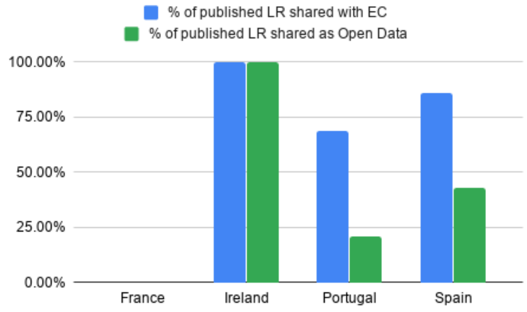


Figure 8: Percentage of published resources shared beyond Member States (2019/09)

Overall, 48 bilingual resources, amounting to 816,553 translation units, have been transferred beyond the national level during the initial resource collection phase in the last six months of the project.[22]

6. Sustainability and Expansion

A key objective of the ELRI project was the development of an infrastructure that would be sustainable beyond the lifetime of the EU-funded initiative. On technical and financial grounds, the outcome of the project is a solution that requires minimal management and associated resources, thus

[22] Although comparisons with other initiatives are difficult to establish, given the available information and differences in reporting methodology, indications regarding these numbers may be drawn from the results reported in (ELRC, 2017). Between 2016 and 2017, the authors report 225 collected resources, covering all official EU languages, plus Icelandic and both variants of Norwegian, out of which 138 were bi-/multi-lingual corpora. Information on the amount of translation units and on the proportion of resources gathered from direct crawling of public websites (an approach not undertaken in ELRI), are not available in the report, making further comparisons difficult.

providing a solid basis for its durable maintenance. The benefits provided by the ELRI infrastructure, from minimal management to integrated support for LR creation and management, play an important role in the decision of the different institutions in the different Member States to sustain its services after completion of the project, as has been the case since then with continued collection, preparation and sharing of LRs. Although the infrastructure provides the means to facilitate resource management, a sustained commitment by each institution in charge of an NRS is required to involve dedicated personnel for resource reviewing and publication. Future dedicated funding support for the National Relay Stations at the national or European level may help consolidate the sustainability investments already made by each institution.

As part of its activities, the ELRI project had also designed a structured plan that enables new countries to join the network and deploy its services with minimal efforts and costs. Managing an ELRI National Relay Station requires a Managing Body, with the following main characteristics and responsibilities:

- Be a public institution of the Member State/EEA country or an institution endorsed by a public body.
- Commit to maintain the NRS operations independently of associated project funding.
- Coordinate with the bodies in charge of similar projects and related initiatives.
- Oversee and execute the relevant activities to deploy, adapt and manage the NRS.

The candidacy of a Managing Body should be approved by the appropriate bodies, part of the governance structure, to be determined by the European Commission. This state of affairs is motivated by the fact that there should be only one National Relay Station per Member State/EEA country, to avoid conflicts and confusion on the part of end-users. Since the ELRI framework was developed within the Connecting Europe Facility programme, the integration of new countries should also be controlled to ensure the expected standards of representation and activity oversight.

The ELRI Advisory Board was established, with the seven entities who led the development of the infrastructure and whose role is to provide members of the network with their expertise on the infrastructure, including requirements, technical knowledge, best practices and overall experience in managing National Relay Stations. The Board is also meant to provide assistance to the European Commission in relation to new candidacies for countries willing to join the network, in an advisory capacity.

The inclusion of new countries is meant to be both facilitated and controlled. Thus, on the one hand, a detailed list of required activities and expected costs was prepared to assist potential new Managing Bodies, supported by the relevant documentation. On the other hand, the established governance structure, which requires approval by the relevant EU bodies, ensures that the deployment of an NRS in a new country would be controlled and in accordance with the established goals of the ELRI framework.

By the end of the project, several Member States and EEA countries had expressed their strong interest in deploying their own National Relay Station, and discussions are under way to follow through on this expansion of the network.

7. Conclusions

We have described the main achievements of the ELRI initiative, which has led to the development of a functional, tested and deployed infrastructure in all four Member States that participated in the CEF Action, namely France, Ireland, Portugal and Spain. The ELRI infrastructure is composed of independent National Relay Stations that facilitate the collection of language resources from public institutions joining the network, providing them with fully automated data processing services that allow the efficient creation of useful resources from raw data, such as translation memories from multilingual documents. The prepared resources can then be used to optimise translation services, provided either by professional human translators or by automated translation systems such as eTranslation.

ELRI services offer flexible means to share language resources and provide data holders, who dedicate time and effort to sharing their data, with prepared resources as an immediate benefit that has been a key feature of the initiative. Thus, the project aimed to benefit all stakeholders equally, as a means to build a community of interest and a positive dynamic around the sharing of quality language resources. Dissemination activities and direct contacts with stakeholders have led to positive feedback and strong interest in joining the ELRI network, from members of public institutions as well as representatives from new Member States willing to host their own National Relay Station.

The adopted bottom-up approach to LR collection, via National Relay Stations reserved for public institutions of a given country, is a unique feature of ELRI that provides a pragmatic solution to the actual difficulties in directly sharing resources outside the national realm. With a majority of collected resources having been shared beyond the national level, to repositories with wider access such as ELRC-SHARE, the ELRI network has demonstrated its potential to act as an effective relay for resource sharing, while also providing a framework adapted to needs and constraints of public institutions at the Member State level.

The collection and preparation of resources within the project was initiated in 2019 and led to the publication of an initial batch of resources in the independently deployed National Relay Stations. Overall, 71 institutions had registered to the network by the end of the project and contributed more than 800,000 translation units within the first months of activity. Although preliminary, and with different volumes collected in each country, the established community of users and dynamic are paving the way for continued and increased sharing of language resources across the board. As a sustainable solution, with National Relay Stations being maintained after the lifetime of the project, ELRI has provided additional building blocks to the global effort towards increased efficiency for translation services in the European Union.

8. Acknowledgements

The ELRI project was co-financed by the Connecting Europe Facility of the European Union, who we thank for their support. We would like to thank all stakeholders from public institutions in France, Ireland, Portugal and Spain who have participated in the data collection activities, for their time and effort in helping collect and share language resources. We also wish to thank the ELRC consortium and the Innovation and Networks Executive Agency for their assistance and support during the project, and the three anonymous IWLTP 2020 reviewers for the insightful comments.

9. Bibliographical References

Azpeitia, A. and Etchegoyhen, T. (2019). Efficient Document Alignment Across Scenarios. *Machine Translation*, 33:205–237.

ELRC. (2017). European language resource coordination: Final report. Technical report.

Etchegoyhen, T. and Azpeitia, A. (2016). A Portable Method for Parallel and Comparable Document Alignment. *Baltic Journal of Modern Computing*, 4(2):243–255. *Special Issue: Proceedings of EAMT 2016*.

Etchegoyhen, T., Gaspari, F., Dunne, J., McHugh, H., Vale, P., Fonseca, J. L., Fonseca, P., Melero, M., Branco, A., Gomes, L., Neto, R., Arranz, V., and Choukri, K. (2019). ELRI: Final Report. Technical report. `http://www.elri-project.eu/resources/D1.3_ELRI_Public_Final_Report.pdf`".

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 177–180. Association for Computational Linguistics.

Lösch, A., Mapelli, V., Piperidis, S., Vasiļjevs, A., Smal, L., Declerck, T., Schnur, E., Choukri, K., and Van Genabith, J. (2018). European language resource coordination: Collecting language resources for public sector multilingual information management. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1339–1343.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing (RANLP 2005)*, pages 590–596.