EUROPEAN LANGUAGE RESOURCE INFRASTRUCTURE

FINAL REPORT

Project Name	European Language Resource Infrastructure
Funding Instrument	Connecting Europe Facility - Telecommunications Sector
Grant Agreement	INEA/CEF/ICT/A2016/1330962
Project website	www.elri-project.eu
Deliverable	D1.3
Activity	A1
Dissemination Level	Public
Version	1.3
Date	2019-09-30
Main author(s)	Thierry Etchegoyhen
Co-author(s)	Federico Gaspari (DCU), Jane Dunne (DCU), Helen McHugh (DCU), Paulo Vale (AMA), José Luis Fonseca (AMA), Patricia Fonseca (AMA), Maite Melero (SEAD), Antónion Branco (FCUL), Luís Gomes (FCUL), Rui Neto (LINKARE), Victoria Arranz (ELDA), Khalid Choukri (ELDA)
Status	Final version

Co-financed by the European Union
Connecting Europe Facility

ELRI Consortium

Legal Disclaimer

The information in this document is provided “as is”, and no guarantee or warranty is given that the information is fit for any particular purpose. The ELRI Consortium members shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials subject to any liability which is mandatory due to applicable law.

Table of Contents

Executive summary **7**

1 Introduction **8**

2 Project overview **10**

2.1 Consortium 10

2.2 Objectives and benefits 11

2.3 Activities 12

2.4 Timeline 13

3 Infrastructure overview **14**

3.1 National Relay Stations 15

3.1.1 Web application 15

3.1.2 Automated data processing 17

3.1.3 Component assembly 19

3.2 Beyond Member States 20

3.3 Software and deployment 20

4 Using a National Relay Station **21**

4.1 What counts as a language resource? 22

4.2 Who has access? 23

4.3 What can be downloaded? 23

4.4 What can be uploaded? 24

4.5 Who is sharing with whom? 24

5 Language resource collection **26**

5.1 Resource collection process 26

5.2 Collected resources 28

6 Dissemination **34**

6.1 Digital presence 34
6.1.1 Project website 34
6.1.2 Social media 35
6.2 Stakeholders conference 35
6.3 Workshops 36
6.3.1 Spain 36
6.3.2 Portugal 37
6.3.3 Ireland 39
6.3.4 France 40
6.4 Other events 41
7 Governance 42
7.1 Governance structure 42
7.2 Activities 43
7.2.1 IT Operation 43
7.2.2 NRS Adaptation 44
7.2.3 LR management 46
7.2.4 Promotion 47
7.3 Summary 48
8 Sustainability 48
8.1 Institutional sustainability 49
8.1.1 France 50
8.1.2 Ireland 50
8.1.3 Portugal 50
8.1.4 Spain 51
8.2 Technical sustainability 51
8.2.1 Stability 51
8.2.2 Maintenance 52
8.2.3 Resource management 52
8.3 Financial sustainability 53
8.3.1 Infrastructure costs 53
8.3.2 Personnel costs 53

8.3.3 Effort estimates 54
8.4 Summary 55
9 Frequently Asked Questions 56
10 Conclusions 58

Abbreviations and acronyms

Acronym	Definition
ELRI	European Language Resource Infrastructure
AT	Automated Translation
CEF	Connecting Europe Facility
DGT	Directorate-General for Translation
DSI	Digital Service Infrastructure
EC	European Commission
INEA	Innovation and Networks Executive Agency
LR	Language resource
NRS	National Relay Station
ELRC	European Language Resource Coordination

List of Figures

Figure 1 Project timeline and milestones 13
Figure 2 Overview of the ELRI network 14
Figure 3 National Relay Stations in Ireland, Spain, France and Portugal (clockwise from the top left-hand corner) 15
Figure 4 Processing different types of data 17
Figure 5 Communication between principal NRS components 19
Figure 6 ELRI process overview 21
Figure 7 Example of TMX content 23
Figure 8 Main steps of LR collection, preparation and sharing 26
Figure 9 Main validation steps 27
Figure 10 Number of registered institutions and active users 29
Figure 11 Number of published resources across Member States 29
Figure 12 Number of published translation units across Member States 30
Figure 13 Percentage of published resources shared beyond Member States 30
Figure 14 Number of resources transferred to ELRC-SHARE 31
Figure 15 Number of translation units transferred to ELRC-SHARE 31
Figure 16 Distribution of DSI domains for collected resources 32
Figure 17 Project website front page 34
Figure 18 Governance structure 42

List of Tables

Table 1 ELRI consortium 10
Table 2 List of activities 12
Table 3 Summary of NRS production releases 20
Table 4 List of Digital Service Infrastructures (DSI) 32
Table 5 ELRI workshop in Spain 36
Table 6 ELRI @ VIII Encontro de Tradutores da Administração Pública 37
Table 7 ELRI hands-on workshop in Portugal 38
Table 8 ELRI workshop in Ireland 39
Table 9 ELRI workshop in France 40
Table 10 NRS adaptation main tasks with relevant tools 45
Table 11 Main sustainability effort estimates 54

Executive summary

The European Language Resource Infrastructure (ELRI) project is an initiative funded within the Connecting Europe Facility (CEF) Programme, under Grant Agreement INEA/CEF/ICT/A2016/1330962, which started in October 2017 and ended in September 2019. Its main goal has been the development of an infrastructure to help collect, process and share language resources in the European Union and provide data relevant to the development of the Digital Service Infrastructures (DSI). The project involved seven partners representing four Member States: France, Ireland, Portugal and Spain.

One of the main accomplishments of ELRI 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 National Relay Station is available to members of public institutions in the Member State and its user interface is completely localised into 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 facilitate the collection of language resources from members of 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.

Additionally, ELRI offers flexible means to share language resources, thus taking into account the constraints that may be tied to sharing specific resources. In all sharing scenarios, ELRI provides the data holders, who dedicate time and effort to sharing their data, with the prepared resources as an immediate benefit. Thus, the project has developed a mechanism which aims to benefit all stakeholders equally, as a means to generate a community of interest and a positive dynamic around the collection and sharing of language resources.

The various dissemination events organised during the project have demonstrated the strong potential of this approach, with highly positive feedback and strong interest in joining the ELRI network, as users from public institutions or as representatives from new Member States willing to host their own National Relay Station. To support this newly created dynamic, one of the outcomes of the project is a detailed governance plan for new countries willing to join the network, and an infrastructure that has been optimised to minimise deployment costs.

The collection and preparation of resources within the project was initiated in 2019 and led to the publication of an initial set of resources in the independently deployed National Relay Stations. Overall, 71 institutions have registered to the network and contributed more than 800,000 translation units (TU) within the first months of activity. Although preliminary, and with different volumes collected depending on the country, the established community of users and strong dynamic is paving the way for continued and increased sharing of useful 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.

In the four Member States that participated in the project, ELRI is now seen as an important component to support digital advancement and language equality in the European Union, and the main outcomes of the project can thus be considered positive overall.

1 Introduction

The European Language Resource Infrastructure project (ELRI) is an initiative funded within the Connecting Europe Facility (CEF) Programme[1], under Grant Agreement INEA/CEF/ICT/A2016/1330962, which started in October 2017 and ended in September 2019. Its main goal has been the development of an infrastructure to help collect, process and share language resources (LR) in the European Union and provide data relevant to the development of the Digital Service Infrastructures (DSI)[2]. The project involved seven partners representing four Member States (MS): France, Ireland, Portugal and Spain.

Language resources encompass all linguistic data that are relevant to provide linguistic services in a given language or group of languages, such as texts translated in multiple languages or translation memories, and are of critical importance for both human translation professionals and for machine translation systems, whose development is dependent on important quantities of language resources. Supporting language equality in the European Union and an efficient multilingual European infrastructure is thus highly dependent on the availability of quality resources for the languages of the Union. Services such as eTranslation,[3] an automated translation service provided by the Directorate-General for Translation (DGT) to Public Administrations, can for instance greatly benefit from language resources produced on a daily basis across the European Union.

The ELRI initiative sought to enhance the collection of high-quality language resources, by mitigating obstacles identified during the data collection efforts of companion initiatives such as the European language Resource Coordination project (ELRC)[4]. One of the identified difficulties was the reluctance of data holders to make their data available due to perceived concerns related to IPR issues, Member State regulations, and the lack of internal expertise or manpower, especially within public institutions, to properly take the steps needed to provide appropriately prepared language 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.

One of the main accomplishments of ELRI 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 Member State and its user interface is completely localised into 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 manual reviewing and processing of file collections whilst also providing stakeholders with fully prepared resources. The integrated processing notably allows the creation of translation memories from raw user data in the form of document collections in multiple languages.

Another feature of ELRI services is the provision of a group-based policy for sharing, where users can select the group(s) with which they aim to share their resources. Default groups are provided to cover sharing at the national, European or Open Data levels, and ad-hoc groups can be

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

[2] https://ec.europa.eu/digital-single-market/en/news/connecting-europe-facility-cef-digital-service-infrastructures

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

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

created to handle specific cases, such as data sharing restricted to a given institution of the Member State and the DGT. This flexible sharing policy provides the means for different ways of sharing resources, acknowledging that different sets of constraints may be tied to specific resources.

ELRI provides both immediate and longer-term benefits, which is another important aspect of the initiative. Data holders who have joined the network can benefit in the short-term from the fully processed LRs that are generated automatically by the system, once they have been validated and published by the dedicated ELRI team in the corresponding Member State. This feature of the network, which acts as an LR processing centre for the Member States and beyond, has an important impact in terms of providing a short-term return that can benefit stakeholders for their own translation services. In the mid-term, LRs shared with the DGT for the improvement of eTranslation services also benefit stakeholders via improved automated translation services that are available not only to EU officials, but also to Public Administrations in the Member States. In addition to this, the availability of high-quality LRs is also beneficial to professional human translation, in particular translation memories that facilitate the timely production of professional translations.

Although the ELRI Action was bound by its start and end dates, one of the main goals of the initiative was to provide a sustainable infrastructure that could be maintained and used after the lifetime of the project itself, in the countries where the infrastructure is deployed. Such a sustainable solution can help support the continued collection of language resources that are necessary to further improve European translation services, be they automated or performed by human translators.

2 Project overview

2.1 Consortium

The ELRI Consortium consists of 7 participants from four different countries, with the required background and expertise to achieve the objectives of the project. The partners of the project are listed in Table 1.

Name	Short name	Country
Fundación Centro de Tecnologías de Interacción Visual y Comunicaciones Vicomtech[5]	VICOM	Spain
Linkare TI – Tecnologias de Informação, Lda.[6]	LINKARE	Portugal
Administrative Modernization Agency[7]	AMA	Portugal
Evaluations and Language Resources Distribution Agency S.A.S[8]	ELDA	Portugal
Faculdade de Ciências da Universidade de Lisboa[9]	FCUL	Portugal
Kingdom of Spain, represented by the Secretary of State for Digital Advancement[10]	SEAD	Spain
Dublin City University School of Computing[11]	DCU	Ireland

Table 1 ELRI consortium

Vicomtech, a research centre located in Spain, was the coordinator of the Action and provided its experience in project coordination as well as its expertise in the development of efficient language technology solutions.

Linkare TI is a software company located in Portugal, who provided their technical expertise in software development and their strong experience in the development of successful CEF solutions.

The Administrative Modernization Agency is the public Body in charge of digital advancement and administrative modernisation for Portugal. AMA provided the institutional support needed for the development of the Action in Portugal and is in charge of managing the National Relay Station in this Member State.

ELDA is a company located in France with strong experience in the collection, preparation and distribution of language resources in Europe, who plays a central role in European initiatives dedicated to the collection and sharing of language resources. ELDA is in charge of managing the National Relay Station for France.

The Faculty of Sciences at the University of Lisbon is a research institution who provided their strong technical expertise in the development of quality language technology and their

[5] www.vicomtech.org
[6] http://www.linkare.com
[7] https://www.ama.gov.pt
[8] http://www.elra.info/en/about/elda/
[9] https://www.ulisboa.pt/en/unidade-organica/faculty-sciences
[10] https://avancedigital.gob.es/en-us/Paginas/index.aspx
[11] https://www.dcu.ie/

experience in the successful development of European projects, including those concerned with infrastructures for language resources.

The Secretary of State for Digital Advancement is the public body in charge of digital advancement in Spain, overseeing the national plan for language technology in this Member State. SEAD is in charge of managing the National Relay Station for Spain.

Finally, Dublin City University, as a leading research centre in Europe, provided their extensive technical expertise in language technologies, as well as their strong experience in conducting successful European projects and work on the promotion of the Irish language. DCU is in charge of managing the National Relay Station for Ireland.

2.2 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.
- Automate the creation of translation memories and other resources from raw data provided by public institutions and translation centres.
- Provide National Relay Stations as flexible means for sharing resources, at the national, European and Open Data levels.
- Prioritise resources that are relevant to Digital Service Infrastructures (DSI).[12]

  Contribute to improve the EU automated translation services that are freely available to all 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 sustainable infrastructure.

These objectives were aligned with the identified challenges regarding the collection of quality language resources, by aiming to develop a sustainable solution that is 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.

Achieving the above objectives aimed to provide the following benefits:

- The provision of flexible means of sharing resources establishes a clear process where compliance with the relevant sharing 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, the Public Sector Information Directive and DSI-specific needs.
- Language resources can be shared as deemed appropriate by stakeholders, with return benefits for providers as well as users of translation services.

[12] https://ec.europa.eu/digital-single-market/en/news/connecting-europe-facility-cef-digital-service-infrastructures

- 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 respective Member States.

This set of expected benefits are at the core of the ELRI project and the development of the infrastructure was designed to support their achievement.

2.3 Activities

To achieve its stated objectives, ELRI was organised into six main activities, listed in Table 2.

ID	Activity Title
A1	Project Coordination
A2	Analysis, Requirement gathering and preparation
A3	Deployment of Machine Translation (MT) -related data network
A4	Identification and Consolidation of the network
A5	Exploitation & Governance
A6	Dissemination

Table 2 List of activities

The goals and scope of these activities can be briefly summarised as follows:

- **A1** centred on project coordination activities;
- **A2** focused on the design and specification of the network;
- **A3** and **A4** focused on the development of the network, by first providing the necessary infrastructure to collect, process and share LRs (**A3**) and then processing the data collected from the participants (**A4**);
- **A5** centred on establishing the adequate governance and sustainability plans for a successful adoption and continuation of the ELRI infrastructure;
- **A6** included all dissemination activities related to the project.

This structure responded to the needs of the project and assured an efficient coordination of the work and an adequate distribution and organisation of the Consortium expertise. The planned activities also aimed to provide the necessary means for a successful deployment and adoption of the resource sharing infrastructure by:

- dedicating specific activities to the technical integration of existing tools and data into the appropriate processing pipelines, and
- facilitating the integration of data holders within and beyond the scope of the action.

Overall, the project activities supported the three main lines of development listed below:

- Project coordination and monitoring (**A1**)
- Technical design, development and data processing (**A2-4)**
- Project dissemination and exploitation (**A5-6**)

The tasks of the project allowed for separate specifications, development and validation phases, for each activity, ensuring the definition of attainable goals and proper means of validation throughout the life of the project.

2.4 Timeline

The timeline of the project and its main milestones are described in Figure 1.

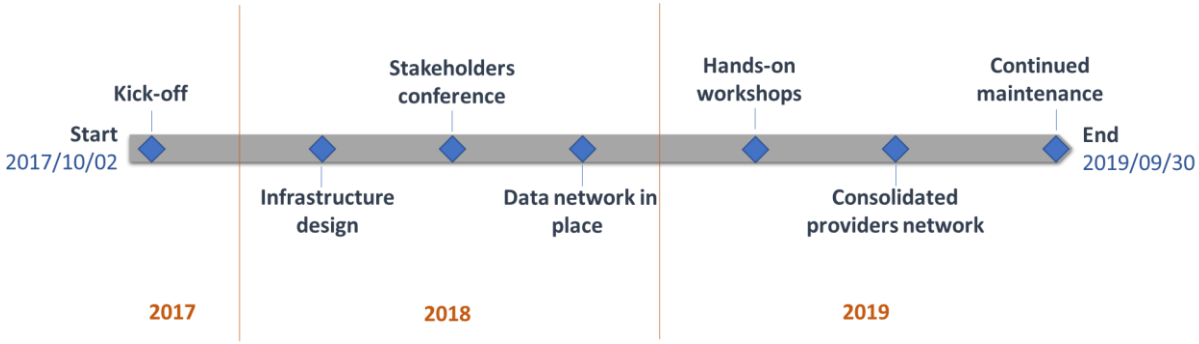


Figure 1 Project timeline and milestones

Overall, this timeline corresponded to three main stages of project development. The first phase, culminating in the Infrastructure design milestone, involved the design of the infrastructure including the specification of system requirements, software architecture and legal and administrative requirements.

The second phase involved the development of the infrastructure itself, from the automated data processing engines to the complete software underlying the National Relay Station application, resulting in a network in place for its use in production.

The third phase, set in 2019, involved active resource collection via the integration of stakeholders as users of the network. These activities were to be supported by the organisation of events promoting the initiative, such as the Stakeholders Conference and dedicated hands-on workshops in each of the four Member States participating in the project.

At the end of the project, the infrastructure entered continued maintenance, with all four National Relay Stations maintained in their corresponding Member States and continued resource collection activities as a key objective.

3 Infrastructure overview

ELRI is a decentralised network composed of National Relay Stations, i.e. the web applications designed to help collect, prepare and share language resources. Figure 2 illustrates the current deployed infrastructure.

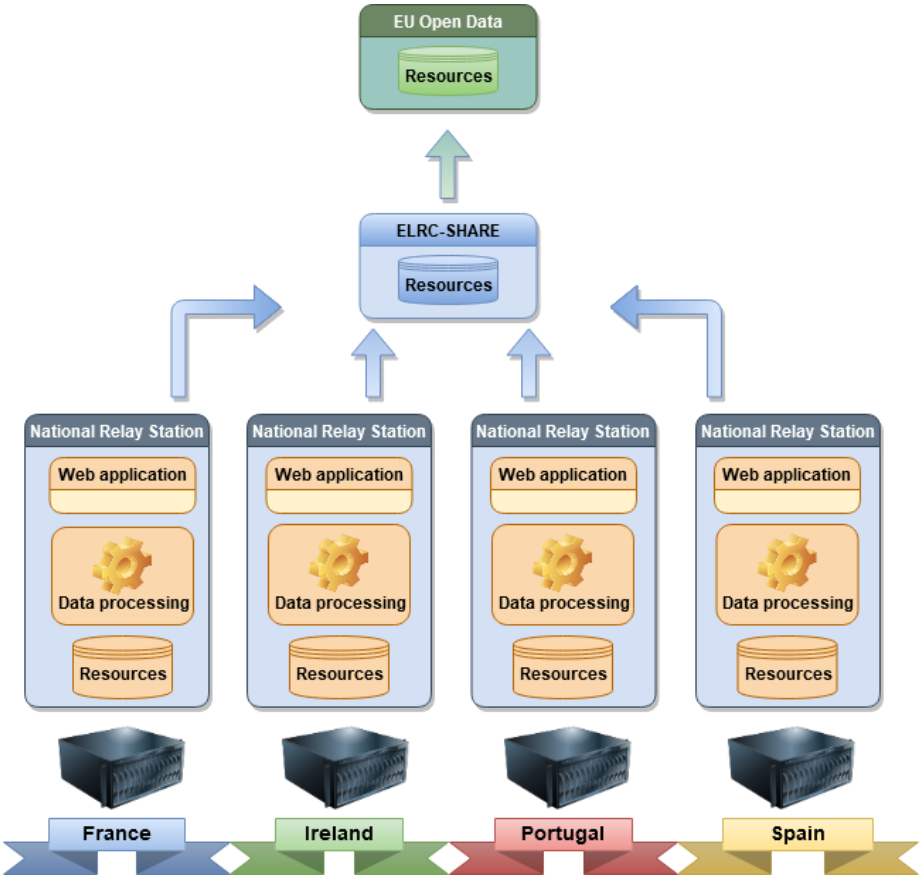


Figure 2 Overview of the ELRI network

Each Member State thus deploys, on a specific server, an instance of a National Relay Station, which comprises a Web application, data processing engines and a database of language resources. New Member States or EEA countries could join the network by preparing their own NRS to enable the collection of resources in their language(s).

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 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 ELRI personnel in each Member State.

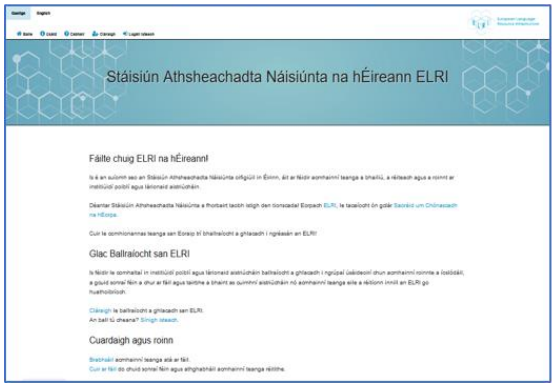
Prepared resources that are deemed valid are then published in the NRS of the Member State, thus becoming directly available to the users who contributed them, as well as to the other users of the groups with which the data contributors were willing to share the resources.

Resources that are shared with the European Commission are then transferred to the ELRC-SHARE repository[13]. Additionally, resources that have been shared as Open Data are deposited to the EU Open Data Portal[14], via links to ELRC-SHARE.

3.1 National Relay Stations

This section provides more details on the components comprising a National Relay Station. Each NRS is accessible via a Web site, available in the language(s) of the Member State where it is deployed. Figure 3 shows the four National Relay Stations deployed during the project, in France, Ireland, Portugal and Spain.[15]

https://elri.dcu.ie



https://elri.plantl.gob.es

https://etraducao.gov.pt

https://elri.elda.org

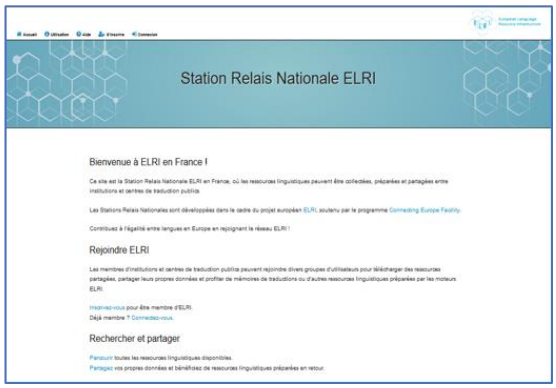


Figure 3 National Relay Stations in Ireland, Spain, France and Portugal (clockwise from the top left-hand corner)

An NRS consists of two main components: a Web application and an automated data processing engine; we provide an overview of each component and their interaction in the following sections.

3.1.1 Web application

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

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

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

[15] The NRS for Ireland is available in two languages, Irish and English; Figure 1 shows the Irish version.

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[16], itself based on the META-SHARE[17] software. The core functionality of the web application is indicated below:

- 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
- Email communication with users of the service

Even though modifications have been made to the look-and-feel of the original 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. 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. Localisation was performed in full for the front-office, i.e. all elements of the Web interface that are accessible to end-users; the back-office was also localised, although in this case aspects that required technical knowledge for proper translation may have been left in the original English version. 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 functionality 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 done via 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 to by default:

- *NationalOrganisations*: This group includes all registered users of the NRS and resources shared with this group will thus be accessible to all registered users of the NRS.

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

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

- *NationalOrganisations+EuropeanCommission*: This group includes all registered users of the NRS and the European Commission, who may then utilise the shared resources to improve the automatic translation services provided freely to public administrations of the EU Member States. Resources shared with this group will thus be accessible to all registered users of the NRS and will also be shared with the European Commission by being placed on the ELRC-SHARE repository.
- *OpenData*: This group includes all registered users of the NRS and all users of the free Open Data portal of the European Union. Resources shared with this group will thus be placed in the Open Data repository in addition to the NRS.

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 dedicated ELRI team in the relevant Member State.

3.1.2 Automated data processing

As previously indicated, each National Relay Station includes data processing engines which can handle different types of content and file formats. Figure 4 describes the main processing steps for the four major types of data handled by the engines.

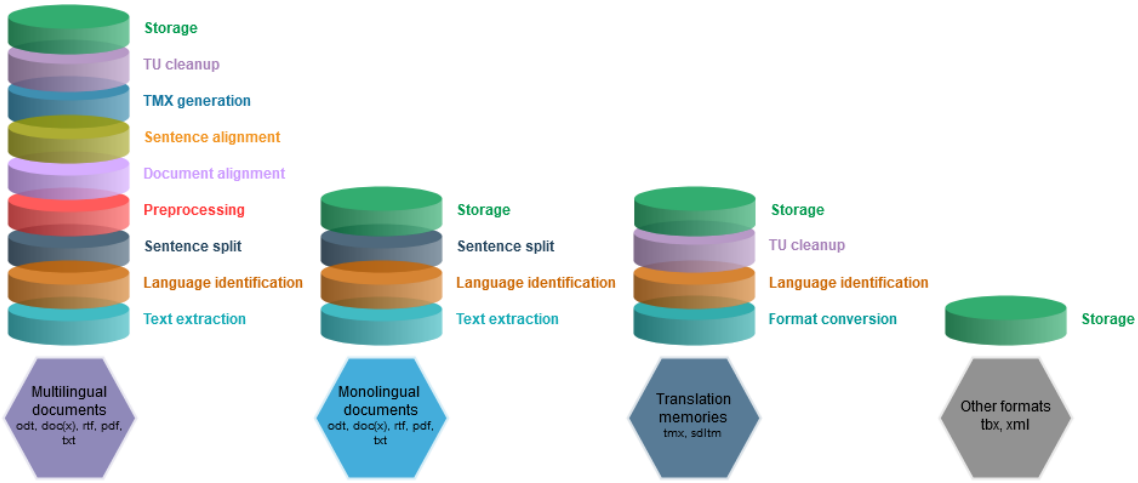


Figure 4 Processing different types of data

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 handled by the engines, and its main steps are summarised below:

- 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.
- 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, i.e. the process which mainly separates words and punctuation, and truecasing, i.e. a process which assigns the most frequent capitalisation to sentence-initial words (which are usually capitalised by default).

- All document pairs with content in different languages are then automatically aligned, i.e. a score is assigned to each pair of documents depending on the similarity of their content.
- 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.
- From the aligned sentences a translation memory in TMX format 1.4b is then generated, 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 applies, 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 useful 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 from the left involves existing translation memories as input. In this case, the first step is to convert the format, 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. After this, the translation memory undergoes the same clean-up operations described in the first main case, 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 though, 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.[18]

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.[19] The ability of the NRS software to ingest raw data in multiple file formats and generate quality structured

[18] The second toolchain shares the initial processing steps in multilingual and monolingual scenarios, as can be seen in Figure 3; despite its name, the output of this toolchain for monolingual input data is a text file, not a TMX.

[19] See Azpeitia, A. & Etchegoyhen, T. *Efficient Document Alignment Across Scenarios*. Machine Translation (2019). https://doi.org/10.1007/s10590-019-09234-9

resources in an automated manner is one of the key features of the implemented ELRI infrastructure.

3.1.3 Component assembly

The communication between the principal components of a National Relay Stations is illustrated in Figure 5.

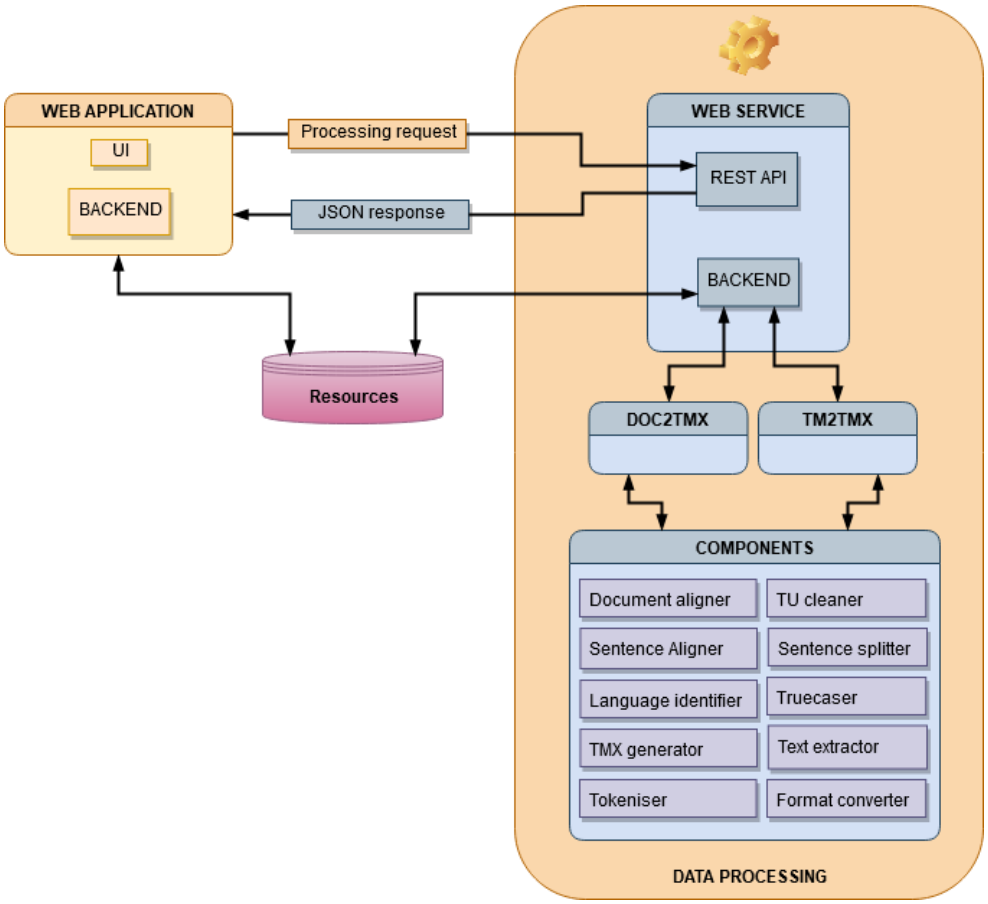


Figure 5 Communication between principal NRS components

Although National Relay Stations are hosted on a single server in the four environments where they have been deployed so far, the components could reside in different servers, as communication is mediated via a web service which wraps the data processing component. Requests are sent by the Web application to this web service via a REST API, with responses provided as JSON objects. The initial user data as well as those generated by the data processing engines are stored in a shared storage repository, accessible to both the Web application and the data processing engines.

The main components of the NRS software listed below are provided as Docker[20] containers, assembled via docker-compose[21]:

- web application: user and data management
- nginx: web server/load balancer/reverse proxy

[20] https://www.docker.com/

[21] https://docs.docker.com/compose/

- solr: search server
- postgres-server: relational database
- toolchain: data processing

Further documentation on the ELRI Docker images is available at the following address: https://github.com/ELDAELRA/ELRI/tree/master/docker

3.2 Beyond Member States

Resources that have been shared with a group that extends beyond the NRS need to be ported to the aforementioned ELRC-SHARE repository. Two methods are available for this data transfer:

A. Manual upload: ELRI personnel registered on ELRC-SHARE upload the relevant resources via the ELRC-SHARE GUI, manually ingest and then publish the resources.

B. API transfer: ELRI personnel use the *Upload resource to ELRC-SHARE* action added to the NRS user interface, which transfers resource data via the ELRC-SHARE API; transferred resources then need to be manually published on the ELRC-SHARE repository.

The first method has been used to transfer all the resources from the different National Relay Stations at the moment. The second method has been implemented in the NRS codebase as an additional means to transfer the core of a resource, its data and description, from a National Relay Station to ELRC-SHARE.

3.3 Software and deployment

The NRS software is available in two different repositories:

- The Web application code and Docker files are available in the following Github repository: https://github.com/ELDAELRA/ELRI/
- The Docker images of the different tagged releases are available in the following repository: https://hub.docker.com/

The NRS software was developed following standard procedures, with different versions tested iteratively. Releases that complied with all mandatory testing requirements were deployed in production. Table 3 summarises the information on deployment of the software in production, in all four Member States involved in the project.[22]

Host	Country	URL
AMA	Portugal	https://etraducao.gov.pt
ADAPT	Ireland	https://elri.dcu.ie
ELDA	France	https://elri.elda.org
MINCOTUR	Spain	https://elri.plantl.gob.es

Table 3 Summary of NRS production releases

[22] *Host* indicates the institution whose IT department is responsible for the deployment and maintenance of the software service.

4 Using a National Relay Station

Although the infrastructure features a certain degree of complexity, as described in Section 3, the interaction with a National Relay Station is meant to be as intuitive as possible for its users. Figure 6 illustrates the overall process when using ELRI services.

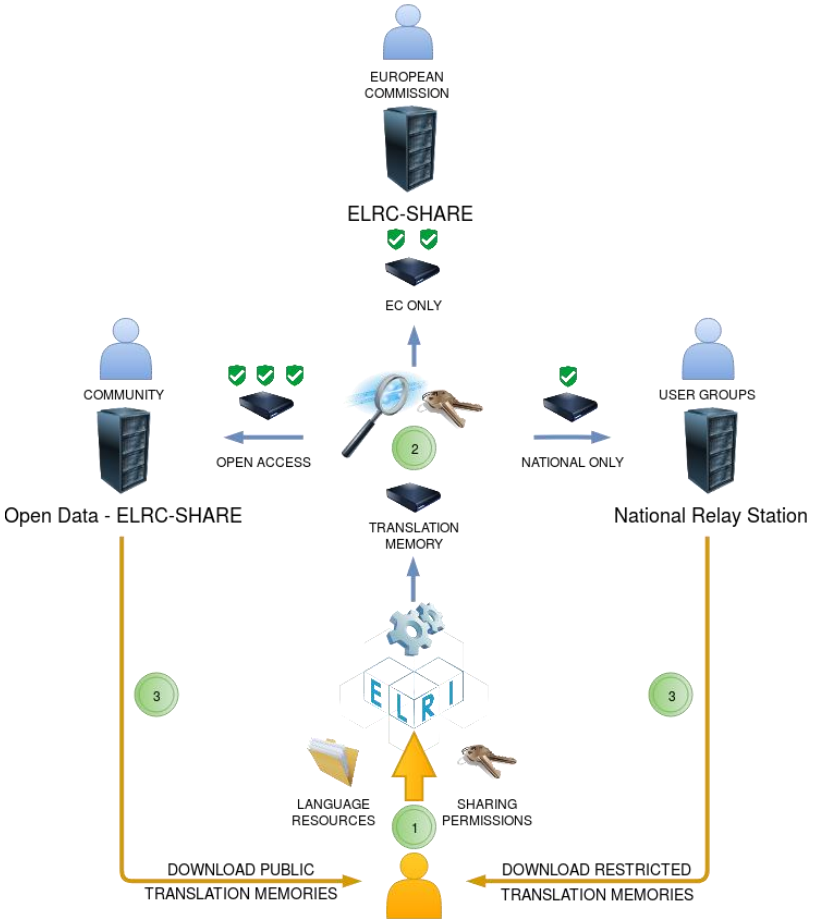


Figure 6 ELRI process overview

The process starts with registered users of an NRS uploading their data, in the form of documents which may be in a number of different formats, and select the sharing permissions they deem appropriate for this specific resource by selecting one of the available sharing groups previously described. Once submitted, their data are automatically processed by the integrated ELRI engines, resulting in a prepared resource, such as a translation memory.

Depending on the specified sharing scope, and after it has been validated by ELRI personnel, the prepared resource may then be available in different repositories. All prepared resources are thus available on the NRS itself, where all users of the selected group can directly download them. Those resources that have been shared beyond the Member State are also available on the ELRC-SHARE repository, where they are accessible to the European Commission and available as resources to further improve the automated translation services developed by the DGT. Finally, resources shared as Open Data will be available on the EU Open Data Portal, where they will be accessible without restrictions.

Although relatively straightforward, this process refers to notions that may not be self-evident to users at first. To alleviate these difficulties, each NRS provides user guidelines that detail the

different steps of the process, along with the main concepts around language resources.[23] We provide a brief summary of these concepts in the next sections.

4.1 What counts as a language resource?

The term *language resource* encompasses all data that is relevant to provide linguistic services in a given language or group of languages. This includes the following main types of resources:

- Corpus: this is the main type of LR and is essentially a prepared set of documents with text in one or more language(s). A corpus can be *monolingual*, i.e. containing text in one language only in a single file, or *bilingual*, i.e. containing translated texts in two languages and in separate files. Corpora can be created from collections of documents in different formats (e.g., MS Word or RTF), by extracting and processing the content of the documents.
- Translation memory: one of the most useful resources for both human translators and machine translation services is the translation memory. The translation memory is a structured document with the standard TMX file extension, where translations between languages are paired in terms of translation units (usually sentences aligned in pairs, i.e. a source/original and its translated equivalent in the target language). An example of TMX content is provided in Figure 7.
- Terminology exchange files: terms which correspond to each other in different languages can also be stored in a structured way, typically in TBX files that are similar to the TBX files described above but contain paired equivalent terms instead of paired sentences.

Other types of language resources exist, such as grammatical descriptions of a language or thesauri. However, for practical purposes, the three types of LRs described above are the most valuable to improve translation services.

Translation memories and terminology files, in particular, are extremely important resources for translation services: human translators rely on them to improve their productivity, by being able to retrieve and reuse previously translated material, whereas modern machine translation systems learn to translate from the human translations stored in bilingual corpora and translation memories.

[23] These guidelines are available in the languages of the four Member States of the Action and are also provided at the following address: http://www.elri-project.eu/resources.html

```
[...]
  <tu tuid="1" creationdate="2018-09-07T12:01:58Z" creationid="ELRI">
   <tuv xml:lang="en">
    <seg>This is an example.</seg>
   </tuv>
   <tuv xml:lang="es">
    <seg>Esto es un ejemplo.</seg>
   </tuv>
  </tu>
  <tu tuid="2" creationdate="2018-09-07T12:01:58Z" creationid="ELRI">
   <tuv xml:lang="en">
    <seg>A translation memory contains translation units.</seg>
   </tuv>
   <tuv xml:lang="es">
    <seg>Una memoria de traducción contiene unidades de traducción.</seg>
   </tuv>
  </tu>
[...]
```

Figure 7 Example of TMX content

4.2 Who has access?

A National Relay Station is freely accessible to all members of a public institution of the Member State, including entities such as public administrations, universities or public translation centres. Members of public institutions can register on the NRS website of their Member State and will be provided access by ELRI personnel of the NRS.

Resources shared at the NRS level are thus only available to registered users from public institutions. However, resources shared as Open Data are made available to all citizens via the EU Open Data Portal.

4.3 What can be downloaded?

The resources that are shared come in a prepared format, automatically created by the ELRI engines from the data uploaded by users. In other words, the original data contributed by users (e.g., collection of documents in multiple languages) is not shared "as is". Instead, the original data undergoes a series of processing steps that lead to the creation of structured language resources, as described in Section 3.1.2, which are then shared according to the permissions specified by the users who uploaded the original data.

These structured LRs come in the following main formats:

- Translation memories: files in standard TMX format, where translations are paired at the sentence level, following an alignment process (as shown in Figure 7).

- Monolingual files: when a document in a given language does not have a translation counterpart in another language, the ELRI engines still produce a processed file from the original file, with the file content split into separate sentences, one sentence per line.
- Terminology database files: files in standard TBX format, where equivalent terms in different languages are paired.

All downloadable resources are available in a compressed file (*archive.zip*), which contains one or more files in the format(s) specified above, along with license information that specifies the conditions of use for the downloaded resource.

4.4 What can be uploaded?

As previously indicated, the NRS integrates the ELRI data processing engines. The function of these engines is to create structured language resources from raw data provided by users in different languages and in a large number of formats, including MS Word doc(x), ODT or RTF.[24]

Users can thus simply upload different documents in one or more languages and let the ELRI engines prepare structured resources in the background. After the resources have been prepared and validated by the ELRI personnel in charge of the NRS, they will then be published on the NRS and become visible according to the sharing restrictions specified by the user who contributed the resources.

The main scenarios regarding data uploading are the following:

- *Upload a collection of documents that contain the same content translated in different languages (e.g. Irish/English, Portuguese/French, Spanish/English, etc.)* The ELRI engines will then produce translation memories (TMX files) from the translated content by automatically pairing sentences from the original documents.
- *Upload existing translation memories*. In practice, translation memories can contain errors, such as translation units in the wrong language or misaligned sentences. In that case, the ELRI engines provide the functionality to clean up existing translation memories automatically.
- *Upload terminology database files or other resources*. Users can also provide files that do not require automatic processing. These files will also be examined and validated by the ELRI team in DCU prior to publishing and sharing the resources.

4.5 Who is sharing with whom?

ELRI aims to provide flexible means to share data, acknowledging that sharing restrictions may need to vary for specific resources.

Sharing via the NRS is done via **groups**, where users can browse and download only those resources that are shared with a group that they belong to. Each NRS offers the three default groups described in Section 3.1.1, repeated here for convenience:

- **NationalOrganisations:** This group includes all registered users of the NRS, and resources shared with this group will thus be accessible to all registered users of the NRS.

[24] The complete list of accepted file formats is provided in the *Contribute* page of the NRS.

- **NationalOrganisations+European Commission:** This group includes all registered users of the NRS and the European Commission, who may then utilise the shared resources to improve the automatic translation services provided freely to public administrations of the EU Member States. Resources shared with this group will thus be accessible to all registered users of the NRS and will also be shared with the European Commission by being placed on the ELRC-SHARE repository.
- **OpenData**: This group includes all registered users of the NRS and all users of the free Open Data portal of the European Union. Resources shared with this group will thus be placed in the Open Data repository in addition to the NRS.

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 dedicated ELRI team in their Member State.

As a simple example, users from a given public organisation (e.g. *OrganisationX*) may need to restrict the sharing of a resource to members of *OrganisationX* and the European Commission only, to specifically help improve the machine translation services of the European Union, without further sharing because of specific distribution restrictions.

This may be achieved by requesting the creation of a specific group joining the two entities, e.g. *OrganisationX+EuropeanCommission*. Users from *OrganisationX* would then be able to contribute their resource and select this specific group to match their desired level of sharing.

5 Language resource collection

The active collection and preparation of resources was initiated in 2019, as part of Activity 4, and led to the publication of an initial batch of resources in the different deployed National Relay Stations. It is worth noting that the ELRI infrastructure is a sustainable solution that aligns with the respective missions of the institutions in charge of the maintenance of the ELRI services. The number of collected and shared language resources is thus planned to increase after completion of the CEF Action, via the sustained activity of the ELRI network.

In the following sections, we first summarise the resource collection process and indicate the main steps that lead to prepared and published resources. We then present the main statistics for the collected resources as of September 2019, towards the end of the CEF Action.

5.1 Resource collection process

Language resources were collected via the four National Relay Stations deployed in France, Ireland, Portugal and Spain, and described in Section 3.1. The complete process for LR collection, preparation and sharing is summarised in Figure 8.

Figure 8 Main steps of LR collection, preparation and sharing

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 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. This involves dedicated work, performed by ELRI personnel on the basis of strict guidelines which involve the main steps described in Figure 9. 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.

Figure 9 Main validation steps

The initial review is meant 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.

The quality review involves manual examination of samples of the processed data, examining for instance the quality of the translation units in the case of translation memories generated by the automated language processing engines. Poor alignment quality would result in the resource not being validated and the user being notified of the issue. This may happen for instance when processing files in PDF format, for which the extraction of textual content is typically more difficult, resulting in the alignment of fragments of sentences that do not properly reflect translation content.

The third main step 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[25], the ELRI validation process involves a specific step to help determine if the contributed data may nonetheless include such data. For this

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

purpose, a specific tool, the ELRI Data Checker (EDC), was developed within the project. The tool processes the data under validation and generates a report on the following main aspects:

- Cases where patterns of sensitive data have been detected. These include patterns for national identification numbers, passport numbers, words and phrases indicating confidential material or typical formulations related to personal information, etc. The patterns are easily configurable by means of text configuration files and can be extended at will by the reviewing team.
- A detection and classification of named entities, i.e. names of individuals, organisations, dates or places, among others.

The retrieved cases are reported in the textual context where they were detected, to help reviewers assess the automatically recognised patterns with additional surrounding information.

The tool is meant only as an aid to detect personal, confidential or sensitive information, and no guarantee is given that it would capture such information in all cases. It is also not expected from ELRI personnel that they guarantee the complete absence of personal, confidential or sensitive after the examination of the EDC reports. However, the tool may help capture the presence of potential personal, confidential or sensitive information, in which case the validation process would be placed on hold until the matters are resolved and abandoned if no resolution is reached.

The next 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[26] licenses. ELRI 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 by ELRI personnel in charge, to properly 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.

5.2 Collected resources

As of September 2019, the National Relay Stations have registered a number of active users from different institutions of the Member States where they are deployed. The number of institutions and authorised users is shown in Figure 10. With 71 participating institutions and 101 authorised users at the time of this writing, the National Relay Stations can be considered to have attracted significant interest among public institutions of the Member States participating in the Action.

[26] https://creativecommons.org/

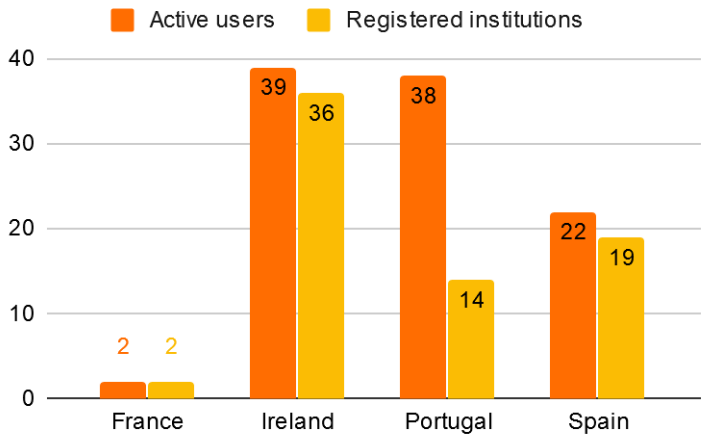


Figure 10 Number of registered institutions and active users

The registered users have contributed a number of initial resources, several of which have been fully validated and published on the corresponding NRS. Figure 11 shows the published resources in the four National Relay Stations as of September 2019.

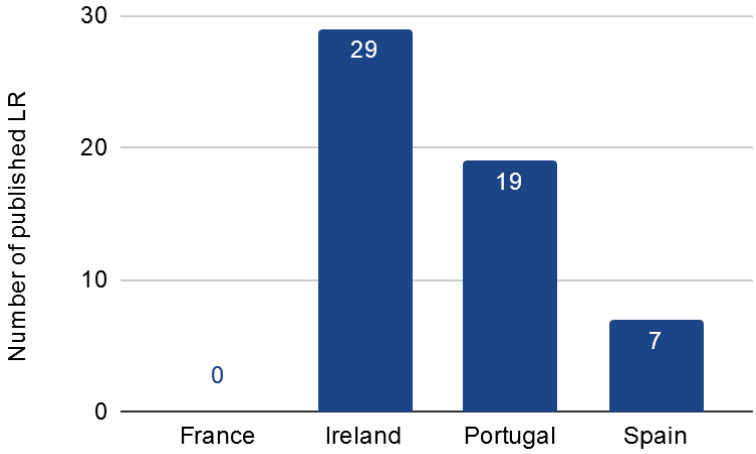


Figure 11 Number of published resources across Member States

The collection of resources has thus been initiated, with first sets of resources in all but one Member State. Regarding the French NRS, it should be pointed out that, despite the fact that no data has been contributed yet, several discussions are currently ongoing with institutions willing to participate and share data. Meetings are already planned for that purpose in the first weeks after completion of the project itself in September 2019 and initial resources are expected at that stage. 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, will be a building block 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 12 illustrates the number of translation units for published resources.

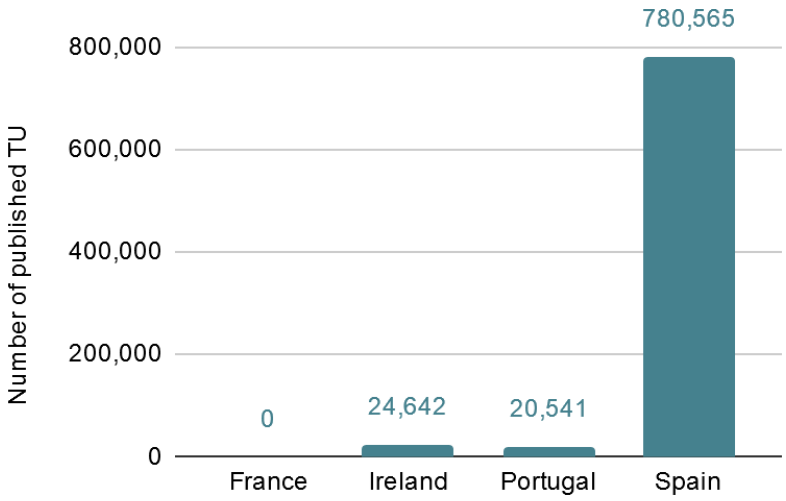


Figure 12 Number of published translation units across Member States

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, resources may be shared beyond the national level. Figure 13 indicates the percentage of resources shared with the European Commission or as Open Data.

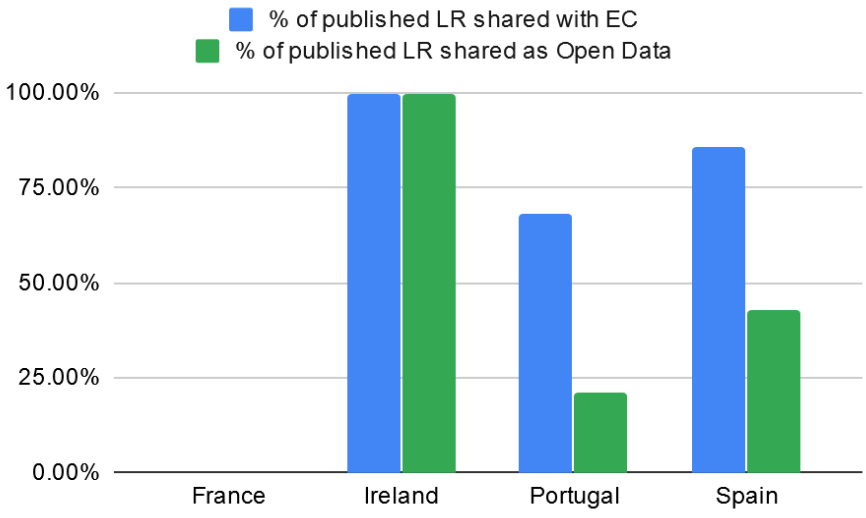


Figure 13 Percentage of published resources shared beyond Member States

As shown by these figures, the National Relay Stations play an important transmission role, with a large majority of the data being shared beyond the Member States. 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 14 and Figure 15 indicate the number of published resources and translation units, respectively, which have been transferred to ELRC-SHARE as of September 2019.

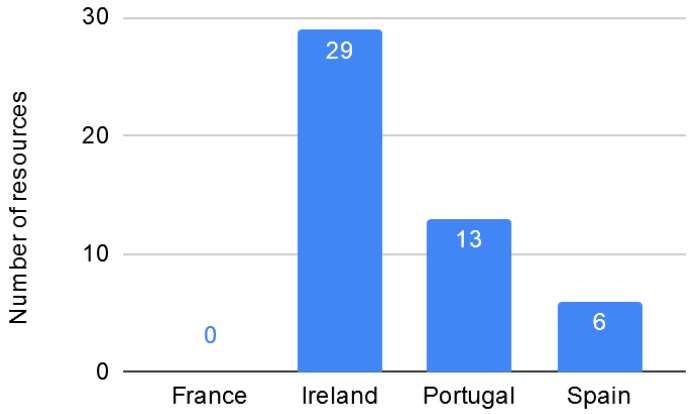


Figure 14 Number of resources transferred to ELRC-SHARE

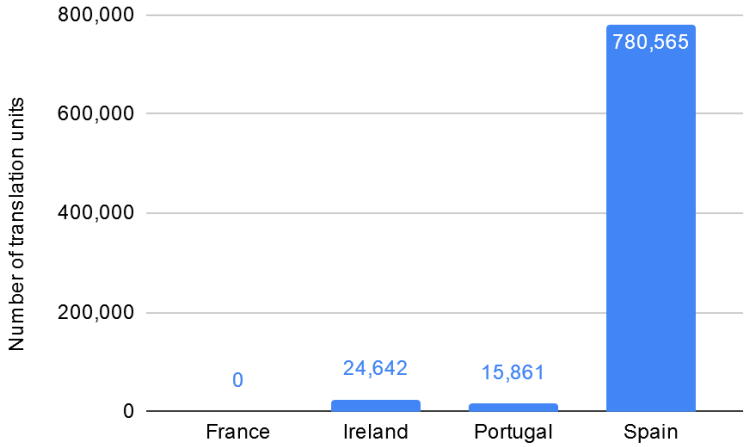


Figure 15 Number of translation units transferred to ELRC-SHARE

Overall, 48 resources, amounting to 816,553 translation units, have been transferred in the initial resource collection phase, which represents a satisfactory initial batch of resources for a resource collection network that is entering continued maintenance.

Finally, an important piece of information relates to content of relevance for Digital Service Infrastructures (DSI), as the collected resources may provide content associated directly to one of the DSI domains indicated in Table 4 below.

Digital Service Infrastructure	Short name
Electronic Exchange of Social Security Information	EESSI
eHealth	eHealth
eJustice	eJustice
eProcurement	eProcurement
Public Open Data	OpenData
Safer Internet	SI
Business Registers Interconnection System	BRIS
eCulture	eCulture
Cybersecurity	Cybersecurity
Unknown/Other	Other

Table 4 List of Digital Service Infrastructures (DSI)

The distribution of DSI-specific content for the resources transferred to ELRC-SHARE is shown in Figure 16 Distribution of DSI domains for collected resources.

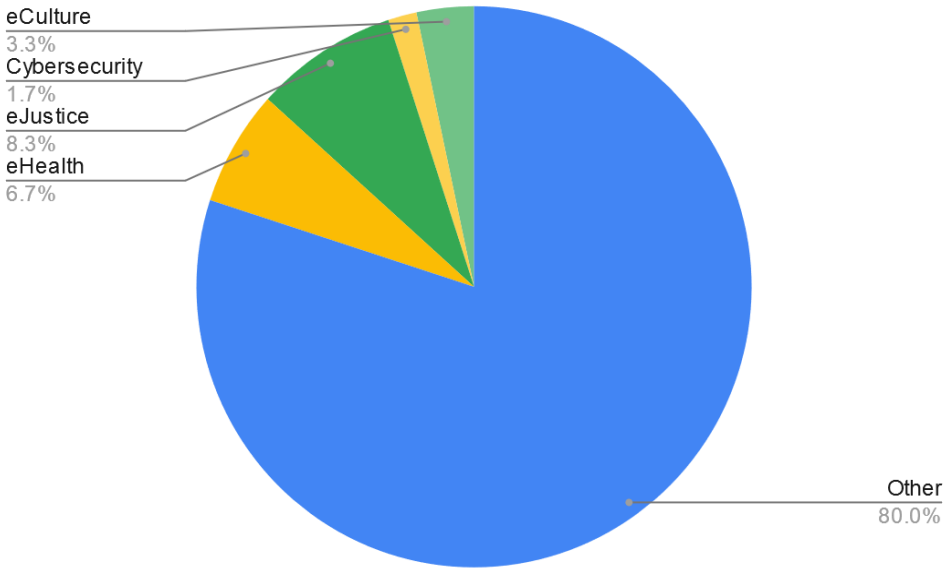


Figure 16 Distribution of DSI domains for collected resources

Although the majority of resources are not specified as relating to a particular DSI domain, several of the contributed resources do provide content of relevance for Digital Service Infrastructures. As the number of contributed resources increases via the sustained ELRI services, it is expected that more DSI-specific content will be made available for both human translators and automated translation services.

As the ELRI infrastructure is a sustainable solution that will be maintained beyond the lifetime of the project, the number of collected and shared language resources is expected to increase after completion of the CEF Action. The processes in place will guarantee similar standards of quality for the collection of language resources within the continued ELRI services.

As of this writing, additional collection of language resources is already planned for the weeks following the completion of the project, with a significant number of potential contributors having expressed a strong interest in contributing their resources via the National Relay Stations. The community of ELRI users that has been put in place in the different Member States is one of the important contributions of the project, with strong potential to significantly increase the amounts of quality resources collected via the National Relay Stations.

6 Dissemination

A number of dissemination activities have taken place during the project, via its digital presence and a number of events summarised in the next sections.

6.1 Digital presence

Part of the dissemination activities were performed via the digital presence of the project, which was mainly carried out via the project website and social media activities.

6.1.1 Project website

Project information is available on the project website, accessible at the following address: www.elri-project.eu. The website was launched in November 2017 and was updated regularly as the project developed. Its main page is shown in Figure 17.

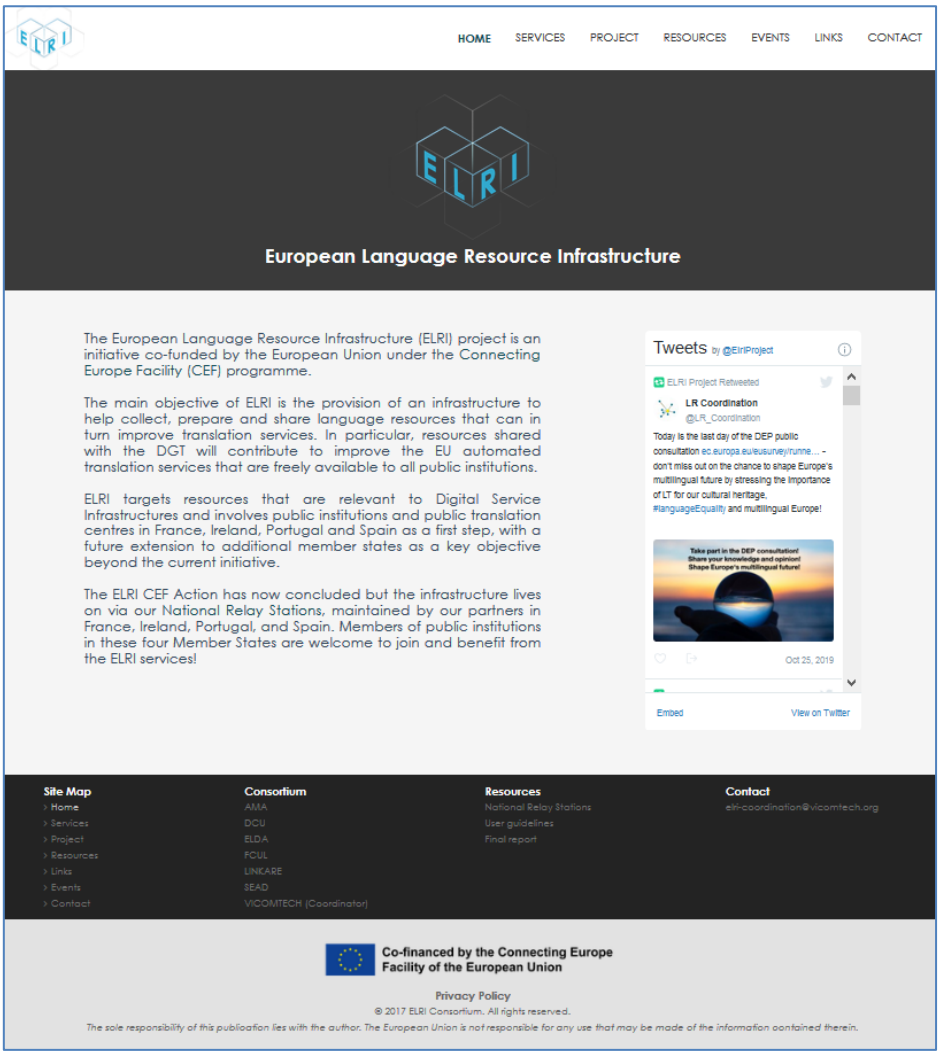


Figure 17 Project website front page

In addition to a view of the latest news disseminated via Twitter, the website provides public information via the following sections:

- Services: Information on the ELRI services and benefits.
- Project: Information on the project timeline and development.
- Resources: Links to useful resources, including NRS URLs and user guidelines.
- Events: List of events where ELRI was represented.
- Links: A collection of useful links, notably to companion initiatives.
- Contact: A list of contacts to reach the developers of the project.

The website plays an important role in centralising information around the project.

6.1.2 Social media

Information and news related to the ELRI project are disseminated on social media via the Twitter platform, under **@ElriProject**. The account was created in November 2019 and has provided regular updates on the project. As of September 2019, the account has 169 followers and 155 tweets were published about the project.

The dissemination of project news and information via Twitter has been an important instrument due to its dynamic and reactive nature, helping promote awareness of the Action and connect it to related initiatives. Additionally, activity data has provided important information on the impact of the different project announcements and events.

As the National Relay Stations will be maintained after completion of the Action, the Twitter channel will be maintained active to help promote the continued usage of the ELRI services.

6.2 Stakeholders conference

ELRI organised its Stakeholder Conference in Paris on November 20th, 2018. The conference had two main objectives. On the one hand, it aimed to raise awareness on the importance of sharing language resources among public institutions and on the meaningful advances that can result from it. On the other hand, the aim was to present the ELRI project itself, its goals, approach and benefits, including a presentation of the National Relay Stations that are at the core of the project and enable flexible sharing of language resources with immediate benefits for stakeholders. [27]

The conference provided an open forum for participants to exchange their views on sharing and using language resources, in addition to presenting the ELRI initiative. In accordance with these goals, the program included the following four main sessions:

- Presentation of the project and challenges of multilingual Europe.
- Use cases related to using and sharing language resources.
- Practical issues when sharing language resources.
- Panel and discussions on data sharing, language resources and translation services.

The conference thus put an emphasis on practical issues for stakeholders in public institutions and translation centres. There were 37 registered participants, from 19 institutions, of which 35 attended the conference.

[27] More information and a complete list of presentations is available at the following address: http://www.elri-project.eu/StakeholdersConf.html

Representatives of the public entities attending the event were very interested in the benefits ELRI could bring to their respective institutions, both in the short term in the form of translation memories and in the long-term via improvements to the eTranslation service. Exchanges with the public institution representatives were instrumental in helping clarify potential issues that could arise during the LR sharing process. These exchanges took place during the sessions themselves, where specific questions were raised regarding the LR sharing process and the ELRI approach, and during the breaks, where fruitful conversations took place between attendees and project partners, and contacts were made for follow-up beyond the event.

The conference was thus impactful in terms of contacts and exchanges with members of key public institutions, with strong potential for information regarding ELRI and its benefits to spread internally at these organisations and their network of related public institutions.

6.3 Workshops

As part of its dissemination activities, workshops were also organised in each of the four Member States represented in the project. The following sections provide summaries of these events, with more information on the content of the sessions available on the project website.

6.3.1 Spain[28]

The first workshop of the series took place in Madrid on March 12th, 2019. As shown in Table 5, it was a highly successful event in terms of attendance, drawing a large number of participants from nearly 60 different institutions.

Date	March 12th, 2019
Location	Representation of the EC in Spain, Paseo de la Castellana, 46, 28046 Madrid. Room EUROPA
Organisers	Maite Melero (SEAD) Thierry Etchegoyhen (VICOM)
Number of registered participants	100
Number of participants	85
Number of speakers	6
Number of participating institutions	56

Table 5 ELRI workshop in Spain

The purpose of this event was twofold: on the one hand, to inform the public servants who have translation needs on the benefits of translation technologies and the demand for multilingual data to feed the technological solutions; and, on the other, to present them with the ELRI National Relay Station, as a useful means to collect, prepare and share their data.

The ELRI initiative is strongly supported by the Secretary for the Digital Advancement, who considers it a fundamental piece of the infrastructure of translation technologies that it is deploying for the benefit of the Public Administration. During the event, two related CEF initiatives were also showcased, which are also a part of this infrastructure.[29]

[28] More information at: http://www.elri-project.eu/ELRI_Workshop_SEAD.html
[29] iADAATPA (https://iadaatpa.com/) and NEC-TM (https://www.nec-tm.eu/).

The morning sessions featured presentations on the ELRI project, from companion initiatives and on the importance of language resources for human and automated translation. The afternoon was fully dedicated to a hands-on session on accessing and using the National Relay Station.

The feedback that we were able to gather in situ was tremendously positive, with respect to the different presentations and the ELRI platform itself.

In terms of impact, the event can be considered highly successful, generating a lot of interest among the relevant stakeholders, as reflected in the number of registered participants which was limited only by the seating capacity of the venue. The event helped open the doors to several public entities, with whom potential data provision is being actively discussed.

6.3.2 Portugal

Portugal hosted two separate events during the month of May 2019, summarised in the next two sections.

6.3.2.1 ELRI @ VIII Encontro de Tradutores da Administração Pública

The first event, the 8th Meeting of Translators of Public Administration, took place on May 6th, 2019 in Lisbon and featured the participation of representatives from a large number of key institutions.

Date	May 6th, 2019
Location	Assembly of the Republic, Lisbon
Organisers	Assembly of the Republic, Attorney General's Office and Administrative Modernization Agency
Number of registered participants	55
Number of participants	53
Number of speakers	7
Number of participating institutions	22

Table 6 ELRI @ VIII Encontro de Tradutores da Administração Pública

The main objective for this public event was to create awareness of the existence of tools that can help the community of translators from the Portuguese public administration in their daily work. Previous assessments of actual practice had made apparent that the use of translation support tools has a low usage rate in public administration. This event was intended to disseminate and raise awareness about tools and ways of collaboration that can benefit the work of human translators.

To achieve this goal, among other topics, the presentation of the ELRI platform was integrated into an event that annually brings together translators from the public administration. This was the eighth event in this series of events with an established community and target audience and so it created a partnership with the Parliament and the Attorney General of the Republic, who were responsible for organising the event in previous years.

This workshop had the following main objectives:

- Introduce the ELRI National Relay Station eTradução

- Explain the use and advantages of ELRI
- Engage the Public Administration institutions with the initiative
- Collect contacts for future engagement

The event ran for the full day and included networking breaks. The different sessions featured active participation, and several participants offered contributions during the discussion sessions.

Overall, the feedback was very positive and allowed the Portuguese ELRI team to compile an initial list of entities who were interested in the initiative and showed potential in terms of sharing resources.

As a strategy and considering the profile of the public administration in Portugal that is concerned with the use of translation support tools, this event was viewed as an important first step prior to holding an ELRI platform hands-on workshop. It was viewed as important to start with an event intended to explain the advantages of using tools and the main reasons why the ELRI project was implemented in Portugal.

In terms of impact, the message was delivered in a positive way and participants became more aware of this topic, as evidenced by the volume of registrations that we considered very positive for the second event, where the use of the ELRI platform was demonstrated in a practical way.

6.3.2.2 Hands-on workshop[30]

The second event was organized in Lisbon on May 29th, 2019, in AMA offices, with the support of FCUL.

Date	May 29th, 2019
Location	Lisbon
Organisers	Administrative Modernization Agency
Number of registered participants	33
Number of participants	28
Number of speakers	2
Number of participating institutions	14

Table 7 ELRI hands-on workshop in Portugal

The main objective for this public event was to present a demonstration of the usage of the ELRI platform and explain how ELRI can impact the eTranslation tool offered by the European Commission. The reason to also include demos of the eTranslation tool was to inform the participants about one immediate advantage of sharing language resources, given that the resources being shared in ELRI can contribute to the improvement of the eTranslation tool.

This workshop had the following main objectives:

- Identify and demonstrate the resources that can be shared
- Create individual accounts on ELRI/eTradução for the participants
- ELRI/eTradução hands-on usage

[30] More information at: http://www.elri-project.eu/ELRI_Workshop_AMA.html

- Gather contacts for future email, phone, or face-to-face interactions to help share existing resources in the institutions of the participants

From the interest expressed by the participants in the activities, the workshop can be considered a great success. Given the highly practical nature of the proposed activities, more time would have been needed for the event, justifying that perhaps in the future AMA may develop new events with similar characteristics.

As a direct impact of this workshop, registrations to the platform and first resources were collected on the day of the event, as participants were also asked to bring resources that they could share. Several contacts were also established for follow-up activities after the event.

6.3.3 Ireland[31]

The first ELRI workshop in Ireland was organised by DCU and took place in Galway on May 8th, 2019.

Date	8th May, 2019
Location	Tribeton, Galway city
Organisers	DCU
Number of registered participants	64 (via Eventbrite)
Number of participants	58 participants were present on the day
Number of speakers	10
List of participating institutions	42

Table 8 ELRI workshop in Ireland

Galway, in the west of Ireland, was chosen as a preferred location for the ELRI Ireland workshop due to its proximity to a number of Irish language organisations and Irish speakers. The workshop was billed as the official launch of the National Relay Station and gathered a significant number of participants from different key institutions. The workshop provided an opportunity to introduce the ELRI NRS for Ireland, not only to those working in public administration who had not yet been met with by the ELRI team at DCU, but also to those with whom there had been extensive communication in the lead up to the event. It also gave people an opportunity to hear from partners in the project, thus driving home the message that this project, while focusing on the Irish language in this country, is part of a wider European project to overcome language barriers in the European Union overall.

The event was viewed as a success, considering the lively participation and attendance. There was an expected difficulty in ensuring good attendance to such an event as it featured concepts that are new to most, with language data collection not being generally considered. However, the attendees came from a variety of different government departments and institutions and the ELRI Ireland team were particularly impressed by the level of engagement and general support for the initiative. The speakers gave relevant and interesting presentations and the organisers of the event were appreciative of the endorsements from the Department of Culture, Heritage and Gaeltacht (DCHG). While some of the attendees had been previously met through outreach activities in the previous months, the event provided an opportunity to meet a large

[31] More information at: http://www.elri-project.eu/ELRI_Workshop_DCU.html

number of potential users for the Irish NRS face to face and provide answers to any questions they had.

Feedback questionnaires were also distributed to (1) get an understanding of current translation practices within public administration and (2) evaluate the launch. 26 questionnaires were completed, with predominantly positive feedback. Of particular interest was the answer to the question: *I understand the importance of sharing my language data and I intend to use the National Relay Station*, to which 70% strongly agreed and 30% agreed.

The high attendance and engagement at the event confirmed that there is genuine interest and support to collect language data for the benefit of the Irish language. A report on the event with an interview with Helen McHugh was broadcasted by the Irish language national broadcaster, TG4 on the daily news bulletin. Micheál Ó Conaire from the DCHG also gave an interview with the national Irish language radio station, Raidió na Gaeltachta, on the morning of the event. In the following weeks, there was a large number of registration requests from users (many of which had attended the event) with subsequent uploads of LRs.

6.3.4 France[32]

The last in the series of ELRI workshops was held in Paris, on June 18th, 2019.

Date	18 June 2019
Location	KLUSTER BUSINESS CENTER 5-9 rue Van Gogh, 75012 Paris, France
Organisers	ELDA
Number of registered participants	25
Number of participants	23
Number of speakers	6
Number of participating institutions	14

Table 9 ELRI workshop in France

The ELRI workshop held in France proved to be an informative and practical gathering where participants expressed their interest in exploring the ELRI platform themselves and processing some of their data. The hands-on session was particularly relevant to some of them, as they were experiencing the same legal and/or technical problems that were described and tackled during the workshop. Some of these users showed great interest in participating more actively in ELRI. However, subsequent discussions and decisions have had very low impact and ELDA continues to work with some of them on how to set up this collaboration in a way that meets both their expectations and those from ELRI and the EC.

Constraints regarding confidentiality and personal data were often brought up by the participants. Nevertheless, feedback was positive during the event and attendees found they had a relatively good idea of what the platform could offer them and what it could help the community with.

At present, several institutions are interested but they need to discuss internally how to advance in this collaboration, as the concept of sharing data is not something they had foreseen before

[32] More information at: http://www.elri-project.eu/ELRI_Workshop_ELDA.html

and legal fears still remain. Furthermore, when compared to the workshops held in other countries, there is a sense that data sharing in France is still a practice that needs to be adopted more dynamically. However, the fact of allowing ELRI users to approach sharing in a flexible manner helps them feel more at ease and we hope that this will end up being the beginning of a fruitful collaboration for data sharing.

6.4 Other events

Over the course of the Action, ELRI was also represented at a number of different events, listed below:

- ELRI was represented at the 3rd ELRC Conference[33], held in Brussels on November 7 & 8 2017. The project coordinator was also invited to participate in a panel on ensuring sustainable data.
- ELRI was invited to present the initiative at the 2nd ELRC workshop in Spain[34].
- As a generic service, ELRI was invited to be present at the 6th LRB (Language Resource Board of ELRC) meeting in France[35].
- The project was also presented at the 21st Conference of the European Association for Machine Translation[36].
- Finally, ELRI was invited to participate in the Workshop on Showcasing LT in H2020 and CEF Telecom projects[37].

These events represented important opportunities to present the initiative and discuss its benefits with diverse audiences, who regularly expressed their interest in the ELRI approach.[38]

[33] http://www.lr-coordination.eu/3rdelrc7
[34] http://www.lr-coordination.eu/l2spain_agenda
[35] http://www.lr-coordination.eu/6LRBMeeting
[36] http://eamt2018.dlsi.ua.es/
[37] https://ec.europa.eu/cefdigital/wiki/display/ETCOMMUNITY/Workshop+Showcasing+LT+Agenda
[38] The ELRI Consortium wishes to thank the organisers of the events listed above for their kind invitations.

7 Governance

The integration of new countries into the ELRI network requires both the establishment of an appropriate Governance structure and the execution of several minimal activities. We discuss each aspect in turn below.

7.1 Governance structure

The proposed Governance structure for new countries joining the ELRI network is described in Figure 18.

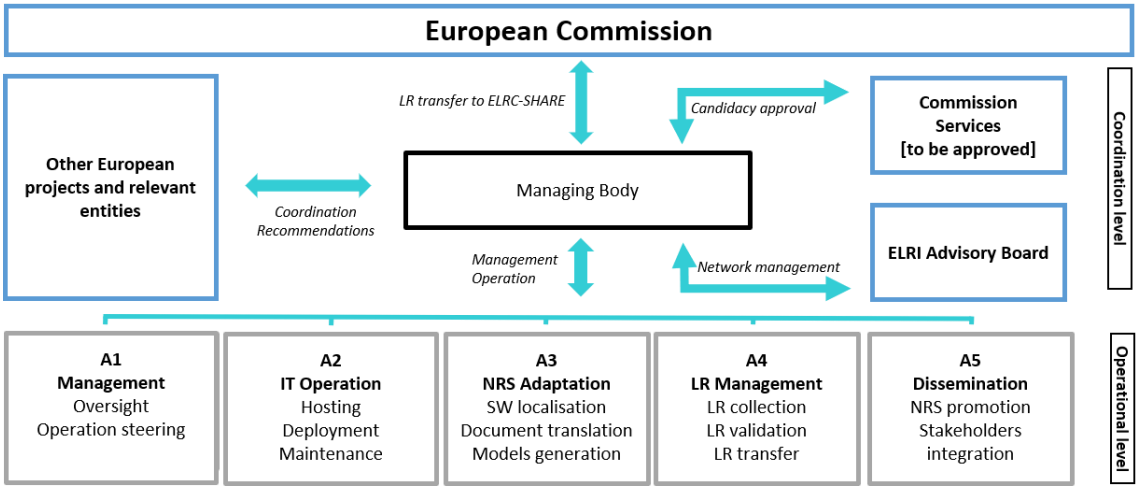


Figure 18 Governance structure

Managing an ELRI National Relay Station requires a Managing Body, with the following characteristics and responsibilities:

- Be a public institution of the Member State/EEA country or a non-public 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 first point is important to emphasise the need for such a platform but also in terms of building trust with the other public institutions of the country, including public administrations, and to be able to support NRS-related activities as part of a public mission.

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 following considerations:

- There should be only one National Relay Station per Member State/EEA country, to avoid conflicts and confusion on the part of end-users. Therefore, there should be some form of control over which institution hosts an NRS for a given country.
- The NRS codebase is freely available in open repositories, but the data processing pipelines are licensed specifically to Innovation and Networks Executive Agency (INEA) and their access requires specific credentials, as they are in a private password-

protected repository. This is meant to provide the relevant Bodies with the means to control the deployment of the ELRI network, in particular its extension to new countries.

- The ELRI framework was developed within a programme of the European Union, namely the Connecting Europe Facility programme, and the integration of new countries to the network should be controlled to ensure the expected standards of representation and activity oversight.

Of special importance is the ELRI Advisory Board, constituted of the seven entities who have led the development of the ELRI infrastructure and managed its deployment in the four Member States of the CEF Action, namely: AMA, DCU, ELDA, FCUL, LINKARE, SESIAD and VICOMTECH. The role of the Board is to provide all members of the ELRI 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 Member States/EEA countries willing to join the network, in an advisory capacity.

Although not an absolute necessity per se, since a given institution may request to be the Managing Body of an NRS in a new country and approved by EU Bodies independently of EU funding, it is recommended that the integration of a new country takes place within the framework of an EU-funded initiative. This is mainly motivated by the fact that the deployment of an NRS requires a set of initial activities for which there may not have been the required funding in full on the part of the candidate Managing Body. The following section describes the core activities that are to be overseen and executed by the Managing Body.

7.2 Activities

As described in Figure 1, the integration of a new country to the network involves a set of activities, summarised below:

- IT Operation (mandatory): server allocation, application deployment and service maintenance.
- NRS Adaptation (mandatory): localisation of the application to the official language(s) of the considered country, translation of the relevant documentation for end-users, and generation or integration of language processing models for the new language(s) of the considered country.
- LR Management (mandatory): collecting, preparing and validating language resources, and transferring LRs shared at the EU or Open Data level to ELRC-SHARE.
- Promotion (recommended): promoting the goals and function of the National Relay Station, contacting public institutions, organising dissemination events.

We consider each of these aspects in more detail below. In each case, we describe the estimates of required resources, in terms of personnel and infrastructure, as well as the efficiency considerations that motivated the decisions undertaken to establish the governance framework.

7.2.1 IT Operation

IT operation involves an initial deployment phase and continuous maintenance of the service. We summarise each part of the activity below.

7.2.1.1 Deployment

As web applications, National Relay Stations need to be hosted on a server. The requirements in terms of hardware can be considered minimal, given the characteristics of modern servers, e.g. there are no requirements for costly components such as Graphical Processing Units.

Once a server is available, deployment involves configuring and running the Docker containers available for the service. Expert IT personnel will be mostly required for this phase, since, although the service is available as a composed image that only requires the installation of the Docker environment, the following aspects can involve dedicated IT supervision:

- Hosting configuration: the service needs to be made available at a URL managed by the public institution, which requires obtaining the relevant permissions and configuring the hosting environment.
- Mail server configuration: the default deployment provides a generic email handling environment for the service, but some configuration may be required to use institution-specific email services.
- Security review: the ELRI software has been reviewed for vulnerabilities prior to its current deployments, but IT departments of specific institutions may need to run their own security testing prior to launching the service within their installations.

After the initial configuration steps, the service will be able to enter a continued maintenance phase.

7.2.1.2 Maintenance

Maintenance of the service involves the following main processes:

- Release update: in the event of continued developments of the service, new releases will need to be deployed. This part of the process involves executing a short series of actions, which comprises stopping the Docker containers, upgrading to the latest versions of the service, and restarting the service.
- Service monitoring: as with all IT services, the ELRI NRS needs to be regularly monitored to ensure its persistence and proper usage.
- Data backup: regular data backup is highly recommended to ensure user data safeguard during the maintenance of the service.

The required maintenance operations are fully documented and require minimal IT personnel involvement.

7.2.2 NRS Adaptation

To provide a user experience that is appropriate for users of a given country, the NRS web application needs to be localised to the official language(s) of the considered country. This includes the main tasks indicated in the following table, which also lists the relevant available and recommended tools:

Task	Mandatory	Tools
Web application localisation	YES	Poedit[39] (Free)
End-user documentation translation	YES	Various (Free)
Dissemination material translation	NO	Various (Free)
Lexical translation model generation	YES	FastAlign[40] (Free)
Tokenisation model generation	YES	Moses[41] scripts (Free)
Truecasing model generation	YES	Moses scripts (Free)
ELRI Data Checker adaptation	YES	spaCy[42], text editor (Free)

Table 10 NRS adaptation main tasks with relevant tools

The specifics of each task, as well as an estimate of the overall effort and costs, are described in the next sections.

7.2.2.1 Localisation and translation

The localisation of the user interface requires translation of the default English strings into the language(s) of the new countries, with dedicated tools such as the freely available Poedit. As the codebase is rather large and may require knowledge of technical translation, the mandatory localisation effort is reduced to translating only those portions of the interface that are related to end-user interaction. The decision to extend localisation to the back-office of the application, were this to facilitate the operations of the system in a given country, is left to the Managing Body in charge of the NRS in the country at hand.

Regarding the translation of the documentation, only end-user NRS documentation is meant to be translated, which mainly includes the following: user guidelines, terms of service, privacy policy. Information related to back office activity, e.g., the NRS maintainers manual or LR validation guidelines, is provided in their current English version, as these documents are neither meant for, nor helpful to, end-users, and English is a common language for technical activities.

Although not mandatory, it is also recommended that NRS adaptation includes the translation of existing dissemination material, such as infographics, or the creation of new material adapted to the specifics of the Member State joining the network.

7.2.2.2 Model generation

The adaptation of the automated processing pipelines is necessary to cover additional languages. This includes the generation of new lexical translation tables, capitalisation and tokenisation models, and the adaptation of the ELRI Data Checker to increase its coverage to new patterns related to the detection of potential personal, confidential and sensitive data identification. This part of the process requires a certain technical know-how, as it involves the use of different toolkits, the proper selection and preparation of training data, and the integration of the newly generated models within a new release of the service.

Although not mandatory, it is also recommended that the adaptation of the service to the new language(s) of a given Member State joining the network is executed within a dedicated project

[39] https://poedit.net/

[40] https://github.com/clab/fast_align

[41] https://github.com/moses-smt

[42] https://spacy.io/

that includes partners with the relevant expertise in the matter, to optimise development times. Alternatively, new Managing Bodies may perform all required model generation activities based on the existing ELRI documentation.

7.2.3 LR management

The goal of a National Relay Station in a given country is to help collect, prepare and validate language resources that can be used to improve human and automated translation services at the national and European levels. To achieve this goal, the main tasks described below need to be undertaken by new Managing Bodies.

7.2.3.1 Data holders' integration

This task involves direct contacts with data holders in public institutions with an interest in contributing resources and benefiting from the ELRI services.

New Managing Bodies will need to establish lists of potentially interested parties, organise meetings with interested parties to describe the objectives, implications and benefits of ELRI, and provide support for the integration of interested parties into the network.

Additionally, new Managing Bodies will need to coordinate with the relevant institutions and actors in their respective countries to ensure an optimal integration of ELRI activities into existing or developing EU-wide or national plans for language technology, language equality or digital advancement. The goal of this effort is to ensure the integration of ELRI as a useful building block that completes current or developing activities at the national level in the targeted countries.

7.2.3.2 Resource management

Registered users of the ELRI service -- once they have been approved by the dedicated team managing the national NRS, following appropriate screening -- upload their resources in raw form, which are then processed by the integrated data processing pipelines. Once processed, resources need to be carefully reviewed prior to being published and shared. A significant portion of the work for new Managing Bodies is thus to manage the resources, following the validation process established within ELRI. This process involves several steps that examine the different stages of a language resource as it is processed by the system.

Within the validation process, resources are first examined in their raw form, to obtain a general view of the contributed data and ensure that no unusual data form part of the contribution. After this general check, resources can be ingested, an action triggered by ELRI reviewers which launches the automated processing of the raw data. The outcome of this process is then checked to examine if the system has produced error reports or if the automatically processed resource generated by the system is consistent. The ELRI Data Checker can then be applied to the resource, to assess if potential personal, sensitive or confidential data may have been included in the contributed resource; if so, the data holder is contacted to clarify the matter as necessary before proceeding further. Finally, distribution information associated to the contributed data is examined, with eventual additional exchanges with the data holder if licensing or IPR issues require clarification. Resources that pass the established validation criteria will then be published and made available to all members of the specific groups with which data holders wish to share them.

Validation reports that correspond to the results of the main steps of the process need to be produced for each resource, allowing for the tracking of potential issues and the delivery of relevant information for further usage. The different steps of the process are supported by existing ELRI documentation.

7.2.3.3 Effort and costs

Although the goals and expected actions would be similar for any new Managing Bodies, the associated costs may vary as they depend on several factors.

First, the size of the country itself needs to be considered, as resource collection activities may imply face-to-face meetings with stakeholders, which may require more personnel time and overall costs in larger countries.

Second, the projected ease of contacting public institutions in a given country also needs to be considered, as some Managing Bodies may be more directly connected than others with potential data holders from other public organisations, thus reducing the costs for the required outreach and subsequent data collection activities.

Finally, the need to coordinate the ELRI activities with existing initiatives at the Member State/EEA country level also needs to be taken into account, as this aspect may induce additional efforts to reach a comprehensive ecosystem for language resources and language promotion in the considered country.

Overall, the effort and associated costs for the LR management task are thus expected to vary between countries. Additionally, since the activity is likely to undergo different phases of varying intensity, as part of specific actions or not, and sustained over an extended period of time, a precise cost estimate is difficult to establish. A minimum of 6 person months within a 12-month period can be considered a reasonable starting point for this activity.

7.2.4 Promotion

As with any activity centred on gathering resources from data holders, promotion activities are necessary to showcase the benefits of the ELRI network for public institutions in the new countries joining the network.

General dissemination activities include the preparation of relevant material, participation in related events, and organisation of events. We discuss the main tasks in the two sections below.

7.2.4.1 Material preparation and ELRI promotion

The preparation of dissemination material is highly recommended to attract data holders who could contribute new language resources. This includes the adaptation of existing ELRI material for the new countries and languages, as well as the generation of new material as deemed appropriate by the new Managing Bodies.

Additionally, to promote the launching of ELRI in a new country, Managing Bodies should also take part in activities related to the promotion of the initiative, such as the participation in relevant events and the dissemination of ELRI objectives and benefits within their respective networks, via social media activity and exchanges with the relevant institutions, for instance.

7.2.4.2 Event organisation

To promote the benefits of the ELRI infrastructure for the public institutions of a new country, dedicated events such as hands-on workshops are recommended. Such events have demonstrated their key importance within the ELRI initiative so far, drawing the interest of public institutions and providing a framework for direct exchanges that help clarify the use of the infrastructure.

As with all dissemination initiatives, assuming these are not required within a specific funded action, the organisation of events for new Managing Bodies is not mandatory to launch the ELRI services in a new country but is highly recommended.

7.3 Summary

The Governance plan described here provides the basis for the integration of new countries into the ELRI network. As part of its activities, the ELRI project has dedicated resources to the design of a structured plan that enables new countries to join the network and deploy its services with minimal efforts and costs. The plan outlined in this document is the result of these activities, and the fruit of the experience acquired during the execution of the ELRI Action.

The inclusion of new countries is meant to be both facilitated and controlled. Thus, on the one hand, the detailed list of required activities and actions, available to potential new Managing Bodies and supported by the relevant documentation produced during the project, provides a clear set of requirements to deploy a National Relay Station in a new country, as well as estimates of expected efforts and costs. On the other hand, the established Governance structure, which requires approval by, and key credentials from, 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.

As previously mentioned, although it is possible for a public institution to request the credentials to act as Managing Body of an NRS in a country where ELRI is not yet deployed, it is recommended that extensions to new countries take place within funded initiatives in order to benefit from the technical expertise of the network developers and thus optimise deployment times in new countries.

8 Sustainability

The collection of language resources is, by definition, a long-term activity, given the continuous production of relevant data in the domains of interest for the Digital Service Infrastructures. This process requires sustainable solutions that can help overcome important blocking issues in data sharing, including reluctance to share data, in general and outside national jurisdictions in particular. The ELRI sustainability plan is based on the main aspects summarised below and described in more detail in the next sections.

The first component of the overall sustainability plan relates to the necessary commitment by institutions in charge of an NRS to maintain the service outside the scope of a funded Action such as the ELRI CEF project. For this to take place, ELRI needs to provide sufficient benefits that balance the costs of the maintenance of the service without external funding. In Section 8.1, we describe the institutional sustainability factors, with particular emphasis on the four Member States where ELRI National Relay Stations are currently deployed.

A second essential component is technical sustainability, in the sense that the ELRI service needs to be maintainable in a given country from a technical point of view with minimal effort and costs. In Section 8.2., we review the main technical features of ELRI that support its sustainability.

Finally, the third major component is financial sustainability, wherein the costs incurred from the maintenance of the service beyond the lifetime of the EU-funded ELRI project should be deemed assumable by the institution in charge of maintaining an NRS. Section 8.3 reviews the main financial aspects related to the maintenance of the ELRI services beyond the funded project.

8.1 Institutional sustainability

For the ELRI services to be sustained, institutions in charge of hosting and maintaining National Relay Stations need to be willing and able to maintain the service beyond the EU-funded ELRI project. This commitment is expected to follow from the public mission of the institutions in charge of an NRS in their country, provided there are enough benefits to outbalance the costs incurred with the maintenance of the service. Although one of the goals of ELRI is to contribute to the sharing of language resources with the largest possible number of stakeholders at the national and European level, or as fully open data, maintaining a National Relay Station in a given Member State is incumbent on the national institutions in charge of the service. Therefore, the benefits for these institutions need to be clear and accessible for a sustainable solution to be viable. ELRI provides such benefits in several ways.

First, language resources are essential to foster the digital presence of the language(s) of a given Member State. In a digital world with ever increasing content to be quickly translated, both human and automated translation services rely on such resources to reach optimal translation quality as well as cost efficiency and timely delivery. ELRI provides dedicated means to accelerate the collection and preparation of structured resources, via its integrated data processing pipelines, a benefit that constitutes a strong incentive for the maintenance of the service.

Additionally, ELRI provides a country-based central repository for public data, which benefits all relevant actors in the country, including the institution in charge of maintaining the NRS. This allows all public institutions in a given Member State to benefit from the prepared resources, irrespective of their being shared beyond the national level. This bottom-up approach is an additional benefit, which helps accomplish the public mission of the national institutions in charge of an NRS, helping to centrally managing national resources and fostering their maximum use in translation services.

Finally, ELRI provides flexible means to share resources, which can be shared at the national, European or Open Data levels. This feature of the infrastructure enables national institutions to benefit immediately from shared resources at the national level, while also providing a relay station to promote their resources under a larger sharing scope and benefit from improved European services for their respective languages.

The ELRI National Relay Stations are now considered important building blocks in the respective language technology ecosystems of the Member States where they are currently deployed, thus demonstrating that the infrastructure is meant to be sustained after completion of the EU-funded initiatives that support its development.

In Spain, for instance, it is one of the building blocks of the translation technology ecosystem that is being developed by the Secretary of State for Digital Advancement. In Ireland, the ELRI NRS is viewed as a central component for the collection and preparation of language resources that can help advance the presence of Irish in the digital age. In France, ELRI is meant to act as an important tool to increase the culture and practice of LR sharing. Finally, in Portugal ELRI is also viewed as an important piece in the strategy for language technology and digital advancement that is carried by the Agency for Administrative Modernisation, under the Secretary of State for Administrative Modernisation.

The benefits of the ELRI National Relay Stations have thus led the institutions in charge of the service in their respective countries to support its maintenance beyond the lifetime of the EU-funded CEF Action. A short summary of the sustainability plan from each of these institutions is provided below.

8.1.1 France

ELDA is one of the partners of the ELRC Consortium that is behind the main EC action related collection of LRs for eTranslation. Even though ELDA is incorporated as a company, it plays a not for profit role in France and beyond. It has established strong cooperation activities with some of the major Public Bodies that care about language promotion but also about multilingualism and translation services (General Delegation for the French language and the languages of France (DGLFLF), ministry of culture; Centre des liaisons européennes et internationales de sécurité sociale (CLEISS), Service de traduction du Ministère de l'Économie et des Finances, etc.). Such cooperation already led to sharing multiple resources with the EC via the ELRA-Share platform. French public bodies were very supportive of the ELRI project and we are very confident that they will continue such support and endorsement. ELDA has committed to support the operations of the NRS beyond the lifetime of ELRI, under some of its internal funds. While doing this, ELDA will advocate for a partnership with the DGLFLF to join forces in such endeavour. The expected cost to ensure continued use of the French NRS by existing registered users and new ones is worth it as it helps provide Language Resources usable by the Language Technology community at large, which is part of the core mission of ELDA.

ELDA does expect that the under discussion national strategy to initiate a language research program will support initiatives such as ELRI and hence will offer another sustainable alternative to the plans of ELDA. In the meantime, the platform maintenance costs will be borne by ELDA after project completion.

8.1.2 Ireland

DCU has had a collaboration with the Department of Culture, Heritage and the Gaeltacht (DCHG) for a number of years, which started initially with joint work on the European Language Resource Coordination (ELRC). This collaboration has involved researchers at DCU delivering technical support for the Irish language through projects such as the bespoke MT system *Tapadóir* for translators working within government departments. Researchers in DCU are also heavily involved in the National Digital Strategy for the Irish language, which is currently being finalised.

The DCHG has been extremely supportive of the ELRI project, with officials from the Department speaking at the Irish National Relay Station (NRS) launch and promoting the use of same. The DCHG is aware that the ELRI project will finish and has expressed its commitment to guarantee the continuation of the Irish NRS (https://elri.dcu.ie). DCU, together with the DCHG, is exploring how it will be possible to support the seamless running of the NRS beyond the lifetime of ELRI, with a view to ensuring continued use of the NRS by existing registered users, while also extending the network of contributing data holders.

8.1.3 Portugal

Currently, in Portugal, there is no national strategy to support initiatives that encourage the collection and sharing of language resources or the development and use of machine translation or translator support tools.

The ELRI project is the first step and an important opportunity to bring together the Public Administration translators' community in a common goal which is the development of

Portuguese language support technologies and the modernization of their working methods through the integration of support tools for translator's work.

The Administrative Modernization Agency (AMA) as the national entity in charge for digital transformation assumes a leading role in this area and looks forward to maintaining this service beyond the project end date. Accordingly, AMA is committed to keep this service active and promote it within the Portuguese Public Administration, maintaining communication and engaging new contacts with different stakeholders, especially those with the greatest potential for contribution.

This effort should continue as long as this issue is relevant to Portugal and the Portuguese Public Administration and as long as the service continues to demonstrate its merits and keep the community interested.

The platform maintenance costs will be borne by AMA after project completion.

8.1.4 Spain

Since 2015, the Secretary of State for Digital Advancement (SEAD) has the mandate to implement a Plan of Impulse of Language Technologies, with the goal of boosting language technologies in Spanish, Catalan, Basque and Galician, as well as take advantage of these technologies to improve public services in Spain.

In the context of this Plan, the SEAD is setting up an ecosystem of translation technologies, with the aim of facilitating access to these technologies to the public administration. Given the fact that data collection is at the very heart of the process, the ELRI NRS in Spain constitutes a fundamental element of this infrastructure. The NRS makes it possible to collect multilingual texts, process them into actionable translation memories and share them with other public institutions.

The translation memories generated by the ELRI station can then be directly ingested by a CAT tool, used to train a domain-specific MT engine or used to negotiate better contracts with external LSP providers.

The SEAD is currently hosting and providing maintenance to the Spanish national relay station and plans to keep on doing so for the coming future, as an integral part of its translation technologies infrastructure.

8.2 Technical sustainability

As a decentralised network where national nodes can be maintained independently, ELRI offers a solution that is not dependent on the persistence of a central entity overseeing the maintenance of the network to guarantee the technical sustainability of the service in the countries where it is deployed. This aspect increases the overall robustness of the technical sustainability plan, where eventual discontinuing of the ELRI services in a given country does not affect the other countries where ELRI is maintained.

Additionally, the different aspects described below are of relevance for the technical sustainability of the ELRI service.

8.2.1 Stability

The latest releases of the service at the end of the project can be considered stable, having undergone both in-depth security reviewing and extensive use-case testing in all four Member States where it is currently deployed, without notable issues after several months of live usage.

The requirements defined for the service have been met and no new critical requirements are expected at this stage, since the main scenarios for the collection, automated preparation and sharing of language resources are covered by the ELRI infrastructure that has been deployed. It is therefore expected that the service will be able to sustain user requirements after the completion of the project, without the need to generate new releases.

8.2.2 Maintenance

A backup procedure has been established for both data and software, available to the maintainers of the service in all four Member States where the solution is deployed. Interruptions or mismanagement of the service may thus be recovered from in terms of user data and ability to restart a stable version of the service.[43] On these grounds as well, the ELRI solution can be viewed as sustainable beyond the lifetime of the project.

Additionally, the procedures for service management have been designed with sustainability in mind, putting special emphasis on ease of deployment and reduced expert intervention. This was accomplished via the provision of a fully dockerised service whose management requires minimal intervention by dedicated IT personnel. This aspect is critical for actual sustainability of a service that is not part of an ongoing project, as it cannot be expected that significant resources would be allocated to the maintenance of a service without the corresponding financing of significant expert personnel monitoring or active intervention.

One important related aspect is the need for critical modifications of the service codebase, which can arise for instance in case of new detected software security vulnerabilities. Unless a viable solution is offered in such cases, the sustainability of the proposed solution would be dependent on the occurrence of this type of vulnerability issues. To facilitate the fixing of such vulnerabilities, the code for the infrastructure is available in an open repository,[44] along with detailed documentation that notably includes the steps to generate new versions of the Docker images used for the deployment of an NRS. If important issues were to arise on the technical side once the project is completed, institutions in charge of maintaining the service will have the means to solve the issues independently of an ongoing project, thus ensuring continuity of the service.

As software evolves over time, some of its components may become deprecated. Although the service may remain functional nonetheless, operating on deprecated software components is not recommended and may actually be prohibited under the operating guidelines of a given institution. For this type of issue, the available open software components would allow in theory any institution running the service to upgrade the relevant components, as needed. However, in practice, for large software codebases such as ELRI, which is based on the large METASHARE/ELRC-SHARE codebase, the cost of such component upgrades may become prohibitive without the support of a specific project. It is therefore recommended that the needed upgrades of the infrastructure are performed within a supporting project with dedicated financing.

8.2.3 Resource management

Finally, as part of the technical considerations for sustainability, the technical support for the personnel in charge of LR management after the completion of the project needs to be taken into account as well. After completion of the project, the dominating activity will be the usage of the service to collect, prepare and share language resources. The service itself was designed

[43] The relevant information is provided in the ELRI *Maintainer Manual*, provided separately.

[44] https://github.com/ELDAELRA/ELRI

to require minimal IT operation after its launch, and similar considerations were taken into account to simplify, as much as possible, the operation of the service for LR management.

First, a significant sustainability issue in terms of LR preparation relates to the need for expertise in natural language processing methods and tools to be able to produce structured LRs of adequate quality. For instance, the process of generating translation memories from raw document collections provided in different file formats requires a sequence of processing steps with appropriate tools which are non-trivial and involve the participation of expert personnel. The ELRI solution was to integrate fully automated processing pipelines for the main use cases, thus allowing for a sustained usage of the service with minimal involvement of expert personnel for the resource generation part of the overall process.

Similarly, different parts of the system were designed and developed to facilitate the overall process of LR management, including for instance the provision of the ELRI Data Checker tool to assist LR management personnel in their assessment of the presence of personal, confidential or sensitive data in the resources provided by the end-users.

As part of the technical support for a sustainable service, dedicated documentation was produced during the project to assist the personnel in charge of LR management after the completion of the project.[45] Specific documentation was also created for end-users, to reduce the need for technical assistance once the service is provided outside the EU-funded project itself.

8.3 Financial sustainability

Although specific emphasis was placed on ensuring technical sustainability and reducing as much as possible the need for technical intervention on the system, maintaining any type of service generates costs, which need to be considered in any sustainable plan. We provide below a review of the main financial aspects in maintaining the ELRI infrastructure after the completion of the project, in a given Member State.

8.3.1 Infrastructure costs

In terms of hardware, ELRI services require minimal maintenance costs, which boil down to maintaining a single dedicated server to serve the NRS, as the initial cost of server acquisition is not part of the sustainability time period.

In terms of infrastructure, the main costs within the sustainability time period are thus mainly electric consumption for the 24/7 maintenance of the dedicated server, bandwidth usage costs and possibly additional disk storage space to be acquired. These costs are fixed for the most part and can be considered minimal for a present-day IT operation structure.

8.3.2 Personnel costs

Maintenance of the service also involves personnel costs, on two main grounds.

First, IT personnel will be needed to verify the proper functioning of the service and operate punctual redeployment if needed as new releases become available. As previously described, though, special attention was paid during the development of the ELRI infrastructure to the provision of a software service that requires minimal maintenance efforts. In particular, the service is offered within a self-contained virtual environment that can be easily maintained and rapidly redeployed as needed with minimal costs in terms of time and expert personnel efforts.

[45] The relevant information is provided in the ELRI *Reviewer Manual*, provided separately.

The costs for the normal operation of the system are therefore expected to be minimal, and have been in our experience within ELRI.

The main identifiable source of potential significant IT personnel cost relates to the potential need for critical patching of the codebase, if new security issues are identified. In this case, significant effort would be required from IT personnel, although the extent of said effort and associated costs would be highly dependent on the specific issue to be fixed.

Secondly, dedicated personnel will be needed to validate new users of the service and perform the data validation process. The process itself is automated for the most part, thus also reducing the costs of preparing quality structured language resources. Nonetheless, contacting data holders, continuing outreach efforts to new potential data holders and reviewing the resources are an unavoidable part of the process prior to publishing and sharing resources, one which requires significant efforts by qualified personnel.

The costs for this activity are thus the most significant ones when considering a sustained maintenance of the service. Although this activity may be viewed as part of the commitment from each institution maintaining a National Relay Station, it may in some cases be supported by dedicating funding made available from national plans supporting language technologies and digital advancement. A mitigating factor is that resources are usually not contributed in a continuous flow and resource validation is thus expected to happen at specific points in time, in particular after the initial phases of active LR collection executed during the lifetime of funded initiatives such as the ELRI CEF Action.

8.3.3 Effort estimates

The previous descriptions centred on the different aspects to be considered in terms of financial sustainability. Although specific cost estimates cannot be provided, given cost differences between different Member States/EEA countries, the following table provides effort estimates for maintenance and usage of the service in terms of person-month and likelihood of each foreseen activity.

Activity	Effort estimate (person month)[46]	Likelihood
IT maintenance	< 0.1	High
Codebase fixes	> 0.5	Low
Codebase upgrades	> 6	Low
LR management	> 3	High

Table 11 Main sustainability effort estimates

Although this table merely provides estimates of the efforts related to each one of the main activities, which may be translated into actual costs, the provided figures are in line with the assessment of the main foreseen costs for the maintenance and usage of the service after the completion of the project, with resource management as the dominant factor and a low risk of requiring significant IT personnel involvement for codebase modifications.

[46] Over a 12-month period.

8.4 Summary

The sustainability aspects and plans described here provide the basis for the continued provision of the ELRI services in the four Member States where the network is currently deployed. A key objective during the design of the project was the development of an infrastructure that is sustainable beyond the lifetime of the EU-funded ELRI initiative.

On technical and financial grounds, the outcome of the project is a solution that requires minimal management and associated resources, thus 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. The usage of the service itself for resource management will require a sustained commitment by each institution in charge of the NRS, to involve dedicated personnel in resource reviewing and publication. This commitment can be made in agreement with the public mission of each institution in charge of a National Relay Station and future dedicated funding support at the national or European level may help consolidate the sustainability investments made by each institution.

9 Frequently Asked Questions

Over the course of the project, a number of questions were frequently raised, at the public events or via the frequent interactions between stakeholders and members of the ELRI project. A summary of the most frequent questions is provided below.

Q: Why should I contribute my data?

There are several benefits in sharing your data. First, you gain prepared resources such as translation memories, which are of high importance to optimise translation processes, be they performed in-house or outsourced. Secondly, by contributing resources that are also shared with the European Commission, you contribute directly to improve the eTranslation automated translation services, which is freely available to all public administrations in the European union. Finally, by sharing resources in your language(s), you contribute directly to promoting language equality in Europe and building a truly multilingual Europe.

Q: Can I upload more than two documents at the same time?

Yes, the ELRI engines will automatically pair those documents that are translations of each other. All documents related to the same topic should be uploaded together to create large resources on that topic.

Q: Are the original documents shared?

ELRI only shares the prepared language resources generated from your original documents, such as translation memories. Only resources that do not undergo processing, such as terminology files are shared as is.

Q: Do you perform anonymisation?

Anonymisation requires dedicated work to be performed correctly, via dedicated tools and often with on-site assistance. You may contact the ELRC consortium[47] for assistance on your specific anonymisation needs.

Q: Is access restricted to public administrations?

In addition to public administrations, members from all public institutions are welcome to join the relevant NRS, thus including for instance members of public universities.

Q: Why is access restricted to users of public institutions?

The initiative is meant to support the general effort towards the collection of language resources coming from public institutions. Although ELRI offers flexible ways of sharing resources, they are oftentimes shared as Open Data, which are available to all citizens on the EU Open Data Portal.

Q: Can I share my data only with the European Commission to contribute to eTranslation?

Yes, by requesting the creation of a specific sharing group between your institution and the European Commission, your data would only be shared with the DGT.

Q: Which languages are supported?

To provide the necessary data processing engines, we currently support those official EU languages that have been determined to cover most of the translation scenarios for the four Member States of the Action, namely: English, French, Irish, Italian, German, Portuguese and

[47] http://www.lr-coordination.eu/helpdesk

Spanish. Additionally, we provide support between Spanish and co-official languages of Spain: Basque, Catalan and Galician.

Q: Should I upload PDF files or the files in their original format (e.g., odt, docx, rtf)?

Files in their original format are preferred rather than PDF, as the quality of the automatically created resource will be much higher and is thus more likely to be validated by ELRI reviewers.

Q: I have data in one language only, is it interesting to share it?

Although less useful in general than bilingual or multilingual documents, monolingual data can be used to create artificial translations that are useful to train modern machine translation systems, especially if the data are from a specific DSI domain.

10 Conclusions

The ELRI initiative concluded on September 30th, 2019 and has completed its goals at the end of the project. Among its main achievements, a functional infrastructure has been developed, tested and deployed in all four Member States that participated in the CEF Action: France, Ireland, Portugal and Spain. This infrastructure consists of a decentralised network, composed of independent National Relay Stations that provide important services in the Member State where they are deployed.

Each National Relay Station facilitates 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.

Additionally, ELRI offers flexible means to share language resources, thus taking into account the constraints that may be tied to sharing specific resources. In all sharing scenarios, ELRI provides the data holders, who dedicate time and effort to sharing their data, with the prepared resources as an immediate benefit. Thus, the project has developed a mechanism which aims to benefit all stakeholders equally, as a means to generate a community of interest and a positive dynamic around the collection and sharing of language resources.

The various dissemination events organised during the project have demonstrated the strong potential of this approach, with highly positive feedback and strong interest in joining the ELRI network, as users from public institutions or as representatives from new Member States willing to host their own National Relay Station. To support this newly created dynamic, one of the outcomes of the project is a detailed governance plan for new countries willing to join the network, and an infrastructure that has been optimised to minimise deployment costs.

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 have registered to the network and contributed more than 800,000 translation units within the first months of activity. Although preliminary, and with different volumes collected depending on the country, the established community of users and strong dynamic are paving the way for continued and increased sharing of useful 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.

In the four Member States that participated in the project, ELRI is now seen as an important component to support digital advancement and language equality in the European Union, and the main outcomes of the project can thus be considered positive overall.