European Language Resource Coordination

Connecting Europe Facility

ELRC WHITE PAPER

Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe

WHY LANGUAGE DATA MATTERS

European Language Resource Coordination

Imprint ELRC White Paper

German Research Center for Artificial Intelligence (DFKI)
Multilinguality and Language Technology
Stuhlsatzenhausweg 3
Saarland Informatics Campus D 3 2
66123 Saarbrücken
Germany

Contact:
Phone: +49 681-85775 5285
E-mail: info@lr-coordination.eu
Web: www.lr-coordination.eu

Publisher:
ELRC Consortium:
DFKI
ELDA
ILSP
Tilde

Design: Svetlana Gurti

Print:
OVD.eu | Verlag & Eventagentur
OVD.de | Druck- & Werbeservice
Johanna-Wendel-Str. 13 | 66119 Saarbrücken

Images: IRLS

Second online edition 2019

First print edition published November 2019
First online edition published November 2019
Second online edition published December 2019

ISBN: 978-3-943853-05-6

Work underlying the White Paper has been created under the ELRC contract with the European Union (SMART 2015/1091 Lot 2). The opinions expressed are those of the ELRC consortium and Language Resource Board and do not represent the contracting authority's official position.

Authors
Contributions from the following authors in alphabetical order

The ELRC Consortium:

Aivars Berzins, *Tilde*
Khalid Choukri, *ELDA*
Maria Giagkou, *ILSP*
Andrea Lösch, *DFKI*
Hélène Mazo, *ELDA*
Stelios Piperidis, *ILSP*
Mickaël Rigault, *ELDA*
Eileen Schnur, *DFKI*
Lilli Smal, *DFKI*
Josef van Genabith, *DFKI*
Andrejs Vasiljevs, *Tilde*

The National Anchor Points:
Andero Adamson, *Ministry of Education and Research*
Dimitra Anastasiou, *Luxembourg Institute of Science and Technology (LIST)*
Natassa Avraamides-Haratsi, *Press and Information Office, Ministry of Interior, Republic of Cyprus*
Núria Bel, *University Pompeu Fabra*
Zoltán Bódi, *Institute of Hungarian Research*
António Branco, *University of Lisbon*
Gerhard Budin, *University of Vienna*
Virginijus Dadurkevičius, *Vilnius University*
Stijn de Smeytere, *Chancellery of the Prime Minister, Belgium*
Hristina Dobreva, *Ministry of Transport, Information Technology and Communications*
Rickard Domeij, *Swedish Language Council*
Jane Dunne, *Dublin City University*
Kristine Eide, *The Language Council of Norway*
Claudia Foti, *Ministry of Justice, Italy*
Maria Gavriilidou, *Institute for Language and Speech Processing (ILSP)*
Thibault Grouas, *General Delegation for the French Language and the Languages of France*
Normunds Grūzītis, *University of Latvia*
Jan Hajič, *Charles University*
Barbara Heinisch, *University of Vienna*
Véronique Hoste, *Ghent University*
Arne Jönsson, *Linköping University*
Fryni Kakoyianni-Doa, *University of Cyprus*
Sabine Kirchmeier, *Danish Language Council*
Svetla Koeva, *Bulgarian Academy of Sciences*
Lucia Konturová, *Ministry of Culture of the Slovak Republic*
Anna Kotarska, *Centre for Integrated Health Care and eHealth*
Jürgen Kotzian, *Austrian National Defense Academy*

Authors

Contributions from the following authors in alphabetical order

Simon Krek, *Jožef Stefan Institute*
Gauti Kristmannsson, *University of Iceland*
Kaisamari Kuhmonen, *Prime Minister's Office, Finland*
Krister Lindén, *University of Helsinki*
Teresa Lynn, *Dublin City University*
Armands Magone, *Culture Information Systems Centre*
Maite Melero, *Barcelona Supercomputing Center (BSC) and SEAD*
Laura Mihăilescu, *European Institute of Romania*
Simonetta Montemagni, *Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) - CNR*
Micheál Ó Conaire, *Department of Arts Heritage and Gaeltacht*
Jan Odijk, *Utrecht University*
Maciej Ogrodniczuk, *Institute of Computer Science, Polish Academy of Sciences*
Jon Arild Olsen, *National Library of Norway*
Pavel Pecina, *Charles University*
Bolette Pedersen, *University of Copenhagen*
David Perez, *Cabinet of the Secretary of State for the Digital Advancement (SEAD)*
Andraz Repar, *Jožef Stefan Institute*
Ayla Rigouts Terryn, *Ghent University*
Eiríkur Rögnvaldsson, *University of Iceland*
Mike Rosner, *University of Malta*
Nancy Routzouni, *Hellenic Ministry of Interior*
Claudia Soria, *Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) - CNR*
Alexandra Soska, *Federal Ministry of the Interior, Building and Community*
Donatienne Spiteri, *Office of the Attorney General, Malta*
Marko Tadić, *Croatian Language Technologies Society*
Carole Tiberius, *The Dutch Language Institute*
Dan Tufiș, *Romanian Academy of Sciences*
Andrius Utka, *State Commission of the Lithuanian Language*
Paulo Vale, *Administrative Modernization Agency (AMA)*
Piet van den Berg, *RINIS Foundation*
Tamás Váradi, *Hungarian Academy of Sciences*
Kadri Vare, *Ministry of Education and Research*
Andreas Witt, *Leibniz-Institute for the German Language*
François Yvon, *CNRS-LIMSI*
Jānis Ziediņš, *Culture Information Systems Centre*
Miroslav Zumrík, *Slovak Academy of Sciences*

Table of Contents

Foreword

Ms Jill Evans
MEP for Wales (1999 - present) and rapporteur of the Report on Language Equality in the Digital Age

I welcome this White Paper by the European Language Resource Coordination (ELRC) and its clear recommendations on how we can achieve sustainable language data sharing to assist in language equality in the digital age.

Language is more than just a tool of communication; language is deeply linked to our culture and identity. The Welsh language is very important to me and to Wales as a nation. There is a Welsh proverb, “cenedl heb iaith, cenedl heb galon” - a nation without a language is a nation without a heart. I have campaigned as an MEP for equality for all of Europe’s languages because just as language is at the heart of a nation, linguistic diversity is at the heart of the EU.

The ever-increasingly digital age presents a threat to many official EU languages but lesser-spoken languages in particular. At the same time it is a huge opportunity. That is why in 2018 I authored an own-initiative report adopted by the European Parliament on language equality in the digital age.

The report recognises that true multilingualism is a significant challenge for the EU, with its twenty-four official languages and more than sixty regional and minority languages in addition to migrant languages and sign languages. It also recognises that language technology can help us meet this challenge.

Language technologies are behind virtually every digital product we use. The ELRC has assisted in the development of language technologies through developing and maintaining language resources. Increasing the quality and quantity of language data is critical for the much-needed development of multilingual technologies and platforms. There is an increasing demand for automated translation for European-wide industries, for instance, but there is currently an insufficient amount of data for such a service.

The EU could become a world leader in language technologies with all the cultural, social and economic benefits that would bring. For that to happen, we must combine initiatives, share resources, ideas, and work together to put them into practice. We have the talent and the experience. Pioneering work is taking place across the EU, and that work must be supported.

That is why one of the recommendations in my report “calls on the Commission to establish a large-scale, long-term coordinated funding programme for research, development and innovation in the field of language technologies, at European, national and regional levels”. This funding programme would take us a step closer to achieving our aims, just as this White Paper and a number of other initiatives do too.

My thanks to the ELRC Consortium and National Anchor Points for this valuable contribution towards language equality. Diolch yn fawr.

Executive summary

Languages are a key part of the rich tapestry of our European culture(s) and identities. While language diversity is a blessing and opportunity, it can also be an obstacle and a challenge. In Europe, national borders are open, but linguistic borders still exist. They hinder the free flow of ideas, knowledge, commerce, people and communication. They may foster language silos and contribute to digital extinction, for example when online digital services only cater to a small number of languages.

Modern language technologies (LT) such as machine translation (MT) can help overcome language barriers while at the same time supporting linguistic diversity (see Section 2). Modern language technologies are based on machine learning (ML). Machine learning is a process where machines improve by learning from high-quality training data. For this, sufficient amounts and the right kind of training data are required. A machine translation system that is trained on texts from the legal domain will therefore not perform well on a news article about a sports event, simply because it has never "learned" to translate such texts. Therefore, training data not only needs to be from the required domain (in-domain data), but also cover all language pairs the MT system is designed to support.

Public administrations and services across Europe have a large need for translation and produce a lot of language data in multiple languages that is public information. This type of language data is invaluable for training MT systems and - according to the Open Data Directive (Directive on open data and the re-use of public sector information 2019/1024/EU) - should be available for reuse. However, to date, public translation data is often not shared nor made available for reuse. This results in a lack of language data for many domains and languages, restricting the quality of translations generated through machine translation.

The European Language Resource Coordination (ELRC) was initiated in 2015 to address this situation. ELRC is successfully collecting language data – language resources – in all official European languages, as well as Norwegian Bokmal, Norwegian Nynorsk and Icelandic, from public services with special focus on bi- and multilingual language data from various domains. The initial purpose was to collect language resources to train eTranslation (former MT@EC), the machine translation service of the European Commission that can also be used free of charge by all public administrations and public services in the EU Member States, Norway and Iceland. The usefulness of language data, however, goes far beyond training eTranslation: Language data is the driving force behind all data-based language technologies. Whenever possible, data collected by ELRC is also made available to the wider public for research and commercial applications: more than 80% of the language resources collected within the ELRC-SHARE repository are re-usable outside ELRC.

In the past years, ELRC investigated the key stakeholders and mechanisms for sharing of language data in EU Member States, Norway and Iceland, including (i) data creators and data holders in each country (ii) translation practices in public services on the national level, (ii) language data sharing infrastructures such as the national Open Data Portals, (iii) language and digital policy, (iv) other relevant stakeholders.

In the course of this investigation, ELRC identified best practices as well as important obstacles (see Section 3) that currently limit language data sharing in Europe, in particular:

Executive summary

- Lack of appreciation of the value of language data
- Structural challenges
- Disposition towards CAT tools and digital skills
- Inadequate language data management practices
- Limited access to outsourced translations
- Legal concerns

Lack of appreciation of the value of language data was identified as the main challenge by far.

Based on the ELRC findings as well as on the best practices observed, corresponding recommendations for the European and national policy level, and the organisational level of the public administrations and services engaged in language data creation and management, were formulated (see Section 4). The main recommendations for *the European and national policy level* include:

- An update of the Open Data Directive (2019/1024/EU) that should reference language data as a high-value data category.
- The conduct of a study on the value of language data to identify and quantify the value of language data for citizens, public administrations and businesses.
- An update of national policies (e.g. national Open Data policy, digital agenda or strategy for Artificial Intelligence) to explicitly support the sharing of language data and language technologies.
- The inclusion of obligatory (language) data management plans in all relevant national funding policies and calls for proposals if not yet included.
- The conduct of national surveys assessing the translation practices in public administrations on all administrative levels.

With regard to the *organisational level*, the most important recommendations address:

- **Translation and Data Management** (in particular the designation of Open Data officers in all public administrations and services, the introduction of general rights management in the data management process, the adoption of translation data management plans, the centralisation of translation workflows, and the adaptation of procurement contracts for translations).
- **Human Capital** (including in particular the provision of technical and legal training for translators and translation managers).
- **IT Infrastructures, Equipment and Tools** (including in particular the provision of computer-assisted translation (CAT) tools, machine translation, data anonymisation methods and tools as well as language data management tools).
- **Translation process / workflow** (including the appropriate licensing of translation data, the identification (and where necessary exclusion) of confidential and personal data and the maximal automisation of the process of translation / language data creation, curation and collection.

In order to enable the successful implementation of the recommended actions across European countries and hence the sustainable sharing of language resources in Europe, future funding schemes should be aligned with the proposed activities provided in this document.

Abbreviations

API	Application Programming Interface
CAT	Computer-Assisted Translation
CEF	Connecting Europe Facility
DGT	Directorate General for Translation
DSM	Digital Single Market
ELRC	European Language Resource Coordination
HLT	Human Language Technology
L1	First language
L2	Second language
LSP	Language Service Provider
LR	Language Resource
LT	Language Technology
ML	Machine Learning
MT	Machine Translation
NAP	National Anchor Point
PSI	Public Sector Information
TM	Translation Memory

Glossary and definitions

Computer-assisted translation (CAT)[1]:
Translation performed by human translators with the help of a variety of computerized tools (synonym: computer-aided translation).

(Human) Language Technology:
Language technology (LT), often also referred to as human language technology comprises computational methods, computer programs and electronic devices that are specialised for analyzing, producing, modifying and translating text and speech.[2]

Language data:
Refers to any textual, audio or audio-visual data produced using human language or data about human language (such as grammars, language models etc.).

Language data creator:
The person(s) or organisation(s) that generate text or speech in digital form. In the context of translation, the author of the source text and the author of the target text (the translator) are the language data creators.

Language Resource:
Sets of language data and descriptions in machine-readable form, including written and spoken corpora, grammars, and terminology databases. Language resources can be used to build, improve, or evaluate natural language systems such as machine translation engines.[3]

Intellectual Property Right (IPR) holder:
The person(s) or organisation(s) that hold(s) the right to benefit from the protection of moral and material interests resulting from authorship of scientific, literary or artistic productions. In the context of translation, the IPR holders are the authors of the source and target text, unless otherwise stipulated by specific agreements/contracts.

Less resourced or low-resource language:
A language can be considered a less or low-resource language when it is less studied, a minority language, a less privileged language or a language for which few linguistic resources such as training data are available.[4]

Meta data:
Data about the data, i.e. structured description of a dataset with its properties (e.g. title, author/publisher, description of the content, size, topic, IPR holder etc.)

Open Data:[5]
Refers to data which is open in terms of: access, redistribution, reuse, absence of technological restriction, attribution, integrity, no discrimination.

Open Government Data:
Is data produced or commissioned by public bodies or government controlled entities which is made accessible, can be freely used, reused and redistributed by anyone.

Public Sector Information:
Is information generated, created, collected, processed, preserved, maintained, disseminated, or funded by or for the Government or public institution.

Textual data[6]:
This term refers to systematically collected material consisting of written, printed, or electronically published words, typically either purposefully written or transcribed from speech or from other modalities (e.g. sign languages).

Translation Memory (TM):
A database of previously translated text segments (i.e. sentences, paragraphs, headings etc.). A TM stores the source segment and its corresponding translation, the target segment, in pairs. These pairs are called “translation units” (TUs).

1 Azzano, Dino: Placeable and localizable elements in translation memory systems, A comparative study, 2011, https://edoc.ub.uni-muenchen.de/13841/2/Azzano_Dino.pdf.
2 Uszkoreit, Hans: *What is LT?*, 2010.
3 European Commission: *eTranslation Definitions*.
4 Palmer, Alexis: *Computational Linguistics for Low-Resource Languages*, 2011.
5 European Data Portal: *Analytical Report 9: The Economic Benefits of Open Data*, p. 7f., 2017.
6 Benoit, Kenneth: “Data, Textual”, in: *International Encyclopedia of Political Science*, 2011.

Readers' guide

The ELRC White Paper "Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe – Why Language Data Matters" provides an analysis of European practices for sharing language data and the corresponding challenges, as well as clear recommendations for policy-level decision makers on how to overcome these challenges.

Each challenges section corresponds to a specific recommendations section. Since some recommendations address several challenges, they will be reiterated in each section to cater to readers that are only interested in the recommendations for a specific challenge. An overview of all recommendations is available in the "Conclusions and Outlook" chapter.

The Annex of this White Paper contains a Country Profile for each participating CEF country, which details vital information about the translation practices in public administrations, its language policy, a short description of the stakeholders relevant to the sharing of language data, the challenges of sharing language data in public administrations, and a corresponding action plan to address and overcome these challenges. Each Country Profile is a self-contained document supplemented by the main body of the White Paper and the other Country Profiles. The level of detail may vary from one profile to another. Unless otherwise stated, all information refers to the situation at the national level of the particular country.

Disclaimer:

Please note that the information is based on the experiences of the ELRC consortium and Language Resource Board including individual investigations and expertise as well as information derived from public reports, national strategies and other types of publications. The solutions and actions suggested in this report reflect the expertise of the ELRC consortium and the Language Resource Board and are not national initiatives unless clearly indicated. The information provided cannot be considered complete. Even though great care was taken to provide valid information and references, their accuracy cannot be guaranteed nor responsibility taken for them.

1. Introduction

In April 2015, the European Language Resource Coordination (ELRC) action was launched by the Connecting Europe Facility (CEF) programme of the European Commission. ELRC is a core service of the CEF eTranslation building block and acts as a facilitator for identifying and collecting language resources that can enhance the performance of the European Commission's machine translation service CEF eTranslation. The language resources collected will allow eTranslation to cover the variety of languages and domains relevant to public administrations and digital services in all EU Member States, Norway and Iceland, and thus overcome existing language barriers in public services across Europe. One major goal of ELRC is to highlight the value of language data for the development of language technologies (LT), with a strong focus on multilingual technologies. ELRC has also developed and it maintains the ELRC-SHARE repository, where currently over 1.300 language resources are stored, of which 80% are reusable beyond eTranslation, e.g. by language technology researchers or industry.

The already collected language resources, however, are only the tip of the iceberg and a lot of existing language data is still not shared. In the course of the ELRC initiatives, a number of circumstances were identified by the ELRC Consortium and the ELRC National Anchor Points (NAPs)[7] that limit the sharing of language data and hence prevent contribution of relevant language resources.

The analysis in the White Paper is based on this collected pool of information and knowledge provided by the NAPs and the ELRC Consortium, national reports and other types of publications. After the identification of the main challenges, these were ranked by the NAPs. In addition, the NAPs specified the current translation practices such as in-house translation vs. outsourcing of translations and the use of CAT tools and MT in the translation process in the public administrations in their country. The results are presented in dedicated figures in Section 3.

The aim of this document is to illustrate these challenges and to provide clear recommendations to national and European decision-makers on how to optimize the sharing of language data in Europe. Language data hold an indispensable asset for the European language technology industry, which represents a significant share of the global language industry market worth 20.7 billion EUR[8], as well as for European citizens and public services, who are the actual users of language technologies.

7 The ELRC National Anchor Points constitute the Language Resource Board: http://lr-coordination.eu/anchor-points.

8 Slator: *Slator 2019 Language Industry Market Report*, 2019.

2. Language data: Value and Status quo

In a world that becomes more and more interconnected, it is crucial to be able to communicate with one another even when there is no common language. This is particularly true for Europe. The European Parliament states in its resolution on "Multilingualism: an asset for Europe and a shared commitment" that "linguistic and cultural diversity have a significant impact on the daily life of citizens of the European Union", recalling "that the importance of multilingualism is not confined to economic and social aspects and that attention must be paid [...] to the importance of translation, both literary and technical, in the lives of citizens and for the EU's long-term development",[9] thus acknowledging the value of European languages as an inherent part of our cultures and at the same time addressing that active efforts need to be made towards keeping Europe multilingual. Linguistic diversity constitutes a challenge for cross-border communication that affects equally citizens, public administrations, and businesses. Language technologies are able to overcome these barriers through a variety of different approaches to and applications of Artificial Intelligence (AI). The report on *Language equality in the digital age*[10] issued by the European Parliament in September 2018 confirms that the availability of language resources and language tools is crucial in this situation, insisting "on the need to make better use of new technological approaches, based on increased computational power and better access to sizeable amounts of data, in order to foster development of deep-learning neural networks, which make human language technologies (HLTs) a real solution to the problem of language barriers".[11]

While smaller or minority languages are the ones to gain most from language technologies, tools and resources are particularly limited for them. This is also true for languages with different varieties, for example Austrian German or Cypriot Greek. Usually, the language variety representing the majority of speakers has a higher number of language data, and is therefore better covered by machine translation. However, the use of (administrative) terminology may differ significantly between these language varieties and may cause misunderstandings. The report states that "common European values of cooperation, solidarity, equality, recognition and respect should mean that all citizens have full and equal access to digital technologies, which would not only improve European cohesiveness and well-being but also enable a multilingual Digital Single Market"[12].

2.1 Language Technologies in the context of Artificial Intelligence

Over the past few years, Artificial Intelligence and language technologies have made huge strides: Computers can translate between languages, understand and produce spoken language, answer questions, engage in conversational interactions, read and produce texts, sense and see their environment, drive cars, play games, analyse X-ray and tomography scans, make predictions from data, and make decisions in uncertain environments. In large part, this progress is due to improvements in machine learning including deep neural networks.

Machine translation (MT) is one way of addressing the huge demand for cross-border communication the European Union and its Member States are facing every day. MT systems can translate large amounts of text in a

9 European Parliament: *Multilingualism: an asset for Europe and a shared commitment (2008/2225(INI))*, p. 63.
10 European Parliament: *Report on Language Equality in the Digital Age (2018/2018(INI))*, 2018.
11 European Parliament: *Report on Language Equality in the Digital Age (2018/2018(INI))*, 2018, p. 8.
12 European Parliament: *Report on Language Equality in the Digital Age (2018/2018(INI))*, 2018, p. 6., p. 9.

2. Language data: Value and Status quo

very short time in multiple languages. They can be used to understand the theme of a text in a context where human translated quality (or: a perfect human translation) is not crucial, or be integrated as an application programming interface (API) in a website or online service, or be used as a tool for professional translators, to name just a few examples.

The European Commission's machine translation service eTranslation, for instance, is already successfully used in public online services such as the Internal Market Information (IMI) System or for the translation of tender notices in Tenders Electronic Daily (TED). It is also used to translate metadata in the European Data Portal to allow for cross-lingual searches or by professional translators of the European Commission. Overall, almost 80 information systems are actively using CEF eTranslation. However, as the report on Language equality in the digital age[13] and the recent CEF Market Study[14] stress, there is still a significant number of scenarios, domains and services where automated translation is needed, but the quality of the output is not yet adequate due to insufficient amounts of training data.

Furthermore, the possibilities and actual application areas for language technologies go far beyond pure machine translation: Search engines make use of language technology through information extraction, cross-lingual searches and web crawling. Automatic Speech Recognition (ASR) and speech synthesis are used in many scenarios such as customer service, intelligent assistants and any kind of verbal human-machine interaction. Even question and answering applications, chat bots, Internet of Things (IoT) and robotics are based on language technology.

However, as indicated earlier, modern language technologies require data to learn how to perform complex tasks (such as translations, dialogues, speech recognition, etc).

2.2 The value of language data

With regard to machine translation, the required data consists of translations produced by human experts. To be able to translate well, computers need sufficient (often large) amounts of the right kind of high-quality language data. For example, to be able to translate weather reports well, the computer needs to have been trained on weather report data that are previous translations of weather reports. The same holds for all other possible translation domains and topics: public service documents, customer reviews, travel reports, legal documents etc.

Language data refers to any textual, audio or audio-visual data produced using human language or data about human language (such as grammars, language models etc.). The collection of one or more language data sets grouped together according to certain criteria constitutes a language resource. In the context of ELRC, the following types of language resources are distinguished:[15]

- **Corpora, a set of documents or a text in one or more languages such as:**
  - Official documents in the official administration (decisions, legal acts etc.)
  - Paper and newspaper articles, reports, magazines, newsletters, etc.

[13] European Parliament: *Report on Language Equality in the Digital Age (2018/2018(INI))*, 2018.
[14] Directorate-General for Communications Networks, Content and Technology (European Commission): *Final study report on CEF Automated Translation value proposition in the context of the European LT market/ ecosystem,* 2019.
[15] The basic structure is based on Aguado de Cea et. al: *Inventario de Recursos Lingüísticos de la Administración Pública para Traducción Automática, p. 16*, 2016.

  - Translated documents with the original document, (aligned/comparable)
  - Translation Memories (i.e. aligned text segments in the source and target language)
- **Grammars and models**
  - Grammars (e.g. a set of rules that formalize a language)
  - Language and translation models
- **Lexical and Conceptual resources**
  - Terminologies
  - Glossaries
  - Thesauri, Wordnets, Ontologies

Already in 2005, Sweden passed a bill addressing its language policy. The bill states that the development of language technologies is crucial for the Swedish language and that large text and speech databases are central to promoting good development of language technologies.[16] Most recently, the Maltese government announced as part of their AI strategy that:

"the country will make crucial investments in the development of Maltese language resources and tools. The investment will enable computers to be able to process, understand and generate Maltese text and speech, and AI solutions to be developed and accessible in both of Malta's national languages and become a part of everyday life. The Maltese language resources and tools will also have a ripple effect on many sectors, including education and health. It will be possible to create more advanced software to process data in Maltese in a more efficient and accurate manner and to create education tools for the Maltese lan-guage that make use of these underlying language technologies."[17]

Similarly, the French and the Danish government are fully aware of the value of data: The French "AI for Humanity" strategy confirms data as a key competitive advantage in the global AI race and hence pursues the development and implementation of an aggressive data policy.[18] The Danish government made the collection and provision of "more and better data", including in particular a large Danish language resource, one of four focus areas in their national strategy for Artificial Intelligence.[19]

2.3 Open Data: The value of sharing language data

The actual value of language data and the open sharing of such data (e.g. as Open Data) extends far beyond their importance for developing language technologies: A comparative study[20] conducted to assess the Open Data potential for Germany estimates that the potential of Open Data lies between 2.5 billion EUR per annum (conservative estimation by DotEcon, 2006) and 131.1 billion EUR per annum (optimistic estimation by McKinsey, 2013).[21] Moreover, the overall " value of the European data economy may increase to 739 billion EUR by 2020, representing 4% of the overall EU GDP"[22]. In addition, several indirect

16 Cf. Sveriges Rigsdag: *Kulturutskottets betaȁkande 2005/06:KrU4*, 2005.
17 Parliamentary Secretariat for Financial Services, Digital Economy and Innovation, Office of the Prime Minister: Malta: *The Ultimate AI Launchpad, A Strategy and Vision for Artificial Intelligence in Malta 2030,* 2019, p. 48; see also Maltese country profile.
18 French Strategy for Artificial Intelligence, AI for Humanity: https://www.aiforhumanity.fr/en/.
19 The Danish Government: *National Strategy for Artificial Intelligence,* 2019, p. 33 ff.; The language resource shall consist of spoken language, written language and word and term bases (cf. ibid. p. 36).
20 Konrad Adenauer Stiftung: Open Data. *The Benefits, Das volkswirtschaftliche Potential für Deutschland*, 2016.
21 Konrad Adenauer Stiftung: Open Data. *The Benefits, Das volkswirtschaftliche Potential für Deutschland,* 2016, p. 56 ff.
22 European Commission: *Building a European Data Economy.*

2. Language data: Value and Status quo

economic benefits are associated with Open Data, including the development of new goods and services, increased efficiency in public services, time savings for users of applications using Open Data, and growth of related markets.[23] Similarly, in the Spanish Plan for the Advancement of Language Technology, "the extraordinary potential value" of language resources created by the public sector is seen as "outstanding opportunity" for the language technology industry.[24]

In order to facilitate the sharing of any data deriving from the public sector in Europe and to enable additional re-use and added value, the Open Data Directive (Directive on open data and the re-use of public sector information 2019/1024/EU[25]) provides the legal framework enabling all Member States to make public sector information available. Public sector information is defined as any data "generated, created, collected, processed, preserved, maintained, disseminated, or funded by or for the Government or public institution" [26].

As such, European public services play a dual role in helping to overcome language barriers in Europe. First, by translating and providing multilingual content and second by sharing the language resources they create. In doing so, they can help to develop language technologies for areas where the demand for translations is much bigger than the translating power of human translators. For example, in the context of the EU Council Presidencies, journalists need access to national news of the hosting country. With the help of the EU Council Presidency Translator, an online machine translation system, text snippets, documents or entire websites of local news can be translated within seconds.[27]

However, despite the attested value of language data generated by the public sector, and despite the fact that the Member States have acknowledged the value of Open Data and have set up infrastructures for making them publicly available (i.e. through their national Open Data Portals), many Member States are still unable to share a significant part of their language data or language resources, because the full potential of Open Data can only be used when the right framework and infrastructure is in place.[28] The following section will hence provide details of the challenges identified in the course of ELRC that were found to hinder the sustainable sharing of language resources within and from European public services.

---

23 Cf. European Data Portal: *Analytical Report 9: The Economic Benefits of Open Data*, https://www.europeandataportal.eu/sites/default/files/analytical_report_n9_economic_benefits_of_open_data.pdf, 2017, p. 11s f.

24 Plan for the Advancement of Language Technology, 2015, p. 7.

25 The Open Data Directive entered into force on 16 July 2019 and replaced the Public Sector Information Directive from 2003 (2003/98/EC) which was amended in 2013 (2013/37/EU).

26 European Data Portal: Analytical Report 9: The Economic Benefits of Open Data, p.7, 2017.

27 EU Council Presidency Translator: https://presidencymt.eu/#/text.

28 Cf. Konrad Adenauer Stiftung: *Open Data. The Benefits, Das volkswirtschaftliche Potential für Deutschland*, p.55, 2016.

29 The challenges/objectives are: Undervalued perception of language data; structural challenges; legal concerns, limited access to outsourced translations; inadequate language data management practices, and if applicable a sixth country specific challenge.

3. Challenges for sustainable language data sharing in European public services

Overall, the following circumstances were found to negatively impact or limit the sharing of language data from and among European public services:

- Textual data is not perceived as a valuable asset
- Structural challenges
- Disposition towards CAT tools and lack of digital skills
- Lack of adequate language data management
- Access to translation memories of outsourced translations
- Legal constraints

3.1 Textual data is not perceived as a valuable asset

One of the biggest challenges for collecting and sharing language data is the common perception both by organisations and policy makers that textual data has no added value. If something has no or little value, no efforts will be made and no resources will be allocated to managing and curating it. The belief that language data has no added value not only holds true for translators, who, rightly so, consider translation as a mental process whose product is per se only valuable as an instance of creative writing. But it also holds true for Open Data officers and managers who do not regard textual data produced by the public sector as Open Data and therefore make no or very little effort to collect textual data. This is exacerbated by the fact that language data is not declared an important data category in the new Open Data Directive.

To identify the most pressing challenges, the ELRC National Anchor Points were asked to rank a number of challenges[29] and respective objectives that constitute the biggest challen-

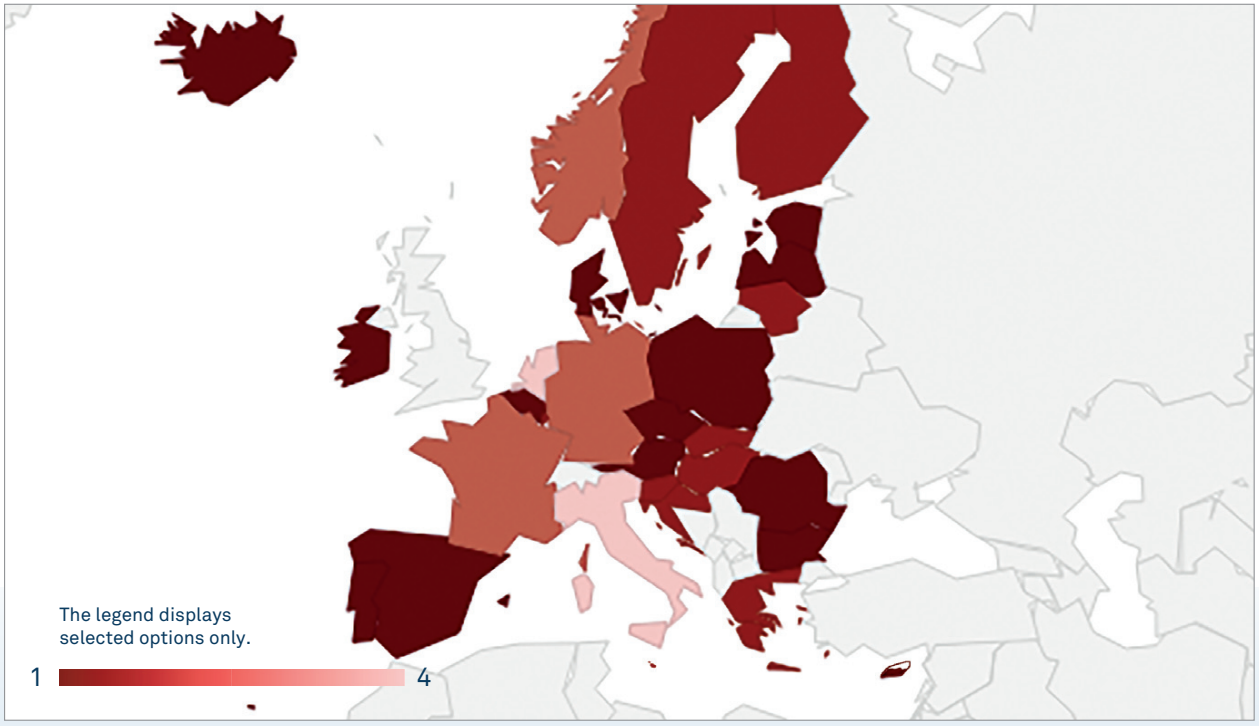


Figure 1: Undervaluation of language data as main challenge across Europe
1 = Undervalued perception of language data is considered the main challenge
6 = Undervalued perception of language data is considered the least important challenge

3. Challenges for sustainable language data sharing in European public services

ges for language data sharing in their country. As shown in Figure 1, 14 out of 29 countries identified the issue of undervalued language data as the main challenge and 10 countries as the second most important challenge that needs to be overcome in order to create sustainable national infrastructures for sharing language data. The figures indicate that the perceived low value of language data is not a country-specific challenge, it is a challenge for Europe (CEF countries) as a whole, for highly digitized countries such as Denmark, technology-affine countries such as Ireland, countries that have a financial plan for the uptake of language technologies such as Estonia and Spain, countries that are founding members of the European Union such as Belgium or countries that joined the EU in this millennium such as Bulgaria or Romania. Especially for countries with a small number of speakers such as the Baltic countries, Malta, Cyprus, Denmark, Greece, Iceland and Norway, where consequently less digital data is being produced, this issue can be a serious stumbling block for the further development of language technologies.

3.2 Structural challenges

In the course of ELRC, several structural issues were identified that complicate the flow of language data. As shown in Figure 2, although 19 out of the 29 participating countries have a national language policy, only 11 countries are addressing language technologies explicitly and only 8 of those countries have a dedicated financial plan for supporting language technologies (Estonia, Iceland, Latvia, Norway, Slovenia, Spain, Sweden and Ireland). While in Ireland, language technologies are only addressed to a limited extent at the moment, the Irish government will launch a Digital Strategy for Irish in 2020.

In many countries, the lack of a language policy addressing multilingualism and translation needs leads to a number of associated

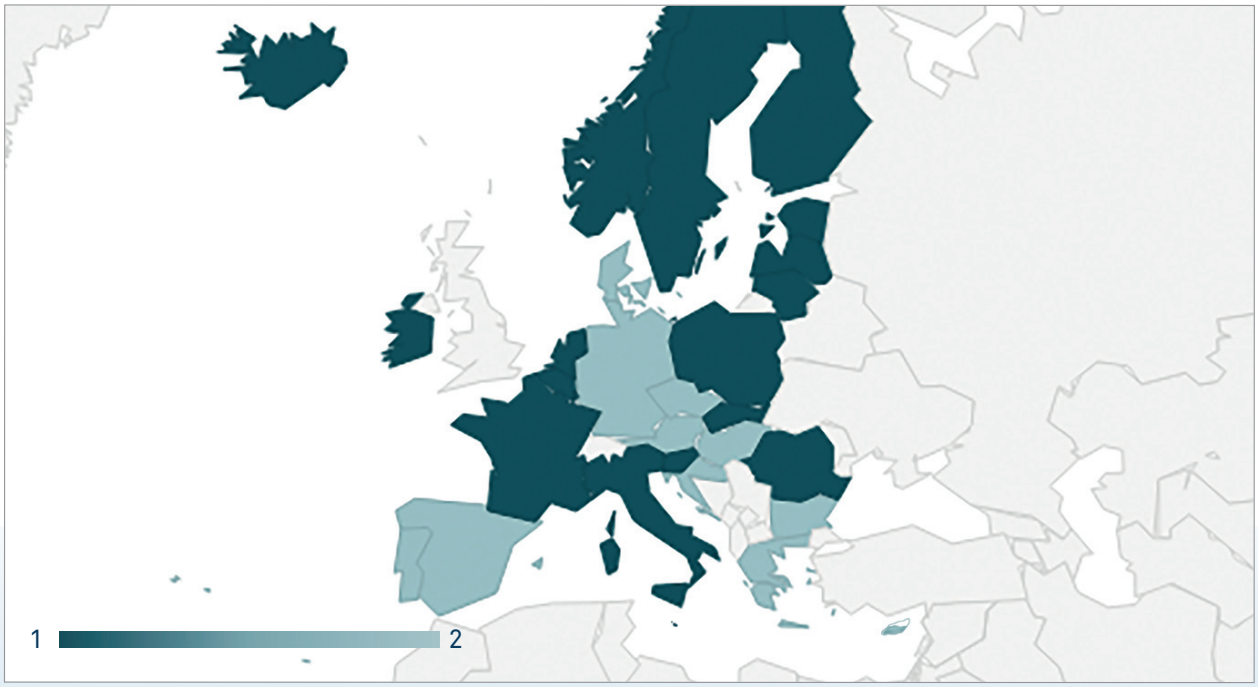


Figure 2: Language policies across Europe
1 = Country has a language policy; 2 = Country has no language policy

structural difficulties that negatively impact the sharing of language resources:

- Investing in language technologies and/or in managing and sharing language data is not a priority in that respective country. Similarly, without a dedicated policy to support multilingualism, public administrations do not see the need to invest in multilingual digital online services.

- Languages, multilingualism and translation may be addressed by various public bodies, but no institution or ministry is coordinating the endeavours as part of their portfolio. Most countries, with the exception of Latvia, Norway and Sweden, have not mandated an institution to collect language data from the public sector. Consequently, there is no decision maker at the national level who can initiate necessary changes at that level, which poses a significant challenge to many countries (see e.g. Austria, Belgium, France, Italy). In Hungary, for instance, the matter of releasing data could not be solved at a hierarchical level below the level of the Secretary of State, which in the end made the endeavor of sharing language data impossible.

- At national level, there is no coordinated approach to meet the translation demands. Only very few exceptions (e.g. Finland, Germany) were found.

As such, the general lack of regulations concerning multilingualism actually leads to a number of subsequent issues at the organisational level of public administrations and services that produce language data, namely the lack of efficient translation and procurement processes, insufficient language data management and legal uncertainties. These issues will be discussed in greater detail below.

3.3 Disposition towards CAT tools and lack of digital skills

It is widely accepted that the use of computer-assisted translation (CAT) tools in the translation process significantly improves the translator's performance, speed, consistency and accuracy. CAT tools, however, are not only indispensable for the translators themselves. They are also a critical prerequisite for the creation of high-quality bi- or multilingual language resources, hence they are an asset for the LT community as a whole. The output of CAT tools is translation memories which are the desired and required language data input for MT systems, as they are, in LT jargon, "MT-ready", meaning that they require little or no pre-processing before they are fed into the MT system.

Figure 3 shows that with regard to the use of CAT tools in public administrations in Europe, only two countries (namely Germany and Slovenia) out of 29 stated that “All translations are carried out with the aid of CAT tools (language services, translation professionals and other translating staff members)”. This is in great contrast to the practice adopted by language service providers (LSPs) and freelance translators: 15 out of 29 countries indicate that all LSPs and freelance translators use CAT tools.

3. Challenges for sustainable language data sharing in European public services

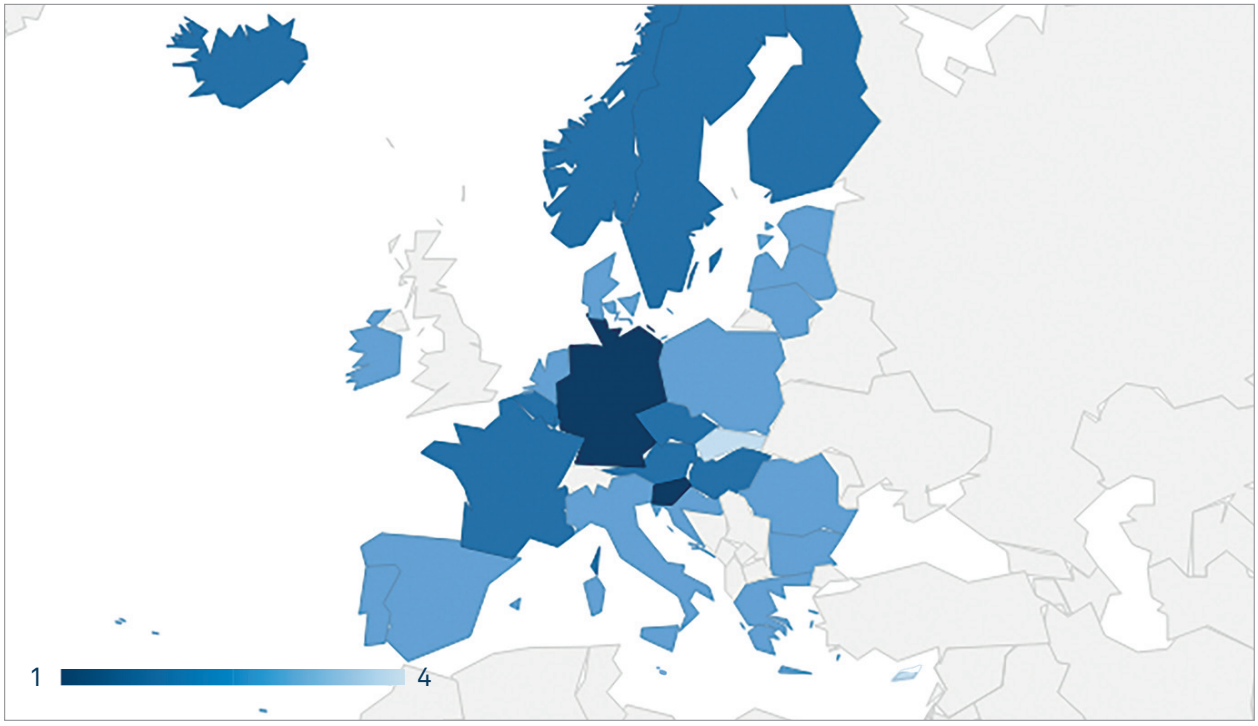


Figure 3: Use of CAT tools by public administrations in Europe

1 = All translations are carried out with the aid of CAT tools (language services, translation professionals and other translating staff members)
2 = It is common practice that translation services/translation professionals use CAT tools
3 = Only single translation services or translators use CAT tools
4 = No use of CAT tools

Similarly, 8 countries indicated that only one or several public administrations have an MT API (Figure 4). To have a machine translation API integrated into the translation process is not a common practice in any country yet, whereas the use of machine translation in general is very frequent.

19 out of 29 countries stated that free online machine translation services were used despite the potential security breaches. In practice, this figure might be even higher. To avoid this, some ministries in Germany introduced rules and recommendations for the use of free online translation tools.

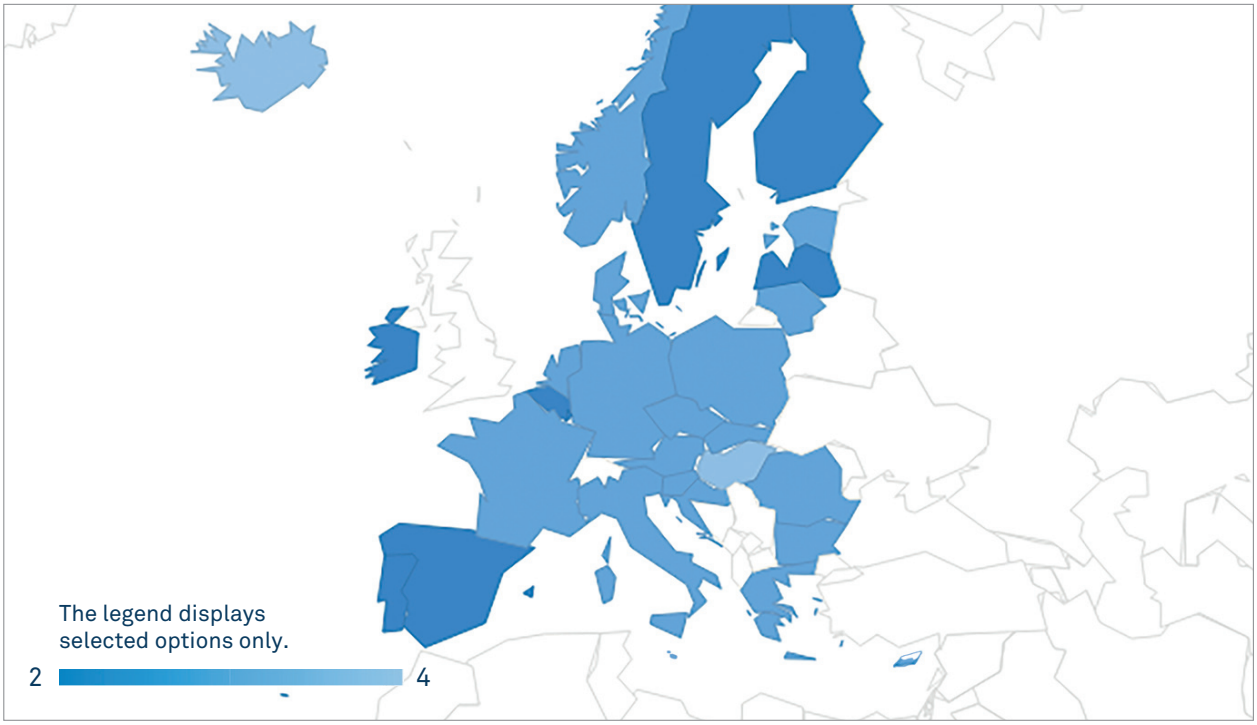


Figure 4: Use of machine translation (MT) by public administrations
1 = Most translation services have a MT API integrated into the translation process
2 = One or several translation services have a MT API
3 = No MT API but use of freely available MT online services
4 = No use of MT

As such, the situation with regard to the use of CAT tools and MT and the possibility to create high-quality language resources is very diverse in Europe: While some public administrations like the Department of Culture, Heritage and the Gaeltacht in Ireland or the Culture Informations Systems Centre in Latvia even developed their own machine translation system and/or multilingual e-service platform (including speech recognition and synthesis), other public administrations in the same country have not adopted the use of computer-assisted translation tools in their translation process.

Several challenges were found to prevent the use of CAT tools in public administrations as part of the translation process, and hence to significantly limit the creation of high-quality language resources:

- A negative disposition towards technology in general and the use of CAT tools in particular prevents the application and adoption of new language technologies (and hence the creation of language resources) in public administrations. While in most countries a fear of human translators becoming obsolete due to language technologies was stated, in some countries (e.g. in Bulgaria and Italy), a general resistance to new technologies was found. In other cases (e.g. in Lithuania), insufficient knowledge about MT (and how language data trains the system) lead to reluctance to share translations because of fear of defective translations.

3. Challenges for sustainable language data sharing in European public services

- Many translators are not sufficiently educated about computer-assisted translation tools and machine translation or language technology in general, and hence lack the required digital skills and understanding. For instance in Croatia and Slovakia, little awareness of the availability of technical solutions that could improve the translation process was demonstrated as well as a generally slow adoption of technologies. In other countries (e.g. Ireland, Greece and Romania), an unskilled or non-use of CAT tools was observed. The Irish Technology National Anchor Point (T-NAP) held a training workshop on the use of CAT tools and MT post-editing in June 2019. Less than 20% of the translators (public servants and freelance translators) had used CAT tools before, clearly indicating the big demand for this kind of training. In Italy, the average age of public servants in ministries is over 54 according to latest figures.[30] Some of these public servants and translators have never received training on how to integrate the latest technologies in their translation workflow. Furthermore, the use of word processing programmes for translating (e.g. in Romania) was reported. Some other public administrations simply need to be trained on how to integrate CAT tools in their translation workflow (e.g. in Ireland). Last but not least, several countries that are familiar with using CAT tools simply lack the awareness of the secure use and secure training of MT: The implications and consequences of using free online translation services are not evident to a surprisingly high number of frequent users, translators and other public servants alike and lead to misuse and potential security breaches and infringements, e.g. GDPR or copyright.

- Finally, the lack of technical and, most importantly, financial resources for providing CAT tools, MT or technologies for preparing and processing language resources before sharing them (e.g. in Bulgaria, Cyprus and Portugal) was also identified as a major obstacle. In fact, the acquisition and provision of CAT or MT tools is considered very expensive, and in many countries, public services and administrations lack the necessary resources to acquire them.

3.4 Lack of adequate language data management

A subsequent issue arising from the aforementioned fact that language data is not considered valuable, is the fact that organisations do not have an adequate process for managing translations/language data - or in most cases entirely lack any defined data management plan that covers translations/language data. The result is that the translations/language data produced by these organisations are largely unshareable for various reasons:

- Even when translation services do use TMs and other CAT tools, and therefore are producing language data in the most useful format (i.e. TMX and TBX files), their translation management may not allow for easy data sharing. For example, when a translation memory contains both unofficial and confidential documents and official publications, the whole TM cannot be shared although the official publication falls under the Open Data Directive.

- In other cases, certain translations are not added to TMs in the first place. For example, if the documents are translated for specific occasions and are therefore unlikely to be reused, they are not added

[30] See Italian country profile.

to TMs. This also applies if the format of the text is considered unsuitable for a TM system. Alternatively, documents may not be translated in the typical sense, but rather be transcreated and therefore not added to TMs, as they cannot be aligned on the sentence level.

As such, it is not surprising to find that 9 out of 29 countries consider the establishment of good data management practices in public services as first (3 countries) or second priority (6 countries) (see below, Figure 5).

With regard to management of translations and language data, typical insufficiencies that were observed include:

- Lack of knowledge about adequate language data management practices in translation services/translation departments of public administrations in general (e.g. in France and Romania).
- Lack of a translation workflow or management of translation data that allows for central collection, maintenance and curation of language data (i.e. each translator or even department manages their own translation needs) (applicable to most countries).
- Translations and language data are not produced, managed or filed in a way that allows for easy data retrieval and re-use, in particular:
  - Missing indication(s) about the inclusion of personal, confidential or copyrighted data in the document (applies to most countries).
  - Missing indication(s) that TMs contain unofficial or confidential documents and publications (e.g. in Finland and France).
  - Missing alignment of original texts and their translations (connection between the source and target text in the metadata is not indicated).

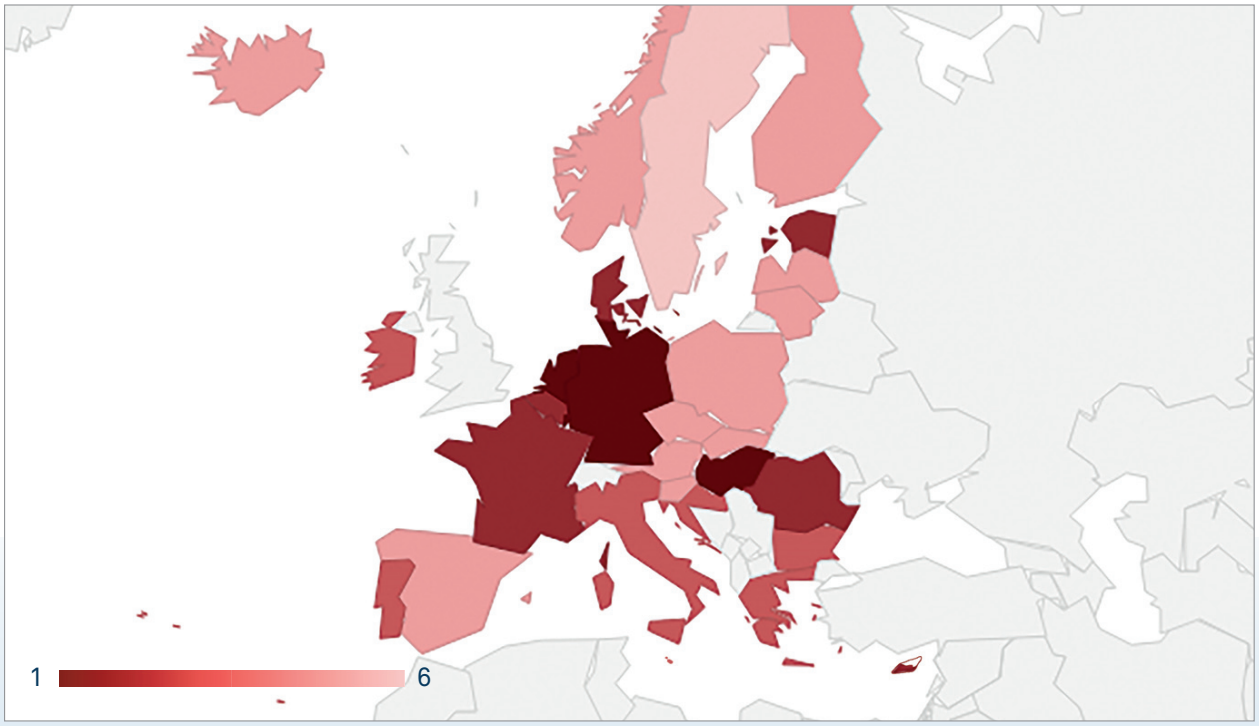


Figure 5: Improving language data management practices
1 = Improving language data management practices is considered the main challenge
6 = Improving language data management practices is considered the least important challenge

3. Challenges for sustainable language data sharing in European public services

– Missing indication(s) of terms and conditions under which the document can or cannot be reused (applies to most countries).

- Public administrations and services in most countries do not have guidelines for translators regarding the nature and treatment of confidential and personal information (one exception is Finland).

- Lack of knowledge of technical as well as translation staff in public administrations and services about anonymisation methods and practices to process TMs in case personal data is contained in the document (e.g. in Belgium and Bulgaria).

3.5 Access to Translation Memories of outsourced translations

The way public administrations deal with their translation demands varies greatly across Europe. Figure 6 below indicates whether translations are outsourced or carried out in-house (either in dedicated translation services and by staff translators or by staff members who only occasionally translate texts).

As Figure 6 shows, public administrations in all countries have one aspect in common: they outsource translations either to language service providers or independent freelance translators.

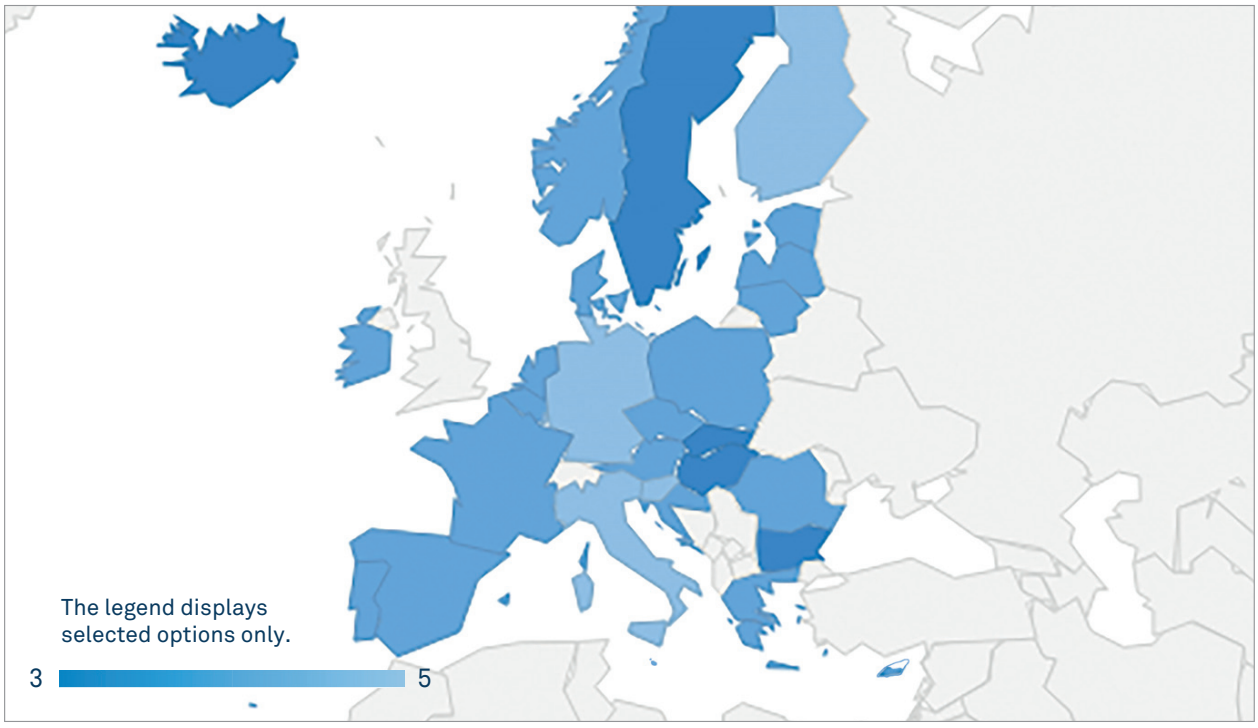


Figure 6: Outsourcing of translations vs. in-house translation

1 = All translations are outsourced via central purchasing body
2 = All translations are independently outsourced
3 = Mostly outsourcing of translations through central purchasing body, only single ministries have in-house translation services.
4 = Mostly independent outsourcing of translations, only single ministries have in-house translation services
5 = More than 50% of translations are carried out by language services and translation professionals in-house; the rest is outsourced

Although recent studies conducted by the NEC TM Data Project[31] have estimated that 300 million EUR are spent on translation services procurement in the Member States of the European Union, the procurement framework does not allow for the reuse of translations by the contracting body.

The predominant procurement practices in public administrations and services in the CEF countries do not foresee that the translation memories and any other by-products of the translation process are shared with the contracting authority along with the translation.

As Figure 7 below indicates, 13 out of 29 countries state that TMs are not requested back at all and 11 out of 29 countries claimed that only some public administrations request back TMs and other by-products of the translation. In Ireland, for example, after the local ELRC and ELRI[32] workshops, an increase of the practice to request the TMs along with the translation was noted.

Overall, only five countries, namely Finland, France, Germany, Luxembourg and Poland indicated that TMs are requested back for most translations.

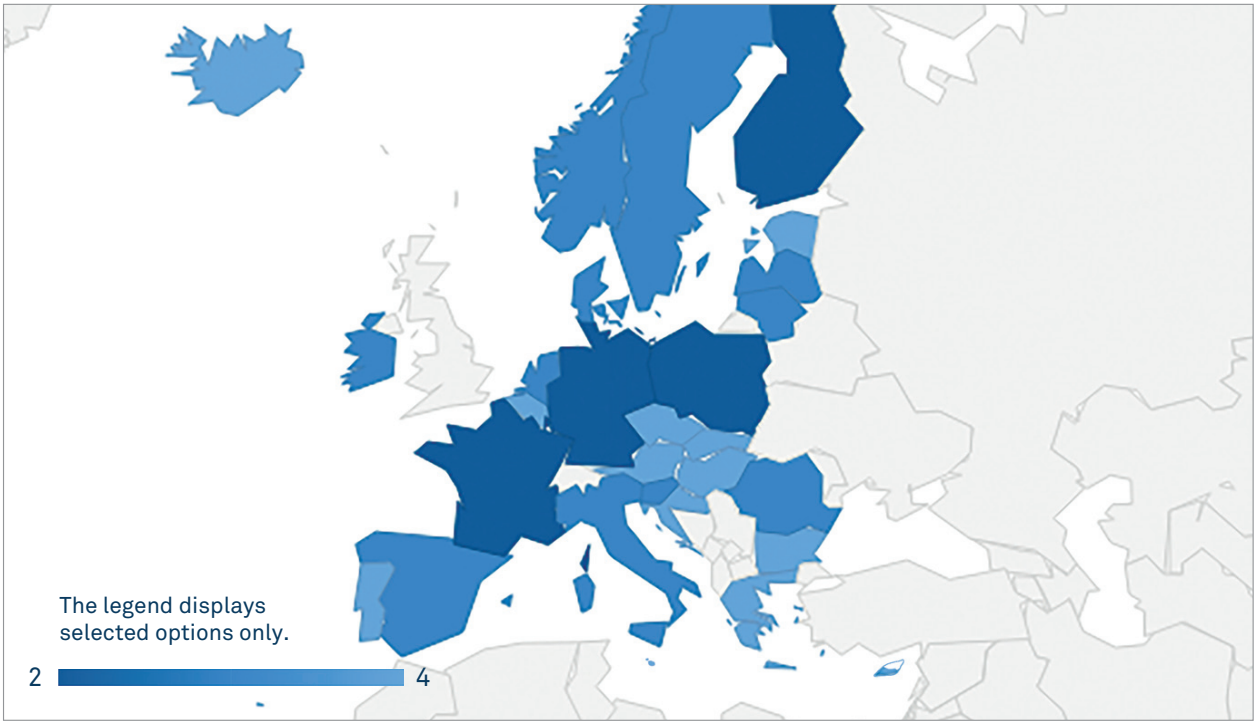


Figure 7: Requesting of TMs for outsourced translations

1 = TMs and any other by-product of translation are requested back by default
2 = TMs and other by-product of translation are requested back for most outsourced translations
3 = Some public administrations request back TMs and/or other by-product of translations
4 = TMS or any other by-product of translation are not requested back

[31] National and European Central Translation Memory Data Project: https://www.nec-tm.eu.
[32] European Language Resource Infrastructure: http://www.elri-project.eu; ELRI in Ireland: https://elri.dcu.ie/en-ie/.

3. Challenges for sustainable language data sharing in European public services

Furthermore, many public administrations do not negotiate the intellectual property rights status of the work with the external translator and unless the intellectual property right is transferred or a necessary license agreed on in the translation contract, the outsourced translations cannot easily be shared or re-used. Similarly, only few public administrations request back associated side products (e.g. terminologies).

ELRC observed the following reasons:

- There is a general lack of expertise in public administrations on how to procure translations in a way that will make future data sharing easier (e.g. in Cyprus and Sweden).
- Even when it is stipulated in the translation contract that TMs should be transferred back to the contracting authority, this clause may be dropped to obtain a better price from the external translator or LSP (e.g. in the Netherlands).
- Translations are not procured by translators and therefore their usefulness and linguistic value is not apparent.
- Some LSPs do not use CAT tools themselves (e.g. in Romania).
- Some public administrations do not have their own in-house translation service that could use the TMs and therefore see no purpose in collecting TMs (e.g. in Germany and Spain).

3.6 Legal constraints

Another aspect of sharing language data that is crucial for making it a safe and law-abiding exercise, is respecting national legislation for sharing public sector information. Despite the above-mentioned PSI directive, there are some overruling laws concerning Intellectual Property and personal information that make data sharing very difficult. Often, information about the IPR holder is not included in the metadata and is difficult to identify a posteriori. This makes the lawful sharing of these textual data nearly impossible. In addition, sometimes licenses are chosen that do not fit the purpose. Even Creative Commons licenses can be subject to national law (so-called ported versions) unless it is an international CC-license.

The recently updated General Data Protection Regulation (GDPR), although an important step, has also lead to increased insecurity and fear of law infringement and has been used as an argument not to share language data. In Slovenia, potential legal violations are the main reason to not share language data and overcoming these challenges is considered the main objective that needs to be addressed to increase the flow of language data.

As illustrated in Figure 8 below, tackling and overcoming legal concerns associated with the sharing of language resources represents one of the top challenges of Europe's public administrations.

The main legislative challenges of sharing language resources include:

- IPR of outsourced translations belongs to the LSP or freelance translator by default unless stipulated otherwise in the translation contract (applies to all countries)
- Privacy and IPR are not clearly documented or transferred (e.g. in Austria, Iceland, Slovenia)
- GDPR is used as an excuse/fear not to share data (e.g. in Greece, Lithuania, Portugal)
- Insufficient legal expertise about LR sharing leads to reluctance (Czech Republic)
- Copyright of the translations belongs to the translator by default and requires expertise for appropriate licenses (e.g. in Germany, France, Belgium, the Netherlands)

Similar to language data management practices, addressing legal issues a posteriori is often either very time-consuming or not possible since relevant information cannot be retrieved. Therefore, it is crucial that all matters related to copyright, intellectual property rights and privacy become an integral part of the translation process from the beginning.

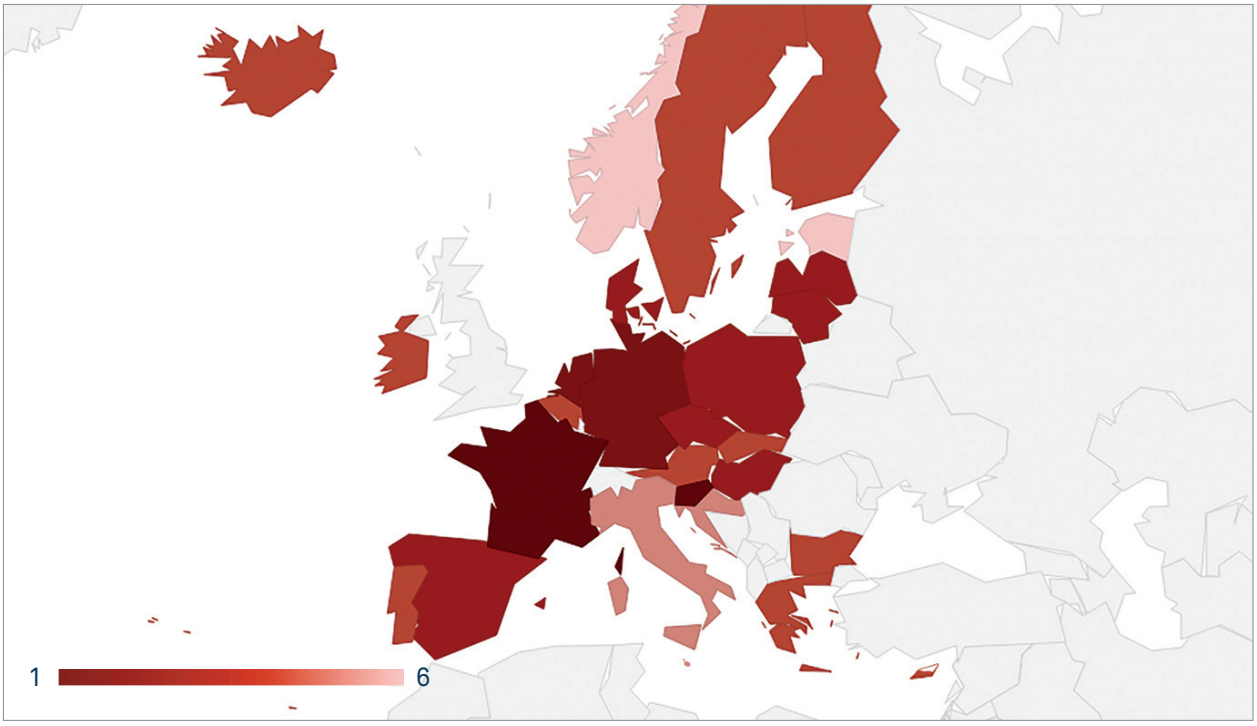


Figure 8: Ranking of challenges: Legal constraints
1 = Legal constraints are considered the main challenge
6 = Legal constraints are considered the least important challenge

4. Recommendations

The following sections provide details of ELRC's recommendations for overcoming the existing challenges of language data sharing in Europe. Our recommendations cover policy, organisational and process level wherever applicable.

4.1 Improving the perception of the value of language data

Many of the challenges identified above are direct or indirect consequences of the fact that language data is often not yet considered a valuable asset and therefore not managed, procured and promoted as such. To overcome these challenges it is important to include language and multilingualism matters and specifically language technologies in both EU and national policy making. Some countries have a dedicated language policy or resolution, others address the collection of language resources in their National Strategy for Artificial Intelligence or Digital Policy. A common European strategy would help to initiate respective changes in national policies. To change the current prevalent perception of the value of language data, the ELRC network recommends the following actions.

Proposed recommendations at the policy level:

- The new Open Data Directive (2019/1024/EU) should consider language resources as one of the high value data sets. The transposition into national law should ensure that public data is released as public domain (or at least under the national government's license that should be as permissive as the CC by 4.0 International license).
- Language data (sharing) should be integrated in the national Open Data policy, digital agenda or strategy for artificial intelligence.
- The role of language data in the digital economy should be emphasized – on all levels – the European, national and organisational level.
- To further leverage on language data, language resource infrastructures should support other types of language data such as speech or audio files or sign language data.
- A comprehensive study on the value of textual data/language data for citizens, public administrations and businesses should be commissioned by the European Commission.

Proposed recommendations at the organisational level:

- Identification or designation of Open Data officers in public administrations and services.
- Raising awareness among translators in public administrations and services of the importance and/or relevance of language data as Open Data in public administrations and services through the respective Open Data officers.
- Raising awareness among translators in public administrations and services on the value chain of language data and on the importance of language resources, especially among translating staff of the organisation.

Proposed recommendations at the process level:

- Websites that only publish public data that falls under the Open Data Directive should be labelled with the respective national Open Government Licence, as this has been defined by the national PSI transposition rules. For instance, the Swedish

Open Data Portal recommends the tag "/psidata"; in Cyprus the standardised Creative Commons Attribution-ShareAlike 4.0 International licence has been selected.

- Develop practical guidelines for translators (and their customers) in public administrations on how to share language data as Open Data.

- Re-design the translation process in public administrations in a way that sharing language data is an inherent part of the creation and curation of language data. Only if sharing language data is enabled and considered at the outset of the translation process, will it allow for easy language data sharing in the end. The overall goal should be to automate the process of language data sharing to a maximum extent in order to minimize the human effort required (see Section 4.4 below for details on increasing the efficiency of the translation process).

4.2 Overcoming structural challenges

As shown in Section 3.2. the lack of a language policy is closely related to several subsequent structural challenges such as:

- the lack of investment in multilingual language technology(ies) and/or language data sharing,

- the lack of a defined decision-maker / clearly defined responsibilities for language data at the national level, and

- the lack of a coordinated approach with regard to language data sharing at national level.

Addressing these challenges hence is an important country-specific endeavour and comprise a variety of recommendations.

Proposed recommendations at the policy level:

- A corresponding policy at national level that supports the sharing of language data should be established. As indicated above (Section 4.1), this could also be done as part of the national Open Data policy, digital agenda or strategy for Artificial Intelligence.

- Matters related to translation and multilingualism including language technologies should be allocated to one ministry or public body that can coordinate and drive the change.

- National infrastructures/repositories for sharing language data of any kind should be created and managed centrally in the particular country.

- A financial plan for funding long-term development in the field of LT is considered crucial to implement changes in the mid and long term.

Proposed recommendations at the organisational level:

- Assigning clear responsibilities with regard to the authorisation of the sharing of language data in public administrations and public services.

4.3 Improving digital skills and technical capacities

As shown in Section 3.3., several factors were found to limit the use and take up of CAT tools, MT and other technologies that enable the creation of high-quality language resources, namely:

4. Recommendations

- A negative disposition of translators towards technology in general and towards the use of CAT tools in particular.

- The fact that many translators are not sufficiently educated about computer assisted translation tools and machine translation or language technology in general, and hence lack the required digital skills and understanding.

- The lack of technical and, most importantly, financial resources within public administrations and services for providing the required technical infrastructures to create re-usable, high-quality translations (e.g. CAT tools, MT or any other technologies for preparing and processing language resources before sharing them).

As such, Europe and its Member States as well as Iceland and Norway need to significantly invest in both technical infrastructure related to the creation of translations and the human capital (namely: translators and managers of translation data) of their public administrations and services.

Proposed recommendations at the organisational level:

- The first recommendation is to significantly invest in the availability and take-up of technologies supporting the translation process in public administrations (in particular CAT tools and MT). Without the consistent use of such tools by translation professionals in Europe's public administrations, the creation of new, high-quality language resources will remain minimal for all languages of countries that do not possess and make use of these technologies.

- Moreover, significant investments in the training of civil servants (in particular translators who are the main producers of language resources) are necessary in order to make them "fit" for the digital revolution. Through training translators and translation managers on the value and use of CAT tools and MT, as well as by raising awareness on the benefits of Open Data (in particular shared language data), consistent improvements could be achieved.

In addition to the purely technical skills that should be acquired as part of the training, it is important to share some key messages. Particular emphasis should be placed on explaining the benefits of computer-assisted translation and machine translation to translators, which is meant to support them in their efficiency (allowing them to translate more in the same amount of time), rather than replacing them. It is also important to raise awareness about the advantages and disadvantages of using free online machine translation systems as GDPR or copyright might be violated. Realistic scenarios of the future work of a translator should be shown and it should be underlined that single translation mistakes will not harm the machine translation output. Last but not least, it is important to illustrate the enormous value that language data have for all applications of data-based language technologies and that good language technology is a prerequisite of many applications of artificial intelligence. Finland for example, is committed to making managers and employees of companies and public institutions in general more familiar with artificial intelligence and offers free and easily accessible online courses "Elements of AI"[33] in Finnish, Swedish and English.

[33] https://www.elementsofai.com

4.4 Improving translation processes to enable language data sharing

As shown earlier (see Section 3.4), the lack of appropriate language data management and a coordinated approach for the creation, collection and curation of language resources produced within the translation workflow has significant impact on the shareability of the produced translation data. Beyond translation data, all language data are valuable, even when their appropriateness for translation memories and machine translation is not obvious, as is the case of data created through transcreation; such data can be used in other language technology applications, for instance cross-lingual summarization or paraphrasing. Although translation practices vary greatly, some suggestions apply to most countries. Recommendations to improve the management of language data include but are not limited to:

Proposed recommendations at the policy level:

- Translation practices in each country on all administrative levels need to be investigated in order to identify and initiate necessary changes: a national survey in all European countries (or alternatively an EU-wide survey covering all European countries) assessing the translation practices on the level of all relevant national public administrations is advisable.

- Following the example of the EU funding instruments, the requirement for a data management plan should be integrated in all national calls for proposals that target the creation of language data and/or improvement of translation processes, including:
  - For any national corpus: the IPR holders should be identified to confirm that the final compiled data set can be made openly available to all users including commercial use to boost the Digital Single Market.
  - Tools: should be released as open sources under a permissive software license.

Proposed recommendations at the organisational level:

- Translation workflows should be centralised in public administrations and services, at least at the institutional level in order to facilitate the collection, curation and sharing of language resources.

- Corresponding plans for the management of translation data should be adopted in all public administrations and/or services that generate language data.

- CAT tools and machine translation should be integrated into the translation workflow in all public administrations and/or services that generate language data (also see above, Section 4.3) Corresponding funding will need to be provided at EU and national level.

Proposed recommendations at the process level:

- Language data management tools, language data management plan templates, and guidelines on how to manage language data should be developed and made available to all public administrations and services with translation needs.

- Protocols for archiving documents, translation data and other types of language data should be defined in terms of:
  - Format
  - Standardized metadata descriptions

4. Recommendations

  - Information about copyright and Intellectual Property Rights
  - Indication of confidential and or personal data

- Anonymisation methods and tools should be developed and tailored to the specific needs and languages of public administrations and services.

4.5 Improving procurement practices for outsourcing translations

As shown in Section 3.5, translation data that is created for public administrations by language service providers and/or external freelance translators through outsourcing, in most cases is not reusable by the contracting authorities because the copyright belongs to the translator by default. Although copyright is subject to national legislation and therefore underlies differences in the Member States, most recommendations provided here are applicable to all countries and will multiply the value and impact of translation data.

Proposed recommendations at the organisational level:

- If and when public administrations outsource translations either through frame work agreements or individual contracts, it should be stipulated that the ownership of the translation, the translation memory and any other by-product of the translation belongs to the public administration. This stipulation should not be negotiable and should be embodied in national law.

- If the IPR ownership cannot be transferred, a corresponding license, ideally an international license, should be stipulated.

- Translations should be centrally procured, at least at the institutional level and the translation data centrally stored (also see above, Section 4.4).

Proposed recommendations at the process level:

- All public administrations should by default request to receive the translation in machine readable formats. Any format is acceptable as long as it allows for re-editing and further processing of the text (e.g.request the DOC file instead of a PDF file). Ideally, any by-product should also be requested, for instance the translation memory files and terminologies, regardless if the public body has an in-house translation service or not, if it uses CAT tools or not, or if reusing these by-products does not seem plausible in the short-term. It is a costless investment for the mid- and long-term.

- The outsourcing practices in general should be tailored to leverage on translation data, in practice, this may require guidelines and training for public servants dealing with public procurement.

- Guidelines and templates for agreements covering the retrieval of translation memories and other by products between public administrations and LSPs / external translators should be developed.

4.6 Adjusting the legal framework to enable language data sharing

In the past, translations were not created with the purpose of making them available to the public for further reuse or even for internal reuse by the in-house translators or to train

an in-domain MT system, for instance. Therefore, necessary actions such as the transfer of copyright or licenses that allow for future reuse were and are often not included in the translation process. With the increasing digitalisation, digital text, like any other type of data, becomes more important as a training base for technologies or as a source of information. To allow for the reuse of language data, legal matters need to be addressed during the whole translation process. ELRC hence proposes the following actions:

Proposed recommendations at the policy level:

- Adopting a standard licence under which translation data in particular and language data in general can be reused.

Proposed recommendations at the organisational level:

- Unless a text is public domain and therefore belongs to "everyone", textual data needs to be licensed, ideally with so-called open licences that allow for commercial and/or non-commercial reuse of the data.

- Rights management should be introduced in the data management process in general, providing legal support for (language) data sharing (also see above, Section 4.5).

Proposed recommendations at the process level:

- Any confidential information needs to be excluded from reusable TMs. Translation data that contains confidential information should clearly be marked (see above, Section 4.5).

- Similarly, information contained in the documents that could link to any identifiable person, needs to be deleted (anonymisation). Translation data that contains any personal information should clearly be marked (see above, Section 4.5).

- Corresponding tools and methods for anonymisation shall be made available to public administrations (see above, Section 4.5)

- Guidelines on safe handling of confidential and personal data should be developed, communicated and followed (see above, Section 4.5).

5. Conclusions and Outlook

At this stage in the information age, the digital revolution is penetrating all areas of our lives and as with any other revolution, it has significantly changed the work and personal lives of people. There are always stakeholders who use the driving changes to their advantage as well as others who do not. But at this moment in time, each and every country, every institution and every individual can still decide to which stakeholder group they want to belong.

The Danish Minister of Culture, Mette Bock, underlines the importance of language technologies for the Danish language and states: "It is quite clear [...] that technologies containing linguistic components play an increasingly important role. It is vital that the Danish language is able to follow this development at a time where language technology gradually becomes part of more and more areas of our lives".[34] We would like to add, it is vital for every country to follow this development and to invest in their own languages and culture.

As part of this White Paper, several important factors limiting the sharing of language resources have been identified (see Section 3) and corresponding recommendations on how to overcome each of the identified challenges have been made (see Section 4). Below we offer recommendations summarizing the actions that were found necessary in order to significantly improve the sharing of language resources in Europe and to unleash the full benefits of language data. As shown below, recommended changes concern both the EU and national policy level. Also, various changes were found necessary at the organisational level of public administrations and public services that create language data, including (i) their translation and data management practices, (ii) their IT infrastructure, equipment and tools, (iii) their translation processes and workflows, and last but not least (iv) their human capital. In order to enable the successful implementation of the recommended actions across European countries and hence the sustainable sharing of language resources in Europe, future funding schemes must be oriented towards supporting the proposed activities listed in the table below. Sharing language data means supporting your language and supporting your language means supporting your country while at the same time building a stronger multilingual Europe.

[34] Cf. European Language Resource Coordination: *New Report on Language Technology for Danish*, 2019.

Recommendations for enabling sustainable sharing of language resources across Europe

PROPOSED CHANGES AT POLICY LEVEL:

European Union

Update Open Data Directive: Consider language resources as one of the high value data sets.

Conduct an EU-wide study commissioned by the EC on the value of language data: Identify and most importantly quantify the value of language data for citizens, public administrations and businesses.

Member State / CEF-affiliated country

Update National policies (Open Data, digital agenda, AI): Integrate language data sharing in national policies.

- The transposition of the Open Data Directive (2019/1024/EU) into national law should ensure that public data is released as public domain (or at least under the national government's license that should be as permissive as the CC by 4.0 International license)
- Emphasize the role of language data in the digital economy.
- Create a financial plan for funding long-term development of language technologies (LT).

Define responsible body for LT / LR sharing at the national level: Allocate matters related to translation and multilingualism (including LT) to one ministry or public body, which coordinates and drives this topic.

National language resource infrastructures: Create and centrally manage national infrastructures/repositories for sharing language data for public administrations and services. Adapt them to include all types of language data (including also speech, audio files or sign language data).

Adjust national funding policies: Integrate the requirement for a data management plan in all national calls for proposals targeting the creation of language data and/or improvement of translation processes, including:

- For any national corpus: the IPR holders should be identified to confirm that the final compiled data set can be made openly available to all users including commercial use to boost the Digital Single Market.
- Tools: should be released as open sources under a permissive software license.

Conduct national survey(s) on the translation practices in public administrations: Assess the translation practices on the level of relevant public administrations (data creators) and identify necessary changes.

Recommendations for enabling sustainable sharing of language resources across Europe

PROPOSED CHANGES AT THE OPERATIONAL LEVEL OF PUBLIC ADMINISTRATIONS:

Translation and Data Management

Designate Open Data officers in all public administrations and public services.

Introduce general rights management in the data management process:

- Define responsible person / department for the authorisation of language data sharing in public administrations and public services.
- Provide legal support for language data sharing to all creators and users of language data in the organisation.
- Label websites that only publish data falling under the PSI Directive/Open Data Directive with the respective national Open Government License.

Adopt translation data management plans in all public administrations that generate language data (data creators).

- Provide practical guidelines for all translators (data creators) and data managers on how to manage translations and language data.
- Provide practical guidelines on how to handle legal issues relating to the sharing of translations / language data (e.g. safe handling of confidential or personal data, copyright and IPR etc.).
- Define protocols for archiving documents, translation data and other types of language data in terms of format, standardised metadata descriptions, information about copyright and property rights, indication of confidential and personal data.

Centralise translation workflows in public administrations that generate language data (at least at the institutional level) to facilitate the collection, curation and sharing of LR. Store language data centrally. Procure language data centrally and provide corresponding guidelines for procuring translations.

Adapt procurement contracts for translations:

- Request to receive the translation in machine-readable formats by default.
- Request to receive any by-product of the translation, such as memory files or terminologies.
- Stipulate that the ownership of outsourced translations, TMs and any other by-products belong to the public administration that outsourced the translation.
- Stipulate a corresponding license (ideally an international license) if the IPR ownership cannot be transferred.
- Develop corresponding guidelines and templates for agreements covering the retrieval of TMs and other by-products between public administrations and LSPs/ external translators.

Human capital

Provide training of translators (data creators) and translation managers:

- Raise awareness of language data as Open Data through Open Data officers in the organisation.
- Raise awareness on the value and use of CAT tools and MT, the value chain of language data and the importance of LR, and the benefits of Open Data.
- Improve technical skills with regard to CAT tools and MT.
- Raise awareness on the legal constraints associated with the sharing of language data.

IT infrastructure, equipment and tools

Invest in the provision and take-up of language technologies, in particular CAT Tools and MT.

Provide any additionally required language data management tools.

Develop / provide data anonymisation methods and tools tailored to the specific needs and languages of the particular public administration.

Translation process / workflow

License translation data with open licences that allow for commercial and/or non-commercial reuse of the data (unless a text is public domain).

Clearly identify/highlight translation data that contains confidential information.

Exclude confidential information from reusable TM or data sets.

Clearly identify/highlight translation data that contains personal data.

Exclude any personal data, i.e. information that could link to any identifiable person from re-usable TMs or data sets (i.e. anonymise such data sets).

Integrate tools and methods for automated anonymisation in the translation data process.

Integrate CAT Tools and MT into the translation workflow in all public administrations that generate language data (data creators).

Automate the process of translation / language data creation, curation and collection/ sharing to a maximum extent.

Annex
Country Profile Austria

Gerhard Budin, Jürgen Kotzian, Barbara Heinisch, Lilli Smal

State of Play:
Translation practices and information exchange in ministries and public administrations in Austria:

Within Austrian federal government organizations, integrated language services are the exception. Most public administrations either outsource translations or translate in-house but not with staffed translators. Overall, the translation process on the federal level is decentralized, which means that every public administration meets their own translation needs, there is no central management tool for translation requests and no formalized exchange of translation memories or expertise. However, an informal working group ARG GUT (Arbeitsgruppe Gouvernementaler Uebersetzungs- und Terminologiedienste) was initiated by the Language Institute of the Austrian Armed Forces in the Federal Ministry of Defence. The working group consisting of translators and terminologists for Austrian German <> English (and partly also for French), not only exchanges information and expertise but also creates resources such as the administrative glossary that is freely available on the Austrian Language Resource Portal (Sprachressourcenportal Österreichs). This portal was created as an aid for the Austrian EU Council Presidency in 2018. However, it is continuously developed further.

In addition to the Austrian Armed Forces Language Institute (SIB) subordinated to the Ministry of Defence, the Ministry of Interior and the Ministry for Sustainability and Tourism, other public administrations have integrated language services, such as the Austrian Financial Market Authority, the National Bank of Austria as well as the Vienna City Administration. The Austrian Armed Forces Language Institute provides translations, terminology work, and language teaching in English, French, Italian, Czech, Slovak, Hungarian, Slovenian, Russian, Ukrainian and Balkan languages. The Ministry of the Interior has several translation cells in various agencies of the ministry. There is the Language Service of the Criminal Intelligence Service with about 15 translators, interpreters and terminologists and small translation cells both in the Federal Bureau of Anti-Corruption and in the Federal Agency for State Protection and Counter Terrorism. There is no coordination or exchange between those three language services, partly due to secrecy and security reasons, partly because there is no language governance on the superordinate ministerial level. However, translators and terminologists from all three services work together within the informal terminology working group ARG GUT.

Due to international police cooperation, the Criminal Intelligence Service has the strongest need for an internal language service with about 15 translators, interpreters and terminologists. The other ministry that has an in-house translation service is in the Ministry for Sustainability and Tourism (formerly known as the Ministry for Agriculture) with three translators/interpreters. Although it is common practice that the in-house translation services use computer-assisted translation (CAT) tools, including translation memories (TMs), the TMs are not managed in a way that allows for easy language data sharing with e.g. the Austrian Open Data Portal.

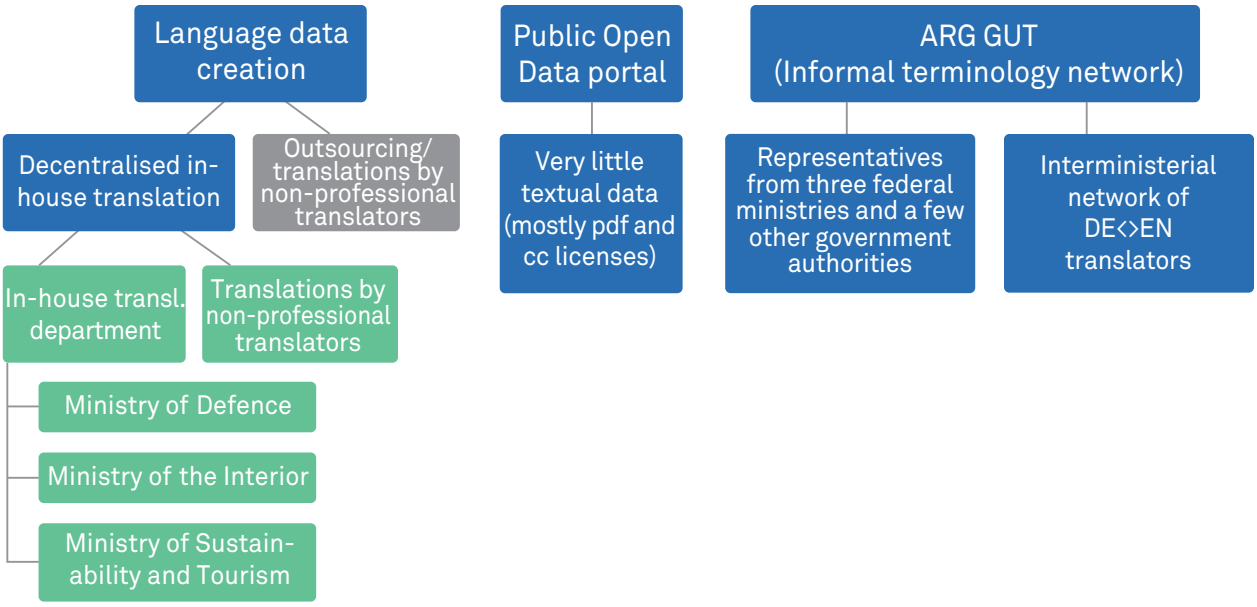
All other administrations meet their translation needs by either outsourcing to freelancers or language service providers or by their own employees who are not professional translators but are making use of free commercial machine translation systems and their foreign language education in school. In these cases, no CAT tools are applied in the translation process and TMs are not requested back from LSP to whom the translations are outsourced.

Interesting fact:
The Austrian Armed Forces Language Institute (SIB) developed apps based on language resources to enable basic intercultural communication, primarily for soldiers in international operations but also for the general public. The apps can be downloaded for free.

Translation needs:

The translation needs in Austrian public administration are threefold. There is a need for international communication and translation, a need for translation between different varieties of German and the need for translation between everyday language citizens use and the language register used by public officials. The Austrian Language Resource Portal mainly addresses the need for appropriate international communication (into and from English) and translation specifically tailored to the Austrian German variety.

The current language data creation and sharing infrastructure in Austrian public bodies looks as follows:



Open Data in Austria:

The Austrian Data Portal data.gv.at has won the United Nations Public Service Award for Open Government Data in 2014 and has over 26,000 data sets, documents and applications available for reuse. It centralizes metadata of decentral data catalogues and is the single point of contact to the European Data Portal. However, there is a strong focus on numerical data exemplified by the fact, that there is no explicit category for language data and textual data currently falls in the category of "documents" available predominantly in PDF format. After the second ELRC workshop in Austria, however, some of the textual data that were PDF files were converted into TMX files and are now available in machine-readable format on ELRC-SHARE. The Austrian Data Portal has a filter for the file format, but TMX files are not available. Currently, a metadata core, optional attributes and a vocabulary for the metadata catalogue for open government data (OGD metadata) are available in German and English.

Digital policy and language policy in Austria:

Austria's digital "roadmap" contains 12 guiding principles. They include a digital education strategy, an ICT strategy, an eGovernment strategy, including big data in public administration, an open government strategy and others. The Open Government strategy is coordinated by the "Cooperation Open Government Data Austria", whose main objective is to create an environment that encourages the sharing of government data as well reaping its benefits. As mentioned above, the focus is currently on numerical data and language data are not yet acknowledged as valuable Open Data resources on the policy level.

Annex
Country Profile Austria

The e-government programmes are coordinated by the Ministry for Digital and Economic Affairs. Digitalisation is seen as a cross-sectional topic which is coordinated by Chief Digital Officers (CDO) appointed for every area of responsibility across all ministries. Together these officers compose the "CDO Taskforce" that is tasked with optimizing the coordination of digitalisation activities. The Ministry of Digital and Economic Affairs is also responsible for the implementation of the Digital Single Market in Austria.[1]

Part of the eGovernment initiative is Austria's central platform for digital public services oesterreich.gv.at. The platform offers citizens services such as an electronic signature, electronic payments, changing residency and others. However, the website is mainly available in German. Some general information is also available in English.

As Austria's sole official language, all government communication is exclusively in German and only partially translated into English, underlining the strong dominance and the status of Austrian German as the only official language. Recognised minority languages are Croatian, Romani, Slovakian, Slovene, Czech and Hungarian. As some regions have a large number of native speakers of Slovene, Croatian and Hungarian, these languages have minority status and school education is offered bilingually in these regions. Additionally, some schools offer mother-tongue teaching in a total of over 26 languages. Article 19 of the Basic Law of 21 December 1867 on the General Rights of Nationals in the Kingdoms and Länder represented in the Council of the Realm specifies that all ethnic entities have a right to the preservation and fostering of their language, including schools, administration and public life.

Stakeholders:

The Federal Ministry for Transport, Innovation and Technology is responsible for the CEF agenda including matters related to eTranslation and therefore an important stakeholder next to the Federal Ministry for Digital and Economic Affairs, who are in charge of eGovernment and Open Data. Still, since translations are carried out in all federal ministries and all areas of responsibility have CODs, all ministries are important stakeholders. So far, about 20 institutions and public administrations, including the Ministry for Digital and Economic Affairs have participated in ELRC events. Language resources in various formats such as plain text, TMX and XML were contributed by the Austrian Armed Forces Language Institute, the Ministry of the Interior, the National Bank of Austria, the City of Vienna and the Centre for Translation Studies at the University of Vienna, among others.
In the area of research, Austria is involved in CLARIN and DARIAH, among others.

Main challenges for sustainable data sharing:

Austria faces a number of challenges when it comes to sharing language data continuously and sustainably. The main challenges are:

- Internal translation workflows are not taking into consideration that the produced translations have linguistic value apart from their inherent purpose, which leads to a number of issues such as:
  - Translation memories are not requested back when translations are outsourced, hence the translations are not available in their most useful file format, i.e. TMX
  - Language data is not managed/filed in a way that allows for easy data sharing
  - Privacy, copyright and IPR are not clearly indicated or transferred
  - It is unclear who can authorize sharing of language resources
  - Superiors do not acknowledge the value of sharing data and therefore do not initiate necessary changes in the workflow or processes
- Language policy or multilingualism is not in the portfolio of a specific ministry (only the Ministry of Education, Science and Research covers language and multilingualism at schools) making it difficult to identify decision makers in relation to multilingualism, language data as Open Data or the translation and procurement processes

[1] N.B.: The responsibilities might change with the new government.

- With the predominant exclusive use of German and no public policy for multilingual content online, public administrations do not see a need to invest in language technologies

Action plan:

In order to address the identified challenges in Austria, the following actions are suggested:

- **The main objective is to improve and establish data management practices that allow for reaping the maximum benefit from language data. Specific actions include:**
  - The identification of data managers
  - Further investigation of data management practices
  - Guidelines for the identification of confidential and personal data
  - Clear indication of confidential and personal data as well as copyright in the translation process to make data sharing in the future easier

As these actions need to be addressed top-down, support and guidelines from the European Commission and ELRC would be very helpful.

- **The second objective is to include machine translation and language technology in the national digi - tal policy and increase the interest in these topics in public services. Specific actions include:**
  - Secure the support of decision makers to include language technology in the national policy
  - Establish synergies with national actions and initiatives related to language technology, machine translation and language data
  - Inform about amounts of language data that are needed to develop language technologies as well as processes that are available to make data sharing safe and secure
  - Stress the importance of the Austrian variety of the German language (in order to receive high-quality machine translation output for Austrian German)

- **The third objective is to gain access to outsourced translation:**
  Gaining access to outsource translations could be a valuable asset, since this data has enormous potential and value.

- **Another objective is to generally raise awareness about the value of language data:**
  This includes its potential when shared and used for machine translation but also more generally in many different areas of artificial intelligence.

References and further reading list:

Austrian Language Resource Portal: sprachressourcen.at.

Bundesministerium Bildung, Wissenschaft und Forschung: *Sprachliche Bildung*, https://bildung.bmbwf.gv.at/schulen/unterricht/ba/sprachenpolitik.html.

Cooperation OGD Österreich: *Infos Cooperation OGD Österreich*, https://www.data.gv.at/infos/cooperation-ogd-austria/.

Open Data Österreich: https://www.data.gv.at.

Federal Ministry Republic of Austria Digital and Economic Affairs: *Verwaltung,* https://www.bmdw.gv.at/Themen/Digitalisierung/Verwaltung.html.

Federal Ministry Republic of Austria Digital and Economic Affairs: *Digital Roadmap Austria,* https://www.digitalroadmap.gv.at/fileadmin/downloads/digital_road_map_broschuere.pdf.

Annex
Country Profile Austria

Federal Chancellery Republic of Austria: *Behörden im Netz, Das österreichische E-Government ABC*, 2017, https://www.digitales.oesterreich.gv.at/documents/22124/30428/E-Government-ABC.pdf/b552f453-7ae9-4d12-9608-30da166d710b, https://www.bmdw.gv.at/dam/jcr:8fc815bb-1dc7-4e45-9610-78d63560944a/E-Government-ABC_2019_EN.pdf.

Heinisch, Kotzian: *ELRC Workshop Report for Austria*, 2018, http://lr-coordination.eu/sites/default/files/Austria/2018/ELRC_Workshop_Austria_Report_public_v1_FINAL.PDF.

Metadata Catalogue for Open Government Data (OGD Metadata): https://www.ref.gv.at/OGD-Metadaten-2-4.3468.0.html.

English translation of the Basic Law of 21 December 1867 on the General Rights of Nationals in the Kingdoms and Länder represented in the Council of the Realm (in German:Staatsgrundgesetz vom 21. December 1867, über die allgemeinen Rechte der Staatsbürger für die im Reichsrathe vertretenen Königreiche und Länder – StGG): https://www.ris.bka.gv.at/Dokumente/Erv/ERV_1867_142/ERV_1867_142.pdf

Stadt Wien: *Digitale Agenda Wien 2025*, 2019, https://digitales.wien.gv.at/site/files/2019/09/20190830_DigitaleAgendaWien_2025.pdf

Open Science Network Austria: https://oana.at/

Annex
Country Profile Belgium

Stijn de Smeytere, Veronique Hoste, Eileen Schnur

State of Play:
Translation practices in ministries and public administrations in Belgium:

In Belgium, each institution is responsible for the translation of their data. Translation needs are often handled on demand and the applied translation practices are diverse. Consequently, there are public administrations, which outsource all their translations, whereas other institutions are solely building on in-house translation. In addition, there are administrations applying a combination of both approaches.

Data exchange and translation are currently not coordinated in Belgium. In public administrations, all outsourced translations are part of call for tenders. Since there is not one call for all administrations, each department has its own tender. Further information is available in the country reports produced by NEC TM.

CAT Tools are used by the vast majority of Belgian language service providers (LSPs) and freelance translators. This is also common practice in Belgian institutions, which use e.g. computer-assisted translation software suites or translation management systems. Although there is a growing awareness that machine translation (MT) can be a valuable asset and facilitate the translation process, it is only rarely used. If the translations were outsourced, the corresponding translation memories (TMs) or any other by-products are usually not transferred back.

> **Interesting fact:**
> **In a few public administrations, data management plans have been developed and integrated. Public administrations are provided with a set of guidelines and handle their data management individually.**

In the academic area, Belgian funding agencies such as the research foundation FWO are now required to submit a data management plan together with their project proposal. Apart from that, in 2017, the DMPbelgium Consortium was founded by a number of Belgian universities, including Ghent University, Hasselt University and University of Antwerp, among others. They developed a shared data management planning tool, offering common data management plan templates, institutional templates and guidance. Further information about data management at Belgian universities is available at https://dmponline.be.

Although there are currently no specific data sharing infrastructures, the focus on data management can be seen as a preparatory step. There is no obligation to share data, but it is increasingly encouraged and corresponding platforms are becoming more and more popular among researchers.

Open Data in Belgium:

In compliance with the PSI Directive, the Belgian Federal Council of Ministers agreed to adopt a federal Open Data strategy with an ambitious roadmap for 2015-2020[2], introducing the principle of open public data by default. According to this strategy, all data collected by the Belgian public administrations has to be freely available and reusable. Exceptions are only acceptable if the data contains private information or content with the potential to harm public security. The federal Open Data strategy includes a set of fifteen practical guidelines to facilitate the reuse of data. The primary goals include[3]:

- The free use of PSI without any reference to the source to facilitate the combination of data sets
- The provision of data in machine-readable formats to facilitate reuse, identification and extraction (e.g. Excel instead of PDF, CSV instead of Excel, etc.) whenever possible.
- The provision of public sector information by the federal government not only upon request, but proactively by 2020.
- The development of an Open Data strategy in each federal public service and the appointment of an "Open Data champion", acting as the Open Data contact point within the organisation.
- The set-up of a web portal providing continuous access to open datasets (https://data.gov.be/en).

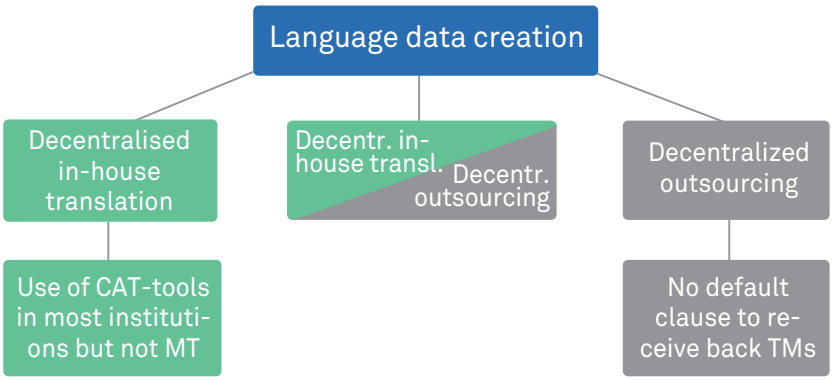
[2] Ferri, Springael: *Federale Open Data-strategie*, 2015.
[3] DLA Piper: B*elgian Government gives green light to a new Open Data strategy*, 2015.

Annex
Country Profile Belgium

The above-mentioned web portal already includes more than 10,000 datasets covering a variety of fields, e.g. public sector, science and technology or environment. It provides a search function and filtering by topic, licence, data format or organisation. According to the Belgian federation for the technology industry Agoria, making public sector information available to non-public entities could also lead to a considerable economic benefit. In more concrete terms, Agoria expects a net gain of around 900 million EUR.[4]

An important aspect, which may prevent public administrations from sharing their translated data is the fact that the translator holds the ownership of the produced text by default.[5] In order to avoid any copyright or GDPR-related issues, it is thus important that public administrations are granted the rights to the translated text when receiving an outsourced translation.

The current language data creation infrastructure in Belgian public bodies looks as follows:



Language policy and digital policy in Belgium:

Multilingualism plays an important role in Belgium. Belgium can be divided into the three regions Flanders, Brussels and Wallonia and has three official languages, i.e. Dutch (approx. 5.2 million speakers), French (4.6 million speakers) and German (72,000 speakers). Language policy in Belgium is based on two principles, the constitutional principle of linguistic freedom and the territoriality principle as described in Article 4 of the Belgian Constitution. According to the latter, Belgium can be divided into four linguistic areas, i.e. the Dutch-speaking region, the French-speaking region, the bilingual region of Brussels-Capital and the German-speaking region.[6]

> **Interesting fact:**
> **Belgium has three official languages and can be divided into four language areas. Official language use is determined by the territoriality principle.**

Each municipality in Belgium is part of only one of the four language areas. The official language use therefore depends on the territorial boundaries and varies from one linguistic area to another. Whereas in Brussels, there are two official languages (Dutch and French), all other regions are monolingual.

However, the linguistic boundary does not quash the constitutional principle of linguistic freedom, stating that "the use of languages spoken in Belgium is optional; only the law can rule on this matter, and only for acts of the public authorities and for judicial affairs."(Article 30). The territoriality principle is restricted to a limited number of domains, including public authority and administration, court, education, social relations (between employer and employees) and official documents. In addition to the three official languages, English

[4] van Tilborg, Luc: *National Initiatives for Digital Public Services*, 2018.
[5] Deene, Joris: *Can data be shared and how?*, 2018.
[6] Hüning, Vogl: *One Nation, One Language? The case of Belgium*, 2010.

also plays an important role in Belgium, especially regarding increased visibility outside the country borders and in multilingual contexts.[7]

Multilingualism is also of utmost importance for Belgian public digital services. At federal level, all public digital services are at least bi-lingual, since their provision in Dutch and French is obligatory. This also applies to Brussels, while in Flanders, all information must be provided in Dutch. In Wallonia, public digital services must be available in French, whereas in the Eastern part of Wallonia, the additional provision of public digital services in German is obligatory, too.

To improve Belgium's position in the digital field, the "Digital Belgium" initiative was launched by the Belgian Federal Government in 2015. It describes Belgium's long-term digital vision and aims to improve the country's position in the digital field. However, language technology is currently not included in the Belgian digital policy. "Digital Belgium" is based on the following five pillars:

- Digital economy
- Digital infrastructure
- Digital skills and jobs
- Digital trust and digital security
- Digital government

One of the main goals of the Belgian Digital Agenda is to bring Belgium in the top 3 of the Digital Economy and Society Index (DESI), where it was ranked 6th in 2017. The best rankings were for connectivity (3rd) and integration of digital technology by businesses (5th).

However, due to the complicated structure of the country, digital public services were considered Belgium's weakness[8]. This point is closely related to one of the key pillars included in the action plan, i.e. the digital government. Digital Belgium aims "to implement a digital transformation of the federal government, aiming at digital-by-default end-to-end interactions with citizens and organisations"[9]. This is to be achieved e.g. with the help of the above-mentioned Open Data Strategy, the introduction of "Mobile ID", a mobile app allowing every Belgian to prove his identity online and the "Government Cloud", a "hybrid cloud using services offered by private companies in public cloud environments and services housed in state-owned data centers".

Stakeholders:

The ELRC National Anchor Points represent two relevant stakeholders, namely the Chancellery of the Prime Minister and Ghent University. The collection of language data is also strongly supported by federal and regional public services, including e.g. the National Bank of Belgium and RIZIV, the National Institute for Health and Disability Insurance. Since 2016, the National Bank of Belgium donated more than 140 term bank entries in all three languages plus English and RIZIV made a translation memory available that consists of more than 30,000 translation units in French and Dutch. Local ELRC events and workshops were attended by representatives of more than 40 institutions, including e.g. FPS Chancellery of the Prime Minister or Nederlandse Taalunie. This clearly demonstrates that many Belgian institutions are already aware of the importance of collecting, managing and sharing language data to facilitate information exchange not only across the four linguistic regions of Belgium, but also across the European Union.

Main challenges for sustainable data sharing:

- Anonymization is often an obstacle to sharing translations and translation memories. Although automatic processes for anonymising data may be helpful to overcome this issue, the output is not 100% reliable. Especially when dealing with unstructured data, this problem can hardly be fixed.
- Another challenge is the authorisation of data contributions by the responsible superiors, because it can be difficult to identify the person who is able to give the permission to share a certain data set. It is often necessary to go high up in the administrative structure to find the right person.

[7] De Smeytere, Hoste, Terryn: *ELRC Workshop Report for Belgium*, 2018.
[8] van Tilborg, Luc: *National Initiatives for Digital Public Services,* 2018.

Annex
Country Profile Belgium

- Apart from that, there are legal issues complicating data sharing, especially when it comes to outsourced translations as mentioned previously. In this context, it is of utmost importance to obtain the correct rights on the translations in order to be allowed to share the data afterwards.
- Technical issues and the required processing of the resources can be time-consuming and complicate data sharing processes.
- In the past, the common lack of awareness of the value of data and the general scepticism towards technology were also two of the main challenges in Belgium.[10] Nonetheless, this seems to have improved over the past few years thanks to the numerous national and European initiatives.

Action plan:

For Belgium, the following objectives could be defined to address the identified challenges. In the order of their priority, they are:

- **To raise awareness of the value of language data:**
  As language data is currently not included in the Belgian digital policy, it is important to promote the benefits of sharing language data. This could be achieved with the help of concrete examples of how data contributions had a positive impact on machine translation systems. In addition, it is important to establish practical guidelines for LR as Open Data and to broaden the definition of textual resources by adding speech data, data for AI and other types of language resources.

- **To establish good data management practices in public services:**
  Although there is already a focus on data management plans in the academic field, the above-mentioned developments will need to be continued and extended to all public services.

- **To increase interest in MT/LT in public services as part of the national digital policy:**
  Concrete examples of how public administrations can benefit from language technologies in their daily operations would raise the institutions' awareness and increase their interest in MT/LT. In addition, the use of MT/LT services could be promoted by identifying and establishing synergies with national projects and initiatives, wherever possible.

- **To tackle legal concerns:**
  Since legal concerns are one of the key challenges when it comes to sharing data in Belgium, it is important to develop and share easy-to-apply guidelines for IPR and privacy issues. Apart from that, the possibility to implement rights management along with data management needs to be investigated.

References and further reading list:

Belgian Government: *Digital Belgium*, 2017,
http://digitalbelgium.be/wp-content/uploads/2017/07/compressed_Brochure_DB_FINAL.pdf.

Deene, Joris: *Can language data be shared and how?*, 2018,
http://www.lr-coordination.eu/sites/default/files/Belgium/2018/S2.5_Language%20Data%20Sharing.pdf.

De Smeytere, Hoste, Terryn: *ELRC Workshop Report for Belgium*, 2018,
http://lr-coordination.eu/sites/default/files/Belgium/2018/ELRC%2B2%20Workshop_Public_Belgium-.pdf.

DLA Piper: *Belgian Government gives green light to a new Open Data strategy*, 2015,
https://www.dlapiper.com/en/uk/insights/publications/2015/12/spotlight-on-belgium-issue-8/belgian-government-approves-open-data-strategy/.

[9] Brochure by the Belgian Government: *Digital Belgium*, 2017.
[10] De Smeytere, Hoste, Terryn: *ELRC Workshop Report for Belgium*, 2018.

Ferri, Springael: *Federale open data-strategie*, 2015,
https://www.presscenter.org/nl/pressrelease/20150724/federale-open-data-strategie?lang=fr.

Hoste, Veronique: *ELRC in Belgium*, 2018, http://lr-coordination.eu/sites/default/files/Belgium/2018/S2.2_ELRC%20in%20Belgium.pdf.

Hüning, Vogl: *One Nation, One Language? The case of Belgium*, 2010,
https:/www.academia.edu/1056036/One_nation_one_language_The_case_of_Belgium.

Research Foundation Flanders (Fonds voor Wetenschappelijk Onderzoek – Vlaanderen, FWO):
https://www.fwo.be/en/.

van Tilborg, Luc: *National Initiatives for Digital Public Services and (Open) Data*, 2018,
http://www.lr-coordination.eu/sites/default/files/Belgium/2018/S1.2_National%20Initiatives%20for%20Digital%20Public%20Services.pdf.

Annex
Country Profile Bulgaria

Hristina Dobreva, Svetla Koeva, Andrea Lösch

State of Play:
Translation practices in ministries and public administrations in Bulgaria:

In Bulgaria, translation services are subject to procurement through the Central Purchasing Body (Ministry of Finance). Each contracting authority shall conduct a small competitive procedure among potential contractors included in the Framework Agreement with the Central purchasing body. As a result of this competitive procedure, a contractor of the translation service is designated among the potential contractors included in the Framework agreement. Any person may submit a request to participate in a competitive procedure with negotiation. The centralized procedure ends with the conclusion of the framework agreement. When the conclusion of the Framework Agreement by the Central purchasing body is postponed or cancelled, the contracting entities shall apply the general rules and award the procurement individually.

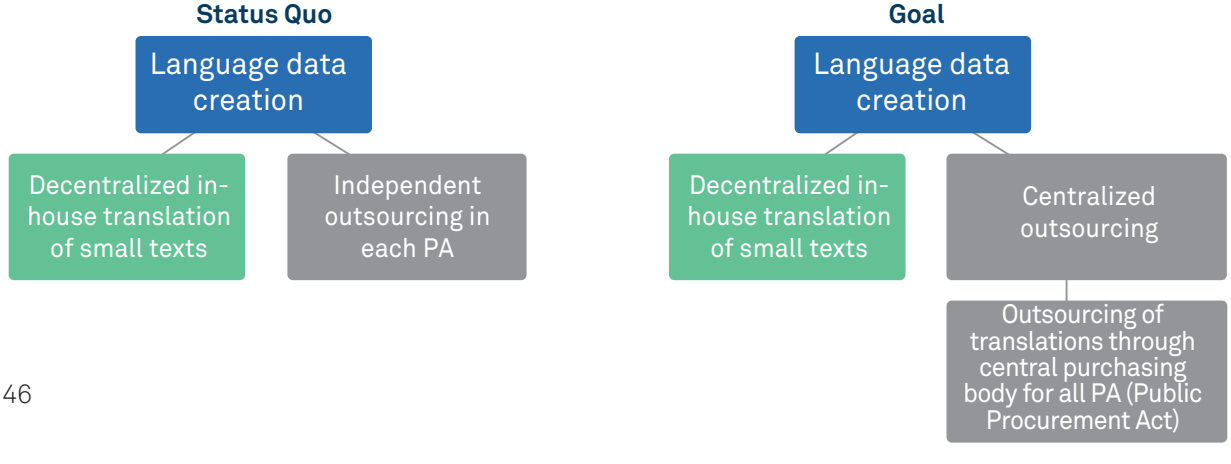
However, the application of the public procurement procedure depends on the value of the contract. There is an existing Framework Agreement in force, but it will be renewed before the end of October 2019. Public procurement data is available through the Public Procurement Agency (http://www.aop.bg/index.php?ln=1) and the Public Procurement Portal of Bulgaria. Currently, there is no coordination of the procurement of translations between the different ministries and public bodies although it is the goal to centralize the translation services in the future.

As regards the provision of translation memories (TMs) as part of the contracts, agreements with the language service providers (LSPs) must be negotiated on a case-by-case basis. A template or a framework prescribing and covering the hand-over of TMs does not exist yet. Similarly, there is no coordinated exchange of translations or language data among ministries (and/or other public bodies) in place at the moment.

The use of computer-assisted translation (CAT) tools or machine translation (MT) is currently not a common practice in administrations and ministries: They mainly use the spelling tool in their Office package and some of them use the European Commission's eTranslation MT system as well as the translator of the Bulgarian EU Council Presidency.

Interesting fact:
The ELRC National Anchor Point of the Public Administration has made a proposal in the new Central purchasing body's framework agreement in the field of translation services to include a requirement for translation contractors to provide translation memories, if applicable, to public administrations and transfer them the ownership rights. Moreover, the Ministry of Transport, Information Technology and Communications has made general registration of all its employees so they can use the CEF Automated Translation (AT) platform without the need of individual requests or registrations. Last but not least, there is a working process for the integration of the eTranslation platform with the Single Information Point of Bulgaria.

The current language data creation infrastructure in Bulgarian public bodies looks as follows:



Open Data and eGovernment in Bulgaria:

The State e-Government Agency (SEGA) is responsible for the e-Government strategy and the faster implementation of e-governance in Bulgaria. Core issues covered in the State e-Government Agency's domain are semantic, legal, technological and organizational interoperability. Data is considered the new most valuable resource and there is an explicit intent to make as much data available as possible for re-use through the public services Bulgaria offers people and businesses. Last but not least, Bulgaria has also implemented the secure electronic delivery system "eDelivery", which is also part of the single application model for payment and provision of electronic public services at national level.

The State e-Government Agency (SEGA) also develops and maintains the Open Data Portal "data.egov.bg". As of September 2019, the Open Data Portal was completely renewed covering 505 registered organizations and 9600 datasets across Bulgaria. It is the central repository for the Open Data from the Bulgarian administration and the public authorities, which all have an obligation to prioritize and publish data for re-use.

Overall, the portal is built in accordance with the requirements of the Access to Public Information Act, and the information it publishes is set out in the regulation on standard terms and conditions for reuse of public sector information. SEGA adopts a list of datasets per priority area, which has to be published in an open format and is updated on an annual basis. SEGA also produces a synthesis report on the availability of public sector information every three years. The following key players have a direct influence on the Open Data strategy in Bulgaria:

- The Administration of the Council of Ministers – the existing National Open Data Portal opendata.government.bg and the new Open Data Portal data.egov.bg;
- The Ministry of Transport, Information Technology and Communications (MTITC);
- The State e-Government Agency (SEGA);
- The Institute for Public Administration – regularly organises Open Data Training.

With regard to the legal framework for sharing data, special attention needs to be paid to the Directive on the reuse of public sector information, the legal framework on the reuse of information and in particular the Access to Public Information Act, which introduced the Directive in Bulgaria through a special regime substituting the older access to information regime.

Stakeholders:

Within ELRC, more than 150 potential stakeholders that are involved in the creation or sharing of language resources, related activities and/or policy setting were identified. They also participated in the latest ELRC workshop. Most importantly, the stakeholder base includes the 18 public administrations and services that are potential providers of language resources. So far, 24 language resources have been contributed to the ELRC-SHARE from Bulgaria from the MTITC, the NRA, the Ministry of Justice, from the National Institute of Justice, the Bulgarian Food Safety Agency and the Bulgarian Academy of Sciences.

Digital services and public organisations with multilingual needs that could benefit from the eTranslation platform include in particular:

- The NRA (National Revenue Agency), which is the public administration in Bulgaria with the highest number of electronic administrative services and could benefit from the instrument.
- Taxation and health insurance services;
- SOLVIT;
- National Institute of Immovable Cultural Heritage;
- All organisations providing public services.

Annex
Country Profile Bulgaria

Main challenges for sustainable data sharing:

- Difficulty to identify and convince high-level officials to authorize data sharing
- No established procedures for the translation of documents on administrative level, which leads to:
  - Potential legal issues, e.g. concerning the ownership of the data, personal data issues, copyright issues
  - Technical difficulties relating to data processing
- Resistance to new technologies
- Lack of resources, which are required for supporting the technical and legal preparation and sharing of language data
- Concerns about the quality of translations, which could be shared (feeling that their quality may not be high enough for sharing)

Action Plan:

Key actions for improving the sharing of language resources in Bulgaria are mainly targeted at (i) raising awareness of language data as Open Data and valuable asset, (ii) increasing interest in MT/LT in public services as part of the national digital policy and (iii) improving access to outsourced translations.

Regarding the awareness raising of language data as Open Data and valuable asset, several activities are planned and/or on the way:

- **Integrating language data in the national Open Data policy, digital agenda, etc.:**
  The National Anchor Points (NAPs) sent letters to all Ministries and their second level spending units on the benefits of sharing their translations through the ELRC-SHARE. A corresponding high-level meeting to stimulate the process of sharing language resources was organised involving the different administrations. The Commissioner for Digital Economy and Society and the Prime Minister in charge in this field are aware of the ongoing efforts.

- **Increasing collaboration with the Open Data officer on national level & establishing practical guidelines for LR as Open Data:**
  Following the launch of the renewed national Open Data Portal in September 2019, the idea to organize a meeting with representatives of the Council of Ministers and the State eGovernment Agency to discuss the possibilities for the acceptance of the language resources as Open Data emerged.

- **Promoting the value and benefits of sharing language data for language activities:**
  Language data sharing should be embraced by a wider audience, while the data contribution processes should be fine-tuned. The Bulgarian NAPs are planning to reach out to the municipal authorities in Bulgaria, which also have materials that could be useful for the eTranslation platform.

- **To increase interest in MT/LT in public services as part of the national digital policy, several important efforts are also planned/have already been started:**
  - **Secure support of decision makers to change/adapt national policy:**
    As indicated above, a high-level meeting to stimulate the process of sharing language resources was organised involving the different administrations. The Commissioner for Digital Economy and Society and the Prime Minister are aware of the ongoing efforts.
  - **Ensuring central accessibility to eTranslation:**
    The Ministry of Transport, Information Technology and Communications has made general registration of all its employees, so they can use the CEF AT platform without the need of individual requests or registrations. There is a working process of integration of the eTranslation platform with the Single Information Point of Bulgaria.

- **Regarding the improvement of access to outsourced translations, several efforts are in planning/ implementation:**
  - **Centralizing procurement of translations:**
    There are major efforts, in particular through the Public Procurement Agency, to centralize the procurement process and set common standards. In this respect, procurement of translations should also be considered and coordinated between the different ministries and public bodies.
  - **Establishing practice of receiving any by-product of outsourced translations:**
    The Public Services NAP has made a proposal in the new Central purchasing body's framework agree ment in the field of translation services to have a requirement for translation contractors to provide translation memories, if applicable, to public administrations and transfer them the ownership rights.

References and further reading list:

Access to Public Information Act:
https://www.me.government.bg/en/library/access-to-public-information-act-448-c25-m258-2.html

Bulgarian Open Data Portal: https://data.egov.bg/

Public Procurement Portal of Bulgaria:
http://rop3-app1.aop.bg:7778/portal/page?_pageid=173,1&_dad=portal&_schema=PORTAL

SOLVIT: https://ec.europa.eu/solvit/index_en.htm.

The State e-Government Agency (SEGA): https://e-gov.bg

Annex
Country Profile Croatia

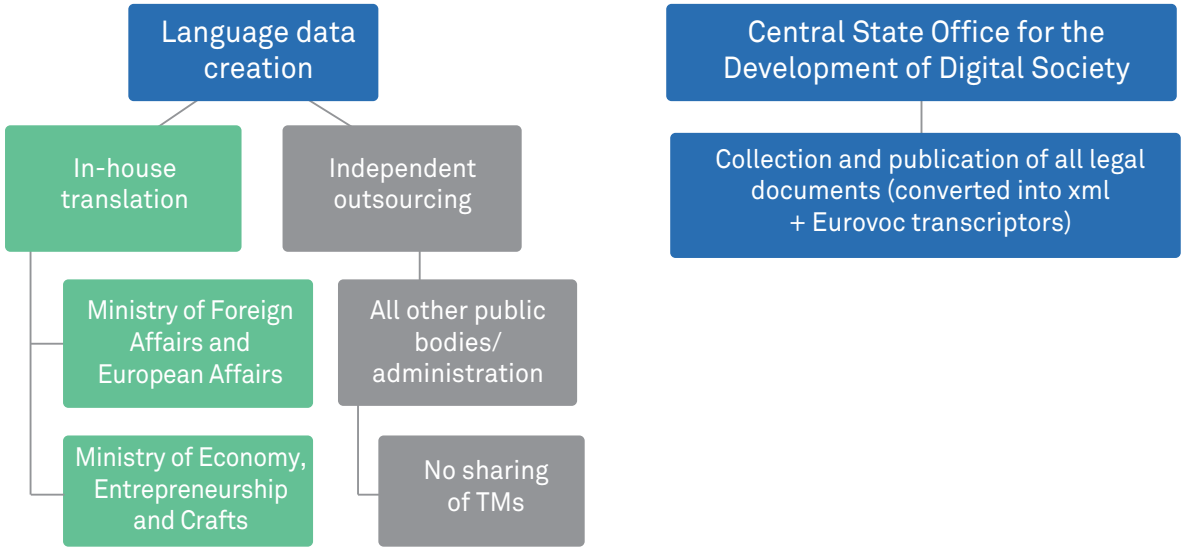
Marko Tadić, Lilli Smal

State of Play:
Translation practices in ministries and public administrations in Croatia:

Translation practices in Croatian public administration are fully decentralized and organized independently. Most public administrations outsource translations to Language Service Providers (LSPs), only the Ministry of Foreign and European Affairs and the Ministry of Economy, Entrepreneurship and Crafts have in-house translation services. There is no regulation that would enforce the usage of computer-assisted translation (CAT) tools or translation memories (TMs) in public administration and its usage is left to an individual initiative and non standardised licensing. Also, there is no central service that is responsible for translations at the level of government, ministries or state offices or agencies. Consequently, when translations are outsourced, TMs are not requested back by any of the outsourcing public administration bodies. Currently, there is no infrastructure in place to exchange translations or glossaries between ministries or to share translations with the national Open Data Portal.

Interesting fact:
The Central State Office for the Development of the Digital Society collects and publishes all legal documents in machine readable format accompanied with relevant EUROVOC descriptors as metadata. They are openly accessible using the language sensitive search engine CADIAL (www.digured.hr).

The current language data sharing infrastructure in Croatian public bodies looks as follows:



Open data and language data in Croatia:

The Croatian Open Data Portal "data.gov.hr" is administered and governed by the Ministry of Public Administration and the Central State Office for the Development of the Digital Society. The Central State Office for the Development of the Digital Society is also responsible for the collection and publication of all legal documents that are converted into XML format and accompanied with Eurovoc transcriptors as metadata. The implementation of the Public Sector Information (PSI) directive is under the jurisdiction of the Information Commissioner's Officer and is fully transposed in Croatian national legislation.[11] All bodies of the public administration are obliged to make public sector information available and accessible in a digital and open for-

[11] Act Nr. 403/13 of 8 March 2013 on the right of access to information (Zakon o pravu na pristup informacijama) and extracts from the General Administrative Procedure Act (Zakon o općem upravnom postupku - Act Nr. 1065/2009 of 1 April 2009)

mat with appropriate metadata. This data is published on the national Open Data Portal, which is the main access point for re-use of public sector information. However, there is very little language data available on the Open Data Portal as it currently mostly holds geolocation data, transportation data, meteorological data, environmental data, and other types of statistical data.[12]

Language policy and digital policy in Croatia:

The usage of the Croatian language and Latin script as the official language and script are regulated by Article 12 of the Croatian Constitution. Other languages and the Cyrillic or other scripts may be used together with the Croatian language and Latin script in individual local units and "under conditions specified by law"[13]. Apart from that, there is no explicit language policy and language technologies are not mentioned in the National Strategy of Education, Science and Technology from 2014. The strategy was developed with the help of more than 130 experts in 19 different working and thematic groups and is under the responsibility of the Croatian State.[14]

Stakeholders:

The Central State Office for the Development of the Digital Society is an important stakeholder as the State Office is the driving force behind the digitalisation process and is responsible for the national Open Data Portal together with the Ministry of Public Administration. CEF Telecom is coordinated by the Ministry of Economy, Entrepreneurship and Crafts and is therefore also an important stakeholder.

More than 30 organisations have participated in past ELRC events and several public administrations have already shared language data with ELRC. Among the data donors are the Ministry of Regional Development and EU Funds and the Ministry of Agriculture.

Main challenges for sustainable data sharing:

- The public sector is much slower in adapting digital infrastructures/innovations than the private sector
- General lack of interest and awareness of the importance of sharing language data among the higher-level officials
- Concerns with respect to:
  - the control of the quality of the language data used in training the systems for machine translation (in terms of both the relevance of the documents and the quality of translations)
  - the impact of the type of the text to the quality of machine translation
  - the accessibility of the eTranslation system to wider audience (universities, translation agencies)[15]

Action plan:

To address the above mentioned challenges, the following recommended actions are considered vital:

- Raising awareness among decision makers is regarded as the most important future step of the ELRC action in Croatia.
- Raising awareness about the importance of sharing language data that originated from public funding.
- Starting the initiative to establish a central translation office that would service the Government, ministries, state offices and agencies with translation to and from Croatian when needed. Such a translation office could use the latest state-of-the-art resources and CAT tools in the translation process (centralised TMs, domain dependent MT, general domain MT, etc.).

---

12 Tadić, Marco: *ELRC Workshop Report for Croatia*, p.10, 2019.

13 Committee on the Constitution: *Standing Orders and Political System of the Croatian Parliament: Constitution of the Republic of Croatia, Consolidated Text*, Article 12, 2010.

14 Government of the Republic of Croatia: *Strategy of Education, Science and Technology, Nove Boje Znanja*.

15. Tadić, Marko: *ELRC Workshop Report for Croatia*, 2019.

Annex
Country Profile Croatia

References and further reading list:

CADIAL search engine for Croatian legal documents: http://www.digured.hr.

Central State Office for the development of Digital Society: https://rdd.gov.hr.

Committee on the Constitution: *Standing Orders and Political System of the Croatian Parliament: Constitution of the Republic of Croatia, Consolidated Text*, 2010, https://www.wipo.int/edocs/lexdocs/laws/en/hr/hr060en.pdf.

General Administrative Procedure Act, Zakon o općem upravnom postupku, Act Nr. 1065/2009 of 1 April 2009: http://digarhiv.gov.hr/arhiva/263/44262/narodne novine.nn.hr/clanci/sluzbeni/2009_04_47_1065.html.

Government of the Republic of Croatia: *Strategy of Education, Science and Technology, Nove Boje Znanja*, https://vlada.gov.hr/highlights-15141/archives/strategy-of-education-science-and-technology-nove-boje-znanja/17784.

Right of Access to Information, Zakon o pravu na pristup informacijama, Act Nr. 025/13 of 8 March 2013: http://digarhiv.gov.hr/arhiva/263/100541/narodne-novine.nn.hr/clanci/sluzbeni/2013_02_25_403.html.

Tadić, Marko: *ELRC Workshop Report for Croatia*, 2019,http://www.lr-coordination.eu/sites/default/files/Croatia/2019/ELRC%2B%20Workshop%20Public%20Report%20Croatia_FINAL.PDF.

Annex
Country Profile Cyprus

Natassa Avraamides-Haratsi, Fryni Kakoyianni-Doa, Andrea Lösch, Eileen Schnur

State of Play:
Translation practices in ministries and public administrations in Cyprus:

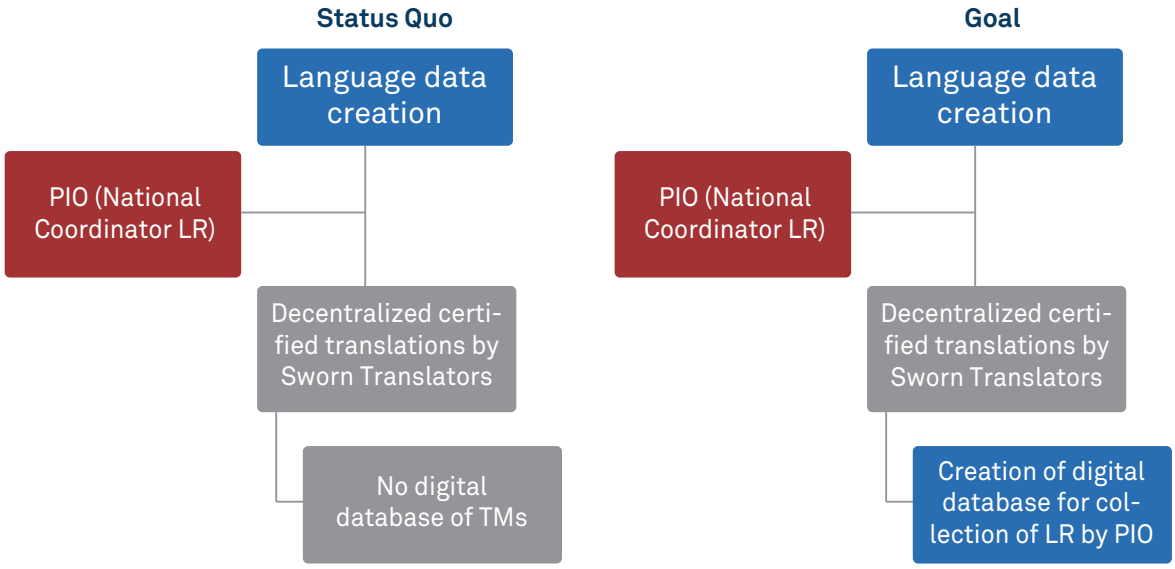
The Press and Information Office (PIO) is the official communication service of the government, subordinated to the Ministry of Interior. Between 1990 and 2019, the PIO was the national agency for certifying translations, recruiting associate translators from the private sector. After the passing of the Law on Sworn Translators in March 2019, the service for certified translations was transferred to the Sworn Translators.

As regards translation practices, there are no specialised in-house translation services in Cypriotic public administrations. If the documents do not require certification, some documents (e.g. press releases) are translated by bilingual or multilingual employees as part of their work within the administration or they are outsourced independently to language service providers (LSPs) or freelance translators. In the case of certified translations, it is obligatory to outsource them to the Sworn Translators.

Currently, there is no central translation/terminology database yet and the use of machine translation (MT) or computer-assisted translation (CAT) tools is not standard (no full digitization of translations). However, with regard to eGovernment, there is an electronic platform called "Ariadni", through which citizens can conduct their transactions with the public administrations electronically on a 24/7 basis. As such, there is an increasing demand to use MT as a tool for enabling multilingualism and an action is underway with the PIO as the coordinating agent in order to facilitate best practices in the collection of language data in Cyprus. The Department of Public Administration and Personnel and Open Data is in close cooperation with the PIO in this regard.

Interesting fact:
The PIO supported translations in 26 languages until 2019.

The current language data creation infrastructure in public bodies of Cyprus looks as follows:



Open Data in Cyprus:

The PIO collaborates closely with the national Open Data Policy Officer and the responsible office for the Digital Agenda in Cyprus. The Open Data policy in Cyprus is based on the Public Sector Information (PSI) directive (2003/98/EC), which was implemented in 2006. In 2015, the national law .205( )/2015 was passed based on the reviewed Directive. The Open Data Portal is available online at https://www.data.gov.cy/?language=en.

Annex
Country Profile Cyprus

Language policy in Cyprus:
Pursuant to Article 3 of the Constitution of the Republic of Cyprus from 1960, the official languages of the Republic of Cyprus are **Greek** and **Turkish**.

Stakeholders:
Within ELRC, more than 20 potential stakeholders that are involved in the creation or sharing of language resources, related activities and/or policy setting were identified including nine ministries as well as the Open Data Portal. So far, several bi- and multilingual language resources were contributed by the Cypriotic police and the PIO, which represents the major provider of language resources for ELRC in Cyprus.

Main challenges for sustainable data sharing:
The central challenge that makes the sharing of language resources currently difficult is the lack of digitization of any translations so far: This includes in particular the lack of a corresponding infrastructure and processes for sharing data on the national level, and directly associated with it the unavailability of translations in formats other than .pdf and .doc (mainly due to the lack of using CAT tools). Also, the availability of an extended qualitative eTranslation service would facilitate the creation of better linguistic data.

Action plan:
In order to overcome the central challenge mentioned above and to enable sustainable data sharing, the following objectives were defined:

- **To develop good data management practices:**
  The establishment of a corresponding infrastructure for sharing language resources (through the creation of a central database) is an important step towards language data sharing. As a direct consequence, it is of utmost importance to develop and establish good data management practices: Existing data management practices need to be thoroughly investigated and updated, e.g. with regard to establishing responsible data managers, introducing a clear separation between confidential/personal data from public sector information, and establishing a practice of receiving any by-product of outsourced translations. Most importantly, the basis for the collection of linguistic data (corresponding system and processes on national level) needs to be established.

- **To increase interest in MT and Language Technology (LT):**
  On policy level, several efforts are needed or already in progress to increase the interest in MT and LT as part of the national digital policy, including in particular:
  - Securing support of decision makers to change/adapt national policy with regard to language data: The PIO has received the support of the Ministry of Interior for the future development of this action.
  - Creating synergies with related national agents: The PIO is already in direct communication with the corresponding partners for further cooperation.

- **To tackle legal issues:**
  In order to pave the way for the sharing of language resources, corresponding legal issues need to be tackled. On the one hand, this includes the development and sharing of easy-to-use guidelines for IPR and privacy issues. On the other hand, further support (in particular training) is needed with regard to the anonymization of textual data.

- **To raise awareness of language data as Open Data and valuable asset:**
  In this context, several important activities, which require support are foreseen or already in progress:

- **Integrating language data in the national Open Data policy and digital agenda:**
  The Open Data officer was contacted. There is an ongoing cooperation between the PIO and Open Data Portal on language resource collection.

- **Broadening the definition of textual resources by adding speech data, data for AI and other types of language resources:**
  First steps need to be taken in this direction.

- **Sharing benefits of sharing language data:**
  The PIO will take care of this activity as the responsible organisation. In particular, this includes the collection of language resources, which has started this year. One important step will be to enable proce dures for the collection of language resources from the Sworn Translators as well as the processing of language resources acquired from other governmental bodies.

References and further reading list:

Cyprus Government Gateway Portal (Ariadni): http://bit.ly/2lRslx1.

Cyprus Open Data Portal: https://www.data.gov.cy/?language=en.

The Registration and Regulation of Certified Translator Services in the Republic of Cyprus Law 2019: http://bit.ly/2lABxGc.

X. Hadjioannou et. al: *Language Policy and language planning in Cyprus*, 2011, https://www.researchgate.net/publication/232995910_Language_policy_and_language_planning_in_Cyprus.

Annex
Country Profile Czech Republic

Jan Hajic, Pavel Pecina

State of Play:
Translation practices in ministries and public administrations in the Czech Republic:

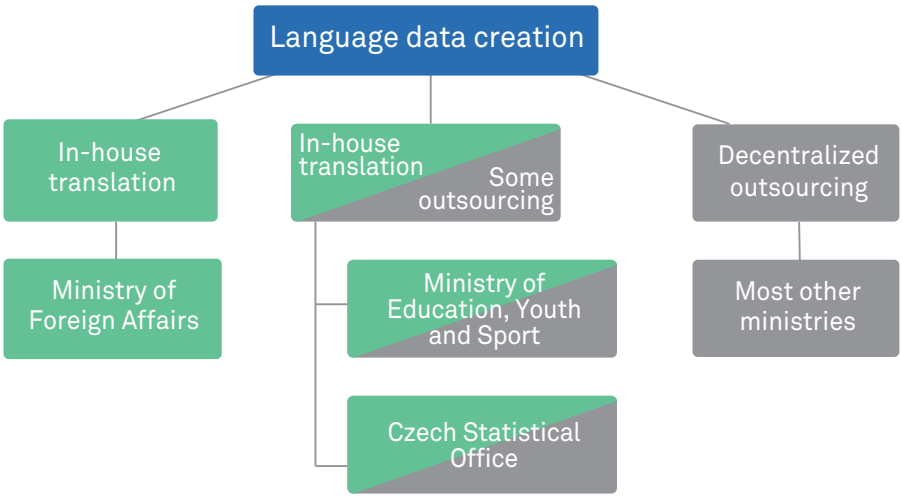
The translation process in Czech public administrations is completely decentralized on the national level. Most public administrations meet their translation needs by outsourcing translations to language service providers (LSPs). The threshold for public procurement is 200,000 Kč (~8,000 EUR). Only the Ministry of Foreign Affairs, the Ministry of Education, Youth and Sports and the Czech Statistical Office have in-house translation services, whereas the latter two also outsource part of their translations.

As regards the applied translation practices, all LSPs and freelance translators are using computer-assisted translation (CAT) tools in their daily operations. The corresponding translation memories (TMs) or any other by-products of the outsourced translations are usually not requested nor automatically referred back, although the use of CAT tools is also common practice in public administrations and ministries. Some public services are also using a machine translation (MT) API or translate their documents with the help of freely available MT web services.

Data sharing infrastructures and Open Data in the Czech Republic:

There is no central repository for translation or language data from public services in the Czech Republic and no infrastructure for continuous language data sharing yet. However, the National Open Data Portal was launched in 2018 encouraging organisations and citizens to share data under open licenses although, currently, very little textual or language data is available on the portal.

The current language data sharing infrastructure in the Czech Republic public bodies looks as follows:



Language policy and digital policy in the Czech Republic:

The Czech language is spoken natively by more than 9 million people, which corresponds to almost 90% of the population in the Czech Republic according to the 2011 census. In descending order, the following languages are spoken natively by minorities in the Czech Republic: Slovak, Moravian, Ukrainian, Polish, Vietnamese and German. Despite the dominant use of the Czech language, Czech is not declared the official or state language in the constitution from 1992 or in a language act.[16] It is however mentioned implicitly as the

[16] Cf. Srpova, Hana: *"Forms of Language Planning and Policy in the Czech Republic"*, in: *Language Planning in the Post-Communist era,* p. 296 ff., 2018.

[17] Cf. Zwilling, Carolin: *Minority Protection and Language Policy in the Czech Republic*, p. 3, 2004.

state language in other laws.[17] Although there is no official language act that regulates the public use of Czech or protects minority languages, the Czech Republic accepted the *European Charter for Regional or Minority Languages* in 2006, granting national minorities rights such as "the facilitation and/or encouragement of the use of regional or minority languages, in speech and writing, in public and private life"[18] including for example the availability of "pre-school education in the relevant regional or minority languages".[19]

Although not constitutionalized, the Czech Language Institute is widely accepted as the regulatory body for the standard Czech language.[20] However, since there is no language act, but several varieties of the Czech Language, there is no officially binding use of a standard variety except for the "codified orthographic and grammatical standard"[21] in Czech language lessons in the education system (only).

The undefined use of the Czech Language permeates in the digital policy, i.e. there is no explicit agenda for language policy or language technology in the digital agenda.

The government Digital agenda governed by the Ministry of the Interior is compiled of four pillars:
- The "Digital Czech" programme
- The "Digital Economy and Society" programme (long-term broad impact and sustainability, "which covers all aspects of government involvement, from legal aspects to direct support to research, development and innovation in the economy" in the digitization process
- The "Information strategy of the Czech Republic"
- The "Czech Republic in the Digital Europe" agenda[22]

The Digital Economy and Society programme has the widest scope as it "covers all aspects of government involvement, from legal aspects to direct support to research, development and innovation in the economy"[23] in the digitization process. Overall, the Czech Republic offers more than 700 public services although only some of them are available digitally or multilingually so far.

Stakeholders:

The Ministry of the Interior is not only mainly responsible for the digital agenda, they also operate the Czech Open Data Portal and are therefore an important stakeholder. More than 30 institutions have participated in past ELRC events, among them the Ministry of the Interior, Social Security Office and Supreme Audit Office who are three of the Czech data donors who already contributed data to ELRC.

Main challenges for sustainable data sharing:

- One of the main challenges for language data sharing is insufficient knowledge, information and expertise related to legal implications of language data sharing, which leads to reluctance to share data.
- Public administrations in the Czech Republic strongly depend on LSPs for translations, who are not eligible to use eTranslation, which potentially undermines the use of eTranslation in public services themselves, but also reduces the incentives to provide and share data to improve eTranslation.

---

18 Council of Europe: *European Charter for Regional or Minority Languages*, p.3, 1992.

19 Cf. Srpova, Hana: "Forms of Language Planning and Policy in the Czech Republic", in: Language Planning in the Post-Communist Era, p. 300; Council of Europe: European Charter for Regional or Minority Languages, p. 4., 2018.

20 Srpova, Hana: *"Forms of Language Planning and Policy in the Czech Republic"*, in: *Language Planning in the Post-Communist Era*, p. 293, 2018.

21 Srpova, Hana: "*Forms of Language Planning and Policy in the Czech Republic*", in: *Language Planning in the Post-Communist era,* p. 296 ff., 2018.

22 Cf. Hajic, Jan; Pecina, Pavel: *ELRC Workshop Report for the Czech Republic*, p. 5, 2018.

23 Hajic, Jan; Pecina, Pavel: *ELRC Workshop Report for the Czech Republic*, p. 5, 2018.

Annex
Country Profile Czech Republic

Action plan:

- **To raise awareness of language data as Open Data and valuable asset by**
  - Identifying an Open Data officer on the national level
  - Establishing practical guidelines for LR as Open Data
  - Sharing the benefits of sharing language data

- **To increase interest in MT/LT in public services as part of the national digital policy by**
  - Diffusing best practices where technology proves cost-cutting and leads to increased productivity
  - Securing support of decision makers and key players

- **To tackle legal concerns by**
  - Developing and sharing easy-to-apply guidelines for IPR and privacy issues with the help of legal experts
  - Finding ideas to implement rights management along with data management supported by legal experts

- **To identify and gain access to outsourced translations by**
  - Cooperating with other projects, like e.g. NEC TM
  - Establishing a common practice of receiving any by-product of outsourced translations (although this may be difficult, since external experts might not be prepared to share such data)

- **To establish good language data management practices**
  - Identification of data managers
  - Investigation of language data management practices
  - Definition of confidential and personal data and separation of confidential/personal data from public sector information in the translation process

References and further reading list:

Council of Europe: *European Charter for Regional or Minority Languages*, 1992, https://rm.coe.int/168007bf4b.

Czech Open Data Portal: data.gov.cz.

Centre for Language Research Infrastructure in the Czech Republic: https://lindat.cz/.

Hajic, Jan; Pecina, Pavel: *ELRC Workshop Report for the Czech Republic*, 2018, http://lr-coordination.eu/sites/default/files/Czech%20Republic/2018/ELRC%2BWorkshopReport_Public_CZ-FINAL.pdf.

Srpova, Hana: *"Forms of Language Planning and Policy in the Czech Republic"*, In: Andrews E. (eds) Language Planning in the Post-Communist Era, 2018, https://link.springer.com/chapter/10.1007/978-3-319-70926-0_12.

Zwilling, Carolin: *Minority Protection and Language Policy in the Czech Republic*, 2004, http://www.gencat.cat/llengua/noves/noves/hm04tardor/docs/zwilling.pdf.

Annex
Country Profile Denmark

Sabine Kirchmeier, Bolette Pedersen, Lilli Smal

State of Play:
Translation needs and practices in ministries and public administrations in Denmark:

Translation needs:

Denmark has a fairly small population of about 5.8 million residents constituting a small linguistic area for Danish - its sole official language. Still Denmark is a multilingual country with Faroese, Greenlandic and German as recognized regional languages, and Swedish also commonly spoken in the area around Copenhagen. Danish Sign Language is not officially recognized, but Danish Sign Language users are supported by the state, and the Danish Sign Language Council is responsible for providing documentation of and information about Danish Sign Language. Among the main immigrant languages are Arabic, Turkish, Polish and Romanian/Romani. Overall, English is the dominant second language, and English language competence among Danish citizens is very elevated. Translation demands, however, are still high for official EU-languages, non-official EU-languages and immigrant languages.

On the national level, for example, the Danish Agency for Labour Market and Recruitment has a strong need for translation from all EU-languages into Danish in the Electronic Exchange of Social Security Information system (EESSI), where electronic documents are exchanged across sectors, and where next to standardized text, a lot of information is given in open text fields. Still, little experience with automated translation has been made so far. Also, the Danish parliament regularly needs translations into and from Greenlandic and Faroese, both not official EU-languages. One of the main needs, especially at the municipal level, is the translation of websites. Very few official websites are available in other languages. Currently, some municipalities have integrated a popular freely available machine translation service into their websites to meet the translation demands from migrant workers and refugees.

Translation practices:

Danish state organs are obliged to outsource 35 % of services that can be made subject to competition to private vendors, and it seems that most public administration outsource their translations. Only a few public institutions have in-house translation services such as the Region South Jutland-Schleswig, the Danish Tax Authorities and the Nordic Council of Ministers. Until the end of 2017, the Ministry of Foreign Affairs had an in-house translation service, which also delivered translation services to other public institutions. In 2018, the translation unit was dissolved and all translations are now outsourced to private vendors. Currently, about 80-90% of translations in the public sector are outsourced to private vendors with a growth tendency. Although from 2012 to 2017 translation services for at total amount of EUR 9.1 million were procured in 11 public tenders (according to opdentender.eu.dk https://opentender.eu/dk/), there is no common approach for procuring language services, although procurement in general is highly regulated, organized and transparent. Translation memories and other by-products of translations are not systematically requested back together with the translation. The lack of a systematic approach to archiving and reusing translated text indicates that the importance and value of language data collection is widely underestimated (unrecognized). In addition, there is currently only one very small Danish provider for translation memory systems.

The lack of a common approach and of recognition of the importance of language services is underlined by the results of a survey, where public servants on the municipal level were asked how they deal with translation demands.[24]

The results about the translation practices on the municipal level in Denmark show that:

- 9,5 % use internal language professionals
- 14,3 % use external private vendors
- 23,8 % use employees who know the language
- 47,6% don't know
- 4,8% other

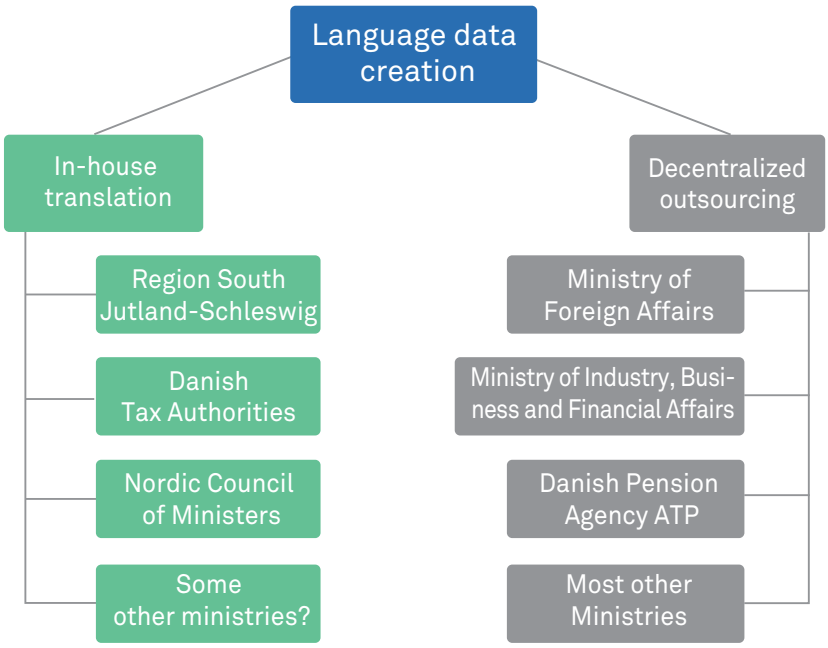
[24] Ingemansson, Meyer, Kjærgaard & Kirchmeier: Sprogarbejdet i danske kommuner. Rapport. Dansk Sprognævn 2017.

Annex
Country Profile Denmark

The results illustrate that here too, no common approach is applied, and that there is little awareness of the value of textual data and the need to collect such data to make it available to the research community, the industry and the public at large.

The current language data creation infrastructure in Danish public bodies:

Restructuring and Goals – 2018 and onward



Data sharing infrastructures in Denmark:

The Centre for Danish Language Technology is the Danish node for DK-CLARIN and is hosting one of the infrastructures/portals in Meta-SHARE. There are also similar initiatives in place at other universities. In fact, all university centres involved in language technology and language data, tend to have their own github repository where Danish and other language data are shared. The Alexandra Institute (an advanced non-profit technology group funded by the Danish Ministry of Science, Technology and Innovation) is planning to collect and share Danish language resources but there is no structured overview of the data available yet. The Danish Research Council and the Ministry of Higher Education and Science have from time to time allocated funds for the collection of language resources but the initiatives and processes have not been coordinated and streamlined so far. The most interesting progress has been made in relation to terminology and concept modelling in the Danish Agency for Digitization.

When it comes to open government data in general, access is provided through The Data Distributor (datafordeler.dk), the Danish Open Data Portal (portal.opendata.dk/), The Data Catalogue (datahub.virk.dk) and several other portals. Based on the Data Agreement from 2012, public sector data must be made available to public bodies, citizens and businesses. Although Denmark is a highly digitized country, it is falling behind when it comes to opening and sharing government data and no commitment to open by default government data has been made yet. The Ministry of Finance who has authority to commit other ministries to use the national data catalogue and update the information continuously, has not exercised this right as of March 2019.[25] Currently, 921 data sets are openly available in 88 different places, however, none of these are textual data sets.

[25] Cf. Folketinget Statsrevisorerne: *Extract from Rigsrevisionen's report submitted to the Public Accounts Committee Open Data*, p.2 ff., 2019.

Digital policy and language technology in Denmark:

Although Denmark is a highly digitized country, not many (financial and human) resources have been allocated for the development of language technologies in the past. Recent developments, however, show a significant shift in the perception of the importance of language technologies for the preservation of the Danish language. To this end, a language technology committee was established in 2018 by the Minister of Culture tasked with the development of a proposition for a national strategy for language technology in Denmark. This committee, headed by the Danish Language Council, produced a report in April 2019 including a list of recommendations and an overview of available resources. The committee simultaneously advised the Ministry of Finance who announced in October 2018 a new strategy for providing world class digital services including world class language technology - recognizing that the two fields cannot be separated from each other. In the "National Strategy for Artificial Intelligence" published by the Ministry of Finance and the Ministry of Industry, Business and Financial Affairs on behalf of the Danish Government in March 2019, one of four key elements are the creation of language resources that can be widely used to develop language technologies for Danish (cf. National Strategy for Artificial Intelligence (Denmark), p. 18). The Danish government has therefore awarded 30 million DKK (4 million Euro) to the development of language resources in order to boost the development of artificial intelligence for Danish.

Interesting fact:
Denmark is a highly digitized country and public digital services are very advanced and well accepted by the general public.

Key stakeholders:

- Danish Agency for Digitization (Ministry of Finance)
- Center for Language Technology/University of Copenhagen (Ministry of Education and Science)
- Danish Language Council (Ministry of Culture)

Main challenges for sustainable data sharing:

The above described translation and data sharing practices indicate that, at the moment, sustainable language data sharing infrastructures and processes are not in place yet. This is mostly due to these main challenges:

- Danish is a small language community, which is why there is generally less language data available and the market for LT for Danish is small
- General lack of coordination of research programs and strategic research projects complicate LR collection
- Little recognition of textual data being valuable in the public sector, and therefore no systematic approach to the curation and sharing of public data resources
- Only a few local developers of language technology products
- Strong tendency to outsource translation projects.

Recommendations from the Danish language committee include:

- A central organization should be established to plan and manage a national Danish language resource bank.
- The language resource bank should contain high quality resources of the following kind:
  - A time annotated Danish speech technology corpus
  - Danish text corpora and annotated gold standards for machine learning
  - A comprehensive lexical database
  - A Danish terminology bank
  - A language technology tool-kit
  - A language portal for the distribution of resources
- More education of training of experts in language technology for Danish
- More research on language technology for Danish (cf. Danish Language Technology report, p. 63 and 72).

Annex
Country Profile Denmark

In the *National Strategy for Artificial Intelligence*, the Danish Government acknowledged the importance of data collection by naming more and better data as one of four focus areas next to: a responsible foundation for AI, strong competences and new insights, and increased investment (cf. *National Strategy for Artificial Intelligence* (Denmark, p. 18). The report on language technology for Danish includes specific recommendations on what language data should be collected and how they should be made available.

Action plan:

The action plan for language technology is managed by The Danish Agency for Digitisation. The project is currently at an early stage, where the focus is on planning and organizing, how to deliver short term wins while still keeping a focus on the long term agenda.

The aim is to develop a technical platform containing various free Danish language resources and functionalities aimed for the NLP-industry.

Next steps:

- Incorporating and upgrading existing Danish digital dictionaries, terms and lexical resources
- The development and implementation of a time-encoded Danish speech recognition corpus in cooperation with stakeholders in Danish media
- Identifying and developing new language technologies in cooperation with stakeholders

References and further reading list:

Danish Government: *National Strategy for Artificial Intelligence*, 2019, https://en.digst.dk/media/19337/305755_gb_version_final-a.pdf.

Folketinget Statsrevisorerne: Extract from *Rigsrevisionen's report submitted to the Public Accounts Committee Open data*, 2019, http://uk.rigsrevisionen.dk/media/2105082/12-2018.pdf

Ingemansson, Meyer, Kjærgaard & Kirchmeier: *Sprogarbejdet i danske kommuner. Rapport. Dansk Sprognævn*, 2017, https://dsn.dk/nyt/nyheder/2017/sprogarbejdet-i-danske-kommuner-ny-rapport-fra-dansk-sprognaevn/?searchterm=danske%20kommuner.

Kirchmeier Sabine: *ELRC Workshop Report for Denmark,* 2018, http://lr-coordination.eu/sites/default/files/Denmark/2018/ELRC%2B%20Workshop%20Report_Denmark.pdf.

Kirchmeier, Diderichsen, Hansen & Henrichsen: *Dansk Sprogteknologi i Verdensklasse*, Rapport fra sprogteknologiudvalget under Dansk Sprognævn nedsat af Kulturministeriet. Dansk Sprognævn, 2019, https://dsn.dk/udgivelser/sprognaevnets-udgivelser/sprognaevnets-rapporter/sprogteknologi-i-verdensklasse

News article by Agency for Digitisation, Ministry of Finance: *New national strategy: Artificial intelligence should benefit individuals, businesses and society as a whole*, 2019, https://en.digst.dk/news/news-archive/2019/march/new-national-strategy-artificial-intelligence-should-benefit-individuals-businesses-and-society-as-a-whole/.

Annex
Country Profile Estonia

Andero Adamson, Kadri Vare

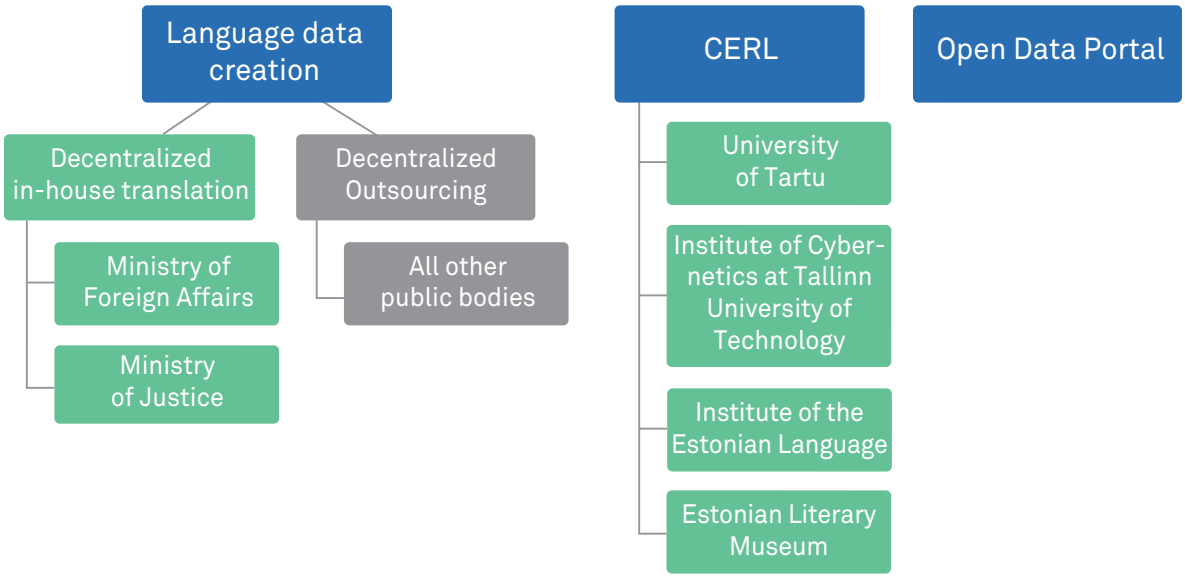
State of Play:
Translation practices in ministries and public administrations in Estonia:

In Estonia, each public sector institution is responsible for its own translation services. Translation needs or procurement services are not centralized although all public bodies outsource at least part of their translations, either independently or through public procurement for order amounts above the threshold of 10,000 EUR. Only the Ministry of Foreign Affairs and the Ministry of Justice have in-house translation services but they also outsource part of their translations. So far, translation memories (TMs) are not requested back, as there does not seem to be a need for it, especially without an in-house translation service that could make use of it and/or maintain the TMs. Still, awareness raising in the past couple of years triggered a shift in public procurement processes as the benefits of re-using TMs have been recognized by state authorities.[26]

Together with the Ministry of Education and Research, the Ministry of Economic Affairs and Communications conducted a survey about the usage of TMs and translation arrangements during 2018 in the public sector in September 2019. One of the survey's objectives was to raise awareness of re-using TMs. 58 public sector organizations participated in the survey. Only a few institutions have in-house translators, who use different TMs and only one institution is using eTranslation officially. Other machine translation (MT) systems are used irregularly and for personal use only. Most translations are to and from English, but also Russian, Finnish and other Baltic languages are being translated. Exact numbers of translated pages cannot be determined, but based on the survey, around 1 million pages were reached in 2018 and translation costs sum up to more than 1 million euros per year. According to the survey, public administrations and ministries would be interested in a centralized MT of TM-based domain-dependent systems, which considers their specific terminology.

Data sharing practices:
There is no central language-related data exchange infrastructure for the public sector on the national level in Estonia. However, language technology resources are collected and shared through the Center of Estonian Language Resources – CERL. The CERL also serves as the national node of the CLARIN infrastructure.[27]

The current language data creation and language resource sharing infrastructure in Estonian public bodies looks as follows:



[26] Cf. Luts, Martin: *ELRC Workshop Report for Estonia*, p. 15, 2018.
[27] Cf. Center of Estonian Language Resources

Annex
Country Profile Estonia

Open Data in Estonia:
The Estonian Open Data Portal (https://opendata.riik.ee/en) provides a single point of access for the general public to unrestricted public sector data with the permission to re-use and redistribute such data for both commercial and non-commercial purposes.

Language policy in Estonia:
Estonian is constitutionally the state language in Estonia and also one of the official EU languages. The current Estonian language development plan (2011-2020) is used as a basis for the sustainable development of the Estonian language. The strategy is used as a blueprint for planning and financing all four areas covered with a special focus on Estonian as a first language (L1).

Strategic planning for the development of the Estonian language started in 1998. The current strategy covers four areas: Estonian as first language; Estonian as second language; Estonian abroad and multilingualism, including foreign languages. In Estonia, the Ministry of Education and Research is responsible for the development of language policy.

The digital future of the Estonian language highly depends on the state of Estonian language technology (LT). By building resources and investing into technologies required for machine translation, speech recognition and speech synthesis, the position of Estonian in the digital sphere will be strengthened.

Stakeholders:
Key stakeholders include the Ministry of Education and Research, the Ministry of Economic Affairs, the Ministry of Justice and the Center of Estonian Language Resources. The second Estonian ELRC workshop received 73 registrations, spanning a wide range of ministries and public organizations, but also language service providers (LSPs) and academia. Before the workshop, targeted communication activities were organised to ensure that key relevant public administrations were represented. With almost 50 participants, the workshop was well-attended. Over 40% of the participants were representatives from public services and public administrations, 35% were participants from the technology eco-system and 14% were language service providers. The remaining participants were part of the organising committee and the European Commission.

Main challenges for sustainable data sharing:
- Incoherent data sharing practices
- The low value of language data
- Little awareness of the potential of language data

Action plan:
Based on the identified challenges, the following six objectives were defined:

- Raising awareness of language data as Open Data and making language data as open as possible.
- Establishing good data management practices in public services.
- Identifying and gaining access to outsourced translations.
- Increasing interest in MT in public services.
- Creating a central MT service for public services.
- Addressing legal concerns.

References and further reading list:
Artificial Intelligence for Estonia: https://www.kratid.ee/in-english

Center of Estonian Language Resources: https://www.etag.ee/en/funding/infrastructure-funding/core-infrastructures/center-of-estonian-language-resources/.

Estonian Language Foundation: *Development Plan of the ESTONIAN LANGUAGE 2011–2017*, 2011, https://www.hm.ee/sites/default/files/eestikeelearengukavainglise.indd_.pdf.

Luts, Martin: *ELRC workshop report for Estonia*, 2018, http://lr-coordination.eu/sites/default/files/Estonia/2018/ELRC%2B%20Workshop%20Report%20Tallinn%202018_v1.0.pdf.

Ministry of Education and Research: *The Language Technology Research and Development Program "Estonian Language Technology 2018-2027"*, 2018, https://www.hm.ee/sites/default/files/estonian_language_technology_2018-2027.pdf.

Ministry of Education and Research: *National programme "Estonian language and cultural memory II (2014-2018)"*: https://www.hm.ee/en/activities/research-and-development/research-programmes.

Annex
Country Profile Finland

Kaisamari Kuhmonen, Krister Lindén, Lilli Smal

State of Play:
Translation practices in ministries and public administrations in Finland:

Finland is one of the few countries that has a fully centralized translation service on the ministerial level. In 2015, the translation services from all the 12 government ministries were regrouped in a centrally organized Translation and Language Services Division (TLD) located at the Prime Minister's Office. Its 67 language specialists now provide translation, language and terminology services to all the ministries covering about 50% of the translation needs.

Interesting fact:
Finland has a centralized translation service for all 12 government ministries.

The incoming requests are handled with an internal request management tool called "VIRKKU", a service application, which is also used for the management of other services. TLD uses computer-assisted translation (CAT) tools for all translations and has an internal term bank and translation memories (TMs) as well as the government external term bank "Valter" (www.valter.fi). New term bases published in Valter will also be published in Open Data format on the Finnish Open Data Portal (www.avoindata.fi/en). However, the Government hesitates to share TMs as they also contain all non-public pre-final versions.

50% of the translations are centrally outsourced to language service providers. It is stipulated in the translation contracts that language service providers (LSPs) have to transfer translations, copyrights of the translations and all TMs to TLD as a minimum criterion. This criterion is not negotiable. TLD provides exports of their TMs to LSPs if they are relevant to the requested translation. However, the returned TMs are archived separately.

The TLD coordinates the government level translations and language resource exchange and in addition, supervises and develops language usage in the ministries. The TLD has recently introduced two customised neural machine translators (EU Council Presidency Translator) and is testing a third one (Fiskmö). It would have been impossible to develop these without the open sharing of language data.

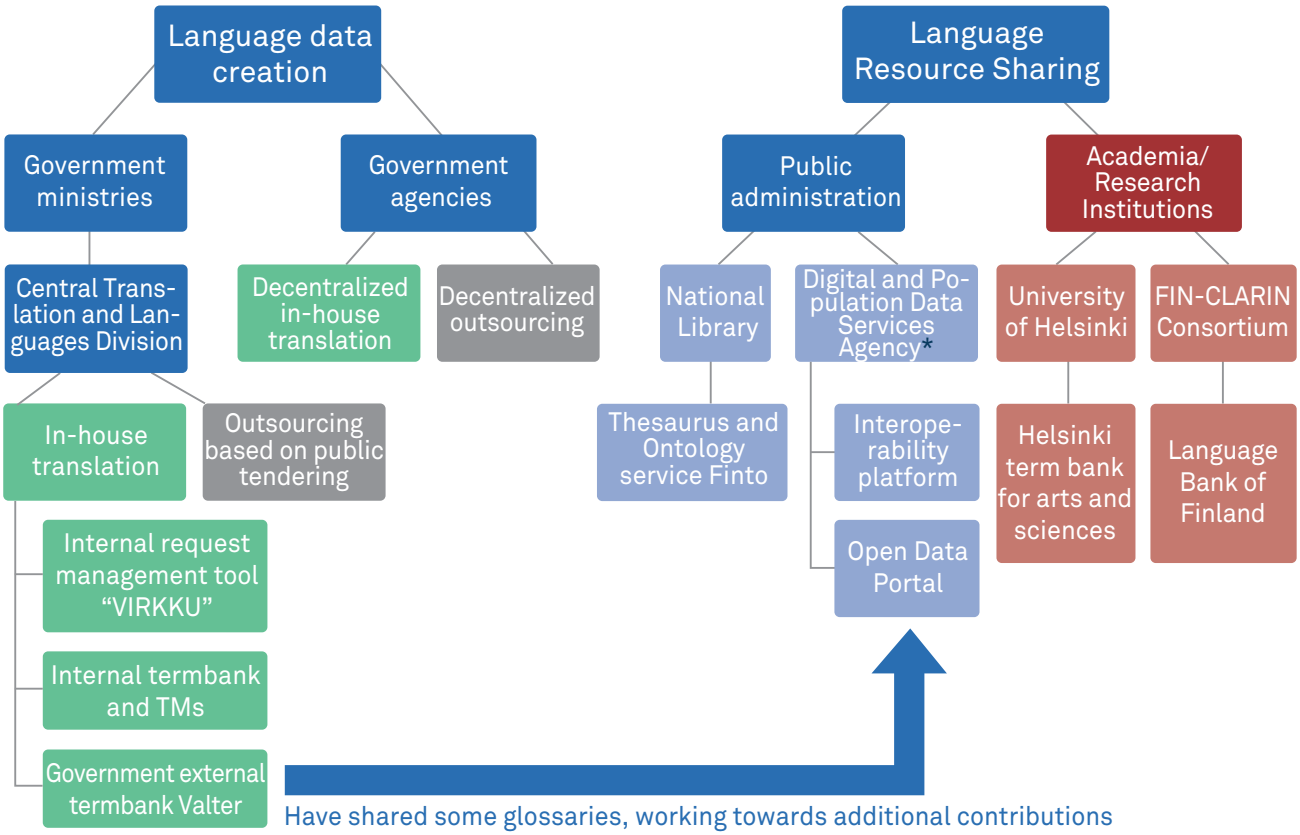
Other government agencies take care of their translations independently with either in-house translation or outsourcing translations to language service providers. The same process holds true for translation needs on the regional and local level. Because of the bilinguality of the country, the demand for translations to and from Swedish is considerable.

Data exchange:
Apart from the access to terminological data through Valter, a new interoperability platform gives public bodies the tools and the method for specifying and managing interoperable data and information content. The platform consists of the terminologies, code lists, and data models needed for flows of data and other forms of information management.

The TEPA term bank maintained by the Finnish Terminology Centre TSK (Sanastokeskus TSK) contains special language terms and definitions in approximately 365 000 terminological entries. Another term bank is the Helsinki term bank for arts and sciences. Additionally, the Finnish thesaurus and ontology service "Finto" is available online.

Occasionally, TLD also exchanges TMs with the Finnish Parliament or other public bodies for specific translations.

The current language data sharing infrastructure in Finnish public administrations looks as follows:



* At the time of writing the report, the Digital and Population Data Services Agency is still known as the Population Register Centre but its transformation and renaming is already announced for 1 January 2020.

Language policy and digital policy in Finland:

The two official languages in Finland are Finnish and Swedish. The Ministry of Justice is responsible for promoting the realisation of linguistic rights by monitoring the implementation and application of the Language Act and by issuing recommendations. The Ministry of Justice has also developed an indicator tool for following-up the implementation of the linguistic rights.

The Strategy for the National Languages of Finland was published in 2012. The strategy is accompanied by a toolkit and an action plan. The Strategy will be revised in 2020 and the new version will also examine the national languages of Finland from the perspective of digitalisation and artificial intelligence.

Finland is one of the few European countries that has an agreed strategy for artificial intelligence including aspects of language technology to make public services available to citizens in their own languages. For instance, the Finnish State Development Company Vake's spearhead projects in Finnish language AI development aims at generating Finnish language AI resources, such as program libraries, ready-to-use language models and training materials, alongside more mainstream language. These resources can be utilised by all parties using Finnish language AI either in adapting hardware or software development.[28]

[28] Cf. The Finnish State Development Company: *Boosting the Finnish Language AI*, 2018.

Annex
Country Profile Finland

Stakeholders:

The key stakeholders related to language policy and language data sharing are:

- The Ministry of Justice who is responsible for monitoring the implementation of the language policy.
- The Translation and Language Services Division at the Prime Minister's Office who has the largest public administration translation service and is therefore the main language resource creator and holder in the public sector in Finland.
- The Digital and Population Data Services Agency (known as Population Register Centre until the end of 2019)[29] that is responsible for the Interoperability Platform and the Open Data Portal, among other things.

The key stakeholders and decision makers for digitalisation and technology are:

- The Ministry of Finance provides preconditions for the digitalisation of the public sector and sets a strong example. This is done for instance, by promoting interoperability, AI and robotisations across administration and enabling the security of authorities' activities.
- The Ministry of Economic Affairs and Employment that is responsible for Finland's innovation and tech nology policy, among other things. The Ministry also promotes business digitalisation.
- The Ministry of Transport and Communications whose key duties include improving access to data, providing opportunities for data-based businesses by means of regulation, drafting legislation concerning data resources and the use of information.

In the past few years, almost 300 Finnish stakeholders could be identified that represent ministries and agencies on the government and the local level, public online services, language service providers and research institutions. Among the stakeholders that already shared language data with ELRC are the Translation and Language Services Division of the Prime Minister's Office, Statistics Finland, the University of Helsinki, the Finnish Terminology Centre TSK and the City of Helsinki.

Main challenges for sustainable data sharing:

Finland is one of the most advanced countries with respect to sharing language data that falls in the scope of public sector information. Still, it faces a few challenges that hinder language data sharing.

- Government hesitates to share TMs as they contain all non-public pre-final versions also. This is also a problem in government agencies and in other public bodies.
- Lacking awareness and knowledge of secure use of MT
- Anonymisation of language data

Action plan:

For Finland, 5 objectives were defined that will help to foster language data sharing infrastructures and awareness of the value of language data. Ranked according to their priorities, these recommended objectives and actions are:

- **Increasing interest in MT/LT in public services as part of the national digital policy, specific actions include:**
  - Establishing synergies with national projects and initiatives
  - Diffusing best practices, where technology proves cost-cutting and increases productivity
  - Securing support of decision makers to adapt the national policy
  - Communicating facts about language data, such as how much data is needed to improve a MT system or details about the anonymization process of data
- **Raising awareness of language data as Open Data. Some targeted actions are:**
  - Including language data in the national Open Data policy and digital agenda
  - Establishing practical guidelines for Language Resources as Open Data
  - Sharing the benefits of sharing data

[29] The Population Register Centre will be renamed to Digital and Population Data Services Agency (short version Finnish Digital Agency) as of 1 January 2020.

- **Raising awareness of secure training of MT engines and use of MT:**
  This objective is targeted at establishing practical guidelines for secure MT training, taking into account information security and the safe use of MT services.

- **Tackling legal concerns:**
  This objective mainly addresses the development and distribution of easy to apply guidelines for Intellectual Property Rights (IPR) and privacy issues in textual data.

- **Establishing good language data management practices in public services:**
  This objective is specifically targeted at the translation process. It includes:
  – A solution for the separation between confidential texts (e.g. working versions of official documents) from public information such as official publications.
  – Defining the best and most efficient process for sharing language data with minimal extra effort for anyone involved in the process.

References and further reading list:

Finnish Interoperability Platform: https://yhteentoimiva.suomi.fi/en/.

Finnish Open Data Portal: https://www.avoindata.fi/en.

Finnish Thesaurus and Ontology Service (Finto): http://finto.fi/en/.

Helsinki Term Bank for Arts and Sciences: https://tieteentermipankki.fi/wiki/Termipankki:Etusivu.

Ministry of Finance: *AuroraAI – Towards a human-centric society,* https://vm.fi/documents/10623/13292513/AuroraAI+development+and+implementation+plan+2019–2023.pdf/7c96ee87-2b0e-dadd-07cd-0235352fc6f9/AuroraAI+development+and+implementation+plan+2019–2023.pdf.

Ministry of Economic Affairs and Employment of Finland: Leading the way into the age of artificial intelligence. Final report of Finland's Artificial Intelligence Programme 2019, 2019, http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/161688/41_19_Leading%20the%20way%20into%20the%20age%20of%20artificial%20intelligence.pdf.

Ministry of Economic Affairs and Employment of Finland: *Finland's age of Artificial Intelligence. Turning Finland into a leading country in the application of artificial intelligence. Objective and recommendations for measures,*https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160391/TEMrap_47_2017_verkkojulkaisu.pdf, 2017.

TEPA Termbank: http://www.tsk.fi/tepa/en/.

The Finnish State Development Company: *Boosting the Finnish Language AI*, 2018, https://vake.fi/en/ohjelmat/boosting-finnish-language-ai/.

Annex
Country Profile France

Thibault Grouas, François Yvon, Hélène Mazo

State of Play:
Translation practices in ministries and public administrations in France:

In France, there is no central translation service or procurement contract and no systematic exchange of translations and/or knowledge between public administrations. Translations for public administration bodies are mainly outsourced to Language Service Providers (LSPs) or freelance translators.

At the central level, two ministries have well-structured in-house services employing translators on a permanent basis with a specific civil servant status. The Ministry of Finances which employs a team of 20 translators, produces approximately 30,000 pages each year using computer-aided translation (CAT) tools and maintains the Minéfiterm, a terminological database of 80,000 terms (in 15 languages, covering 40 domains). They mainly work for Tax and Budget services in the Ministry of Finances, but also provide translations for other administrations such as Customs. They outsource to trained freelance translators. The Ministry of Foreign Affairs' team is tighter with 13 translators working on translating civil status documents, but also speeches and diplomatic reports and documents for the President and the Prime Minister. They resort to freelance translators for all translations over 5,000 words. They also use CAT tools and maintain a terminological database. Some in-house translation departments can be found in other ministries, but they are usually very small and work only for their own administration. The service in charge of international affairs at the Ministry of Interior Affairs for instance has a very limited team and addresses the internal translation requests, mainly confidential.

The only other public administration with in-house translation services is CLEISS (Centre of European and International Liaisons for Social Security), France's single help-desk for international mobility and social security which meets the translation needs of French social security institutions. With 50,000 pages being translated on average each year from 40+ languages, it is France's premier public translator.

Data sharing infrastructure and Open Data in France:

"Data as infrastructure" is the motto of the French Government. Within the legal framework of the Digital Republic Bill voted to foster data openness, the Government has based its data policy on three main objectives: provide high-quality data through the public data service; enable the flow of data by applying the "access by default" principle to all communicable data (through the development of APIs and data platforms promoting the exchange of data between administrations and with the civil society); exploit the data in order to improve the efficiency of public initiatives. Etalab is France's Chief Data Officer and, under the authority of the Prime Minister, coordinates the action of public administration with regard to data inventory, governance, production, circulation and use. As the government's task force for Open Data and data policy, the agency brings support to all French public administrations to facilitate the publication and re-use of public information on the Open Data Platform: data.gouv.fr.

The amount of language data shared on the French Open Data Platform is still limited and the exchange and sharing of data between public administrations has not been formalized. There are collaborative initiatives, such as the Inter-ministerial Working Group on Translation (GIT) co-chaired by the Ministry of Finances and the Ministry of Culture (DGLFLF), which takes place every 6 months for almost 15 years, and gathers all the translation professionals working in the public sphere, including translators, terminologists, academics, decision makers to exchange information and best practices. But this remains quite informal. The translation services in French public administrations remain reluctant to share their data, whether translation memories or translated documents, mainly out of confidentiality concerns.

Language policy in France:

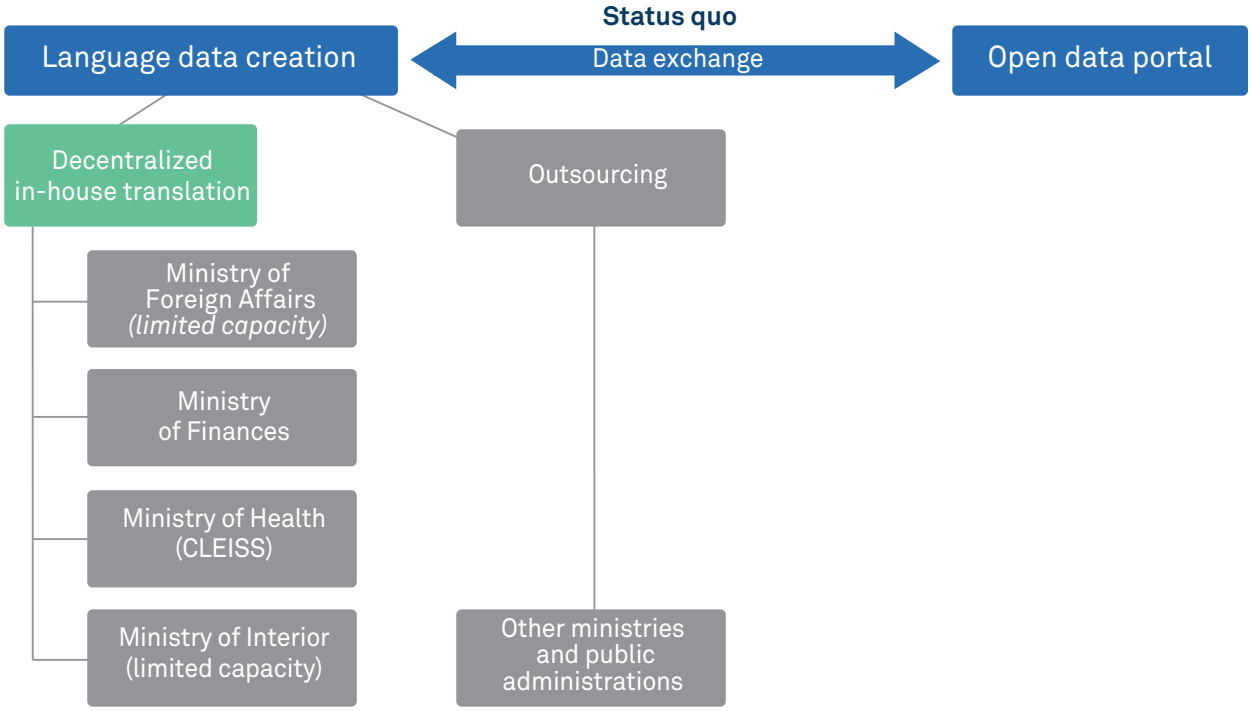
French is the official language of France and appears as such in the Constitution (article 2) since 1992. However, as the white paper on *The French Language in Digital Age*[30] reminds us, in order to take into account the European Charter for Regional or Minority Languages the Constitution (article 75-1) acknowledges regional languages spoken in France as part of French cultural heritage, since 2009. Nevertheless, French re-

[30] META-NET White Paper Series. Mariani et. al: *The French Language in the Digital Age*, 2012.

mains the language mainly used in France, and there are strong constraints for its use in the public sphere (school, administration, television, etc.)

In France, the Délégation générale à la Langue Française et aux Langues de France (DGLFLF), a body of the Ministry of Culture, is in charge of coordinating the French Government's language policy. Currently, DGLFLF plays a prominent role in the implementation of President Macron's plan: "An ambition for the French language and multilingualism" presented on 20 March 2018 at the Académie française, which encompasses measures such as the teaching of two European languages in addition to the native language as well as language training in European and international institutions. As part of the workplan Culture 2018-2022, both translation and digital technology as a tool for multilingualism are the key objectives in the French Government's strategy. Both issues will also be highlighted during the French EU Presidency in 2022.

The current language data creation infrastructure in France public bodies:



Stakeholders:

The French National Points represent two key institutions related to language policy (P-NAP from DGLFLF) and language technology (T-NAP from LIMSI-CNRS), which highlights the interest and importance of this topic in France. The local ELRC events were attended by representatives from many public institutions, including the ANR (research agency), the Ministry of Finances, the CLEISS, Etalab (the Open Data agency), the CNRS, the Banque de France and the French translators' union (SFT). Some of these institutions have already contributed language data to ELRC, namely the Ministry of Finances and the ANR. Other stakeholders like the Banque de France-ACPR have developed their own NMT solution as the confidentiality of the data is key to their activity.

On the French scene, there are a number of French and international institutions based in Paris which, having to deal with the growing volume of translation, are looking into the support that can be provided by language technologies, with a specific interest in the eTranslation developments. For instance, the OECD, based in Paris, is currently developing its own NMT engine for both their official languages French and English. There are also French public services such as OFPRA (French Office for the Protection of Refugees and Stateless

Annex
Country Profile France

Persons) where daily activities involve interviews in up to 127 different languages or the Inter-ministerial Delegation for refugees' reception and integration (DIAIR), a body of the Ministry of Internal Affairs, which helps refugees make their way into the administrative procedures by providing multilingual information on a dedicated collaborative platform.

Main challenges for sustainable data sharing:

The challenges France is facing in sharing data are as follows:

- IPR Issues
  - Privacy concerns prevent ministries from sharing data (Finances, Interior, Foreign Affairs)
  - By law, translators own the translation and the translation memory, resulting IPR issues can hinder the sharing of language data.
- Translation workflow
  - LSPs and translators know how to manage the translation data, but only very few public administrations do.
  - CAT tools are not always used, even in institutions with in-house translators, because they deal with documents or formats for which CAT tools are not considered useful/needed.
  - Not all institutions make sure to collect the TMs along with the translated document(s) and some simply lack technical staff to process the parallel texts into TMs a posteriori.
- Data sharing
  - There is no TM exchange because there is no infrastructure for this at the State level. The Open Data Portal is not used for this purpose in France.
  - In some institutions, the common perception is that since most of their documents are confidential, they cannot share even the reports that are public and could be useful for training the MT engine.
  - In many cases, the sharing could be easily processed if the decision-maker at the administration level could be easily identified/convinced.

Action plan:

The public administrations are invited to share their data, but in practice, there are many steps, which should be taken (IPR clearing, anonymization, formatting, etc.) that they cannot always manage. We have seen that very few public administrations have technical staff in-house.

A set of guidelines should be drafted detailing how to overcome the obstacles in data sharing. These guidelines should be disseminated at the management level as hierarchy is considered very important in French administrations. The support from Etalab in France could be seeked.

References and further reading list:

French Open Data Platform: https://www.data.gouv.fr/en/

Digital Republic Bill: https://www.republique-numerique.fr/pages/digital-republic-bill-rationale

META-NET White Paper Series. Mariani et. al: *The French Language in the Digital Age*, 2012, http://www.meta-net.eu/whitepapers/e-book/french.pdf.

Annex
Country Profile Germany

Alexandra Soska, Andreas Witt, Lilli Smal

State of Play:
Translation practices in federal ministries and public administrations in Germany:

The way translations are carried out in Germany varies depending on the administrative level. All federal ministries have an in-house translation service whereas only some federal authorities and very few state ministries have their own translation service. It is the common practice to use computer-assisted translation (CAT) tools, including translation memories (TMs) and terminology systems (termbases), in the translation process both by in-house translation services of the federal ministries as well as by language service providers (LSPs) and freelance translators. To fully meet their translation needs, all public authorities outsource at least some translations to either freelance translators or language service providers. When it comes to outsourcing, three different scenarios exist: public authorities that have in-house translation services have a smaller demand for outsourcing translations and therefore do not usually call for tenders but vendor small contracts to freelance translators that provide high quality translations and are generally willing to share their translation memories with the contracting authority. Public administrations that do not have their own translation service, outsource translations to language service providers and usually do not request that the translation memories and new terminological entries are returned to them as they do not see the need for it since the TMs will not be reused in-house. In general, all federal authorities may also order translations under a central call for tenders but are not obliged to do so.

Language data sharing and Open Data in Germany:

There is no formalized exchange of information or data between public services unless they are part of the terminology network ("Terminologiedatenbankverbund"). The majority of federal translation services are part of the terminology network. Despite its name terminology is not the main focus of this network. Its main purpose is to exchange information about translation practices and developments in the field. Additionally, all members use the collaborative platform MultiTerm Online granting reading access to terminology databases of other public services. However, there is no active exchange of TMs or textual data.

> **Interesting fact:**
> **In Germany, the copyright (das Urheberrecht) of translations belongs to the translator by default and is not transferable. For rightful reuse of the texts, respective licenses are needed.**

In the Federal Government's National Action Plan to implement the G8 Open Data Charter, (*Nationaler Aktionsplan der Bundesregierung zur Umsetzung der Open-Data-Charta der G8*) it is explicitly stated that: "Unstructured information such as notes, files, studies, reports or other texts do not constitute data in this sense." and therefore do not fall in the category of Open Data.[31] However, in the course of the past years, several contacts were established with the German Open Data community. The German Public Services NAP for example spoke to the Open Data coordinator at the Federal Ministry of the Interior and offered to participate in an Open Data pilot at the ministry to advocate for language data to be included in the pilot project. Some interest was also shown in publishing language resources (LR) on GovData, the German Federal Open Data portal. However, currently there is no function on GovData to search for language data, language resources or textual data on the portal.

According to the E-Goverment Act[32], public authorities ought to make data that is of public interest and can be shared according to the "Informationsverarbeitungsgesetz"[33], the German implementation of the Public Sector Information directive (2003/98/EC), available in machine readable format online. However, data is defined as structured information mainly available in tables and lists.[34] Although this definition does not explicitly exclude language data, it does show the implicit interest of statistical and numerical data and little

[31] Federal Ministry of the Interior: *Nationaler Aktionsplan der Bundesregierung zur Umsetzung der Open-Data-Charta der G8*, 2014.

[32] Federal Ministry of Justice and Consumer Protection: *Online access act.*

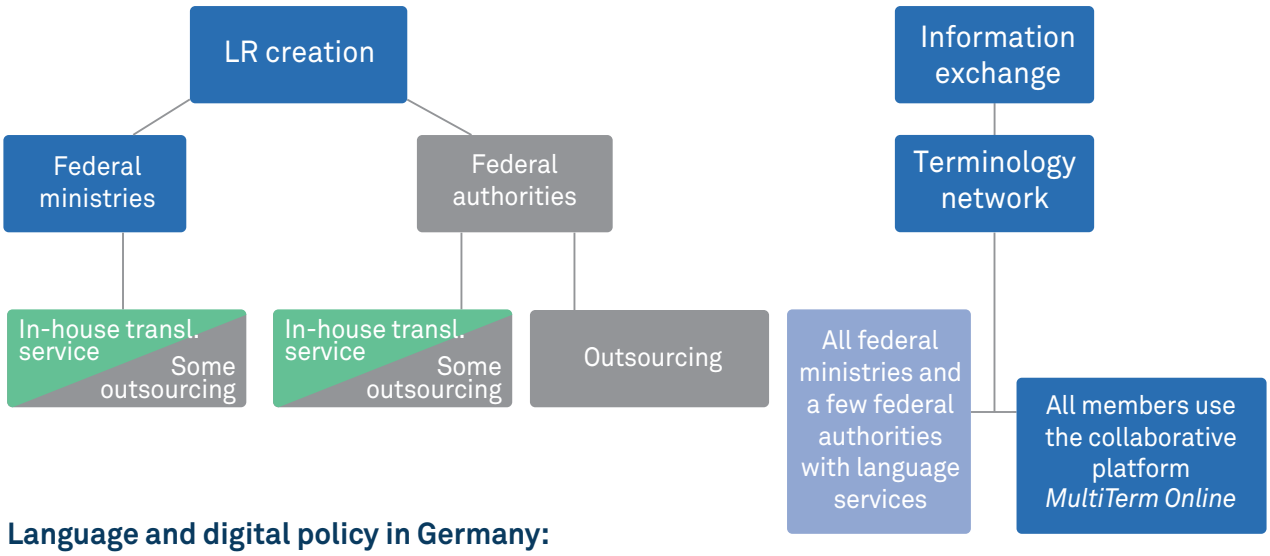
[33] Federal Ministry of Justice and Consumer Protection: *Informationsweiterverwendungsgesetz.*

[34] Federal Ministry of Justice and Consumer Protection: *E-Government Act* (NB: Not mentioned in the English translation.)

Annex
Country Profile Germany

interest in language data. It can be said that sharing data (textual or other) is still not an inherent part of the everyday work in public administrations in Germany and is a field with much unused potential: Germany is ranked 24th on the Global Open Data Index, in 10th place on the Open Data Barometer and comes in 17th place on the Open Data Maturity Dashboard on the European Data Portal (EDP)[35] underlining the fact that although data is an important resource in the 21st century, Germany only barely uses its potential.[36]

The current language data creation infrastructure in German public bodies looks as follows:



Language and digital policy in Germany:

The fact that language data or textual data is not considered a valuable asset may relate to the German language policy or better the lack thereof. Due to the federal organization of the country, there is not a single ministry or public authority that is in charge of digital policy or language policy in Germany. In addition, there is no Language Council per se but a supranational Council for German Orthography (Rat für deutsche Rechtschreibung) representing several countries with German as (one) official language observing the developments of the German language and proposing corresponding adjustments that are then implemented into national legislation.[37]

German is the sole official language in Germany and has a strong tendency towards being the single language for public usage. This is exemplified by the fact that the role of the German language for a number of historic and other reasons is not even mentioned in the German constitution.[38] There are however four autochthonous minority languages and one regional language, namely Danish, Frisian, Sorbian, Romani and Low German, that are recognized as minority languages.[39]

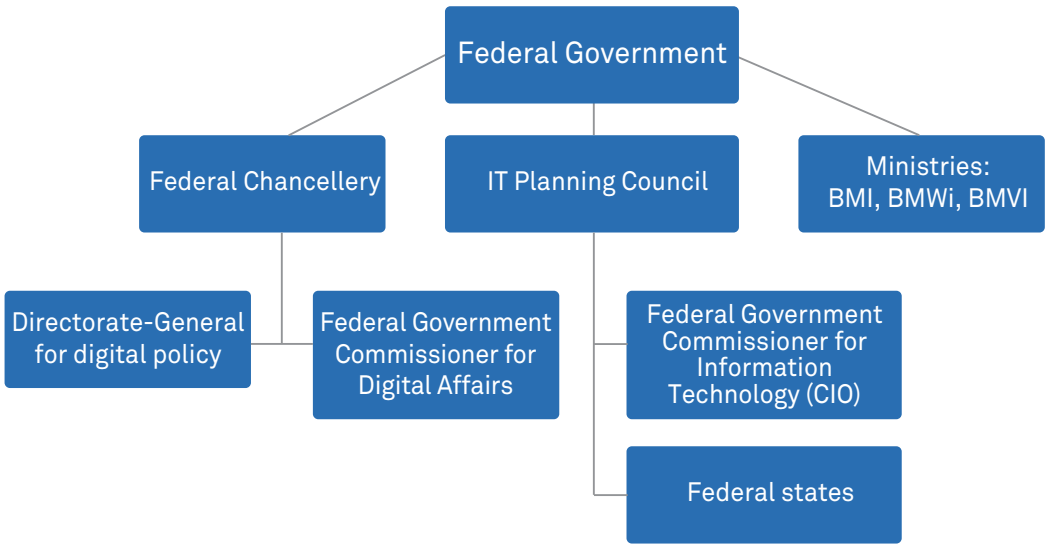
In a survey conducted by the Leibniz Institute for the German Language, 90.2% of the respondents indicated German as their first language (L1), but generally there is very little information about the use of language(s)

35 Information as of 25 July 2019.
36 Cf. Konrad Adenauer Stiftung: *Open Data. The Benefits, Das volkswirtschaftliche Potential für Deutschland*,p. 9, 2016.
37 Cf. Adler, Astrid; Beyer, Rahel: "Languages and language policies in Germany / Sprachen und Sprachenpolitik in Deutschland", in: *National language institutions and national languages. Contributions to the EFNIL Conference 2017 in Mannheim*, 2018, p. 227.
38 Cf. Adler, Astrid; Beyer, Rahel: "Languages and language policies in Germany / Sprachen und Sprachenpolitik in Deutschland", in: *National language institutions and national languages. Contributions to the EFNIL Conference 2017 in Mannheim,* 2018, p. 222 ff.
39 Cf. Adler, Astrid; Beyer, Rahel: "Languages and language policies in Germany / Sprachen und Sprachenpolitik in Deutschland", in: *National language institutions and national languages. Contributions to the EFNIL Conference 2017 in Mannheim*, 2018, p. 222 ff.

in Germany. [40] A decreasing use of regiolects or dialects can be observed and immigration history shows that allochthonous languages are spoken but no figures are available to indicate their use in the population. [41]

Given Germany's federal structure, its digital policy landscape is complex and multi-layered. The simplified diagram below aims at showing the different stakeholders involved in digital policy in Germany.

The Online Access Act of 14 August 2017 states that public services have to be made available electronically within 5 years on the federal and state level. [42] All public services (including services of third parties)[43] are in the process of being described in a dedicated catalogue ("Leistungskatalog") many of which are to be described multilingually.



Main challenges for sustainable data sharing:

- Translators do not see benefits for themselves although their data could be useful for various other administrations
- It is noted that freely available online machine translation services are being used fairly frequently although the exploitation of the results is not known
- Texts are by default copyrighted by the author/translator
- Language data is not considered valuable
- No central coordination of translations or language policy in general

Stakeholders:

In the context of sharing language data, the Federal Government Commissioner for IT as well as the heads of the translation departments across the ministries are the key decision makers. The Federal Office of Administration, an executive agency of the Federal Ministry of the Interior, Building and Community, has a consultancy function for Open Data and is therefore also an important stakeholder. Overall, more than 30 institutions representing federal and state ministries, language service providers and research institutions attended past ELRC events. Some federal ministries actively contributed some of their language resources.

[40] Cf. Adler, Astrid; Beyer, Rahel: "Languages and language policies in Germany / Sprachen und Sprachenpolitik in Deutschland", in: *National language institutions and national languages. Contributions to the EFNIL Conference 2017 in Mannheim*, 2018, p. 221.
[41] Cf. Adler, Astrid; Beyer, Rahel: "Languages and language policies in Germany / Sprachen und Sprachenpolitik in Deutschland", in: *National language institutions and national languages. Contributions to the EFNIL Conference 2017 in Mannheim*, 2018, p. 222.
[42] Federal Ministry of Justice and Consumer Protection: *Online access act.*
[43] Cf. IT-Planungsrat: *Handbuch – LeiKa-plus*, p. 6.

Annex
Country Profile Germany

Among them are the Federal Foreign Office, the Federal Ministry of the Interior, Building and Community, the Federal Ministry of Transport and Digital Infrastructure and the Federal Financial Supervisiory Authority.

Action Plan:

To address the identified challenges, the following objectives and actions are proposed:

- **To tackle legal concerns:**
  - Develop and share easy to apply guidelines for intellectual property rights (IPR) and privacy issues that can be followed by data creators and holders in order to decide whether their data can be shared
  - Investigate the idea to implement rights management along with data management, i.e. legal in-house support for data sharing in general

- **To identify and gain access to outsourced translations:**
  - Establish the practice of receiving any by-product of outsourced translations whenever translations are outsourced, irrespective of whether the contracting authority has an in-house translation service

- **To increase the interest in machine translation (MT) and language technology (LT) in public services as part of the national digital policy:**
  - Establish synergies with national projects and initiatives such as the evaluation of the need for an MT system for the federal administration
  - Secure the support of decision makers to change/adapt the national policy is an on-going activity. The topic of MT and language data is frequently brought up in internal meetings with the language services divisions of the federal ministries and will be continued
  - Anonymization is a central topic, important to many language data holders, however, internal expertise is lacking in this area, therefore support is needed

- **To establish good data management practices in public services:**
  - The identification of data managers is ongoing and is considered important to introduce changes such as:
    - The practice of clear separation between texts that contain confidential and personal data from texts that fall under the public sector information directive (or the “Informationsweiterverwendungsgesetz”)
  - Establish practice that public administrations have the right to share and publish translations although the copyright belongs to the author/translator and is as such not transferable

- **To raise awareness of language data as Open Data and a valuable asset:**
  - To integrate language data in the national Open Data policy and digital agenda is an ongoing process. The language services division of the interior ministry offered to participate in an Open Data pilot to share LR with Germany's Open Data portal Govdata. The federal ministries' Open Data officers have been made aware of the value of LR as Open Data.
  - Establish practical guidelines for LR as Open Data
  - Continue to share benefits of sharing language data is considered crucial to achieve the above-mentioned objectives

References and further reading list:

Adler, Astrid; Beyer, Rahel: “Languages and language policies in Germany / Sprachen und Sprachenpolitik in Deutschland”, in: *National language institutions and national languages. Contributions to the EFNIL Conference 2017 in Mannheim*, 2018, urn:nbn:de:bsz:mh39-78536.

Data Ethics Commission: *Gutachten der Datenethikkommission/Opinion of the Data Ethics Commission*, https://datenethikkommission.de/wp-content/uploads/191015_DEK_Gutachten_screen.pdf; https://datenethikkommission.de/en/.

Federal Office of Administration: *Handbuch für offene Verwaltungsdaten*, https://www.verwaltung-innovativ.de/SharedDocs/Publikationen/eGovernment/open_data_handbuch.pdf?__blob=publicationFile&v=2.

Federal Ministry of the Interior: *Nationaler Aktionsplan der Bundesregierung zur Umsetzung der Open-Data-Charta der G8*, 2014, https://www.bmi.bund.de/SharedDocs/downloads/DE/veroeffentlichungen/2014/aktionsplan-open-data.pdf?__blob=publicationFile&v=1.

Federal Ministry of Justice and Consumer Protection: *Informationsweiterverwendungsgesetz*, https://www.gesetze-im-internet.de/iwg/BJNR291300006.html.

Federal Ministry of Justice and Consumer Protection: *Online access act*, http://www.gesetze-im-internet.de/englisch_egovg/index.html.

Federal Ministry of Justice and Consumer Protection: *Informationsweiterverwendungsgesetz,* https://www.gesetze-im-internet.de/iwg/BJNR291300006.html.

Federal Ministry of Justice and Consumer Protection*: E-Government Act*, article 12 a, http://www.gesetze-im-internet.de/egovg/index.html.

Fraunhofer Fokus: *Leitfaden für qualitative hochwertige Daten und Metadaten*, 2019, https://cdn0.scrvt.com/fokus/e472f1bf447f370f/32c99a36d8b3/NQDM_Leitfaden-f-r-qualitativ-hochwertige-Daten-und-Metadaten_2019.pdf.

Germany in the Global Data Index: https://index.okfn.org/place/de/.

IT-Planungsrat: *Handbuch – LeiKa-plus*: https://www.fimportal.de/download-dokumente.

IT-Planungsrat: *National E-Government Strategy Update*, 2015, https://www.it-planungsrat.de/SharedDocs/Downloads/EN/Entscheidungen/18Sitzung_27_NEGS-Fortschreibung_2015.pdf?__blob=publicationFile&v=2.

Konrad Adenauer Stiftung: *Open Data. The Benefits, Das volkswirtschaftliche Potential für Deutschland*, 2016, https://www.kas.de/c/document_library/get_file?uuid=3fbb9ec5-096c-076e-1cc4-473cd84784df&groupId=252038.

Kompetenzzentrum Öffentliche IT: *Deutschland-Index der Digitalisierung 2019*: https://www.oeffentliche-it.de/publikationen?doc=95167&title=Deutschland-Index+der+Digitalisierung+2019

Open Data Barometer: https://opendatabarometer.org/?_year=2017&indicator=ODB.

Open Data Maturity Dashboard: https://www.europeandataportal.eu/en/dashboard#2018.

Annex
Country Profile Greece

Maria Gavriilidou, Maria Giagkou, Stelios Piperidis

State of Play:
Translation practices in Ministries and Public Administration in Greece:

Greek is mostly used in the country by its inhabitants, while outside the country it is used as heritage language by Greek expatriates; therefore, the need for translation of documents from and into Greek is great. There is a continuous need for translation of all EU-documents, mainly from English (and to a lesser extent from French) to Greek, but an increasing need for translation between Greek and the immigrant languages spoken in the country is also attested.

Very few of the Public Administration institutions meet their translation needs in-house, with dedicated translation departments. These are mainly ministries that have increased translation needs, either due to frequent exchange of documents with EU services or foreign countries or their nature and objectives entail heavy communication with citizens (Greek nationals or immigrants). Some of the bodies that have in-house translation departments are the Ministry of Justice, the Ministry of Foreign Affairs, the Hellenic Statistical Authority, The Bank of Greece, and the Hellenic Army General Staff. Broadly used practices include (a) outsourcing of translations to LSPs and (b) in-house translation for internal, unofficial needs by non-authorized personnel, usually within communication/media and/or publications departments.

However, official certified translations, such as documents that need to be submitted to public authorities, need to be assigned to the Translation Service of the Ministry of Foreign Affairs, whose task is to validly translate public and private documents. Official translation needs can occur either between Public Administration bodies at the national or international level (such as documents from/to foreign governments, embassies, etc., documents exchanged between national services that include text in a foreign language e.g. texts from Europol/Interpol, etc.) or documents exchanged between national or foreign citizens and Greek Public Administration (translation of identity papers/passports, University degrees, etc.). The Translation Service outsources translation tasks to LSPs/translation companies, selected from the Official Registry of Translators; the Registry is the outcome of specialized examinations organized by the Ministry of Foreign Affairs. Lawyers, registered in one of the Greek Lawyers Associations after having succeeded in the relevant exams, as well as graduates of the Department of Foreign Languages, Translation and Interpreting of the Ionion University (the only public University specializing on translation)[44], have also official translation rights and their translations have full validity, even in courts of Justice.

There is no centralized and uniform procedure for the procurement of translation tasks, apart from translations outsourced by the Translation Service of the Ministry of Foreign Affairs, nor official quality evaluation process. A steadily increasing number of LSPs use CAT tools for their translation services they offer. The Public Administration bodies, however, do not request the Translation Memories (or other by-products, such as term lists) to be submitted together with the translated documents; it is quite probable that they are not even aware of this possibility. As a consequence, the degree of usage of CAT tools by the translation companies cannot be documented. What is more important, however, is the fact that Public Administration does not benefit from CAT technology and that similar or exactly identical documents need to be translated anew. The single identified exception is the Bank of Greece, a partially state owned S.A., which is quite advanced in terms of in-house CAT and Translation Memories production.

Data sharing infrastructures and Open Data in Greece:

Access to Public Information: the Transparency Portal

Since October 2010, with the Transparency Program initiative, all government institutions are obliged to upload their acts and decisions on the Transparency portal (https://diavgeia.gov.gr/en) with special attention to issues of national security and sensitive personal data. Following the latest legislative initiative (Access to Information Law 4210/2013) of the (then) Ministry of Administrative Reform and e-Governance, adminis-

[44] It should be noted that several public Universities and private specialized schools offer courses on translation studies; however, these graduates do not have official translation rights.

trative acts and decisions are not valid unless published online; publication in the Transparency Portal overrides the validity of the Official Gazette itself. All the acts and decisions published on the Transparency portal are exclusively monolingual, i.e. in Greek only.

Open Access in Public Administration

While the Transparency portal hosts Public Administration's acts and decisions, the dedicated Open Government Data Portal hosts the central catalog of Open Government Data and offers open access to digital resources of the Greek government institutions to citizens, services and information systems for reuse for any purpose. It implements the Open Data policy adopted following the transposition of the EU Directive 2013/37/EE (Law 4305/2014).

Currently (Sept. 2019) approximately 10.000 datasets are hosted by the Greek Open Data portal. Analytics performed on the traffic of the portal, which shows a significant increase of visits and downloads, indicate its impact to government officials, but also to journalists, researchers and students, private companies from Greece and abroad, but also the interest of the broad public. The users' preferences focus on public administration documents, whose importance is related to transparency issues and government accountability. The second preferred category is economic/business data, followed by geospatial and environmental data and finally statistical data. It is worth noting that the vast majority (79%) of the datasets are numerical data (formats xls/xlsx/csv), whereas textual data are mainly in pdf format or images (15%), and only very few in doc or txt (6%). Worth mentioning is the fact that multilinguality is not catered for by the portal's metadata schema; at the time of writing the country profile it was not possible to investigate if any of the data hosted are available in languages other than Greek.

Training public servants on the value of Open Data and, most importantly, raising awareness on the value of language data has not been an undemanding procedure (as indicated by the numbers above). The National School of Public Administration and Local Government (ESDDA) has the mission to create a body of specialized officials of the Public Administration with comprehensive professional training; Digital Policy and Digital Governance feature among the subjects taught, while e-ESDDA, the digital repository of the School is responsible for preservation and dissemination of the School's digital material.

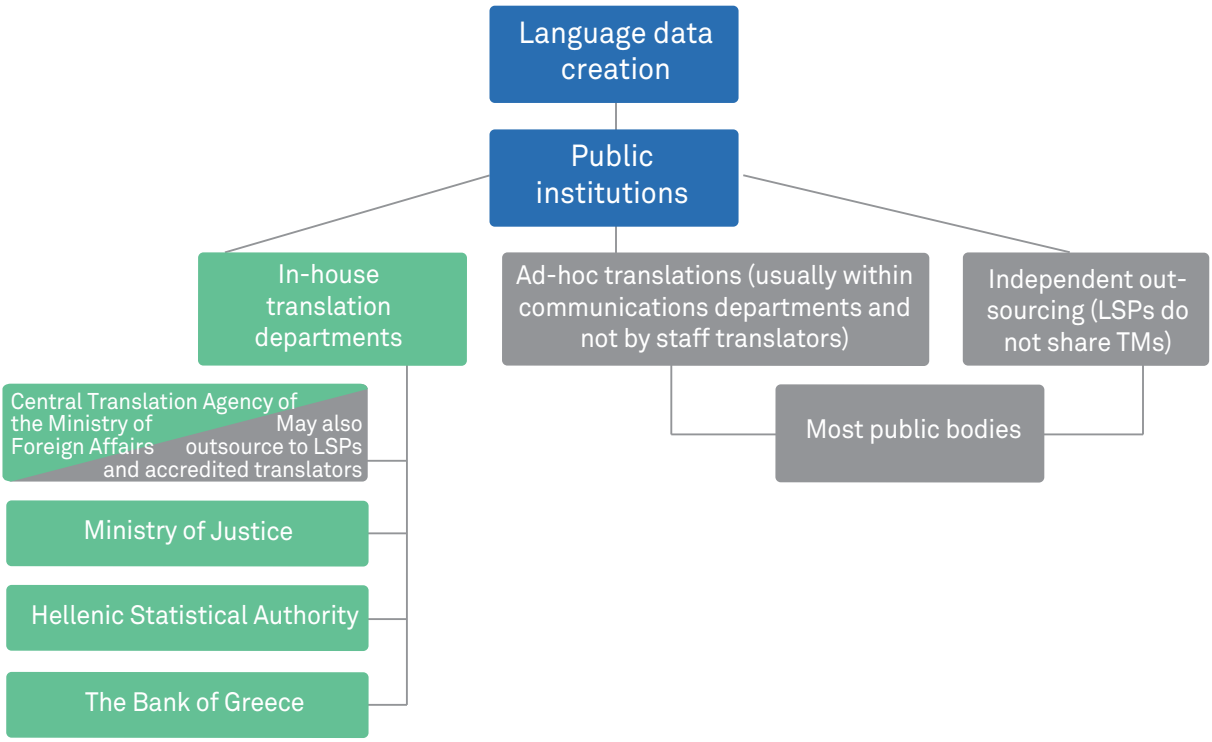
The situation regarding data sharing varies among the Public Administration bodies. Several Ministries have been moving forward as regards the digitization of their services and workflows and are keen on making their data openly available; indicatively, but not exclusively, the Ministry of Environment and Energy with its dedicated Open Data portals on various subjects of its responsibility, the Ministry of Justice with its plan for eJustice in place and its various electronic services for the citizens, the Ministry of Finance with its eServices and the Hellenic Statistical Authority, an independent authority which digests data to produce statistics useful for public policy, the economy, and more broadly the life of the people.

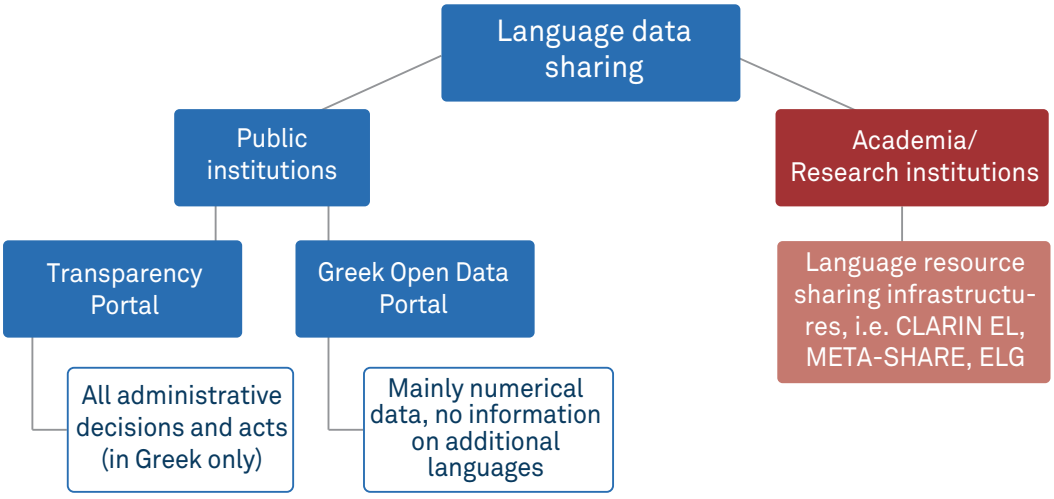
Dedicated language data sharing infrastructures

There are several language resources repositories and research infrastructures in Greece, stemming from R&D activities and initiatives related to Language Resources and Technology, either national or European. The CLARIN:EL infrastructure with its portal and its inventory of language resources, hosted by ILSP/Athena RC which coordinates a distributed network of 10 nodes, caters for language resources and technologies sharing as well as for training and raising awareness on the significance of language technology. Language resources and technologies for Greek (but also other languages) are also being shared by META-SHARE, the open and secure network of repositories for sharing and exchanging language data, tools and related web services and also by the recently created European Language Grid Platform, which aims at listing data sets and language technology services as well as relevant stakeholders, from technology development to research centres, from small and medium-sized companies to large enterprises.

Annex
Country Profile Greece

The current language data sharing infrastructure in Greek public bodies looks as follows:





Language policy in Greece:

Greece is a small country (approximately 10 million inhabitants), whose official language is Greek. The only minority language which is officially recognized is Turkish, spoken by the Muslim minority in Western Thrace. Due to the immigration flow attested in the recent years, there are also numerous immigrant languages spoken in Greece: mainly Albanian, Bulgarian, Russian, but also Chinese, Pashto (Afghanistan), Urdu (Pakistan), Kurdish, Arabic etc.

The language policy of Greece is designed by the Ministry of Education and Religious Affairs, the Ministry of Foreign Affairs (through its General Secretariat for Intercultural Education and Greek Studies Abroad) and the Ministry of Culture and is implemented by the Centre for the Greek Language and the Institute of Educational Policy. It defines the educational goals, the principles and the structure of teaching of Greek as first (L1) and second language (L2) in the country and abroad, but does not set priority axes as regards language and language technology research.

Key Stakeholders:

The key decision makers in Greece for the topics adherent to this white paper are the following:

- For the national digital agenda: Ministry of Digital Government and Ministry of Development and Funds, General Secretariat for Research and Technology
- For Open Data and open government: Ministry of the Interior (former Ministry of Administrative Reform)
- For language policy: Ministry of Education and Religious Affairs, and its affiliated bodies, the Centre for the Greek Language and the Institute of Educational Policy.

64 unique organisations have attended the second ELRC workshop,of which more than 50% represented public sector bodies. A considerable number of public bodies have already shared their language data either under permissive or restrictive licenses with ELRC. Indicatively but not exhaustively, these include:

- Central Bank of Greece
- Ministry of Justice
- Ministry of Environment
- Ministry of Finance

Identified challenges and issues:

- **Language Resources and Technology:**
  its significance is not visibly recognized and reflected in the national language policy. Its importance is acknowledged and manifested through the inclusion of CLARIN:EL, the Language Resources infrastructure, in the national Research Infrastructures Roadmap. However, the limitations this entails (i.e. not secured funding for the future) should not be underestimated.
- **Scarce data:**
  Greece being a small country, the production of digital (language) data is limited, as compared to larger countries with broadly used languages at national and international levels.
- **Lack of centralization:**
  there is no central agency/service for PA translation workflow management, lack of central procurement mechanism, lack of language data management flows and policies; this results in data getting lost, no possibility of re-usability of data, no economies of scale.
- **Lack of experience in CAT/Lack of TMs:**
  repetition of work, no training of systems possible, lagging behind in digital literacy.
- **No translations' sharing culture by LSPs and Translators.** No willingness to change the current workflow, fear it will put them out of business.
- **Issues related to GDPR** render public authorities uneasy to share data or offer a good excuse not to share.

Annex
Country Profile Greece

Action plan:
In order to address the identified challenges, the following actions are proposed:

- Raise awareness on language data as valuable asset with Policy makers, Academia and Public Administration
- Establish good data management policies in public services
- Raise awareness on textual data as valuable Open Data
- Increase interest in MT/LT in public services as part of the national digital policy
- Organised training of Public Administration on CAT
- Need for change of workflows and procedures regarding translation within Ministries
- Tackle legal issues and provide training of Public Administration
- Identify and gain access to outsourced translations

References and further reading list:

Bank of Greece: https://www.bankofgreece.gr/en/homepage.

Centre for the Greek Language: https://www.greeklanguage.gr/.

CLARIN:EL: https://www.clarin.gr/.

Digital Repository e-ESDDA: https://www.ekdd.gr/en/the-school/digital-repository-e-esdda/.

eServices offered by Ministry of Finance: https://www.minfin.gr/web/guest/yperesies

European Language Grid Platform: https://www.european-language-grid.eu/grid/.

Gavriilidou, Giagkou, Pouli, Piperidis: *ELRC workshop report for Greece 2017*, http://www.lr-coordination.eu/sites/default/files/Greece/ELRC2-Workshop-Report_Greece%202017-Public_FINAL.PDF.

General Secretariat for Research and Technology, Ministry of Development and Funds: http://www.gsrt.gr.

Greek Government Open Data portal: http://data.gov.gr/.

Greek Government Open Data Portal datasets: http://data.gov.gr/dataset.

Greek Open Data Policy Law 4305/2014: http://www.et.gr/idocs-nph/search/pdfViewerForm.html?args=5C7QrtC22wEc63YDhn5AeXdtvSoClrL8oeK-AuTKOuiV5MXD0LzQTLWPU9yLzB8V68knBzLCmTXKaO6fpVZ6Lx3UnKl3nP8NxdnJ5r9cmWyJWelDvWS_18kAEhATUkJb0x1LIdQ163nV9K--td6SIubMfH2r_a2DXjO6MJnF-5f9_LW7pRMszX0fGIINVmMIh.

Hellenic Statistical Authority: https://www.statistics.gr/en/mission.

Institute of Educational Policy: http://iep.edu.gr/en.

Ionion University, Department of Foreign Languages, Translation and Interpreting: http://dflti.ionio.gr/en/about.

Legal framework for open access and reuse of public domain documents, information and data: http://data.gov.gr/pages/thesmikoplaisio.

Ministry of Digital Government: https://mindigital.gr/.

Ministry of Environment and Energy:
http://www.ypeka.gr/Default.aspx?tabid=37&locale=el-GR&language=en-US.

Ministry of Justice:
http://www.ministryofjustice.gr/site/el/%CE%91%CE%A1%CE%A7%CE%99%CE%9A%CE%97.asp

Ministry of the Interior: http://www.ypes.gr.

National Digital Strategy 2016-2021:
http://www.opengov.gr/digitalandbrief/wp-content/uploads/downloads/2016/11/digital_strategy.pdf.

National School of Public Administration and Local Government (ESDDA):
https://www.ekdd.gr/en/the-school/esdda-profile/.

Official Translation Service of the Ministry of Foreign Affairs (Mission):
https://www.mfa.gr/en/citizen-services/translation-service/translation-service.html.

Open Data Portals by the Ministry of Environment and Energy:
http://www.ypeka.gr/Default.aspx?tabid=823&locale=el-GR&language=en-US.

Transparency Portal: https://diavgeia.gov.gr/en (access to Public Administration acts and decisions).

Annex
Country Profile Hungary

Zoltán Bódi, Tamás Váradi, Andrea Lösch

State of Play:
Translation practices in ministries and public administrations in Hungary:

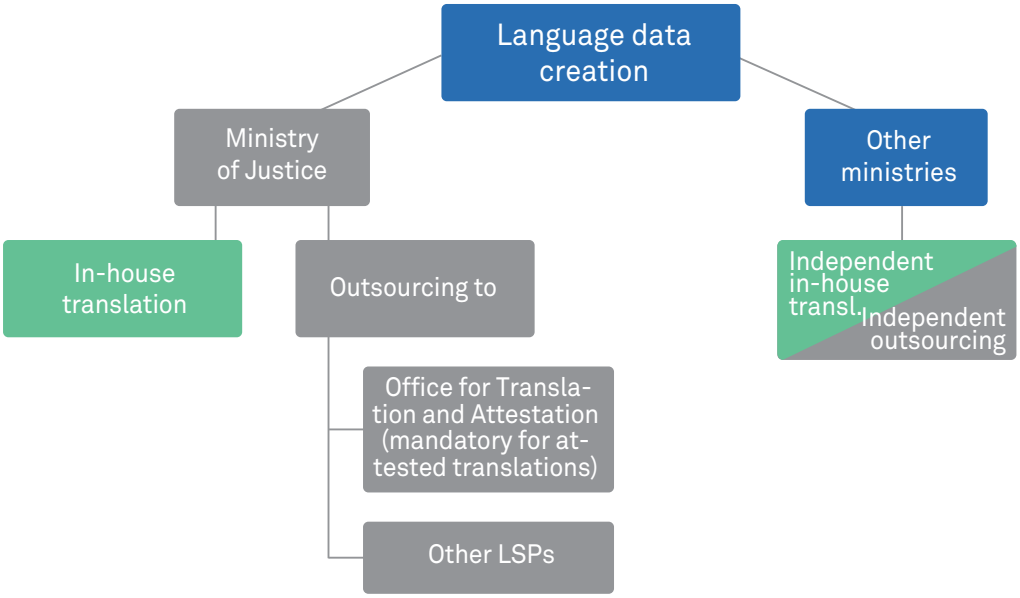
Pursuant to Decree No. 24/1986 (26 June) of the Council of Ministers on Translation and Interpretation and Decree No. 7/1986 (26 June) of the Minister of Justice on the Implementation thereof, attested translation, attestation of translations and making attested copies of foreign language source documents are the exclusive competence of the Hungarian Office for Translation and Attestation Ltd. (OFFI). Apart from meeting its exclusive line of duty and making attested translations, OFFI also provides technical translations, including legal, public administrative translations, translations of laws and revision services with special expertise and terminology expert support. However, it is not obligatory to make use of the OFFI for translations, since other language service providers (LSP) can be contracted as well.

Interesting fact:
The Hungarian Office for Translation and Attestation is a unique institution in Europe in the field of certified (or attested) translations with a history of over 150 years. It has also a unique situation, because since 1994, it is a 100% state-owned shareholding company.

Additional translation activities are also carried out by the ministries and public administration bodies themselves (ad-hoc translations for the ministries' everyday tasks). However, there is no common practice for centralizing these ad-hoc translation activities of the Hungarian public administrations. The coordination of translation activity is usually assigned to one department within the particular ministry or public administration. This includes both in-house translations as well as outsourced translations. In some cases, the translation activity is not assigned to one department, but to a secretary of state or a head of department responsible for international communication. The management of the translated documents then follows the specific practices of public administration bodies.

By Act XL of 1995 on Public Procurement, the Public Procurement Authority of Hungary (PPA) was established as a central budgetary organ. The PPA is subordinated to the Parliament. The existing rules for public procurement are based on the Act CXLIII of 2015 on Public Procurement. The threshold of the national public administration procurement contracts is 15.000.000 HUF (about 45.000 EUR), but if the public procurement value exceeds 1.000.000 HUF (about 3.000 EUR), at least three different offers must be requested.

The current language data creation infrastructure in Hungary public bodies looks as follows:



Open Data and eGovernment strategy in Hungary:

Following the Digital Economy and Society Index from 2017, Hungary has exhibited several significant improvements and successes such as eID cards, the Electronic Health Cooperation Service Space and the Hungarian Municipality Application Service Provider (ASP). The latter provides modern, integrated shared services for local administrative management, ensuring standardised internal operation and a common platform for e-government service provision on the local government level to the end-users.

Open government data in Hungary is shared and collected through the National Open Data Portal http://opendata.hu/, the first free and public Hungarian data directory. The site was created by volunteers and non-governmental organizations with the aim of creating the first Hungarian Open Data collection database. The site is free to upload and organize public or public interest data which is not subject to copyright. However, there are currently no translations and/or language data available through the portal.

As regards the legal aspect of sharing data, the Public Sector Information (PSI) Directive was implemented in the corresponding 2012 LXIII. Law about the Reuse of Public Data. Overall, however, stronger coordination between the Ministry of Interior and the Ministry for Innovation and Technology is intended with regard to Open Data in Hungary.

Language policy in Hungary:

The Prime Minister's Office State Secretariat for the National Policy and its organizations are coordinating the policy and strategy for the Hungarian nation in the Carpathian basin and all over the world. A major contributor of the Hungarian language strategy is the former Institute for Hungarian Language Strategy, the current Institute for Hungarian Studies Research Center for Language Planning.

The main governmental strategies, which contain language policy elements are:

- The National Info-Communication Strategy 2014-2020 (available only in Hungarian).
- Strategies of the Digital Success Programme 2.0 (including the Digital Child Protection Strategy of Hungary, the Digital Export Development Strategy of Hungary, the Digital Education Strategy of Hungary, and the Digital Startup Strategy of Hungary, Digitalization Strategy of Public Collections)

The main legal acts, which contain elements of the national language policy include:

- Fundamental Law (the status of the Hungarian language)
- Act XLVI of 2012 on the Land Surveying and Cartography (geographical names)
- Act LXII of 2001 on the Hungarian Nation Living in the Neighbouring Countries
- Act CXXV of 2009 on the Hungarian Sign Language
- Developing Strategy for the Public Administration and Public Services 2014-2020
- Act CLXXXV of 2010 on the Mass-Media and Mass-Communication
- Act CCI of 2017 on the Rights of Nationalities
- Act XCVI of 2001 on Publication in Hungarian Language of Economic Advertisements, Shop Labels and some of Public Statements
- Act LXIV of 2001 on Protecting Cultural Heritage

There are also important resolutions of the Hungarian National Assembly mentioning language policy elements, including the resolutions made on the Day of the Hungarian Language, on the National Heritage Day and on the Day of the Nationalities.

Most interestingly, as a result of a year of research and development work of a team of hundreds of pedagogical experts, psychologists and practitioners, the Education 2030 Learning Sciences Research Group at Eszterházy Károly University made a proposal for the new National Core Curriculum. The National Core Curriculum contains the full language policy for the education system of Hungary.

Annex
Country Profile Hungary

Stakeholders:
Within ELRC, around 80 potential stakeholders that are involved in the creation or sharing of language resources, related activities and/or policy setting were identified, including nine ministries as well as the Open Data Portal. 30 of these stakeholders participated in the ELRC Workshop in 2019.

In addition to the Hungarian Office for Translation and Attestation (OFFI), the different ministries and additional certified translators represent the major provider of language resources in Hungary. So far, 2 language resources have been contributed to the ELRC-SHARE from Hungary.

Because of its central role in the provision of translation services to public administrations and ministries, the OFFI among the ministries, public administration institutes, the OFFI can be considered as the main beneficiaries of eTranslation.

Main challenges for sustainable data sharing:
- Fundamental internal issues: Public administrations are not aware of the value of language data and do not perceive it as an asset.
- Legal issues related to outsourced translations: Outsourced translations are intellectual property of LSPs, which makes it difficult for public administrations to share outsourced data.
- Continuity issues: Changing government, framework contracts etc.

Action Plan:
Taking into account the main challenges in Hungary, corresponding actions to enable/improve the sharing of language resources in Hungary should focus on:

- **Discovering the document management process of the Hungarian public administration**
- **Discovering the translation practice (tools, competent persons and departments) used and needed during the everyday work in the leading public administrations (ministries, governmental offices)**
- **Raising awareness of language data as Open Data and a valuable asset, including in particular:**
  - Sharing benefits of sharing language data
  - Integrating language data in the national Open Data policy as well as digital agenda
- **Identify and gain access to outsourced translations:**
  - Adapt the procurement process for buying translations in a way that tmx and all usage rights are transferred to the purchasing authority
- **Initiate the institutional organization and collection of multilingual language data, building a national multilingual public administration terminology database.**

References and further reading list:
Act XL of 1995 on Public Procurement: https://www.kozbeszerzes.hu/english/.

Act CXLIII of 2015 on Public Procurement: https://www.kozbeszerzes.hu/cikkek/hungarian-public-procurement-rules.

Annex
Country Profile Iceland

Gauti Kristmannsson, Eiríkur Rögnvaldsson, Eileen Schnur

State of Play:
Translation practices in ministries and public administrations in Iceland:

Since 1990, all Acts falling under the EEA agreement are translated by the Translation Centre of the Ministry of Foreign Affairs. In addition, the Translation Centre is responsible for the translation of texts related to the European Economic Area, other international agreements and legal acts. Approximately 35 translators are grouped according to fields such as society, finance, science or technology. In 2016, the Translation Centre's terminology contained about 70,000 entries and it is continuously growing.

Most other translation and interpreting services for the Icelandic state are procured through the state procurement agency Rikiskaup via a call for tender. Those who are accepted are taken into the so-called framework agreement until the next call for tender has been finished. Most of the signatories are small companies, which are frequently working with freelancers. This makes it difficult to gather data from them, except perhaps from the largest translation agencies. One of them, Skopos, is actually working on a CEF project called "Principle" and is offering bilingual data.

Some other institutions have in-house translation services, such as the RÚV, the Icelandic state broadcaster. Those translations are only for television and are rarely kept in a data base after their use. The University of Iceland has one in-house translator to translate legal regulations and web material into English. The corresponding bilingual data base will be used in the Principle project.

The use of computer-assisted translation (CAT) Tools depends on whether the translations are outsourced or managed in-house. Language Service Providers (LSPs) and freelance translators usually translate their documents with the help of CAT Tools. However, neither CAT Tools nor machine translation (MT) systems are used in Icelandic public administrations and ministries. This may also explain why translation memories (TMs) are usually not transferred back to the public administrations if the translation was outsourced.

Due to the limited number of Icelandic speakers, it is difficult to build and develop costly language technologies. Therefore, the language technology industry in Iceland is relatively small and language technology support for Icelandic is weak.[45] Various companies developed LT software and systems, such as a spell-checking program or a text-to-speech system for Icelandic, but neither of them continued their work in the field afterwards. However, in recent years, the Icelandic government has started various actions to improve their position in the digital world and to raise awareness on the importance of language technology.

Data sharing infrastructures and Open Data in Iceland:

Data privacy in Iceland is legislated by the Data Protection Act. Pursuant to the Icelandic State and Municipal Policy on the Information Society 2013-2016, non-personal information and files stored by the State or municipalities should be accessible to the general public, businesses and stakeholders. The Information Act (No.50/1996) includes conditions on the re-use of public sector information (PSI) and defines both access and restrictions to information. It covers almost all aspects related to the PSI Directive (2003/98/ES), except for the access and re-use of information through electronic means like e.g. databases.

The Icelandic Open Data Portal provides access to a growing list of government data and databases. It is available at https://opingogn.is/.

> **Interesting fact:**
> **In 2018, Iceland signed an agreement with the Nordic Institute for Interoperability to start using Straumurinn. It is based on the Estonian X-Road platform, which will enable standardised, efficient and secure data exchange between public administrations and ministries.**

In order to create synergies between different IT systems of public administrations, Iceland has signed an agreement with the Nordic Institute for Interoperability Solutions (NIIS Institute), which also cooperates with

[45] Rögnvaldsson et. al: *META-NET White Paper Series "The Icelandic Language in the Digital Age"*, 2012

Annex
Country Profile Iceland

Finland and Estonia. By using the Straumurinn data line, processes for data exchange will be streamlined and automated. Together with a comprehensive management plan, the Straumurinn data line is considered the foundation for effective and transparent public services in Iceland.

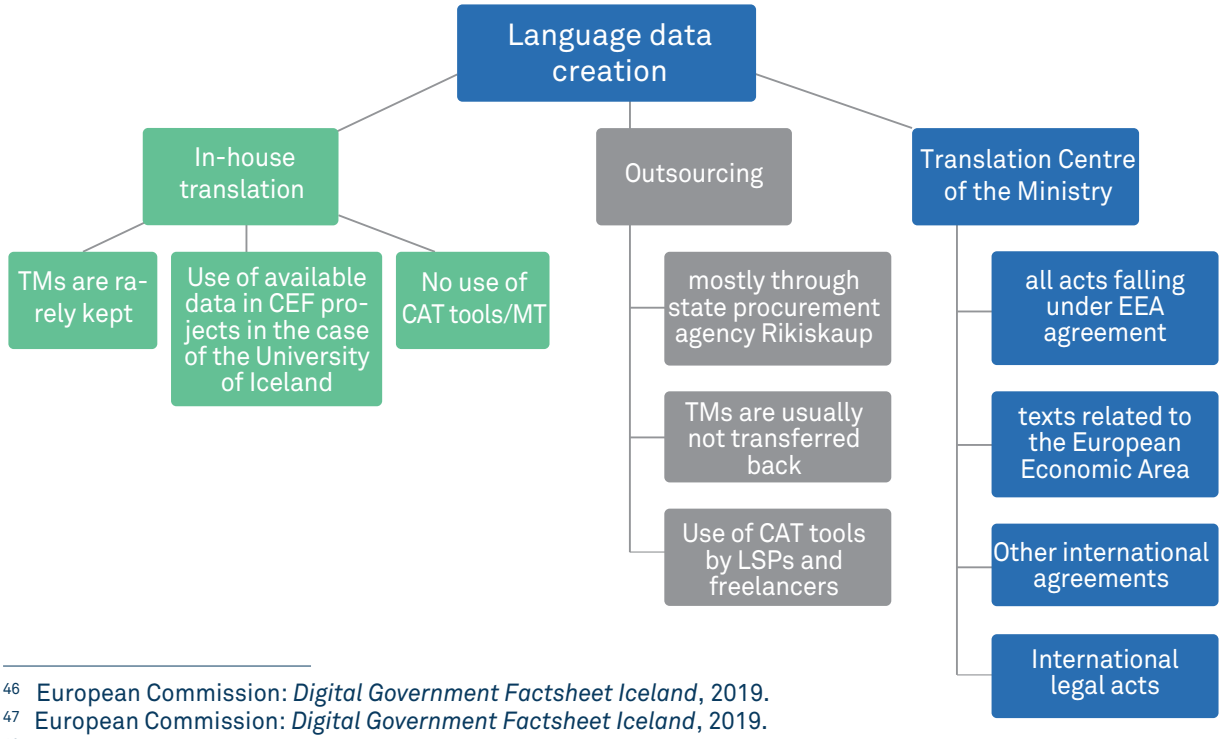
According to the Icelandic financial plan 2019-2023[46], public sector services should be based on information systems that fulfil the needs and the technical requirements of both public institutions and the industry. They shall be able to access open-source data in one place, monitor discussions on issues in the administration and participate in transparent reporting processes for e.g. draft proposals or policy papers. According to the financial plan 2019-2023, public data "will be free of charge and reusable as much as possible"[47].

The LT Project Plan[48] also explicitly states that all resources and tools, which will be developed as parts of the core tasks will be completely open and free. They will be made available through CLARIN-IS and maybe other CEF-funded projects. A number of language resources are already available through CLARIN-IS. The majority of them are free and open (under CC licenses).

The first phase of monolingual data gathering was collected under the wing of the Arni Magnusson Institute. The datasets include e.g. a large dictionary of the most frequent words in Iceland, the MÍM (2012) corpus including 25 million words, a new giant corpus of 1250 million words and some other smaller corpora. In the meantime, a second phase has begun, which also includes bilingual corpora.

However, more parallel language databases will be required to develop accurate machine translation systems [49], which is one of the reasons why Iceland participates in the above-mentioned Principle project together with partners from Ireland, Norway and Croatia. It is dedicated to the gathering of bilingual corpora for the purpose of creating MT engines. The Translation Centre of the Foreign Ministry will provide its TM database of 1.3 million sentence pairs, and other partners will contribute data as well.

The current language data creation infrastructure in Icelandic public bodies looks as follows:



[46] European Commission: *Digital Government Factsheet Iceland*, 2019.
[47] European Commission: *Digital Government Factsheet Iceland*, 2019.
[48] Nikulásdóttir et. al: *Language Technology for Icelandic 2018-2020, Project Plan*, 2018.
[49] Þorláksdóttir, Arnbjörnsdóttir: *ELRC Workshop Report for Iceland*, p.6, 2018.

Language policy and digital policy in Iceland:

Iceland is a small language with approximately 330,000 native speakers. Pursuant to a law passed by the Icelandic Parliament in 2011, Icelandic enjoys the status of official language. It is spoken in government and administration, in the field of education and in the Icelanders' day-to-day life. Due to the limited number of speakers, citizens are concerned about the extinction of their language, since they consider the Icelandic language as the centre of their culture and identity.[50]

Interesting fact: Due to the Icelanders' interest in the preservation of their language, linguistic quizzes or discussions are essential parts of the National Radios' programmes.

Already in the early 1960s, the Icelandic Language Council was built to protect the language. It has multiple advisory and reporting functions, provides the government with suggestions concerning language policy, reports on the status of the language and decides on the spelling rules.

In 2000, a special Language Technology (LT) programme was launched. It aimed to support institutions in creating basic resources for Icelandic language technology, which lead to considerable results, including e.g. a corpus of 25 million words, an isolated word speech recogniser, etc. After the end of the programme, the Icelandic Centre for Language Technology (ICLT) was established by researchers from the University of Iceland, Reykjavik University and the Árni Magnússon Institute for Icelandic Studies. Their work resulted in a number of projects, leading to the development of various Language Technology tools and resources, e.g. the open source IceNLP package.

The publication of the META NET White Papers in 2012 was a landmark for the Icelandic LT development. Iceland was one of the four countries with the lowest scores in all categories, which highlighted the lack of language technology support in the country. This is why after its publication, extensive propaganda for the development of language technology has started in Iceland. The white paper was even discussed in the Icelandic Parliament in the same year, which lead to the Resolution on the Necessity of Making Icelandic Usable in the Digital Domain in 2014.

The current drive in LT is much better funded by the Icelandic authorities with 2.2 billion ISK (15.7 million EUR) until 2022. Consequently, three language technology experts were commissioned to develop a detailed 5-year project plan for Icelandic and its technology in 2016, which resulted in the Language Technology Project Plan 2018-2022. The non-profit organisation Almannarómur has been contracted to be in charge of the execution and coordination of this LT plan. It defines five core tasks[51], i.e.

- Speech recognition
- Speech synthesis
- Machine Translation
- Language and style checking
- Language resources

In addition, the Icelandic government aims to establish relations with major international IT companies to convince them to include Icelandic in their products, e.g. Microsoft, which has added Icelandic to their translator. Iceland has also joined CLARIN ERIC as an observer and aims to become a full member of the infrastructure. This clearly demonstrates that Iceland is dedicated to improving its position in the digital world and that there are numerous efforts to develop, improve and extend Icelandic language technologies.

[50] Rögnvaldsson et. al: The Icelandic Language in the Digital Age, in: *META-NET White Paper Series,* 2012.
[51] Nikulásdóttir et. al: *Language Technology for Icelandic 2018-2020*, *Project Plan*, 2018.

Annex
Country Profile Iceland

Stakeholders:
The ELRC National Anchor Points for Iceland represent a relevant stakeholder, i.e. the University of Iceland, which is involved in the above-mentioned ICLT and other projects that are relevant to ELRC. Other stakeholders that have already contributed data to ELRC include the Central Bank of Iceland and the European Medicines Agency. In summary, more than 35 organisations showed their interest in the ELRC initiative by participating in local workshops and ELRC conferences.

Main challenges for sustainable data sharing:
- In Iceland, open issues regarding access, copyright and privacy often prevent data holders from sharing their data.
- At the same time, there is a general lack of available parallel language resources, making it hard to train and improve already existing machine translation systems. However, with the new CEF project Principle, greater emphasis will be put on acquiring high-quality bilingual corpora and preparing them for MT engines.
- The limited number of Icelandic speakers makes it difficult to create language resources, since this is also associated with high costs.

Action plan:
Based on the status quo in Iceland and the identified challenges, the following three main objectives were defined:

- **To raise awareness of language data as Open Data:**
  This is to be achieved by e.g. further integrating language data in the national Open Data policy and in the digital agenda. The National LT Project Plan is already an important step in this direction, since it highlights the importance of language data by stating that it is impossible to develop language technology if language resources do not exist.[52] At the same time, the project plan aims to ensure that the data created within the program are not only accessible, but also usable for further development in research or business.[53]

- **To increase interest in MT in public services:**
  In order to increase the public administrations' interest in machine translation, synergies will be established through dedicated projects and initiatives. The project Principle already serves as a good example of these efforts. In addition, best practices and good examples of successful use of machine translation will be promoted.

- **To identify and gain access to outsourced translations:**
  It can be a challenge to gather data from procured translations if the tender was awarded to small companies, which are often working with freelancers. However, larger Icelandic companies may be able to contribute data, which is why they should be involved in projects and initiatives as it was the case in e.g. the Principle project.

[52] Nikulástóttir et. al: *Language Technology for Icelandic 2018-2020, Project Plan*, p.13, 2018.
[53] Nikulástóttir et. al: *Language Technology for Icelandic 2018-2020, Project Plan*, p.142, 2018.

European Language Resource Coordination
Connecting Europe Facility

References and further reading list:

Rögnvaldsson et. al: *The Icelandic Language in the Digital Age*, in: *META-NET White Paper Series*, 2012, www.meta-net.eu/whitepapers/e-book/icelandic.pdf.

European Commission: *Digital Government Factsheet Iceland*, 2019, https://joinup.ec.europa.eu/sites/default/files/inline-files/Digital_Government_Factsheets_Iceland_2019.pdf.

Tagged Icelandic MÍM Corpus: http://www.malfong.is/index.php?lang=en&pg=mim.

Þorláksdóttir, Arnbjörnsdóttir: *ELRC Workshop Report for Iceland*, 2018, http://www.lr-coordination.eu/sites/default/files/Iceland/ELRC%20Workshop%20Report%20for%20Iceland_PU_0.pdf.

Rögnvaldsson, Eiríkur: *Language Technology News – Iceland, ELG Conference*, 2019, https://notendur.hi.is/eirikur/ELG_Brussel.pdf.

Nikulásdóttir et. al: *Language Technology for Icelandic 2018-2020, Project Plan,* 2018, http://clarin.is/en/links/LTProjectPlan/.

Rögnvaldsson et. al: *Icelandic Language Resources and Technology: Status and Prospects*, 2009, https://dspace.ut.ee/bitstream/handle/10062/9670/Icelandic%20language%20resources.pdf.

State procurement agency Rikiskaup: https://www.rikiskaup.is/is/english.
CLARIN-IS: http://clarin.is/en/resources/.

Language Resources for Icelandic: http://www.malfong.is/index.php?lang=en&pg= .
Debate about “The Icelandic Language in the Digital Age”in the Icelandic Parliament, 2012, https://www.althingi.is/altext/upptokur/lidur/?lidur=lid20121121T153618.

Resolution on the Necessity of Making Icelandic Usable in the Digital Domain, 2014, https://www.althingi.is/altext/143/s/1076.html.

Annex
Country Profile Ireland

Teresa Lynn, Micheál Ó Conaire, Jane Dunne

State of Play:
Translation practices in ministries and public administrations in Ireland:

The Irish language has been highlighted as an under-resourced and therefore a priority language in the context of data collection for improving the EU's automated translation systems. As a result, the Irish <> English language pair (in terms of data collection) has been the current focus of ELRC-related activities to date in Ireland and thus serves as the focus of this report.

Irish is the first official language of Ireland. It is a minority language, with the most recent census[54] reporting 1.7 million speakers, of whom just over 73,000 speak it on a daily basis outside of the education system. The Irish language is a unique minority language in many ways as it has been afforded significant constitutional and legislative protection by the Irish State since its foundation. In addition to the official status of the Irish language in the Constitution, it was recognised as an official and working language of the European Union in 2007.

The Official Languages Act 2003 was signed into law on 14th July 2003, with the primary objective to ensure the improved provision of public services through the Irish language. The Office of An Coimisinéir Teanga (Language Commissioner) was established under the Act in 2004 to monitor compliance by public bodies with the provisions of the Act and to take appropriate measures to ensure such compliance. Each public body defines their own language scheme, which describe the services it proposes to provide either in Irish only, in English only or bilingually. As a result of the Languages Act, all public bodies are under obligation to translate official documentation into Irish and therefore large quantities of documentation are available in both English and Irish. This means that when it comes to translation needs in Irish public administration, differentiation should be made between translations for Irish <> English and other language pairs. The Language Act was updated in 2012 and the Gaeltacht Act was introduced, giving statutory effect to a 20-year strategy for Irish. With respect to Irish <> English translations on a European level, the status of Irish as an official and working EU language came into effect from the 1st January 2007 under Regulation 920/2005 which included derogation on the use of the language, to be reviewed every five years. In December 2015, the Council of the European Union adopted a regulation aimed at eliminating the derogation on an incremental basis by the end of 2021 to eventually provide services through Irish at the same level as the other official EU languages from this date.

Current Infrastructure vs. Goal

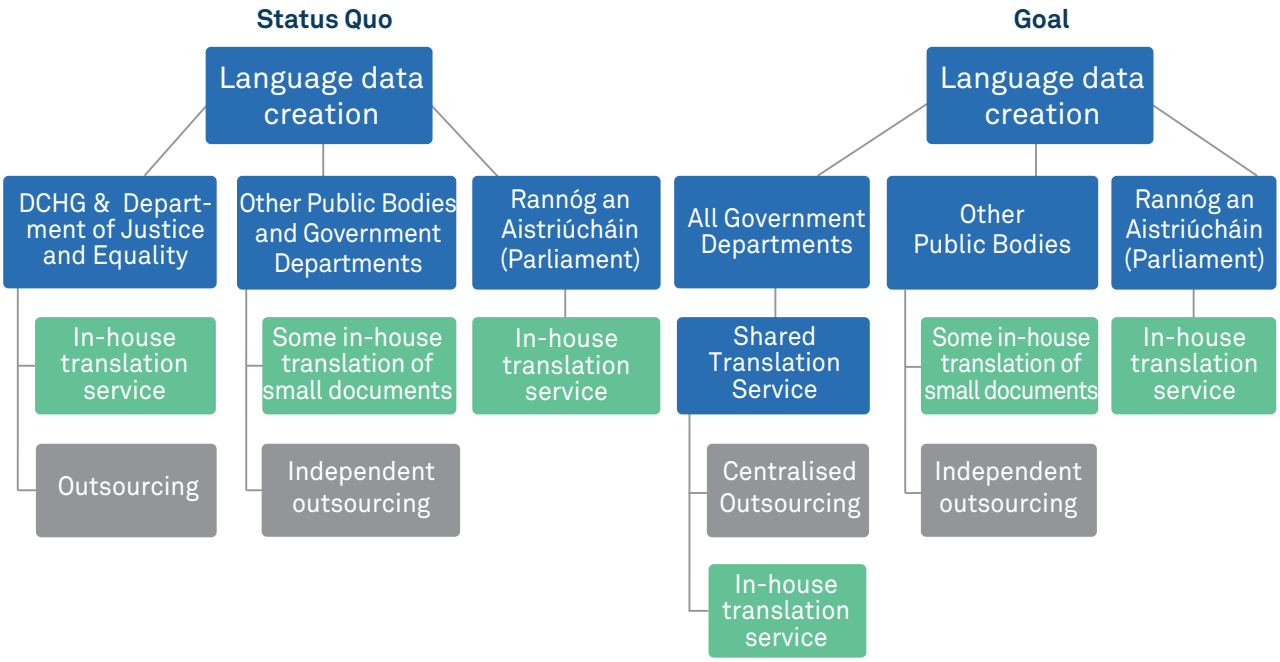
At present, each Government Department is responsible for managing its own Irish-language translation requirements. However, only two of the 17 Government Departments (Department of Culture, Heritage and the Gaeltacht (DCHG) and the Department of Justice and Equality) have in-house staff translators (who use computer-assisted translation (CAT) tools) and much of the other Departments' translation work is done by external translation companies or freelance translators.

Since 2016, a framework for outsourcing translation work was established which includes a small number of selected accredited language service providers (LSPs) with set rates for translation. Most of the documents outsourced by Government Departments for translation are larger corporate documents. The State Examinations Commission have in-house translators and the Houses of the Oireachtas (the Irish Parliament) have a large translation team called Rannóg an Aistriúcháin comprising over 20 staff. The latter team have recently begun using CAT tools, thus generating TMX (translation memory) files for reuse both in-house and at the DGT Irish Language Unit. Many other public bodies, such as Universities and county councils, have dedicated Irish language officers who often carry out small translation tasks in-house, without the use of CAT tools. Larger translation tasks are outsourced to LSPs.

[54] According to the Central Statistics Office (CSO) at a population of 4,761,865 in 2016.

It is not common practice for public bodies or government departments to request the return of TMX files from an LSP (which is a by-product of a translation procurement). Since the launch of the European Language Resource Coordination (ELRC) workshops in 2016, however, some members of public administration have begun to do this, while some departments have since reported the inclusion of such a requirement in their translation contracts. Many Irish language officers or translators in public administration are unaware of technology opportunities or benefits. In June 2019, a Computer Aided Translation Workshop was held in Dublin City University (DCU) to provide CAT and Machine Translation post-editing training to freelance and public administration translators. Only 16% of attendees had previous experience of CAT tools and only one attendee had ever post-edited machine translation output.

The current language data creation infrastructure in Ireland public bodies looks as follows:



The DCHG's establishment of a shared translation service An tSeirbhís Chomhroinnte Aistriúcháin (Shared Translation Service - STS) aligns with the goal outlined below. The proposed model involves developing a centralised point to which all government department translation requests can be submitted and managed, and from which all translation tasks are either handled in-house or outsourced appropriately. Through coordinating a single approach to translation practices and language resource re-use, the STS will assist Departments in complying with their statutory obligations, streamlining and regularising translation services in order to reduce the costs of such services. Other public bodies would continue to translate in-house when possible and outsource larger translation tasks.

Digital Policy and Language Policy

Like many minority languages, the relatively low number of speakers of Irish has resulted in little investment from industry to date. As such, language technology support for Irish is weak, with the availability of most existing tools and resources being made possible only through volunteer activities or short-term projects based in universities. The Action Plan for the Irish language (2018 - 2022)[55] highlights the immediate need

[55] Department of Culture, Heritage and the Gaeltacht: *Action Plan for the Irish language (2018 - 2022)* 20-Year Strategy for the Irish Language 2010-2030, 2018.

for further research and development in the area of language technology for Irish. To address this, a Digital Plan for the Irish language is currently being prepared to outline technological requirements for safeguarding the future of the language. The development of such a plan is crucial to avoid the risk of the language falling behind with regard to technological developments as per 'The Irish language in the Digital Age', compiled by META-NET[56].

Interesting fact:
Researchers at the ADAPT Centre in Dublin City University developed a bespoke machine translation system - Tapadóir - based on a free statistical machine translation engine, which is tuned specifically on public administration data, and is now fully integrated into the translation workflow of the in-house translation unit at the Department of Culture, Heritage and the Gaeltacht.

A national Terminology Committee (an Coiste Téarmaíochta) was established to develop, approve and provide authoritative, standardized Irish language terminology. A national terminology database (www.tearma.ie) is then updated accordingly with decisions made by the committee. The database can be downloaded freely in TBX or TXT format. With respect to generating Irish terminology for the European Union, the DCHG provides funding for the Irish/EU terminology project LEX (GA IATE) and will continue until 2021.

Until recently, Irish public administrations did not exchange language resources such as translation memories or glossaries centrally. Following ELRC promotional and educational activities in Ireland (through workshops[57], outreach seminars[58] and on-site visits), a number of stakeholders began to contribute existing language resources and change internal data-management processes. These stakeholders included some government departments, county councils, universities, the national broadcaster (RTÉ), dictionary publishers and the language commissioner's office. These practices have become significantly more widespread over the past year through the work carried out by the European Language Resource Infrastructure (ELRI) project. As part of this CEF-funded project, Ireland's National Relay Station (NRS), available both in Irish and English was developed (https://elri.dcu.ie): this is a pioneering, online, secure platform where members of public institutions in Ireland can contribute their own language data to a national centralised portal, receive automatically generated TMX versions of their datasets and download shared resources from other public body contributors. According to their sharing licensing or agreements, this data is then "relayed" onto the ELRC-SHARE. There are currently 39 registered users.[59]

The DCHG and the Houses of the Oireachtas (through Rannóg an Aistriúcháin), have also committed to providing the EU institutions with legal text translation memories, containing alignments of English and Irish texts from national legislation. The recently commenced CEF-funded PRINCIPLE project (in collaboration with Norwegian, Icelandic and Croatian partners) will also assist in helping meet this request by processing legacy and archived data into TMX format for reuse at national and EU level.

Sharing public sector information in Ireland

According to Ireland's implementation of the Public Sector Information Directive (PSI) 2003/98/EC, (now known as Open Data Directive (EU) 2019/1024), an individual or a legal entity may make a request in a legible form to a public sector body to release documents for re-use. Every request made in a language other than Irish or English shall be accompanied by a translation of the request into Irish or English. Open Data listed in data.gov.ie is published by over 80 Government Departments and public bodies and is operated by the Government Reform Unit of the Department of Public Expenditure and Reform (DPER). This national Open Data

[56] Judge et. al: *The Irish Language in the Digital Age, in: META-NET White Paper Series*, 2012.
[57] Dowling, Judge, Way: *ELRC Workshop Report for Ireland*, 2016.
[58] Dowling, Lynn, Way: *ELRC Workshop Report for Ireland*, 2017.
[59] Figures as of November 2019.

portal is intended to provide easy access to datasets that are free to use, reuse, and redistribute with many of datasets being individually published and updated by public organisations. Ireland is currently ranked 1 for Open Data maturity in Europe[60]. With regards to sharing language data, users of the Irish National Relay Station can choose to share their data as Open Data and the Open Data Unit in DPER are actively working on linking the LRs uploaded to the NRS (which have been uploaded as Open Data) with the data.gov.ie website.

In the forthcoming Digital Plan for the Irish language, recommendations are made for datasets (language data in digital format, whether bilingual, monolingual, terminology-based, linguistically annotated, etc.) to be made open and freely available where possible (e.g. CLARIN). While some datasets (e.g. the New Corpus for Ireland) may not be released due to copyright reasons, the recommendation is that such resources are at least shared on a restrictive licence basis for research and development purposes (e.g. training word embeddings for neural-based systems, training language models, etc.).

Main challenges for sustainable data sharing:

Feedback surveys conducted during the ELRC and ELRI outreach workshops[61] reported that many people working in the public sector and dealing with language, translation and data are very enthusiastic about using language technology, sharing data and thus supporting the Irish language. Yet, for a number of reasons, language data sharing is still a difficult endeavour. The main reasons for this are the following:

- Each government department and public body manages their own translation needs, either through in-house translation or outsourcing. There is no regulation around the management of translation data or the requirement for LSPs to return translation memory files with the translated documents.
- Until the recent establishment of the National Relay Station, there has been no culture of sharing translation memories or terminologies across departments or public bodies.
- Public servants raise concerns about whether or not they have permission to share their data. This is linked to general lack of awareness or understanding of the Open Data Directive.
- Lack of technical skills with respect to CAT tools amongst translators in public administration.
- General unawareness of the value of language data and leveraging opportunities it presents.
- Unawareness of the need for language data to build translation systems.
- Within public administration, language data management is outside the scope of any specific role and therefore can be difficult to ensure follow-through. In addition, staff changeover/ department changes or merges, result in staff not being able to find legacy data.
- Language Technology and Machine Translation is not on top of the priority list of most in public administration, making the efforts more difficult even when people are generally supportive.
- Misuse of free online translation services which leads to skepticism and wariness of MT in general (without understanding the strengths of domain-specific MT systems).
- There is a widespread lack of awareness and uptake of eTranslation within public administration in Ireland and therefore a lack of full understanding of the long-term benefits.

60 https://www.europeandataportal.eu/en/dashboard#tab-overview

61 Representatives from 41 institutions included but were not limited to bodies such as Department of Foreign Affairs, Department of Culture, Heritage and the Gaeltacht, An Post, County and City Councils, Universities, Health Service Executive, Defence Forces and the Language Commissioner.

Annex
Country Profile Ireland

Action plan:

To address the identified challenges, the following five main objectives were formulated, ranked by their importance for the landscape in Ireland:

- **Raising awareness of language data as Open Data and a valuable asset:**
  Awareness has thus far been raised through both the ELRC and ELRI workshops, TV, radio, social media, youtube, online news articles, public lectures, podcasts and so on. These promotional and awareness-raising efforts require continued support from the DCHG along with the Open Data Unit (currently based in DPER).

- **Increasing interest in MT/LT in public services as part of the national digital policy:**
  The forthcoming Digital Strategy for Irish highlights the immediate need for upskilling current translators and increasing the uptake of the use of translation technology in public administration, in addition to providing technical training in translation courses. It is expected that through using the NRS, the availability of shared TMX files amongst public administrators will encourage an increased use of CAT tools. The EC Representative in Ireland is planning a CAT training workshop for freelance and public administration translators in 2020 with the help of experts at DCU.

- **Identify and gain access to outsourced translations:**
  Since the ELRC and ELRI data collection campaigns, a number of stakeholders have updated their contracts with their LSPs in procurement of translations to stipulate the requirement of the return of a TMX file, terminology database or related glossaries. The Shared Translation Service will also ensure streamlined and centralised access to all outsourced translations from government departments.

- **Establish good data management practices in public services:**
  Outreach workshops, onsite assistance and online training videos are amongst the approaches being taken in Ireland to encourage improvements in data management practices both within Government Departments and other public bodies.

- **Tackle legal concerns:**
  The establishment of the National Relay Station aimed to address concerns of contributors with respect to data sharing licensing and copyright in Ireland. Administrators of the NRS provide a first stop information point to advise a user on the appropriate sharing levels for any given dataset. Any queries that require further detailed examination or investigation are referred to the Open Data Unit or the ELRC help desk.

References and further reading list:

Official Languages Act 2003: http://www.irishstatutebook.ie/eli/2003/act/32/enacted/en/html.

Office of An Coimisinéir Teanga: *Official Languages Act 2003, Guidebook,* 2015, https://www.chg.gov.ie/gaeltacht/the-irish-language/official-languages-act-2003/.

Department of Culture, Heritage and the Gaeltacht: *Action Plan for the Irish language (2018 - 2022) https://www.chg.gov.ie/app/uploads/2019/02/action-plan-2018-2022.pdf*

20-Year Strategy for the Irish Language 2010-2030, 2018, https://www.chg.gov.ie/gaeltacht/20-year-strategy-for-the-irish-language-2010-2030/.

Dowling, Judge, Way: ELRC Workshop Report for Ireland, 2016, http://lr-coordination.eu/sites/default/files/Irleand/ELRC-Workshop-Report_Ireland-Public.pdf.

Dowling, Lynn, Way: ELRC Workshop Report for Ireland, 2017, http://www.lr-coordination.eu/sites/default/files/Ireland2/ELRC%2B%20Ireland%20Workshop%20Report-Public_0.pdf

Judge, Ní Chasaide, Ní Dhubhda, Scannell, Uí Dhonnchadha: *The Irish Language in the Digital Age*, in: *META-NET White Paper Series,* 2012, http://www.meta-net.eu/whitepapers/volumes/irish.

CEF Telecom call - Automated Translation (CEF-TC-2019-1) for PRINCIPLE project: https://euroalert.net/call/3864/2019-cef-telecom-call-automated-translation-cef-tc-2019-1.

European Research Infrastructure for Language Resources and Technology (CLARIN): www.clarin.eu.

The New Corpus for Ireland: http://corpas.focloir.ie/.

ELRI Ireland training video series: https://www.youtube.com/watch?v=xW-sKTTkSzY&t=14s.

Annex
Country Profile Italy

Claudia Foti, Simonetta Montemagni, Claudia Soria, Lilli Smal

State of Play:
Translation practices and needs in ministries and public administrations in Italy:

The translation process in Italian public administrations is highly decentralized, meaning that there is no formalized exchange of know-how or language data between public bodies and the procurement process is not centralized either. Only larger ministries have their own translation service but all ministries take care of their own translation needs. For the most part, translations are executed in a traditional way, meaning there is little or no use of computer-assisted translation (CAT) tools and no official use of safe machine translation (MT). There is an "unconfessed" use of open source MT web interfaces underlining the fact that there is little awareness of related copyright issues and potential security breaches through uploading potentially confidential or personal data to these web services.

Although Italy has only one official language, 12 other languages have co-official status at regional level, out of which only three are official EU languages (French, German, Slovene). Some other immigrant groups form additional linguistic minorities (e.g. Chinese, Arabic) which leads to a high demand for translations. To meet the high demand in translations, all public administrations (PAs) make use of procuring translations. As mentioned above, there is no system or formalized process in place and the produced translation memories (TMs) and any other by-products are not requested back together with the translation.

However, there are also some recent developments where some translators in public administrations (Bank of Italy, Department for European Policies at the Presidency of the Council of Ministers, Ministry of Defence) started to use CAT tools. As of recently, the contracts for outsourced translations were rewarded by the Ministry of Defence and the Department for European Policies to claim ownership of the translation memories and any by-products of outsourced translations.

At the second ELRC workshop in Italy, many participants expressed their wish for an API that can be integrated into websites for automated translation, but concerns were raised with respect to the costs of implementing and maintaining such a service. With respect to a potential eTranslation API, concerns were raised regarding the availability of the service free of charge.

> **Interesting fact:**
> **The average age of civil servants working in ministries, the Presidency of the Council of Ministers and other public bodies is over 54 in Italy. More than 16 % are over 60 years old and less than 3 % are under the age of 30.**[62]

Language data sharing and Open Data in Italy:

Although there is no central repository for language data sharing in Italy and no sustainable infrastructure in place yet, several public administrations and research institutions have already donated language data to the ELRC-SHARE repository. At the time of writing this country profile (October 2019), the data is not yet available on the national Open Data Portal.

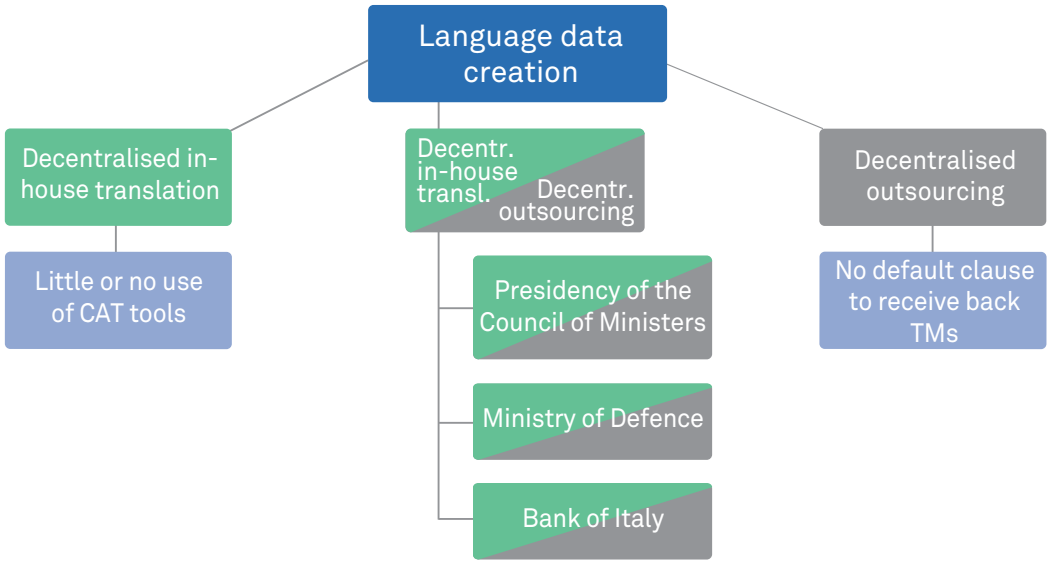
In the past years, considerable efforts were made to improve publishing of Open Data - making Italy one of the trendsetters for Open Data in Europe. On this front, there has been considerable growth, and according to the Digital Economy and Society Index (DESI) 2018, the country has in fact improved its position in the ranking of 11 places, thus exceeding the EU average for Open Data.[63] The initiative "Open Data 200 Italy" aimed at carrying out the first comprehensive, internationally comparable study of Italian companies that are using Open Data to generate business, develop products and services, and create social value. Open Data is widely promoted through public initiatives, but textual data is underrepresented in terms of shared language data and its promotion as a valuable resource.

---

62 Source: ForumPA 2019.

63 Cf. European Commission: *Indice di digitalizzazione dell'economia e della societá (DESI), relazione nazionale sull'Italia per il 2018*, 2018.

The current language data creation infrastructure in Italy public bodies looks as follows:



Language policy and digital policy in Italy:

Italian is spoken by about 58 millions of resident people and is the official language of the Italian Republic, according to National Law no. 482 of 1999. Before this date, an official language had never been explicitly stated. The same law provides the institutional framework for protection of historical linguistic minorities by the Republic protecting the language and culture of the Albanian, Catalan, Germanic, Greek, Slovenian and Croatian populations and of those speaking French, Franco-Provençal, Friulian, Ladin, Occitan and Sardinian.

Italian is largely used for all types of communication in everyday life and is the language of almost all national media, publishing and public administrations of the State. Its use is not explicitly regulated. The use of recognised minority languages is permitted in school education, PA offices and institutions, and public signage.

The European Charter of Minority Languages has been signed, but not yet ratified.
As the Open Data initiatives already indicated, Italy puts a lot of effort into modernizing and digitalizing the public sector. To achieve the objectives of the "Three-Year Plan for Information Technology in Public Administration", Artificial Intelligence (AI) methods and tools shall be applied to public services. The White Paper "L'Intelligenza Artificiale – al servizio del Cittadino" published in March 2018 underlines the role AI shall play in the process. The paper is edited by the AI Task Force and is the result of a consultation, synthesis and analysis process that has involved hundreds of public and private subjects dealing with AI.[64] Machine translation is listed as one of the technologies playing a key role to meet the challenges public administrations are facing with the help of AI at the service of citizens. However, the integration of MT tools into everyday translation work carried out by public administrations is still fairly limited. A notable exception is represented by the Dipartimento per le Politiche Europee della Presidenza del Consiglio dei Ministri (Department for European Policies of the Presidency of the Council of Ministers), which is the first Italian PA to integrate eTranslation into their data management flow.

The Italian national strategy was developed by converting the objectives of the European Digital Agenda into initiatives aimed at the digital transformation of Public Administration. The four key pillars are:

[64] The Agency of Digital Italy: *Artificial Intelligence at the service of citizens*, 2018.

Annex
Country Profile Italy

- Digital ecosystems
- Immaterial infrastructures
- Physical infrastructures
- Cybersecurity

Yet, there is still a gap between the supply of digital services and their actual use by the population.

Stakeholders:

The Agency for Digital Italy is not only implementing the digital agenda in Italy, the agency is also responsible for the national Open Data Portal. As such, they are one of the key stakeholders for sustainable language data sharing who themselves already contributed language data to ELRC.

Over the past years, more than 50 institutions, among them federal and regional public administrations and agencies as well as research institutions, participated in ELRC events. The national ELRC workshops provided the opportunity to involve almost all the public institutions that have translation needs as well as those that are currently using embedded eTranslation in the provision of cross-border services. The increased number of attendees of the second ELRC workshop with new institutions involved (inter alia Bank of Italy, INPS, Ministry of Defence) and of new language resources provided, indicate a growing interest in MT.

Among the active contributors of language data are the Ministry of Interior, the Ministry of Justice, the Province of Bolzano, the Prefecture of Florence, as well as the Universities of Bologna and Pisa and the Institute for Specialised Communication and Multilingualism, EURAC Research, Bank of Italy, and Presidency of the Council of Ministers.

Main challenges for sustainable data sharing:

The challenges Italy is facing in sharing data and restructuring the translation workflow to make it more efficient can be grouped in three categories:

- **Public perception of language data**
  - In general, there is little awareness of the importance of language resources for machine translation and also other applications of artificial intelligence.
  - Language data is not considered a valuable resource and therefore, its management is not regulated through respective policies on the policy level or through appropriate language data management structures on the institutional level.
  - Lack of proper information and education leads to insecurity when it comes to data sharing and open data, which results in little willingness to share translations.

- **Structural issues**
  - Language issues are only considered relevant because they impose costs.
  - Only larger ministries have their own translation service, but when the translation service is centralized at the institutional level, the head of office is usually not a linguist and not familiar with CAT tools and recent developments of machine translation.
  - All ministries take care of their own translation needs independently and there is no exchange of translation memories, terminologies or expertise at the national level.
  - Much translation work is outsourced, but no system or infrastructure is in place to manage these translations, so that the translations are not reused and their linguistic value is lost
  - Little knowledge about automatic anonymization tools results in valuable data remaining unshareable, since manual anonymization is very time consuming.
  - Privacy concerns regarding confidential and personal data (GDPR) are an obstacle for many Ministries (Justice/Interior) to share language data.

  - Ownership of in-house translations lies with public administrations, which requires expertise for appropriate licensing (Art. 11 of Italian Copyright Law).
  - Frequent turnover of top managerial staff in public administration makes the identification and involvement of decision makers very difficult (e.g. identifying the person who knows if data can be shared or someone who can initiate changes).

- **Translation workflow**
  - Little or no use of CAT tools and eTranslation (in some public administrations, translators have only recently started to use CAT tools).
  - Unconfessed use of free online translation services.
  - The original text is often on paper or "digitized" through bad quality scans.
  - Resistance to changing the translation workflow due to average age of civil servants being above 54 in ministries.
  - Investing in CAT Tools is only measured by the cost of purchase, the productivity gain or reuse of language resources is not taken into consideration.

Action plan:

Italy needs awareness raising activities on the value of language data both with translators and decision makers in public administrations. This should be done through examples of how language data management practices can reduce costs and improve quality.

Legal, privacy and ownership concerns should be addressed and best practices in the use of CAT tools and language data management should be developed, preferably by a central body.

The following specific objectives are suggested to address the challenges Italy is facing when it comes to sharing language data:

- **To increase the interest in MT/LT in public services as part of the national digital policy.**
  Specific actions include:
  - Establishing synergies with national projects and initiatives
  - Diffusing best practices
  - Securing the support of decision makers to adapt the national policy

- **To identify and gain access to outsourced translations**
  - Establish practice of receiving any by-product of outsourced translations, especially translation memories

- **To establish good data management practices in public services**
  - Further investigation of data management practices
  - Definition of confidential and personal data that can be used to introduce the practice of clear separation between confidential and personal data from public sector information in the translation process
  - Establish shared language data management practices to reduce costs, improve quality and leverage on existing language assets

- **To raise awareness of the importance of language data as a valuable asset and as Open Data**
  - Raise awareness on the value chain of language data and the importance of LR
  - Share benefits of sharing language data
  - Enhance the publishing of Open Data making Italy one of the trendsetters of Open Data in Europe
  - Integrate language data in the national Open Data policy
  - Emphasize the role of digital texts in the digital economy
  - Establish practical guidelines for language data as Open Data

Annex
Country Profile Italy

- **To tackle legal concerns**
  - Develop, share easy to apply guidelines for IPR and privacy issues

References and further reading list:

European Commission: *Indice di digitalizzazione dell'economia e della societá (DESI), relazione nazionale sull'Italia per il 2018,* 2018, http://ec.europa.eu/information_society/newsroom/image/document/2018-20/it-desi_2018-country-profile-lang_4AA6AC9F-0F0F-0F48-8D21A979E9D5A1B7_52348.pdf.

ForumPA 2019: https://www.forumpa.it.

Presidency of the Council of Ministers: http://www.politicheeuropee.gov.it/it.

Italian Open Data Portal: https://www.dati.gov.it.

The Agency of Digital Italy: *Artificial Intelligence at the service of citizens,* 2018, https://ia.italia.it/assets/whitepaper.pdf.

The Agency of Digital Italy: https://libro-bianco-ia.readthedocs.io/en/latest/index.html.

Annex
Country Profile Latvia

Armands Magone, Jānis Ziediņš, Normunds Grūzītis, Aivars Bērziņš

State of Play:
Translation practices in ministries and public administrations in Latvia:

In Latvia, most translations are independently outsourced, and there are only some ministries and public administrations with in-house translation services. If documents are translated in-house or outsourced to freelance translators, computer-assisted translation (CAT) tools are rarely used. Contrary to that, the use of CAT tools is common practice for Latvian language service providers (LSPs) from whom some public administrations request back translation memories (TMs). However, through the language technology platform HUGO.lv, all public administrations and citizens have access to a free machine translation platform (see section "Data sharing infrastructures and Open Data in Latvia"). As regards procurement of translations, there are no specific procedures regulating translation subcontracts in Latvia. Latvia is subject to the EU regulation. Public procurement is currently regulated by two laws, the 2016 Law on Public Procurement "Public procurement Law" (Latvijas Vēstnesis, 254, 29.12.2016.) transposing the EU Directive 2014/24/EU and the Law on the Procurement of Public Service Providers (Latvijas Vēstnesis, 36, 16.02.2017.).

The procedure on the national level prescribes that procurements above the threshold of 10,000 EUR must be published on the website of the Procurement Monitoring Bureau (including the notification and the results of the procurement and if applicable the winner).

Interesting fact:
Every internet user and all Latvian public administrations have free access to the Latvian State administration language technology platform HUGO.lv offering machine translation, an online CAT tool, speech recognition, speech synthesis and other tools free of charge.

Data sharing infrastructures and Open Data in Latvia:

In Latvia, there are a number of portals and platforms available with the goal to share data and make language technologies available. They include:

- **The Open Data Portal**

The purpose of the Open Data Portal data.gov.lv is to gather and to circulate data collected by Government institutions and Government organizations in one central place for public use and reuse as this data is valuable for the development of innovations in the State. The Latvian Open Data Portal was created by the project Nr. 2.2.1.1/16/I/001 "Public Administration Information and Communication Technology Architecture Management System" (PIKTAPS) co-financed by the European Regional Development Fund.

- **HUGO.lv**

Hugo.lv is a Latvian State administration language technology platform that is freely accessible to everyone. It provides machine translation, an online CAT tool, speech recognition and speech synthesis, as well as a range of tools for supporting multilingual features in e-services. Hugo.lv is customized to the Latvian language and state administration documents, thus, its translation quality is significantly higher than in other online translation services. Furthermore, users of Hugo.lv can enjoy the services of the translation assistant for more convenient translation. One of the functions of the platform is the resource management feature with two main functionalities – "Submit resource" (TMX) and "Search the database". "Submit resource" is an easy and straightforward process - language resources can be submitted by creating a new user and by opening the "Submit resources" section. The user enters the resource name, the e-mail address for communication, describes the areas and languages covered, and uploads files containing text. By clicking the "Upload or drag a file" button, documents can be added or dragged into the enclosed area with a dashed line. The attached files will appear in the list. "Search in repository" - allows the user to search the repository for publicly available language corpora and download the returned results. When the "Resource Search" section is opened, the user can view all available corpora and the languages covered in the drop-down list of the search form. The user enters the keyword, selects the body and language, and then clicks "Search".

Annex
Country Profile Latvia

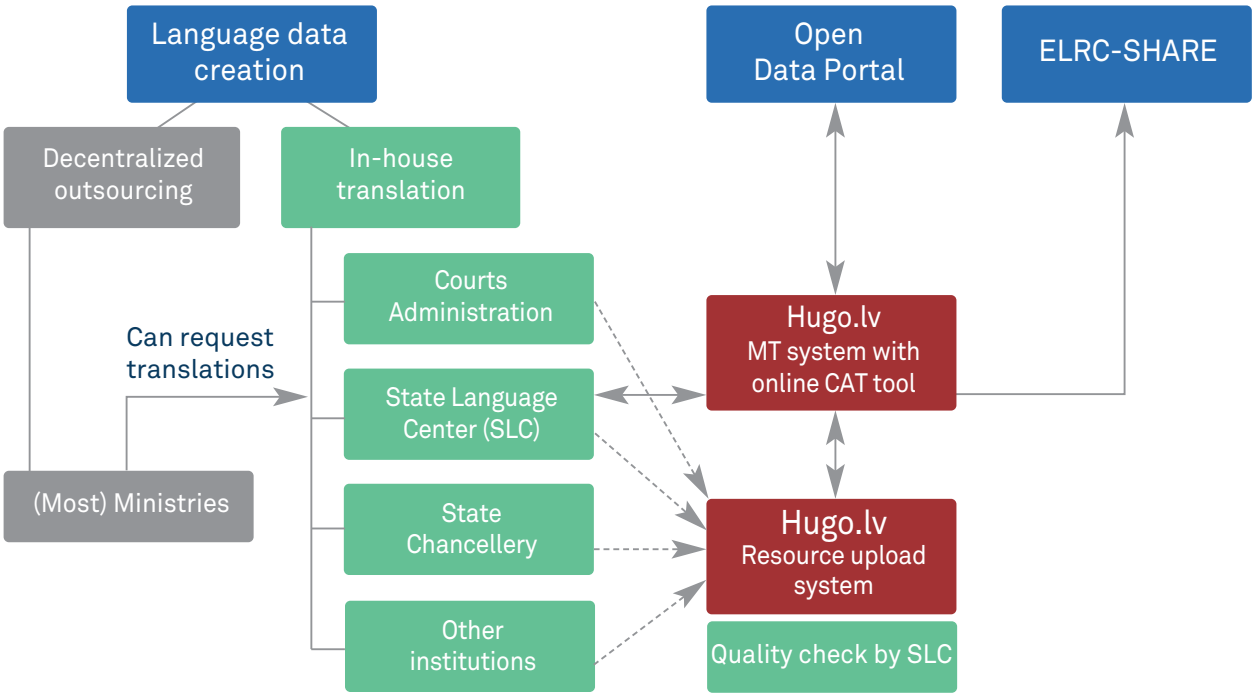
- **META-SHARE**

META-SHARE, the open language resource exchange facility, is devoted to the sustainable sharing and dissemination of language resources (LRs) and aims at increasing access to such resources in a global scale. META-SHARE Latvia node is metashare.tilde.com and it serves as the META-SHARE hub for the Baltic and Nordic region.

- **CLARIN-LV**

The Common Language Resources and Technology Infrastructure (CLARIN) is a research infrastructure that was initiated from the vision that all digital language resources and tools from all over Europe and beyond are accessible through a single sign-on online environment for the support of researchers in the humanities and social sciences. Latvia joined CLARIN ERIC in June 2016. The coordinating centre of CLARIN Latvia is the Artificial Intelligence Laboratory (AiLab) at the Institute of Mathematics and Computer Science, University of Latvia. In addition, AiLab provides Latvian language resources (treebanks, framebanks and other annotated corpora, as well as lexical resources) as Open Data for language technology research and innovation through public Git repositories.

The current language data sharing infrastructure in Latvian public bodies looks as follows:



Language policy in Latvia:

The necessity of language technology(ies) to support Latvian as Latvia's sole official language in digital means has been recognized by the national government and is reflected in a number of language policy documents. Regulatory enactments governing the use of the official language of Latvia include:

- **Laws:**
  - The Constitution of the Republic of Latvia
  - Official language law
  - Law on Submissions

- **Cabinet regulations:**
  – Regulations regarding the amount of knowledge of the official language and procedures for the verification of the proficiency of the official language for the performance of professional duties, receipt of a permanent residence permit and acquisition of the status of a permanent resident of the European Union and the State fee for the verification of the fluency of the official language
  – Procedures by which translations of documents into the official language shall be certified
  – Rules on the provision of translations into measures
  – Rules on the use of languages in information technologies
  – Rules regarding the formation and use of the names of institutions, public organisations, undertakings (companies) and the names of events
  – Provisions regarding the use of foreign languages on stamps and forms
  – Provisions regarding the State fee for the performance of professional duties regarding the attestation of proficiency in the official language
  – Provisions regarding the spelling and use of personal names in Latvian, as well as their identification
  – Information rules for place names
  – By-law of the meeting of senior officials in matters of the European Union

- **Other cabinet regulations:**
  – On priority axes for science in 2018-2021. (Languages and values are one of the nine scientific priority directions for Latvia in 2018-2021)

- **Cabinet instructions:**
  – Procedures for evaluating, harmonising and correcting translations of European Union documents
  – Procedures for requesting and providing translations

- **By-laws of national regulatory authorities related to the official language:**
  – By-law of the Latvian Language Expert Commission of the State Language Centre
  – By-law of the Terminology Commission of the Latvian Academy of Sciences

- **Policy Planning Documents:**
  – National language policy guidelines for 2015-2020
  – Concept for the development of the system of administrative penalties
  – Latvia's Open Data Strategy
  – Language technology is part of the seven directions that are supported by Guidelines of the Information Society for 2014-2020 (Order of Cabinet of Ministers Nr. 468) with specific focus on promoting the distribution of Latvian language applications in the digital environment

Stakeholders:

Overall, more than 60 organisations participated in local ELRC workshops or conferences including high-level officials, indicating strong interest in the topics covered by ELRC in Latvia.

Interesting fact:
In September 2019, the Culture Information Systems Centre contributed open language data, which was generated by the LT platform Hugo.lv to ELRC. The donation includes monolingual corpora with more than 300 million words, parallel corpora with 15 million words and 19 000 Latvian terms.[65]

[65] ELRC News Article: *Latvia contributes language data for eTranslation,* 2019.

Annex
Country Profile Latvia

In addition, several datasets were already contributed to ELRC by the Bank of Latvia and other institutions for example. Other important key stakeholders include:

- **Culture Information Systems Centre (CISC):**
  CISC is the owner and administrator of HUGO.LV. The CISC operates under the supervision of the Ministry of Culture of Latvia. The objective of the Centre is to provide access to information resources and cultural heritage stored in archives, museums and libraries. CISC implements national and international ICT, provides training and supplies public access to projects and programs to enable free and equal access to information, resources and cultural heritage stored in libraries, archives, museums and other cultural institutions.

- **The State Language Centre:**
  The center has two main objectives, i.e. (1) to implement the national policy with regard to supervision and control of the conformity with laws and regulations in the field of the official language use and (2) to supervise the implementation of the Official Language Law.

- **The Latvian Language Agency:**
  The Latvian Language Agency is a direct administration institution supervised by the Minister of Education and Science, and its aim is to enhance the status and promote sustainable development of the Latvian language – the official State language of the Republic of Latvia and an official language of the European Union.

- **Latvian Language Expert Commission:**
  The Latvian Language Expert Commission, on a regular basis,examines the compliance of norms provided for in laws and regulations to the rules of the Latvian language, codifies norms of the literary language, provides opinions on various language issues, for example, the use of capital letters in the names of establishments, the spelling of internationally recognised names of countries and territories, house names and numbers, addresses, languages and language groups in the Latvian language in compliance with the requirements of ISO 639-2 et. al. The commission prepared several draft legal acts and participated in the formation of the normative basis for the Official Language Law.

Main challenges for sustainable data sharing:

- Lack of awareness and distribution of responsibilities in ministries and other state administration institutions.

- Language data is missing as one of the important categories in the PSI Directive.

- Practices and procedures for subcontracting translations including the return of translation memories are not defined by the state administration.

Action plan:

- Dissemination campaign in all ministries to raise awareness about language data as an important asset for language equality and digital presence is considered vital.

- Together with the Ministry of Environmental Protection and Regional Development changes of procurement procedures for translation subcontracts by state administration should be initiated.

- The state administration should develop good internal practices for language data management.

- The importance of language technologies and benefits for the state administration and citizens from introducing language technologies should be promoted.

European Language Resource Coordination
Connecting Europe Facility

Further reading list:

Artificial Intelligence Laboratory: http://ailab.lv/en/.

Berzins, Aivars, Kalnins, Rihards: *ELRC Workshop Report for Latvia*, 2016, http://lr-coordination.eu/sites/default/files/ELRC-Workshop-Report_Latvia.pdf.

CLARIN Latvia: http://www.clarin.lv/en-us/.

CLARIN ERIC: https://www.clarin.eu/glossary#ERIC.

ELRC News Article: *Latvia contributes language data for eTranslation*, 2019, http://www.lr-coordination.eu/news/Latvia-contributes-language-data-to-for-eTranslation.

HUGO.lv: https://hugo.lv/lv.

Latvian Open Data Portal: https://data.gov.lv

META-SHARE: http://metashare.tilde.com/.

Public Administration Information and Communication Technology Architecture Management System (PIKTAPS): http://www.varam.gov.lv/lat/darbibas_veidi/e_parv/piktaps_projekts/.

Annex
Country Profile Lithuania

Andrius Utka, Virginijus Dadurkevičius, Eileen Schnur

State of Play:
Translation practices in ministries and public administrations in Lithuania:

In Lithuania, there is no centralised translation service in the public administrations yet. Consequently, translation practices vary from institution to institution, ranging from decentralised outsourcing to in-house translation. In the case of Lithuania, independent outsourcing clearly dominates in public administrations and there are only single ministries with in-house translation services. If the translations were outsourced, in some public administrations, it is common practice to request the translation memories (TMs) and other by-products of the translations back.

When it comes to translation services, only single language service providers (LSPs) and freelance translators use computer-assisted translation tools (CAT Tools). This also applies to public administrations, where CAT Tools are only rarely used. Public administrations and ministries usually do not use machine translation (MT) APIs, but freely available MT services. However, this is not the case in the Lithuanian Parliament, since its translators are already successfully using the European Commission's MT system eTranslation.

> **Interesting fact:**
> **In 2018, a new EU project was launched that is developing an MT system for Lithuanian, English, French and Russian to be used nationwide. The project is one of five EU-funded projects in the programme "The Lithuanian Language for Information technologies". It is being implemented by Vilnius University and focuses on the modernisation of previously developed MT systems. Almost 4 million EUR have been allocated for the implementation of this project.**

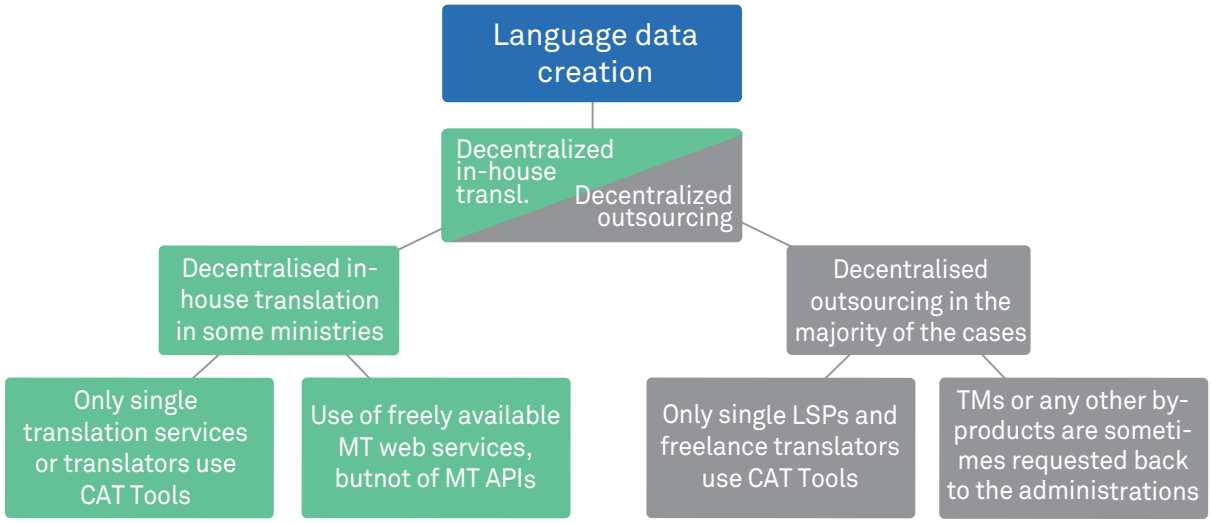
In Lithuania, the Law on Public Procurement of the Republic of Lithuania (LPP) is the main piece of legislation, which governs the implementation of public procurement[66]. The official procurement portal is called the Central Public Procurement Information System. The one-stop-shop portal for public procurement is managed by the Public Procurement Office and its use is mandatory for all public buyers. The portal includes tender notices, publications of awarded contracts and procurement plans and even allows direct electronic communication between the buyer and the economic operators.

> **Interesting fact: The use of Lithuania's official procurement portal "The Central Public Procurement Information System" is mandatory for all public buyers.**

According to the findings of NEC TM, Lithuania spent almost 10 million EUR for translation contracts between 2015 and 2018. Organisations with the highest demand are Lithuania's central purchasing body CPO LT, the Education Exchanges Support Foundation and the State Tourism Department under the Ministry of Economy. This clearly demonstrates that multilingual services are of high relevance in all fields of Lithuanian public services, ranging from finances to education and social affairs.

[66] NEC TM Report: *Report on National Translation Contracts in Lithuania Public Procurement Market Research: Process and Findings for Lithuania,* 2019.

The current language data creation infrastructure in Lithuanian public bodies looks as follows:



Data sharing infrastructures and Open Data in Lithuania:

As regards the exchange and collection of data on the national level, Lithuania is part of CLARIN ERIC. The corresponding CLARIN-LT consortium was founded by three partner universities, i.e. Vytautas Magnus University, Kaunas Technology University and Vilnius University and maintains a repository to collect language data.

According to the Vice-Minister of the Ministry of Economy in Lithuania, the country is known for leading public e-government services. Data openness is mentioned as one of the key objectives of the Digital Agenda for Lithuania from 2014-2020, including the development of legal means for opening data from state and municipal authorities and agencies for example. The creation of effective management structures for opening such data also plays an important role in Lithuania's current Digital Agenda.

In addition, the creation and development of publicly available written and spoken digital content in the Lithuanian language and their implementation in Information and Communications Technology (ICT) and eServices are explicitly mentioned. In order to comply with the Public Sector Information (PSI) Directive, the "Law on the Right to Receive Information from State and Local Authorities and Institutions" was adopted, thus increasing the scope of information intended for re-use to e.g. museums and archives and defining the conditions for the open license to use public sector information based on the Creative Common (CC) License. The latter makes it easier for recipients of information to share, re-use, process or translate any received information.

An important portal is the central electronic services portal "eGovernment gateway". It provides public information and e-services for citizens and businesses by redirecting the portal's visitors to appropriate websites. In recent years, the Lithuanian government has aimed to proceed with the centralised digitalisation of public services. Consequently, the Lithuanian Ministry of Economy and Innovations created a strategy for real-time digital government.

Language policy and digital policy in Lithuania:

With approximately four million speakers, Lithuanian is one of the least commonly spoken European languages. Pursuant to the Constitution of the Republic of Lithuania, the Lithuanian language has the status of state language. In order to enforce this status and to protect the language, the so-called Law on the State language was introduced in 1995. Due to the small number of speakers, the Lithuanian government supports

Annex
Country Profile Lithuania

a number of different programmes, which aim to promote linguistic research and dissemination. In this context, the Institute of the Lithuanian Language plays a key role, since it is one of the most important centres for research and dissemination of the Lithuanian language.

The Lithuanian language policy consists of a set of principal guidelines, which partly go beyond the national borders of Lithuania. It states that the Lithuanian language must be in line with the language policy of the European Union and that it should be developed as a constituent part of multilingual terminologies and resources of the EU. In addition, automated translation is considered highly relevant when it comes to language use within the EU.[67]

The development of the Lithuanian language resources and technologies can be divided into three stages, i.e. the first from 2004 to 2012, the second from 2012 to 2015 and the third from 2016 to 2020.[68] Whereas during the first phase, Lithuanian was considered highly under-resourced and without any (or only weak) language technology support, the period between 2012 and 2015 was labelled as the breakthrough. This was achieved due to three major actions, i.e. the implementation of the national programme "The Lithuanian Language for Information Society", the preparation of political guidelines for the further development of language technologies for Lithuanian 2016 to 2020 and international collaboration of LT communities and infrastructures. However, since there is only a small market for language technologies in Lithuania, the private sector does not show much interest when it comes to the development of LT.

During the third phase, a new national programme was launched in 2018, i.e. "The Lithuanian Language for Information technologies". The programme is funded by EU Structural funds and in total, more than 17 million EUR have been allocated for five language technology projects. The overall aim of the projects is to produce 21 new electronic language services for the general public and public institutions (i. e. machine translation, speech recognition, automatic speech transcription, automatic summarization etc.). On the smaller scale, during this phase, the Research Council of Lithuania has funded a number of important scientific LT projects that are important for the overall picture.

Language technologies are also explicitly mentioned In the Digital Agenda of Lithuania ("The Lithuanian Information Society Development Programme 2014-2020") approved by the Government in 2014. The programme highlights that the Lithuanian culture and its language should be promoted through ICT. Furthermore, the agenda aims to "make as many public and administrative services as possible available digitally"[69]. In addition to international collaborations, two national infrastructures were created, i.e. raštija.lt and LKSSAIS. The latter features main language technologies and language resources and provides public services for institutions and private users.

Stakeholders:

Two relevant stakeholders are represented by the Lithuanian ELRC National Anchor Points, i.e. the State Commission of the Lithuanian Language, Vytautas Magnus University and Vilnius University, who are involved in a number of initiatives that are of relevance to ELRC, such as CLARIN or META-Net. Other important stakeholders are the Office of the Government of the Republic of Lithuania and the Seimas, which has already contributed language data to ELRC. The Research Council of Lithuania and the Institute of Lithuanian language may also play important role in contributing language resources. Overall, more than 30 Lithuanian organisations have participated in previous ELRC events, demonstrating that there is an increasing interest in the topics addressed by ELRC.

67 Vaišnienė, Zabarskaitė: Meta-NET White Paper Series *"The Lithuanian Language in the Digital Age"*, 2012.
68 Utka et. al: *Overview of the Development of Language Resources and Technologies in Lithuania (2012-2015)*, 2016.
69 European Commission: *Digital Government Factsheet Lithuania*, 2019.

Main challenges for sustainable data sharing:

- Potential data contributors are sometimes reluctant to share their language resources due to concerns about their relevance and/or quality;
- Lack of interest at the level of decision makers;
- Lack of effective anonymization procedures, as recent GDPR restrictions made some contributors worry about revealing their sensitive privacy data to outside sources.

Action plan:

Taking the current challenges into account, five objectives could be defined. Ranked by their priority, these are:

- **To increase interest in MT in public services:**
  As it is currently not common practice to use machine translation systems in public administrations, examples of how MT can be a useful asset for institutions and ministries could raise attention and increase interest. These examples should clearly demonstrate how MT can facilitate daily operations and increase productivity. In addition, synergies with other national LT/MT projects and initiatives could have a positive impact on the Lithuanians' general interest in MT.

- **To raise awareness of language data as Open Data:**
  This could be achieved by emphasizing the role of digital texts in the digital economy and by making people aware of the benefits of sharing data. As potential data contributors are sometimes concerned about the relevance and quality of their data, it would also be important to spread the message that single translation mistakes do not have a significant impact on the quality of the MT output.

- **To tackle legal concerns:**
  As legal concerns can prevent potential contributors from sharing their data, we would suggest to create and develop effective anonymisation procedures and tools, which are currently not available in Lithua nia. In addition, it would be important to develop and share easy-to-apply guidelines, which can help data contributors to overcome issues related to privacy or intellectual property rights (IPR).

- **To identify and gain access to outsourced translations:**
  This could be achieved by e.g. cooperating with the NEC TM Data Project.

- **To establish good data management practices in public services:**
  A first step towards improving data management practices in Lithuanian public services could be the appointment of a data manager who is responsible for the data management practices in the respective ministry or public administration. In order to be able to decide on the practices to be applied, it would be important to investigate potentially useful data management practices on the one hand, but also to introduce clear definitions of confidential and personal data on the other.

References and further reading list:

CLARIN-LT: clarin-lt.lt/?page_id=22.

European Commission: *Digital Government Factsheet Lithuania,* 2019, https://joinup.ec.europa.eu/sites/default/files/inline-files/Digital_Government_Factsheets_Lithuania_2019_0.pdf.

Fundings for Lithuanian language in IT:
https://www.esinvesticijos.lt/lt/finansavimas/patvirtintos_priemones/veiksmu-programos-prioriteto-igyvendinimo-priemone-nr-02-3-1-cpva-v-527-lietuviu-kalba-informacinese-technologijose.

MT System developed by Vilnius University:
https://www.xn--ratija-ckb.lt/vu-ma%C5%A1ininis-vertimas/vilniaus-universiteto-ma%C5%A1ininis-vertimas-/16.

Annex
Country Profile Lithuania

NEC TM Report: *Report on National Translation Contracts in Lithuania Public Procurement Market Research: Process and Findings for Lithuania,* 2019, https://www.nec-tm.eu/wp-content/uploads/2019/03/Lithuania-Report.pdf.

Spukiene, Renata: *ELRC Workshop Report for Lithuania*, 2016, http://www.lr-coordination.eu/sites/default/files/Lithuania/ELRC-Workshop-Report_Lithuania_public.pdf.

Spukiene, Renata: *ELRC Workshop Report for Lithuania*, 2019, http://www.lr-coordination.eu/sites/default/files/Lithuania/2019/ELRC%2B%20Workshop%20Report_Public_Lithuania.pdf.

Utka et. al: *Overview of the Development of Language Resources and Technologies in Lithuania (2012-2015)*, 2016,https://etalpykla.lituanistikadb.lt/object/LT-LDB-0001:J.04~2016~1569937265781/J.04~2016~1569937265781.pdf.

Vaišnienė, Zabarskaitė: *"The Lithuanian Language in the Digital Age"*, in: *Meta-NET White Paper Series,* 2012, http://www.meta-net.eu/whitepapers/e-book/lithuanian.pdf.

Annex
Country Profile Luxembourg

Dimitra Anastasiou, Lilli Smal

State of Play:
Translation practices in ministries and public administrations in Luxembourg:

Luxembourg is a multilingual country with three administrative languages (French, German and Luxembourgish) and a multilingual population leading to a considerable need for translation into French, German and English. Luxembourg shows significant efforts into making public services digital and multilingual, the main web portal in this domain being Guichet.lu[70] that is managed by the Government IT Centre[71] (Centre des technologies de l'information de l'État – CTIE). Guichet.lu has its own in-house translation service and regularly exchanges translation memories. In order to fully meet their translation needs, all public authorities outsource at least some translations to either freelance translators or language service providers.

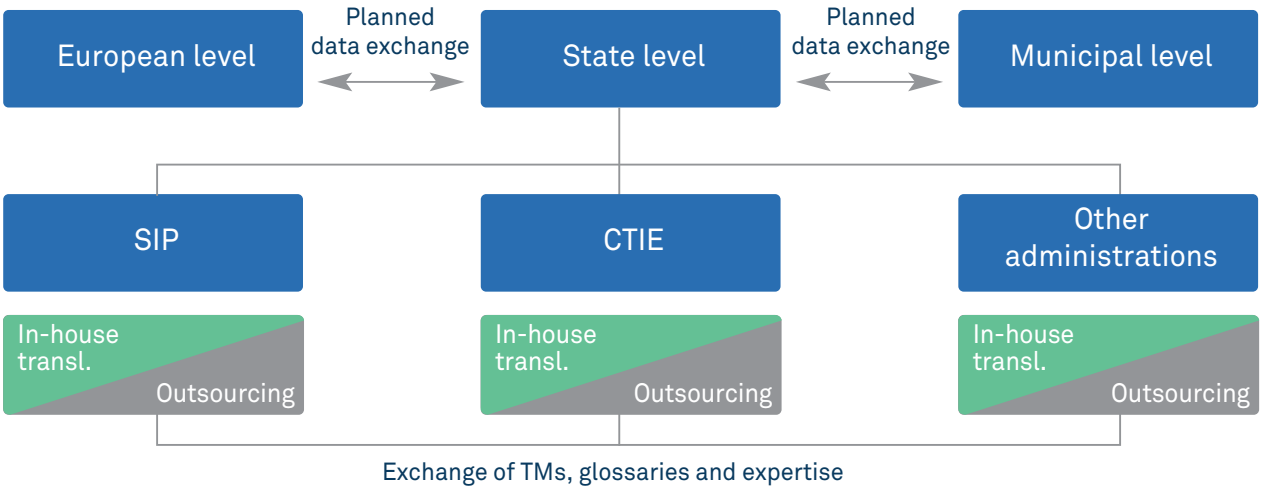
The government.lu[72] website is the information portal of the governmental Information and Press service, SIP. It federates all information and news concerning the Luxembourg government in three languages (German, French, and English) and sometimes in Luxembourgish.

When providing information for its citizens and businesses, the Luxembourg Government follows the principle to adapt the content to a large population by using short sentences and easy to understand, non-specialised language. To ensure that the published information is up to date, a legal team of the competent administration is in charge of checking new laws and procedures, and of adapting and revising the texts accordingly. If the technical solution is available, these changes automatically trigger a new translation job in the CAT tool of the translation unit at Guichet.lu. After the full translation process (human translation, revision, proofreading and validated finalisation), the translated text is automatically sent to the publishing tool.

Time delays between the French and other versions are generally visualised by an "update" hint on the website informing the users about upcoming changes.

Interesting fact:
The Luxembourg Government purposefully uses simple language to provide information for their citizens to allow for equal access to information as well as a solid base for translation into multiple languages.

The current infrastructure of language data creation and exchange in Luxembourg:



70 https://guichet.public.lu/en.html
71 https://ctie.gouvernement.lu/en.html
72 https://gouvernement.lu/en.html

Annex
Country Profile Luxembourg

Goals for future data exchange:

As mentioned above, Guichet.lu exchanges glossaries with the Information and Press Service[73] (Service information et presse, SIP) and the national Open Data portal. The exchange of TMs between ministries and Guichet.lu is to be extended to the Agency for the Development of Employment (ADEM), the Information and Press Service (SIP), the Luxembourg City Municipality and the European Commission (DG CONNECT, UNIT G3: Learning, Multilingualism & Accessibility) in the future.

Open data in Luxembourg:

The Luxembourgish Open Data Portal (https://data.public.lu/en/) was launched in April 2016 and it currently hosts more than 800 published datasets. They are in majority numerical datasets, and not textual. Luxembourg Institute of Science and Technology (LIST) published in 2019 an evaluation of the impact of Open Data in Luxembourg in order to better understand its users and their expectations in terms of content and functionality. The main satisfaction results of the survey follow:

- More communication and advertising about the portal is needed;
- Finding datasets is the main goal of visitors;
- Datasets should be expanded and improved (real time, documentation, harmonisation of data-sets);
- An advanced search tool is requested;
- All the domains are validated (with improvement ideas);
- The socio-economic impact of data.public.lu is real.

According to the survey, the main target of the users was to find “datasets for big data, similar to Kaggle.com”, “Raster data, data about Luxembourg” as well as “Roadworks”.

As far as the impact of Open Data is concerned, users had to rate the economic, environmental, and social feedback. The social impact was ranked as the highest with 68% and the environmental and economic equally with 63%. Regarding the social impact, respondents consider that the portal facilitates citizen science initiatives and provides understanding of population need, demand and use. The economic impact focuses on business activities in ITC and data science market, while the economic impact is about land rezoning and urban planning, among others.

Language and digital policy in Luxembourg:

In the Grand Duchy of Luxembourg, Luxembourgish (Lëtzebuergesch), a Moselle Franconian dialect, is the national language. According to the provisions of the Languages Law of 1984, the three languages of the Grand Duchy, Luxembourgish, French and German, are the languages used in administration and the judiciary system. Legislative documents are in French and an important consequence of this on a judicial level is that only the French language text is deemed authentic for all levels of public administration. The Grand Ducal Regulation of 30 July 1999[74] reformed official spelling in Luxembourgish.

According to the National Institute for statistics and economic studies (STATEC)[75], as of 01.01.2019, the total population of Luxembourg is 613,894 with the native Luxembourgish people 322,430 and total foreigners 291,464[76]. Based on another study published by STATEC in 2013, 70.5% of the population use Luxembourgish at work, at school and/or at home, 55.7% use French, and 30.6% German. On average, 2.2 languages are used. At the same census, 55.8% - a large majority of the country's inhabitants - gave Luxembourgish as their 'principal language'. Portuguese and French followed in second and third positions (15.7% and 12.1% respectively).

[73] https://sip.gouvernement.lu/en.html
[74] http://legilux.public.lu/eli/etat/leg/rgd/1999/07/30/n2/jo
[75] https://statistiques.public.lu/fr/acteurs/statec/index.html
[76] https://statistiques.public.lu/stat/TableViewer/tableView.aspx?ReportId=12859&IF_Language=eng&MainTheme=2&FldrName=1$

Stakeholders:

After the election of the new Government in October 2018, the Government's political programme has placed digitalisation at the centre of its policies. The importance of this topic is thoroughly discussed in the coalition agreement and, for the first time, a Ministry only dedicated to Digitalisation has been created showing that Luxembourg sees digitalisation as a core element of its development. The position of Minister for Digitalisation is filled by the Prime Minister, Xavier BETTEL, who is also Minister for Administrative Reform, which underlines the central importance given to digitalisation and the reformation of the public administration. Marc HANSEN is appointed as Luxembourg's Minister Delegate for Digitalisation and works on a daily basis to enhance Luxembourg Government's efforts to support citizens and businesses on the road to digitalisation. One of the main objectives of the **Ministry for Digitalisation**[77] is to make the lives of the citizens and businesses of Luxembourg easier and to act as a 'facilitator' and a 'coordinator' of all activities related to digitalisation and eGovernment across all ministries.

A central player in this process is the **Government IT Centre** (Centre des technologies de l'information de l'État, CTIE), directly subordinated to the Ministry for Digitalisation and in charge of the setting up and development of eGovernment. Its main mission is to accompany the digital transition of the Grand Duchy's administrations, so that each of them may take full advantage of the opportunities offered by the information and communication technologies (ICTs).

Guichet.lu is managed by CTIE and is a web portal run by the Luxembourg Government to facilitate users' access to information and online services pertaining to any life event or administrative procedure they may have to deal with as private citizens or representatives of businesses, and to simplify administrative procedures. It was launched in November 2008 and offers step-by-step guidance on some 1600 administrative procedures.

Digital Lëtzebuerg[78] ('Digital Luxembourg), founded in 2014, is a multidisciplinary government initiative working with public, private and academic players to harness digitalisation for positive transformation. It approaches digitalisation holistically, focusing on five key areas: skills, policy, infrastructure, ecosystem and government. Executing the Luxembourg government's digitalisation strategy, Digital Luxembourg enables new projects, supports existing ones & boosts the visibility of nationwide efforts.

Main challenges for sustainable data sharing:

- Raising motivation and awareness about language data and machine translation in public administration;
- Small country with many multilingual people – need for translation.

Possible objectives:

Taking the current challenges into account, the most important objective for Luxembourg seems to be to raise awareness that language data should be considered as Open Data and a valuable asset. The role of digital texts in the digital economy is the first step in the terms of this kind of awareness.

As a further objective, even better data management practices could be established in public services. In the European Union, the legislative framework of the Open Data movement is set out in Directive 2003/98/EC and Directive 2013/37/EU on the reuse of public-sector information. In the Grand Duchy, these Directives have been transposed by the Law of 4 December 2007, as amended, on the reuse of public-sector information.

It could also perhaps be helpful to integrate the use of MT and Language Technology more strongly in the national digital policy. As Luxembourg is a small country with many foreigners, there is a clear need for translations of information available to the citizens and businesses.

[77] https://digital.gouvernement.lu/en.html

[78] https://digital.gouvernement.lu/en/dossiers.gouvernement%2Ben%2Bdossiers%2B2014%2Bdigital-letzebuerg.html

Annex
Country Profile Luxembourg

References and further reading list:

Digital Lëtzebuerg:
https://digital.gouvernement.lu/en/dossiers.gouvernement%2Ben%2Bdossiers%2B2014%2Bdigital-letzebuerg.html.

Government IT Centre (Centre des technologies de l'information de l'Etat (CTIE)):
https://ctiegouvernement.lu/en.html.

Government IT Centre (Press Release): *English-language version of the Citizens Portal on Guichet.lu*, 2017,
https://ctie.gouvernement.lu/en/support/recherche.gouvernement%2Ben%2Bactualites%2Btoutes_actualites%2Bcommuniques%2B2017%2B11-novembre%2B07-guichet-anglais-en.html.

Gautier, Martin, Turki: *Impacts of Open Data in Luxembourg and the Greater Region - 2019*:
https://download.data.public.lu/resources/study-impacts-of-open-data-in-luxembourg-and-the-greater-region-2019/20190510-143345/impacts-of-open-data-in-luxembourg-and-the-greater-region-2019-final.pdf

Guichet.lu: *Legal notice*, 2019, https://guichet.public.lu/en/support/aspects-legaux.html.

Luxembourgish Open Data Portal: https://data.public.lu/en/.

Pundel, Lynn: Guichet.lu as a prime example for enabling and sustaining multilingualism, 2018,
https://ec.europa.eu/cefdigital/wiki/pages/viewpage.action?pageId=61932141&preview=/61932141/73543847/2_05_Lynn%20Pundel_Guichet.lu.pdf.

STATEC: *Populations by nationalities in detail 2011-2019,* 2019,
https://statistiques.public.lu/stat/TableViewer/tableView.aspx?ReportId=12859&IF_Language=eng&MainTheme=2&FldrName=1.

Annex:
Country Profile Malta

Donatienne Spiteri, Mike Rosner

State of Play:
Translation practices in ministries and public administrations in Malta:

Each Ministry caters for its own translation needs, with translations being carried out either internally (exclusively) or both internally and through outsourcing (to individual translators or translating companies). The frequency of requests for outsourced translations may vary, with demands being sometimes made on a monthly or quarterly basis. Although most translations are required from MT to EN or vice-versa, some Ministries have also pointed out other language combinations, namely: FR to EN and vice-versa; IT to EN and vice-versa; DE to EN and vice-versa.

The kind of documents required to be translated also varies, with the following examples having been quoted by Ministries: legal documents, press releases, reports, speeches, calls for applications and recruitment calls, official websites, memos, promotional content, letters and forms sent to new or existing pensioners, job descriptions, parliamentary questions, notifications and conference set-ups. No CAT Tools are used when translating in-house, although commercially available translation platforms are sometimes employed. [this overview is based on answers provided by a sample of Ministries during the third quarter of 2019].

There is as yet no government policy for sharing or pooling of translations. Instead there is a tendency for Departments and offices to adopt the same silo mentality that applies to untranslated documents whereby once produced, they remain in situ. As regards pooling of translation resources, there is no contractual obligation for LSPs carrying out translations to deliver translation resources alongside the translations themselves.

Interesting fact:
Maltese is the only EU official language with Semitic roots, which however is written in the Latin alphabet.

Open Data in Malta:

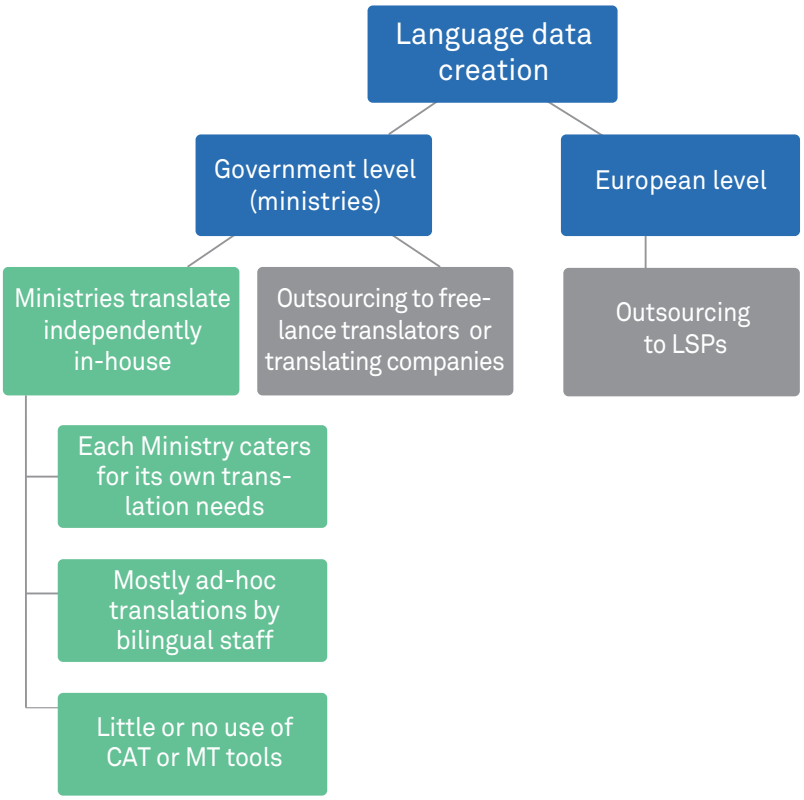
The PSI Directive has been transposed into Maltese law by virtue of Chapter 546 of the Laws of Malta, namely the Re-Use of Public Sector Information Act.

As part of the implementation of the Public Sector Information Re-Use (PSI) Directive, MITA on behalf of the Government of Malta is currently drafting a strategy that provides a holistic and comprehensive vision for the management of data across the whole Public Administration. This is in the context of the Public Administration being one of the pillars of the Digital Malta Strategy together with the Citizen and the Business perspectives.

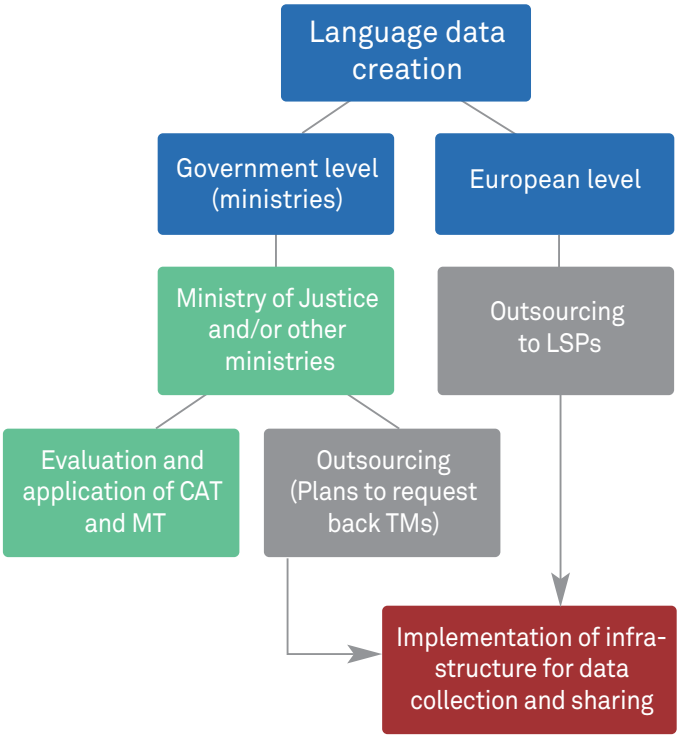
The proposed National Data Strategy is primarily composed of a set of General Principles and Best Practices that should guide future investments in this domain. The document also includes details of the organisational setup required and a set of building blocks that will be used to help Malta to position itself better in the context of Open Data and Big Data, in line with EU direction as part of the Digital Economy and the Digital Single Market. This while also ensuring that the correct and authorized sharing of other internal-use only data for the implementation of the Once-Only principle is also handled in an efficient and effective manner for the simplification of processes and reduction of bureaucracy.

Annex
Country Profile Malta

The current language data sharing infrastructure in Maltese public bodies looks as follows:



The goal is to implement an infrastructure for language data collection and sharing:



Language policy in Malta:

The language policy of Malta is encapsulated in Article 5 of the Constitution which states that the National language of Malta is the Maltese language. In addition, Maltese and English are currently the official languages of Malta and the Administration may for all official purposes use any of such languages. Any person may address the Administration in any of the official languages and the reply of the Administration shall be in such language. The article also stipulates that Parliament may prescribe another language as an official language by means of a law to that effect passed by not less than two-thirds of all the members of the House of Representatives.

The language of the Courts is the Maltese language, provided that Parliament may make such provision for the use of the English language in such cases and under such conditions as it may prescribe.[79] The House of Representatives may, in regulating its own procedure, determine the language or languages that shall be used in Parliamentary proceedings and records.

Besides the constitution, the enactment of the Maltese Language Act (Chapter 470 of the Laws of Malta) established the Maltese Language Council (http://www.kunsilltalmalti.gov.mt/) in April 2005. The Council aims to promote the National Language and provide the necessary means to achieve this aim. The Council was established to adopt and promote a suitable language policy and strategy for the Maltese islands and to verify their performance and observance in every sector of Maltese life, for the benefit and development of the National Language and the identity of the Maltese people. The Council appoints subcommittees in different areas of specialisation including information technology. The aim of each such subcommittee is to observe the situation of the Maltese Language, discuss all necessary measures for its development, and draft the linguistic policy for its respective field.

Maltese in Digital Environments:

In particular the Council's IT subcommittee has produced specifications for the digital use of Maltese including a keyboard layout and a Maltese Locale, as well as tools and apps that facilitate the use of Maltese characters on computers and smartphones. In 2016, the subcommittee published a consultation document entitled Digital Language Resources and Tools for the Languages of Malta: a Roadmap which proposes, inter alia, the establishment of a central repository of language resources and tools related to Maltese, as well as the other languages used in the Maltese islands.[80]

In the recently published document Malta, The Ultimate AI Launchpad, A Strategy and Vision for Artificial Intelligence in Malta 2030, it was noted that English has become the default language choice across most technological devices in Malta, simply because there are more resources available and they are easier to use.

Therefore,

> **to counteract this, the country will make crucial investments in the development of Maltese language resources and tools. The investment will enable computers to be able to process, understand and generate Maltese text and speech, and AI solutions to be developed and accessible in both of Malta's national languages and become a part of everyday life.**
>
> **The Maltese language resources and tools will also have a ripple effect on many sectors, including education and health. It will be possible to create more advanced software to process data in Maltese in a more efficient and accurate manner and to create education tools for the Maltese language that make use of these underlying language technologies.** [81]

[79] See the Judicial Proceedings (Use of English Language) Act, Chapter 189 of the Laws of Malta.

[80] Digital Language Resources and Tools for the Languages of Malta: a Roadmap http://www.kunsilltalmalti.gov.mt/file.aspx?f=309

[81] Cf. Parliamentary Secretariat for Financial Services, Digital Economy and Innovation, Office of the Prime Minister: *Malta: The Ultimate AI Launchpad, A Strategy and Vision for Artificial Intelligence in Malta 2030*, p. 48, 2019.

Annex
Country Profile Malta

Stakeholders:
The Ministries and the Office of the Prime Minister are currently being identified as the main stakeholders, both as contributors of language resources as well as potential eTranslation users, with interest being shown both for "individual to machine" use as well as "machine to machine" use (potential integration of eTranslation in public digital services). "Ministries" are to be understood as inclusive of the most granular level, thus covering specialised agencies, authorities and other entities falling thereunder, that come in contact with specific domain information and that could therefore populate the pool of resources with specialised corpora and terminologies.

The interest is tangible and very promising. The recently organised ELRC Malta seminar reached 140 registrations, 120 of which coming from the Maltese public administration. Most of the registered and attending public officials were high-ranking officials, including Permanent Secretaries, CIOs, Directors General and Directors.

Main challenges for sustainable data sharing:
For Malta, the following challenges for sustainable data sharing could be defined:

- **Lack of awareness of the importance of language resources:**
  The main challenge is to raise awareness and to put the message across that the documents produced by the public administration on a daily basis constitute the language resources that are required, which, like any other resource, should not be left dormant to be wasted but retained with the aim of maximising their value.

- **Rare re-use of translated data:**
  The current situation wherein a document, once translated, dies a natural death, should become a thing of the past. For the purpose of eTranslation, the life of that document is just about to start and may indeed be given an eternal existence if injected into the ELRC-SHARE pool of resources and used to train the engines and fine-tune translation results.

- **Lack of awareness concerning the benefits of data contribution:** Currently, potential data contributors are not aware of how they can benefit from eTranslation. Consequently, it is important to convey the right message: use eTranslation + donate to eTranslation: it's a win-win situation.

The time is also right, since AI is currently at the forefront of Malta's national digital policy, with two documents in this regard having been published during the month of October of this year (2019), i.e. The Ultimate AI Launchpad, A Strategy and Vision for Artificial Intelligence in Malta 2030 and "*Malta: Towards Trustworthy AI, Malta's Ethical AI Framework*. Further information about the documents can be found in the further reading list.

Action plan:
On the basis of the identified challenges, a number of actions could be defined:

- **To increase interest in MT in public services and ministries:**
  Based on the understanding that every public official is a potential eTranslation user and therefore a potential contributor of data, one-to-one sessions with ministries at all levels, including more granular levels (authorities, agencies, departments, units etc.) should be organised. During such meetings, eTranslation will be presented and a live demo will be given on the spot. In the past, this proved to be very effective, triggering off genuine interest and forming the basis for strong, lasting working relationships.

- **To tackle technical and legal issues:**
  Public administrations should be informed about the technical and legal support that ELRC may provide, including advice about data management generally. The plan is simple, but the message will be strong: "Data is power".

- **To update translation policy:**
  Further steps need to be taken, so that the Maltese government recognises the importance of language data. These efforts should result in an updated translation policy prescribing that any by-products of translations, e.g. translation memories and terminological equivalencies should be included amongst the deliverables of all translation contracts between third parties and the government.

References and further reading list:

Digital Language Resources and Tools for the Languages of Malta: a Roadmap
http://www.kunsilltalmalti.gov.mt/file.aspx?f=309.

Draft National Data Strategy: https://mita.gov.mt/en/nationaldatastrategy/Documents/Data-Driven%20Public%20Administration%20(Malta).pdf.

Judicial Proceedings (Use of English Language) Act, Chapter 189 of the Laws of Malta:
https://legislation.mt/eli/cap/189/eng/pdf.

Parliamentary Secretariat for Financial Services, Digital Economy and Innovation, Office of the Prime Minister: Malta: *The Ultimate AI Launchpad, A Strategy and Vision for Artificial Intelligence in Malta 2030*, 2019, https://malta.ai/wp-content/uploads/2019/10/Malta_The_Ultimate_AI_Launchpad_vFinal.pdf.

Parliamentary Secretariat for Financial Services, Digital Economy and Innovation, Office of the Prime Minister: *Malta: Towards Trustworthy AI, Malta's Ethical AI Framework*, 2019, https://malta.ai/wp-content/uploads/2019/10/Malta_Towards_Ethical_and_Trustworthy_AI_vFINAL.pdf.

Annex
Country Profile Norway

Kristine Eide, Jon Arild Olsen, Lilli Smal

State of Play:
Translation practices in public administrations in Norway:

In Norway, each public administration is responsible for adapting its information for different language users. There is therefore no central translation service or procurement contract and no systematic exchange of translations and/or knowledge between public administrations. Translations for public administration bodies are mainly carried out by procuring services from commercial translation companies or in an ad hoc manner and without use of CAT tools by the administrations' own employees. Only very few ministries and public administrations have in-house translation services generating Translation Memories. These include the Ministry for Foreign Affairs that have two translation units: one unit for the European Economic Area (EEA) and trade law and another unit for general translations, from whom the government and other ministries can also request translations. The only other public administrations that have in-house translation services are the Norwegian Maritime Authority, the Norwegian Public Roads Administration and the EFTA Secretariat in Brussels. Overall, these in-house translation services cover about 50 % of the translations carried out for public administrations on the national level. The remaining ca. 50% are outsourced to approximately five dominating language service providers.

Language data sharing:

As indicated above, most Norwegian public administrations do not exchange Translation Memories or expertise about translation procedures with each other. When it comes to sharing textual data with the European Commission with the purpose of improving eTranslation, it is important to remember that unlike the situation in EU-countries, the European Commission does not translate any texts to or from Norwegian. The collection and provision of Norwegian language resources for eTranslation hence is not a supplementary activity, but constitutes the only source for Norwegian language data in eTranslation. Currently, most Norwegian language resources in eTranslation have been provided by the law department in the Norwegian Ministry for Foreign Affairs. These data have been supplemented with translations of anonymised complaints provided by the European Consumer Center Norway. The Norwegian Maritime Authority and the Norwegian Public Roads Administration have also shared their TMs with the National Language Bank (Språkbanken) and ELRC.

The National Library of Norway has also concluded several agreements with commercial language services providers that led to the transfer of Translation Memories derived from public contracts. The agreements permit the reuse of Translation Memories and multilingual terminology lists on the condition that the memories are submitted to random scrambling in order to prevent automatic reconstitution of the translated texts. Additional language resources were generated based on parallel texts published on the web by various Norwegian public institutions that were transformed into Translation Memory files by ELRC.

Digital Policy and Language Policy:

Norway has two official languages, Norwegian and Sami. Norwegian exists in two written varieties, Bokmål and Nynorsk that must be treated as equal. Norwegian Bokmål is employed by the major part of the population, whereas Norwegian Nynorsk is used by approximately 12% of the population. To ensure equal treatment, both official language forms must be represented by at least 25% of publications published by state level agencies. Text books for primary and secondary schools must also be available in both language forms. The legal requirements for the use of Nynorsk in public administrations, however, only apply to published text. Consequently, the amount of Nynorsk in internal documents (e.g. translations) is significantly lower than for Norwegian Bokmål. Still, citizens can address public administrations in either variety and will receive an answer in the same language form they used.

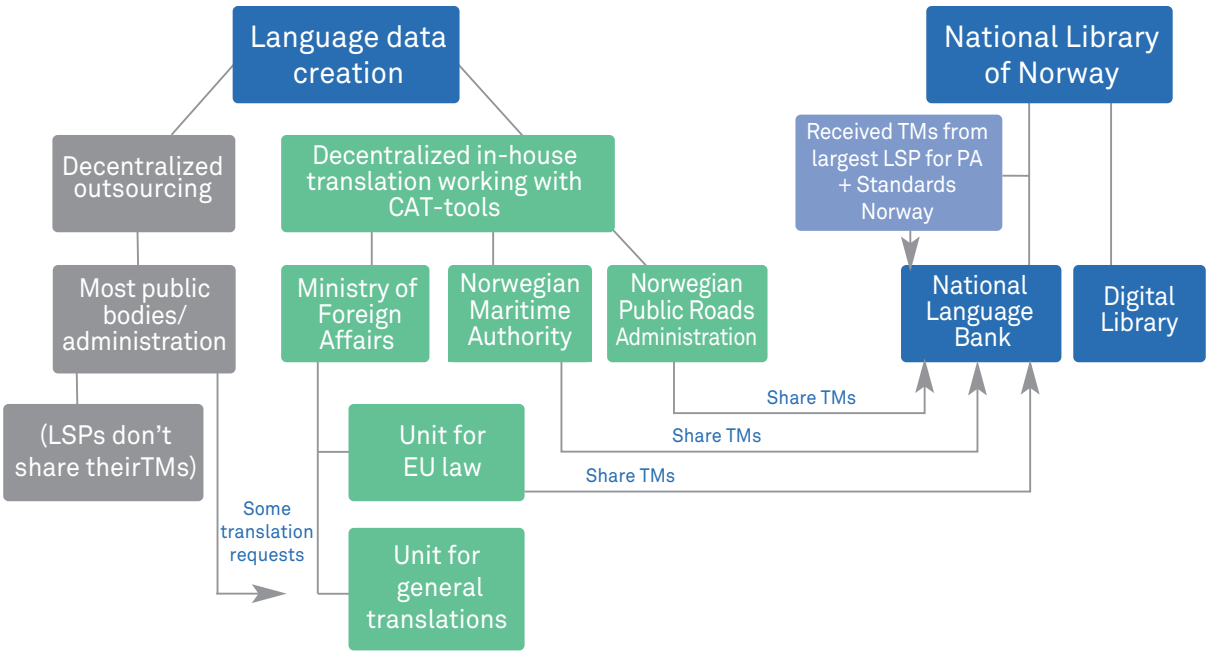
Interesting fact:
Language equality is applied to both written varieties of Norwegian. Norwegian Bokmål and Norwegian Nynorsk have to be treated as equal and therefore both forms must be represented in at least 25% of publications published by state level agencies.

In 2016, a report[82] was published stating the need for automated translation in the Norwegian public sector. Since then, the Norwegian Ministry of Culture, which is in charge of language policy, has decided that both official forms of written Norwegian must be available in eTranslation.

The Language Council of Norway, subordinated to the Ministry of Culture, oversees the Norwegian language policies. The tasks include ensuring that language technology, including machine translation, works for both varieties of the written language. There are ongoing conversations with the Agency for Public Management and eGovernment (Difi) on how to secure data for Nynorsk.

The National Library of Norway, also subordinated to the Ministry of Culture, hosts Språkbanken – the Norwegian Language Bank – an initiative to ensure the development of language technology solutions for the Norwegian language, thereby preventing domain loss of Norwegian in technology-dependent areas. Språkbanken offers digital language resources to the language technology industry, to linguistic research and education, and to public administration. Among the resources are bilingual corpora, mainly English-Bokmål and Bokmål-Nynorsk.

The current language data sharing infrastructure in Norwegian public bodies looks as follows:



Stakeholders:

The Norwegian National Anchor Points represent two key institutions (Norwegian National Library and the Language Council of Norway) related to language policy and language data collection, which underlines the interest and importance of this topic in Norway. This is exemplified by the fact that both commercial language service providers and national public administrations have already contributed language data to ELRC-SHARE. The local ELRC events were attended by representatives from more than 30 institutions showing that a significant number of national stakeholders were informed about the importance of collecting, managing and sharing language data to ensure language equality in Norway and beyond. The collection of language data is also strongly supported by the Agency for Public Management and eGovernment (Difi).

[82] Oslo Economics: *Kartlegging av behovet for automatisk oversettelse I statlig sektor/2016-15*, 2016.

Annex
Country Profile Norway

Main challenges for sustainable data sharing:

- One of the main challenges for using and sharing language data for Norwegian Nynorsk is the fact that there is simply not much language data available at this point.
- Additionally, even less data is available for parallel texts in Norwegian Nynorsk and English as "existing parallel corpora are made up almost exclusively of Norwegian Bokmål and English."
- A third challenge Norway is facing, applies to both varieties of Norwegian and results from the fact that language data "produced by public administrations, whether published online or for internal use, and more specifically the value of translations done internally or outsourced" are not considered valuable, worth managing, processing and sharing.
- "The need for awareness raising also applies to the privacy and confidentiality of documents sent out for external treatment, for example when such information is kept in the form of Translation Memories by external executives without the client being aware of this. Either the documents must be anonymised before dispatch, or the contract with external executives must ensure that Translation Memories are returned and/or deleted." [83]

Action plan:

For Norway, six objectives could be defined that will help to address the identified challenges. In the order of their priority these are:

- Increase the number of bilingual resources in English - Norwegian (Bokmål or Nynorsk), including terminology
- Collect/create a parallel corpus in Norwegian Bokmål and Nynorsk
- Raise awareness of language data as Open Data
- Increase interest in MT in public services
- Establish good data management practices in public services
- Tackle legal concerns

The most important objective for Norway is to increase the number of bilingual resources in both varieties of Norwegian and English in order to ensure equal treatment of both varieties of written Norwegian and to enhance the quality of translations from or into other European languages provided by eTranslation. In order to meet this challenge, the National Library has decided to collect a substantial parallel corpus of Bokmål and Nynorsk. The most important source of parallel texts in Norwegian Nynorsk and Norwegian Bokmål are textbooks and teacher manuals for Norwegian schools, which must be published in both official forms of written Norwegian according to Norwegian law. On the basis of these publications, the National Library of Norway aims to create a parallel corpus. Another important source of parallel texts in Nynorsk and Bokmål is the Nynorsk Press Office, which agreed to contribute a corpus of news texts translated from Norwegian Bokmål into Norwegian Nynorsk covering various subject areas. The Language Council of Norway is also in cooperation with the Brønnøysund Register Centre, where the aim is to collect and publish terminology for the public sector in both varieties of Norwegian, with translations into English where possible.

These ongoing collection tasks and negotiations have a direct impact on the perception of the value of language data. In addition, by opening up language data and making it freely available in the National Language Bank, the visibility of language data and its value is further supported. Next to dissemination activities, the collection and sharing of language data itself helps to raise awareness of language data as Open Data.

The Language Council of Norway works closely with the National Library to ensure that the language policies can be implemented successfully. It has started conversations with the Agency for Public Management and

[83] Cf. Olsen, Jon Arild: *ELRC workshop report for Norway,* 2019.

eGovernment on how it can gather different kinds of language data automatically from the public sector and thereby create sustainable infrastructures for sharing language data and at the same time increase the interest in machine translation in public services. These activities will improve the overall management of language data, including privacy and confidentiality of documents as well.

Further reading list:

Norwegian Ministry of Local Government and Modernisation: *Digital agenda for Norway in brief*, https://www.regjeringen.no/contentassets/07b212c03fee4d0a94234b101c5b8ef0/en-gb/pdfs/digital_agenda_for_norway_in_brief.pdf.

Olsen, Jon Arild: *ELRC workshop report for Norway*, 2019, http://lr-coordination.eu/sites/default/files/Norway/ELRC%2B%20Workshop%20Report_public.pdf.

Oslo Economics: *Kartlegging av behovet for automatisk oversettelse I statlig sektor/2016-15*, 2016, https://www.regjeringen.no/contentassets/61298b7ccab04fddb2c2b3bb9465cf38/automatisk_oversettelse_oe.pdf.

Annex
Country Profile Poland

Anna Kotarska, Maciej Ogrodniczuk, Andrea Lösch, Eileen Schnur

State of Play:
Translation practices in ministries and public administrations in Poland:

Polish ministries, public institutions and state-owned enterprises (such as the Industrial Development Agency, the Polish Press Agency etc.) translate their texts either by using their internal resources or by outsourcing the services to external language service providers (LSPs), with outsourcing being the prevailing trend. The estimated volume of outsourced translations amounts to 80 to 90% of the total volume. There is no central translation authority nor office to coordinate the related activities. As a result, there is no central terminology base nor organized management of translation memories (or other language resources).

Due to their specific requirements (e.g. confidentiality), some ministries, such as the Ministry of Foreign Affairs or the Polish Financial Supervision Authority (KNF) have small in-house units providing translation (2 to 7 people). There are also individual in-house translators working e.g. in the departments responsible for international cooperation or implementing international projects or delegated to foreign representation offices or embassies. Part of the translations hence is delivered internally by the staff having appropriate knowledge of the language(s); however, in the majority of the cases, they are specialists/experts in other fields and therefore have to rely on external LSPs. This policy is also due to the translation volumes and the need for highly specialized (legal, medical, technical) translations and/or certification of documents by a sworn translator.

> **Interesting fact:**
> **The Unit of Sworn Translators and Interpreters at the Ministry of Justice deals with formal licensing of translators and interpreters by e.g. arranging appropriate examinations and keeping a register of persons who have passed the state exam. This unit is also responsible for the recognition of professional qualifications acquired by sworn translators and interpreters in other Member States.**

The CEF eTranslation tool is also used, but its use is not widespread yet due to the low awareness of its existence and its potential applications among the public sector employees. The expected human translation quality is also an important factor.

The ministries as well as their subordinate units outsource translation services individually, which results from both budget regulations and potential limitations concerning the organisation and coordination of joint procurement procedures. Typically, the amount of a procurement contract for translation with a ministry varies between 0.4 million PLN to 0.7 million PLN with the highest ones reaching up to 1 million PLN[84]. The contracts are usually concluded for a period of one year and with a single LSP, which does not allow for developing good practices in cooperation between the contracting authority and the LSP, including e.g. terminology management.

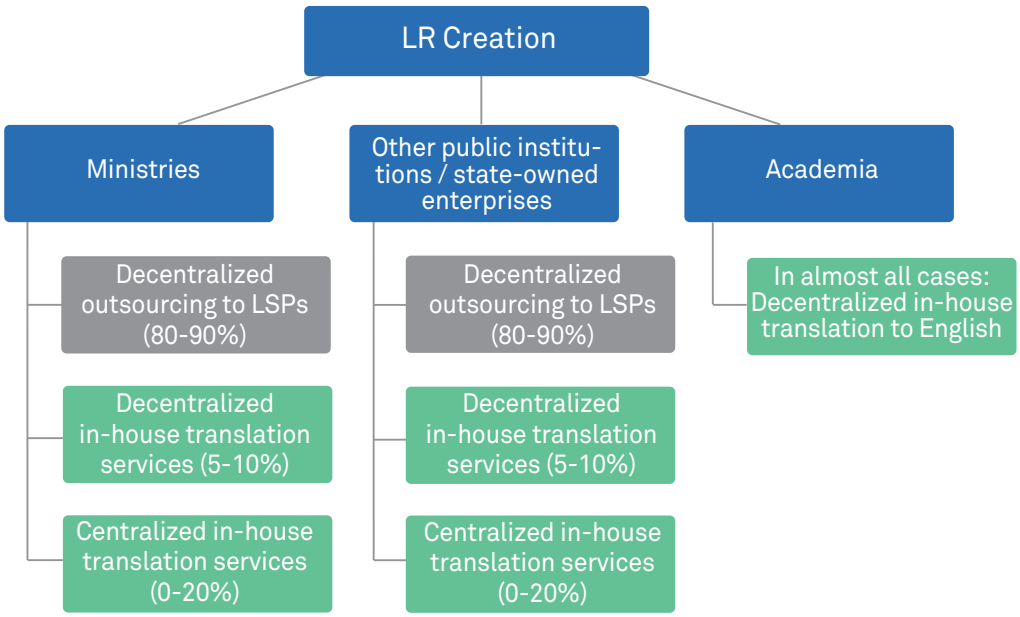
> **Interesting fact:**
> **The translation contracts most frequently include both translation and interpreting services, which is often problematic, as LSPs are typically specialized in providing either translation or interpretation services.**

The procurement process is often deficient as most emphasis is still put on the price criteria while not taking into account (or not taking into account in a sufficient degree) the quality criteria such as e.g. the experience of an LSP of providing translations in a particular subject area. The requirements regarding the translation industry norms, the use of computer-assisted translation (CAT) tools and the delivery of translation memories (TMs) to the contracting authority are therefore frequently omitted. It should be noted here that this is not necessarily a deliberate action, but often a consequence of the limited knowledge of the language service market and language technologies (LT) on behalf of the procurement units.

[84] 1 PLN equals 0.23 EUR (oanda.com, October 2019).

Recently, however, a more positive trend can be observed, including the involvement of the Public Procurement Office (Polish acronym: UZP) and the development of corresponding procurement practices: In January 2019, the UZP published the first set of documents describing good practices in proceedings for outsourcing translation services in its Knowledge Base (Industry Practices Forum). It was prepared by the Employers of Pomerania and written by the linguist-lawyer Wojciech Wołoszyk. A revision is planned with the support of the newly established Polish Association of LSPs POLOT. As a consequence, public entities tend to use non-price criteria more often and pay more attention to the actual expertise of the language services provider and guidelines prepared by industry associations, also with regard to interpreting services.[85] Technological awareness and the use of CAT tools also gradually improve thanks to various actions and assessments by e.g. the Ministry of Foreign Affairs, which currently assesses the suitability of various CAT tools for in-house use or the Polish Air Navigation Services Agency (Polska Agencja Żeglugi Powietrznej), which intends to procure software for computer-assisted translation[86].

The current language data creation infrastructure in Polish public bodies looks as follows:



Open Data in Poland:

The Ministry of Digital Affairs has prepared the Public Data Opening Programme (Open Data Programme) specifying data sharing standards adopted by resolution of the Council of Ministers on 20 September 2016, No. 107/2016. The implementation of the programme is coordinated by the Minister of Digital Affairs (MoDA). Further information can be found online at the Polish Open Data Portal: https://dane.gov.pl/. One of the goals of the Open Data Programme consists in building and coordinating a cooperation network, including institutional plenipotentiaries[87] for Open Data.

The tasks performed by the MoDA also include the development of standards for public data opening with regard to legal, security, technical and API issues, trainings and workshops for administrative staff on data opening, as well as knowledge dissemination. The new project Open Data Plus, which is the successor of the

[85] Leaflet by Polskie Stowarzyszenie Tłumaczy Konferencyjnych (PSTK).

[86] Tenders Electronic Daily (TED), Contract Notice: *Poland-Warsaw: Software package and information systems*, 2019.

[87] Plenipotentiaries are civil servants working in the Ministries, the units subordinated to ministries, the Chancellery of the Prime Minister, and Central Statistical Office appointed for permanent cooperation in the implementation of the Open Data Programme, responsible for the scope and deadlines of data provision by individual offices.

Annex
Country Profile Poland

Public Data Opening Programme, aims to e.g. build new APIs for a number of public databases, opening an analytical central Open Data Laboratory that supports the development of relevant policies in offices and ministries. Related educational activities are carried out by the Open Data Academy.

In July 2019, the Ministry of Digital Affairs also published a corresponding "Data Opening. Good practice guide". The good practice guidelines are part of the project "Open Data – access, standard, education", which aims to increase the availability and quality of Open Data and its reuse. The guide describes the basic framework for the process of opening data by referencing relevant legal acts, identifying desired institutional settings, and presenting practical scenarios for data opening in government offices. It focuses on covering the most important legal regulations affecting the opening and usage of public data and shows inter-institutional and non-institutional cooperation models that have worked best in this context. Furthermore, the guide shows how to implement the data opening process effectively and provides guidance on the standards for data openness. The MoDA organises study visits of the plenipotentiaries to other member states to exchange experiences, good practices and information on opening data of the public sector.

The MoDA also popularizes the idea of Open Data by being a partner and member of the jury of the largest hackathon in Europe, which took place in September 2019. In addition, the MoDA set a task related to Open Data monitoring. In particular, an application/a tool was to be created for the portal administrator that would identify products, services or applications that had used Open Data provided on the portal and that would show where such a corresponding product, service or application has been made available. Moreover, an information campaign is planned for winter 2019/20 to raise the awareness of public Open Data available at www.dane.gov.pl.

Language policy in Poland:

Following the Polish Language Act of 1999 of 7 October 1999, the Polish language is the only official language in the territory of the Republic of Poland. Generally, the provisions of the Act shall apply to the protection of the Polish language, the use of the Polish language in implementation of public tasks and the use of the Polish language in the course of trade and implementation of the provisions on the use of the scope of labour law (in the territory of the Republic of Poland). There is a corresponding Council for the Polish Language at the Presidium of the Polish Academy of Sciences. The Council's main task is to provide valuations and assessments on all matters concerning the use of the Polish language in public communication.

The issue of regional languages is regulated on the EU level by the provisions of the European Charter for Regional or Minority Languages ratified by Poland in 2009. In Poland, the Kashubian ethnolect enjoys the status of the regional language, which is regulated by the Act of 6 January 2005 on National and Ethnic Minorities and Regional Language. In some municipalities of the Kashubian region, officials are legally obliged to respond to the letters of the interested parties in the respective ethnolect, if the interested parties wish to do so. There are also repeated efforts to recognize such status for Silesian.

Stakeholders:

Within ELRC, more than 80 potential stakeholders that are involved in the creation or sharing of language resources (LR), related activities and/or policy setting were identified, including in particular LR holders and creators (public bodies as well as language service providers). Thirty of these stakeholders participated in the last ELRC Workshop.

In addition to certified and specialized translators (via LSPs/translation agencies), major providers of language resources in Poland include, for instance, the Chancellery of the Prime Minister, Ministry of Foreign Affairs, the Ministry of Justice, the Ministry of Culture and National Heritage, the Polish Press Agency, the Ministry of Entrepreneurship and Technology, the Ministry of Health, and the Polish National Bank. These institutions could also be considered as the main beneficiaries of the use of eTranslation. So far, with the support of ELRC, more than 20 data sets have been collected, mostly bilingual Polish-English corpora, but also two terminological resources and the monolingual Polish Court Rulings Corpus. The data covered 10 domains, 2M words in the multilingual part and 178M words in the monolingual corpus.

Main challenges for sustainable data sharing:

ELRC identified several challenges and issues that need to be overcome to enable sustainable data sharing, namely:

- General lack of transparency among public institutions related to opening their language data for further reuse.
- The lack of a public-private sector initiative for developing guidelines and good practices for contracting translation services in the public sector to be approved by the Public Procurement Office.
- The shortcomings of the procurement process (in particular the lack of the requirement to deliver the full rights and TMs to the contracting authorities).
- Individual procurement of translation services by each public entity, usually for a short period of time (one year).
- The lack of awareness-raising actions on the purpose of ELRC in relevant public sector events and focused meetings.
- The lack of technological expertise in the public sector especially with regard to the use of CAT tools. As a result, sharing of language resources by public entities is limited to non-TMX formats in almost all cases.
- Technical problems concerning data delivery and the public institutions' lack of awareness on how ELRC can help to solve such issues.

Action plan:

Based on the identified challenges, the following objectives could be defined for Poland. In the order of their priority, these are:

- **Raising awareness of language data as Open Data and a valuable asset. This is to be achieved with the help of the following actions:**
  - Raising the interest in language data among the members of the relevant public bodies and sharing benefits of sharing language data. This is an ongoing activity, since the Polish PS NAP held more than ten public presentations/presentations at ministries and public bodies in the past year. Further presentations and publications are already planned[88]. In addition, more workshops should be conducted (targeting also respective IT and procurement personnel).
  - More OSA (on-site assistance) cases funded under successive tenders as well as greater support for the promotional work and communication activities of the NAPs.
  - Reaching out to the MoDA and integrating language data in the national Open Data policy, digital agenda etc.
  - Proposing to the MoDA to appoint a plenipotentiary for open language data and an eTranslation national contact point at the ministry.
  - Establishing practical guidelines for LRs as Open Data. In this context, the Ministry of Digital Affairs has established a corresponding Open Data Programme and Good practice guide.
  - Emphasizing the role of digital texts in the digital economy (data as the main source for Artificial Intelligence), also illustrating the value and application of Natural Language Processing tools (e.g. for preventing online violence, fake news/disinformation).
  - Emphasizing the role of digital texts as part of national cultural heritage.

- **Increasing interest in MT/LT in public services as part of the national digital policy by:**
  - Establishing synergies with national projects/initiatives
  - Promoting the eTranslation API
  - Providing best practices and use cases of MT applications in public administrations in other EU countries, particularly in trans-border DSIs like EESSI
  - Educating about LT and CAT tools and their benefits
  - Co-organizing Info Days of LT projects like NEC TM Data

[88] Including, for instance, Translating Europe Workshops in cooperation with the DGT, an article in a professional journal, a presentation at a conference for IT professionals etc.

Annex
Country Profile Poland

  - Organising regular communication activities (e.g. a newsletter)
  - Educating NAPs (enabling them to participate in related events of educational value, such as META FORUM or other EU-funded events related to eTranslation, such as the CEF AT Tools and Services workshops and conferences)
  - Cooperating with the National CEF Contact Point at the MoDA

- **Tackling legal concerns by:**
  - Developing and sharing easy-to-apply guidelines for IPR and privacy issues
  - Investigating an idea to implement rights management along with data management (in collaboration with the MoDA)
  - Providing clear anonymisation guidelines
  - Informing and involving the Personal Data Protection Office in the above actions
  - Including presentations on IPR and privacy issues in the agenda of national Translating Europe workshops and other events for translators

- **Identifying and gaining access to outsourced translations by:**
  - Establishing cooperation between translators/translation units and public procurement units to include relevant clauses for retaining TM and corresponding rights
  - Promoting clear licensing guidelines
  - Promoting clear recommendations on procurement of translation services by public sector bodies in cooperation with the Public Procurement Office.
  - Obtaining financial support (e.g. more OSA cases or Generic Services projects in this area as well as financial support of communication activities such as writing articles or newsletters)

- **Establishing good data management practices in public services by:**
  - Extending the scope of available public Open Data categories to language data (responsible: Minister of Digital Affairs)
  - Involving the plenipotentiaries in activities supporting their sharing of language data on behalf of their mother institutions
  - Identification of data managers
  - Investigation of data management practices
  - Establishing Data Management Plans based on ELRC findings
  - Establishing cooperation with COTSOES on the national level

References and further reading list:

Act on National and Ethnic Minorities and Regional Language, 2005, http://prawo.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20170000823.

Complete list of plenipotentiaries: https://www.gov.pl/web/cyfryzacja/pelnomicnicy-ds-otwartosci-danych.

Complete list of state-owned enterprises: https://nadzor.kprm.gov.pl/spolki-z-udzialem-skarbu-panstwa.

HackYeah: https://hackyeah.pl/.

Information Campaign to raise awareness on public Open Data: https://www.wirtualnemedia.pl/artykul/resort-cyfryzacji-chce-popularyzowac-w-zakresie-otwartych-danych-publicznych.

Leaflet by Polskie Stowarzyszenie Tłumaczy Konferencyjnych (PSTK), 2019, http://pstk.org.pl/wp-content/uploads/2019/01/PSTK-Ulotka-A4.pdf.

Open Data Practice Guide: https://dane.gov.pl/media/ckeditor/2019/07/04/open-data-good-practice-guide.pdf.

Open Data Programme, Ministry of Digital Affairs: https://www.dane.gov.pl/media/resources/20171201/Program-EN.doc.

Polish Association of Language Service Providers: http://www.polot.org.pl/.

Polish Language Act of 1999: http://prawo.sejm.gov.pl/isap.nsf/download.xsp/WDU19990900999/U/D19990999Lj.pdf.

Polish Open Data Portal: https://dane.gov.pl/.

Polish Personal Data Protection Office: https://www.uodo.gov.pl/

Project “Open Data Plus”: https://www.gov.pl/web/cyfryzacja/otwarte-dane-plus.

Project ”Open Data – access, standard, education”: https://www.gov.pl/web/cyfryzacja/otwarte-dane-dostep-standard-edukacja2.

Regulations of the Polish Language Council: http://www.rjp.pan.pl/index.php?option=com_content&view=article&id=194&catid=40&Itemid=73.

Tenders Electronic Daily (TED), Contract Notice: *Poland-Warsaw: Software package and information systems*, 2019, https://ted.europa.eu/TED/notice/udl?uri=TED:NOTICE:496078-2019:TEXT:EN:HTML&src=0.

Wojciech Wołoszyk: *First set of documents describing good practices in proceedings for outsourcing translation services* (published by UZP, prepared by Employers of Pomerania), https://www.uzp.gov.pl/baza-wiedzy/dobre-praktyki/forum-praktyk-branzowych/uslugi-biznesowe-prawnicze,-marketingowe,-konsultingowe,-rekrutacji,-drukowania-i-zabezpieczania/dobre-praktyki-w-postepowaniach-na-uslugi-tlumaczen-pisemnych-pracodawcy-pomorza.

Annex
Country Profile Portugal

Paulo Vale, António Branco, Eileen Schnur

State of Play:
Translation practices in ministries and public administrations in Portugal:

In Portugal, most translations are outsourced independently by public administrations, but there are also institutions with small in-house translation services. Given the specific areas of activity in each institution, in many cases, public authorities tend to hire the same companies or freelancers in order to maintain consistency between translations. The efforts of recent governments to reduce public spending are reflected in the reduced number of public officials and increasing difficulties in hiring new staff. Consequently, it is often easier for public administrations to contract services than to create a permanent team of translators.
If the documents are translated in-house, computer-assisted translation (CAT) tools are only rarely used in public administrations. Contrary to that, the use of CAT tools is an important part of the work of language service providers (LSPs) and freelance translators. As regards language technology (LT), the META-NET white paper series stated that despite steady improvements, there is still no or only weak language technology support for Portuguese.

Some public administrations are using free translation services that are available online, but translation practices vary from institution to institution. In the Portuguese Republic Assembly for example, documents are frequently translated using computer-assisted translation software. In addition, terminological databases including Portuguese, English and French terms are created.[89] According to representatives of the Portuguese Republic Assembly and the Ministry of Foreign Affairs, their institutions frequently use the European Commission's machine translation (MT) system eTranslation (formerly known as MT@EC) instead of other online translation services. Contrary to that, the Portuguese Institute of Registration and Notary Affairs have their texts automatically translated by a private system that was obtained by the institution. The translations are revised and post-edited by the employees afterwards.

Due to the above-mentioned restrictions in hiring public servants, the reduction of translation costs is not always a strong argument for those managing translations in the public services. Political changes, staff instability and other priorities in terms of digital transformation make it increasingly difficult to establish a national strategy for procuring and outsourcing translations.

In Portugal, public procurement data is available on the BASE Portal. The portal gathers all relevant information on public procurement in Portugal and makes it available to citizens in an open and transparent way. Further information is available on www.base.gov.pt/Base/en/Homepage.

> **Interesting fact:**
> **The Portuguese public procurement portal is called "BASE" and collects information about all contracts concluded under the Public Contracts Code (PCC).**

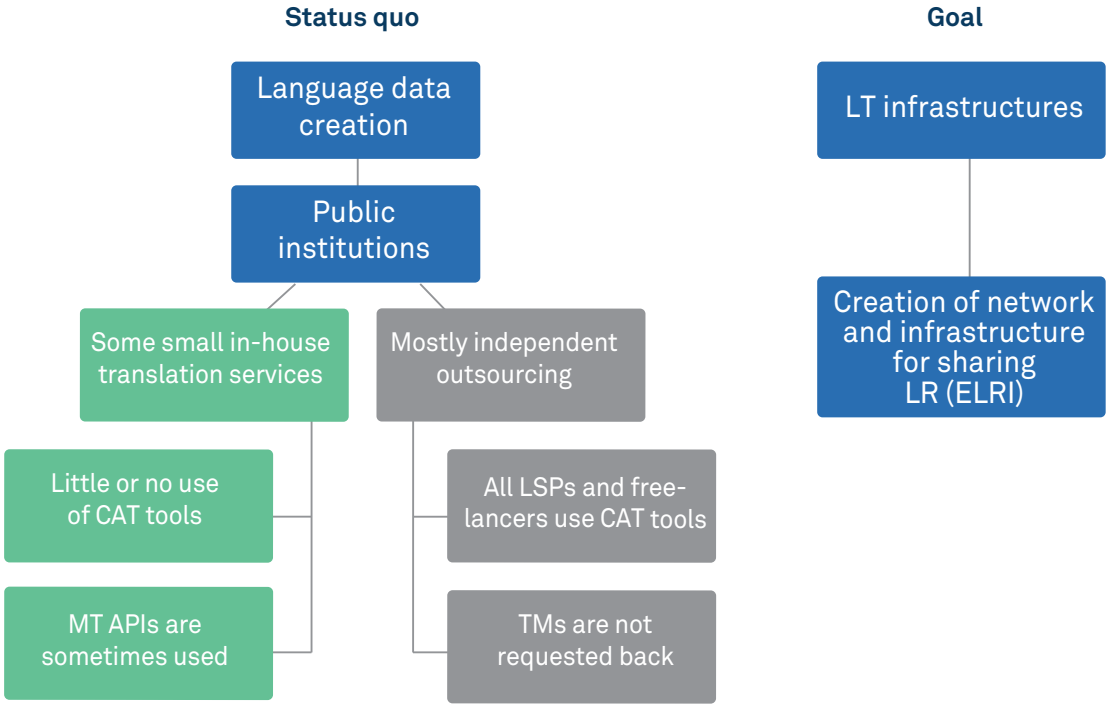
When it comes to outsourcing translations, according to the NEC TM Country Reports, there was a significant increase of awarded contracts for translation and interpretation services from 2015 (63 awarded contracts) to 2018 (110 awarded contracts) in the public sector. This could be an indicator of the increasing importance of multilingual content in Portuguese public administrations. The report also states that the Portuguese market is very fragmented and that contracts are awarded to Portuguese companies, but also to independent freelance translators. In the majority of the cases, the documents' source language was Portuguese, which had to be translated into English, followed by Spanish. The areas for which most translation services were requested include Education and Communication, Social Questions, but also Finance and International Relations.

In Portugal, there is currently a low utilization rate of translation support technology, resulting in the perception that there is no obvious benefit in adding a clause to new contracts, requiring the delivery of translation memories along with the translated work. The almost non-existent use of CAT tools also contributes

[89] Querido, Carvalho: *ELRC Workshop Report Portugal*, 2016.

to the fact that even within each organisation, there is no common practice for sharing resources. Translated work is considered relevant only for the purpose for which it was produced, but the potential benefits of its reuse are currently not taken into account.

The current language data creation infrastructure in Portuguese public bodies looks as follows:



Data sharing infrastructures and Open Data in Portugal:

The Portuguese Republic shares reusable data under the 2013/37U Public Sector Information (PSI) directive, which has been transposed into the Portuguese legislation by the so-called "Lei de Acesso aos Documentos Administrativos" (LADA, Lei n.º 26/2016 de 22 de Agosto) in 2016.

In Portugal, there is a central Open Data Portal available, which is called "Dados.gov" ("Dados" is the Portuguese equivalent to "data"). It is managed by the Administrative Modernisation Agency (AMA), the public institute responsible for the promotion and development of administrative modernization and digital transformation in Portugal. The Dados.gov portal gathers and harvests data from different sources among the Portuguese public sector[90] and contains more than 2163 data sets and 5714 resources. Although this is a significant number of available data, in 2018, there were no linguistic datasets included.[91] In general, the number of reuses and users is rather small and there is only a limited number of organisations represented. However, when comparing the statistics of 2018 and 2019, it becomes evident that the portal and its usage are growing. Whereas in 2018, there were 91 users and 60 organisations, according to the Dados.gov dashboard, there are now already 651 users and 84 organisations.

When it comes to sharing Open Data, the reuse license is probably one of the essential conditions. The license which is used by default and which is of recommended use at the dados.gov portal is the Creative Commons Attribution 4.0 – CC BY 4.0. In some public administrations, such as the Lisbon municipality, the Creative

[90] European Commission: *eGovernment in Portugal*, 2018.
[91] Branco, Oliveira: *ELRC Workshop Report for Portugal*, 2018.

Annex
Country Profile Portugal

Commons Zero (CC 0), which is the most permissive CC license, is used. Data sets with a CC 0 license can be used as if they were public domain.

With respect to Language Technology (LT) and ELRC, AMA aims to support citizens by facilitating translations and by setting up a national repository of digital services available for all services of the public administration. As an outcome of the ELRI project, AMA made available the National Repository for Translation Resources, known as eTradução, where language resources are collected, prepared and shared between public institutions and research centres. This web platform has been active since May 6, 2019 and currently, there are 38 registered users representing 14 institutions of the Portuguese public administration. The registered users have already contributed 19 resources (more than 20.000 translation units), several of which have been fully validated, published and uploaded to ELRC-SHARE repository. Around 25% of eTradução resources were uploaded as Open Data and AMA plans to make these resources available at the national Open Data portal aiming to raise awareness to the importance of language datasets. The eTradução initiative represents a milestone by becoming the first initiative capable of creating a cooperation and sharing environment related to language resources in Portugal.

Portugal is represented in 18 EU-funded projects, which cover the three building blocks eID and eSignature, eInvoicing and eTranslation. Three projects are related to Open Data, i.e. the Open Waste Compliance, the Cross-Nature project and the Urban Co-creation Data Lab project. The latter started in October 2019 and aims to develop a new generation of public services in the context of smart cities exploiting supercomputing facilities and public and private data to analyse complex combinations of large datasets in areas of public interest. In addition, the country is actively involved in two initiatives of the eTranslation building block, since AMA and Lisbon University are not only part of the ELRI consortium, but also represented by the National Anchor Points for Portugal in ELRC.[92]

Language policy and digital policy in Portugal:

Portuguese is the official language throughout the country and also the official language in nine other countries globally. With 10 million speakers in Portugal and around 220 million speakers in total, it is the third most spoken European language in the world.[93]

In 2017, the Portuguese government approved the ICT2020 Strategy, which is also known as the Portuguese Digital Transformation Strategy. It aims to facilitate the cooperation between public administrations and focuses on the creation of new eGovernment services and the reduction of public sector costs.

The strategy is built on three main pillars[94]:
- Promotion of integration and interoperability;
- Innovation and competitiveness;
- Resource sharing and investment in digital competences.

In 2018, Portugal also launched two policy initiatives on digital competences and digitisation of the economy. One of them is INCoDe.2030, the National Initiative on Digital Competences, which aims to enhance and foster digital competences by educating young people and requalifying available human resources.
In the same year, the so-called “Indústria 4.0” was launched, which focuses on the development of industry in the digital area.[95]

Portugal has made significant progress over the past years in the field of digital public services, but there is still room for improvement when it comes to Open Data[96] and the availability of language resources. There

92 Branco, Oliveira: *ELRC Workshop Report for Portugal,* 2018.
93 Branco et. al.: *The Portuguese Language in the Digital Age*, in: *META-NET White Paper Series*, 2012.
94 Hillenius, Gijs: *Modernisation the focus of new ICT strategy Portugal*, 2017.
95 European Commission: *Digital Transformation Monitor, Country: Portugal “Indústria 4.0”*, 2017.
96 European Commission: *Digital Economy and Society Index (DESI) 2018 Country Report Portugal,* p. 12, 2018.

are already several digital services available like automatic tax declaration or electronic authentication through the public administration web portal, but the provision of multilingual digital services is not a common practice yet.[97]

Stakeholders:

The ELRC National Anchor Points for Portugal represent two relevant stakeholders, namely the Administrative Modernisation Agency (AMA) and the University of Lisbon. Both institutions are also involved in the ELRI project, which aims to provide an infrastructure to help collect, prepare and share language resources, which can improve translation services. The collected data can be used locally, thus supporting the data collection activities of ELRC.[98] Another important stakeholder is the Portuguese Parliament, since it has already contributed a significant amount of language data to ELRC.

ELRC events like the local workshops and the annual conferences were attended by more than 20 different institutions from Portugal, which clearly demonstrates that there is an interest in the topics dealt with by ELRC.

Main challenges for sustainable data sharing:

- **Lack of awareness and information flow:** According to a representative from the Ministry of Justice, a compelling message on eTranslation needs to be conveyed to be able to convince public administrations to share their data and to use the eTranslation services. Currently, potential mutualisation is hindered by the lack of information flow among the entities that would have the required competences. There is generally a lot of willingness to participate, but impediments remain.

- **Lack of professionals:** Whereas Open Data initiatives are usually perceived positively, the biggest barrier that prevents Portuguese administrations from sharing their data is the lack of skilled staff (i.e. computer scientists), who are capable of identifying, preparing and uploading data on the CEF platform.[99]
- **Financial issues:** Preparing the data for eTranslation does not only require skilled personnel, but also financial resources, which hinders Portuguese institutions from sharing their data.

- **Lack of available resources to train MT systems:** More language data will be required to achieve high-quality machine translation output. At the same time, dissatisfying translation results may lead to growing scepticism among Portuguese public institutions.

- **Political changes:** Political changes can complicate the establishment of a national strategy for procuring and outsourcing translations.

- **Legal concerns:** Many Portuguese public administrations are concerned about the protection of personal information. They fear that it might still be possible to identify people even after anonymisation.

Action plan:

In order to tackle the identified challenges mentioned above, five objectives could be defined. In the order of their priority, these are:

- **To raise awareness of language data as Open Data and valuable asset:**
  As mentioned above, public administrations and ministries are currently not aware of the value of language data. Therefore, it would be important to emphasize the role of digital texts in the digital economy and to clearly illustrate the benefits of sharing and reusing language resources.

- **To increase interest in MT/LT in public services as part of the national digital policy:**
  Since MT and LT are currently not in the focus of Portugal's digital transformation strategy, it would be

[97] Branco, Oliviera: *ELRC Workshop Report,* p.6, 2018.
[98] Branco, Oliviera: *ELRC Workshop Report,* p.9, 2018.
[99] Branco, Oliviera: *ELRC Workshop Report,* p.7, 2018.

important to further promote the use of machine translation and to demonstrate how public administrations can benefit from MT/LT in their daily operations. Furthermore, additional information about MT should be shared, including details about e.g. the anonymisation process or the amount of data required to improve an MT system.

Another step towards increasing the interest in the topics would be the collaboration with national projects and initiatives on the one hand, and the support of key decision makers on the other.

- **To establish good data management practices in public services:**
  As mentioned above, there is currently no common practice for data management in Portuguese public administrations and ministries, because the value of re-using data has not been recognised yet. However, the goal of establishing common data management practices could be achieved by appointing a data manager, who takes care of the data management in the respective ministry or administration. In addition, it would be important to define confidential and personal data and to clearly differentiate between confidential/personal data and public sector information.

- **To tackle legal concerns:**
  As already indicated, legal issues such as e.g. concerns about anonymisation and the protection of personal data may prevent public services from sharing data. Easy-to-apply guidelines for IPR and privacy issues could support potential contributors in overcoming such challenges. This could also be facilitated by investigating ideas to implement rights management along with the above-mentioned data management.

- **To identify and gain access to outsourced translations:**
  As the majority of the translations are currently outsourced by Portuguese administrations, identifying and accessing them may lead to an increase in available language data. The outsourced translations could be identified with the help of the NEC TM Data Project and by establishing a common practice of receiving any by-product of outsourced translations back from the LSPs and/or freelancers.

References and further reading list:

Administrative Modernization Agency (AMA): https://www.ama.gov.pt/web/english.

Branco et. al.:*The Portuguese Language in the Digital Age*, in:
META-NET White Paper Series, 2012, http://www.meta-net.eu/whitepapers/e-book/portuguese.pdf.

Branco, Oliveira: *ELRC Workshop Report for Portugal,* 2018, http://www.lr-coordination.eu/sites/default/files/Portugal/ELRC%2B2%20Workshop_Public_Portugal.pdf.

eTradução: https://etraducao.gov.pt/pt-pt/.

European Commission: *Digital Economy and Society Index (DESI) 2018 Country Report Portugal*, 2018, http://ec.europa.eu/information_society/newsroom/image/document/2018-20/pt-desi_2018-country_profile_eng_B440E073-A50F-CF68-82F6A8FB53D31DE5_52232.pdf.

European Commission: *Digital Transformation Monitor, Country: Portugal "Indústria 4.0"*, 2017, https://ec.europa.eu/growth/tools-databases/dem/monitor/sites/default/files/DTM_Ind%C3%BAstria%204.pdf.

European Commission: *eGovernment in Portugal,* 2018, https://joinup.ec.europa.eu/sites/default/files/inline-files/eGovernment_in_Portugal_2018_0.pdf.

Guides to Implementation of the (Revised) PSI Directive in Portal: https://www.europeandataportal.eu/sites/default/files/country_portugal.pdf.

Hillenius, Gijs: Modernisation the focus of new ICT strategy Portugal, 2017, https://joinup.ec.europa.eu/collection/egovernment/news/modernisation-focus-ne.

ICT strategy 2020: https://tic.gov.pt/pt/web/tic/-/estrategia-tic-2020.

INCoDe.2030: https://www.incode2030.gov.pt/en/incode2030.

Portuguese Open Data Portal “Dados.gov”: https://dados.gov.pt/en/dashboard/.

Querido, Carvalho: *ELRC Workshop Report Portugal*, 2016, http://www.lr-coordination.eu/sites/default/files/Portugal/WorkshopELRCPortugal_PublicReport.pdf.

Annex
Country Profile Romania

Laura Mihăilescu, Dan Tufiș, Lilli Smal

State of Play:
Translation practices in ministries and public administrations in Romania:

In Romania, there are different scenarios how translation needs are met in the public administration. In general, translations are regarded as a secondary activity and most translations are outsourced when a need arises. The Ministry of Culture, the Superior Council of Magistracy and the Ministry of Justice for example outsource most of the needed translation to language service providers.

When other public administrations need translations, they are either done in-house by whoever knows a foreign language and without CAT tools, or they are outsourced. Only very few institutions in Romania have dedicated translation departments or employees whose main task is to translate, these institutions are: the European Institute of Romania – EIR – (4 translators, 3 legal revisers, 1 terminologist), the National Bank of Romania (7 translators), the Constitutional Court and the Romanian Standards Association (ASRO). The most common language combination for in-house translations is Romanian <> English and sometimes Romanian <> French (French is mainly used for the legal field). Although the EIR operates under the coordination of the Ministry of Foreign Affairs, translations in the ministry itself are mainly outsourced because of the volume and the variety of languages needed. At the ministerial level, there is no administration with their own in-house translation service.

Most translations are done from English and French into Romanian, and from Romanian into English and Hungarian.

Public procurement and the use of CAT tools:

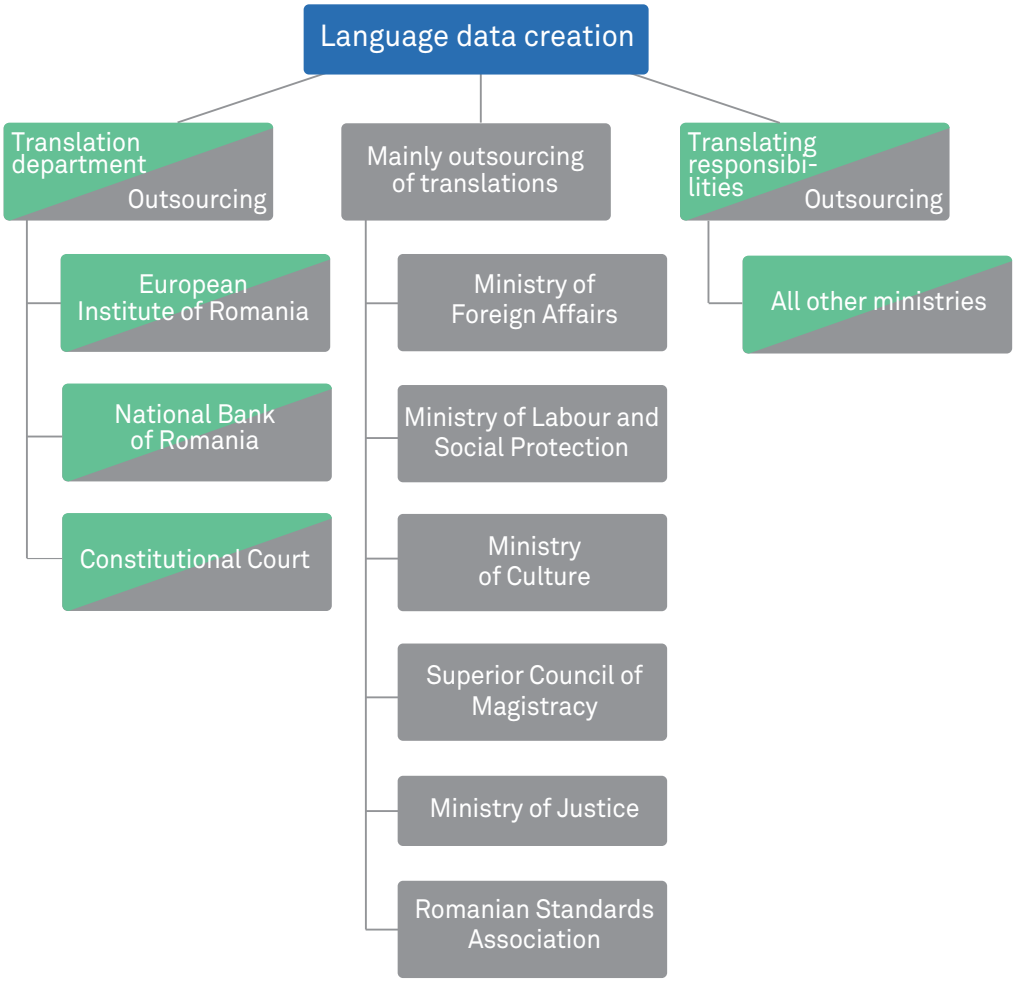
Public institutions that outsource translations were obliged by law to search the electronic public procurement system first (SICAP: http://sicap-prod.e-licitatie.ro/pub). If public institutions could prove that they could not find a suitable offer or if the offers were more expensive on SICAP than on the open market, they were free to choose any LSP on the market. However, this is now no longer an obligation but a recommendation.

A regulation that is still in effect is a fix price for the translation of one page in the legal field. According to Order no 772/2009, the Ministry of Justice and other institutions operating in the legal field pay a certified translator with 33,56 lei per page (that is approx. 6,71 euro) – at this price, the translation is usually accepted by inexperienced language service providers and the result usually lacks in quality. Even when the translator can provide a certificate for translating legal texts, this is not necessarily a seal for quality because obtaining certificates for translating legal texts is a formality and does not require specific legal training. The certificates can be obtained upon request to the Ministry of Justice by any person who graduated from any language university regardless if the respective student has followed any course for translation of legal texts. In addition, by taking an exam organised by the Ministry of Culture, a person without a university diploma in foreign languages may obtain such a certificate. This causes a serious issue when it comes to the translation quality of legal texts.

> **Interesting fact:**
> **It is not yet standard practice for language service providers to use CAT tools. Usually, only LSPs which work for foreign clients invest in CAT tools following suit to clients' requests.**

Another challenge arises from the fact that the majority of those who request translations (public institutions or not) usually do not request additional services like revision/review, terminology lists, glossaries or translation memories. The awarding of contracts is based on the number of pages, pair of languages, sometimes the domain, the deadline and certificates of the translators. CAT tools are considered to be very expensive and sometimes too complex. Therefore, only very few LSPs can afford them. LSPs that work with European institutions or foreign clients however, usually do use CAT tools. CAT tools are also used by the translators in the in-house translation services mentioned above, whereas only the European Institute of Romania uses a server-based translation memory system.

The current language data sharing infrastructure in Romania public bodies looks as follows:



In Romania, there is no proactive exchange of translation memories, terminology or expertise on the national or inter-ministerial level. During the translation of the acquis, a terminology network was put in place and although it was considered very useful, it proved to be difficult to manage for a number of reasons. The terminology experts in the various ministries were not paid for this activity and therefore could only allocate limited time to this task. In addition, a high fluctuation of human resources in the ministries and their constant reorganising led to information loss. Another initiative that was started by the Romanian Language Department in the DGT was the RO+Network, a Linguistic Network of Excellence for Institutional Romanian that allowed for information exchange between the DGT and experts of the Romanian language regarding linguistic or terminological questions. The experts provided advice pro bono but the network is no longer active.

Currently, there are no concrete plans for the organisation of a terminology network on the national level.

Open Data in Romania:

Most people in public institutions do not consider or do not know that language resources can also be Open Data. Hence there is no bi- or multilingual data sets in TMX format provided by public institutions on the national Open Data Portal (http://data.gov.ro/). Language data is also not sought after as only numerical data is considered useful, especially for decision makers.

Annex
Country Profile Romania

Language policy in Romania:

Romania's official language is Romanian, which belongs to the group of Romance languages. It is spoken by about 25 million people in Romania and abroad. The most spoken minority languages are Hungarian, German and Romany although 20 languages have minority status in Romania. Education is also provided in the languages of the minorities and learning foreign languages is included in the compulsory school curriculum. According to Law no 500/2004, the Romanian language is to be used in all official documents. The law does not address the use or planned use of language technologies to protect and support the Romanian language in the digital age. Some other provisions state that any technical manual or instructions regarding the use of a foreign product must also be translated into Romanian and that all TV productions in a foreign language must be subtitled into Romanian.

Stakeholders:

Although the use of machine translation and computer aided translation tools are not a common practice in Romanian public administration and in the public sector yet, an increased interest in past ELRC events was shown. The second ELRC workshop was attended by almost 100 participants representing various ministries and language service providers. Among the data donors of multilingual language data to ELRC are the Romanian Parliament, the European Institute of Romania, and RACAI. The latter two institutions are represented by the ELRC National Anchor Points and are critical institutions for language data collection in Romania.

Main challenges for sustainable data sharing:

- Legislation regarding the use of the Romanian language is not observed by all public institutions (according to Law no 500/2004, public institutions have to use the Romanian language in all official documents, that is with special characters (ă, â, î, ș, ț) and observing the directives of the Romanian Academy ("sunt/sînt", writing with "â/î ").

- There is a general tendency to disregard the quality of the Romanian language for a number of reasons (e.g. the message is considered more important, lack of time, low speed when typing with Romanian special characters), which also affects the quality of the Romanian language on the internet. This low quality also affects the collection of language resources in various ways. Ignoring the official orthography, i.e. not using diacritics, or using non-standard ones (or combining these practices) can result in a different meaning which turns textual documents into low-quality data that is not accurately representing the Romanian language and is thus avoided by data collectors.

- **Educational issue:** Poor use of CAT tools (they are considered to be very expensive and more of a luxury) and of computers (still not a norm to use Spell check, Track Changes, advanced text formatting); Proper (education) on data management is a major challenge

- **Quality issue:** It is still not a standard for public institutions to ask LSPs to also provide revision/review for their translations and to return the translation memories (the decision criterion for contracting LSPs is usually the lowest price not quality).

- **Financial issue:** CAT tools and the respective training are very expensive for the public sector.

- I**nteroperability issues:** For example, Romanian characters were not initially supported by CAT tools.

- **Fundamental issue:** Language data is not considered a valuable asset and is not managed adequately.

- **Continuity issue:** Decision makers change frequently, however, proposed changes must be top-down, creating an even bigger challenge.

- **Political leadership:** The Secretariat General of the Government could/should lead the reform consequently, as the institutional level is not very relevant.

Action plan:

To support the Romanian language, to improve the translation workflow and to make data sharing in the future easier, the following actions are recommended:

- **Provide comprehensive access to CAT tools:**
  This objective addresses the need to raise awareness of the productivity gain through CAT tools and MT but also the procurement of necessary funding to make them available to staff translators. Special attention needs to be paid to facilitating training for efficient and purposeful use of CAT tools, including managing TMs in a way that allows for uncomplicated future language resource sharing.

- **Raising awareness of language data as Open Data and a valuable asset:**
  To achieve this objective, EU legislation would be most effective and would help to increase interest in language technology related issues.

- **Establishing good data management practices in public services:**
  This goal includes actions such as creating databases with translated documents and their metadata (e.g. date of translation, information whether the document was translated in-house or outsourced, IPR holder etc.).

- **Identifying and gaining access to outsourced translations:**
  This objective could be addressed through a high level decision with a clear mandate for public institutions to collect language data and make it available respectively.

References and further reading list:

ELRC Workshop Report for Romania (2018):
http://www.lr-coordination.eu/sites/default/files/Romania/2018/ELRC%20Workshop%20Report%20ROMANIA_Public_FINAL.pdf

Survey on the translation needs in public institutions (2017):
http://ier.gov.ro/wp-content/uploads/newsletter/newsletter_noiembrie_2017_en.pdf

ELRC Workshop Report for Romania (2016):
http://lr-coordination.eu/sites/default/files/Romania/ELRC-Workshop-Romania-Public_Report.pdf

Annex
Country Profile Slovakia

Miroslav Zumrík, Lucia Konturová, Eileen Schnur

State of Play:
Translation practices in ministries and public administrations in Slovakia:

In Slovakia, there are no centralised operations for translations and each ministry has its own translation practices. The applied translation practices are diverse, ranging from in-house translation to decentralised outsourcing or a mixture of both.

While language service providers (LSPs) and freelance translators build on the support of computer-assisted translation tools (CAT tools) for their translation activities, their use is not common practice in public administrations. Since in the majority of Slovak public institutions, there are no specialised language and translation services, the use of outsourcing is often inevitable.[100] Public procurement data is openly available on the national procurement portal (https://www.uvo.gov.sk). According to the NEC TM Report[101], more than 1.3 million EUR were spent for outsourced translation services between 2015 and 2018. Most of these outsourced translations were contracted by the Slovak administrative sector. The results of the above-mentioned report demonstrate that there is a need for multilingual content within public administrations and ministries.

According to a representative from the National Agency for Network and Electronic Service (NASES), the need for more multilingual digital services is one of the biggest challenges in Slovakia. However, according to the META-NET White Papers published in 2012, language technology industry is not sufficiently developed and the quality of Slovak language technologies and resources is not satisfactory yet.[102] Machine translation (MT) systems are hardly used in Slovakian public administrations and ministries. However, a Slovak institution that has already translated documents with the help of machine translation is the Social Insurance Agency. They used the European Commission's machine translation system eTranslation, but not all results were satisfactory, since some of the texts had special requirements concerning term accuracy, which could not be fulfilled by the MT system.

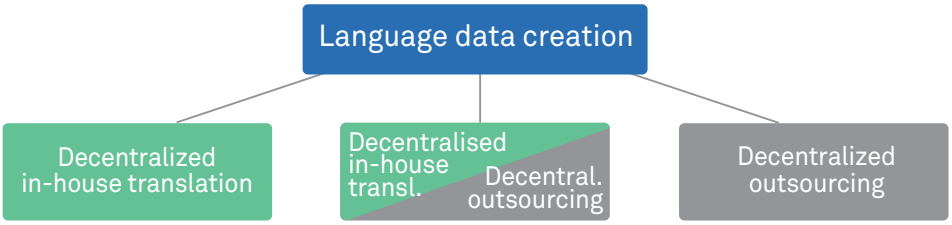
Data sharing infrastructures/Open Data in Slovakia:

In Slovakia, the Act on Free Access to Information is an important legal document when it comes to sharing data. Pursuant to this law, public institutions and ministries are required to provide information upon request. When it comes to Open Data, the Slovak government has already put a lot of effort in maintaining the national Open Data portals. The two main national Open Data portals are Data.Gov.sk and Slovensko.sk. They are maintained by the National Agency for Network and Electronic Services and include information about all state/public institutions and all Open Data. The electronic services offered by Slovensko.sk are diverse and aim to address the Slovaks' needs in their daily lives. In 2018, there were approximately 1700 services available in the national Open Data portal. It was used by about 480,000 registered users primarily from Slovakia, but also from other EU countries. Besides Data.Gov.sk and Slovensko.sk, there are also other local and regional Open Data portals such as Crime Map, which is based on criminal statistics of the Slovak police.
In addition, Slovakia supports the Free Flow of Data initiative by the European Commission, which turned into effect in May 2019. The initiative "aims at removing obstacles to the free movement of non-personal data across Member States and IT systems in Europe."[103]

According to Martina Slabejová, the Slovak Digital Leader, collecting Open Data needs to go hand in hand with raising awareness of how the gathered data can be used within the public administrations. There are several projects focusing on improving eGovernment services and making data available to the public. However, these initiatives primarily focus on the provision of e-services and not on sharing language data. One of them e.g. aims at integrating Slovak public bodies into the government cloud, where both public and private information can be stored. Nonetheless, the public part could deliver valuable Open Data material that is usable for e.g. training and improving the European Commission's machine translation service eTranslation.[104]

[100] Zumrík, Levická: *ELRC Workshop Report for Slovakia*. 2016.
[101] NEC-TM Report: *Slovakian National Contracts Report: Process and Findings*, 2019.
[102] Šimková et. al: *The Slovak Language in the Digital Age*, in: *META-NET White Paper Series*, 2012.
[103] Cf. European Commission: *Free Flow of Non-Personal Data*.
[104] Zumrík, Miroslav: *ELRC Workshop Report for Slovakia*, 2018.

The current language data creation infrastructure in Slovak public bodies looks as follows:



Language policy and Digital Policy in Slovakia:

Pursuant to the Act on the State Language of the Slovak Republic published in 1995, Slovak is the official language of Slovakia. It is currently spoken by about 4.5 million inhabitants of the country, followed by Hungarian with more than 450,000 speakers.

> **Interesting fact:**
> **In 2009, the Slovak language policy was modified and preferential use of the state language was mandated. This was criticised by the Hungarian community, which makes up 10% of Slovakia's population.**

The Deputy Prime Minister's Office for Investments and Informatisation (UPVII) is the key organisation for digitalisation of public services in Slovakia. Its goal is the centralisation of informatisation. Slovakia aims to cooperate with European institutions, which is why the UPVII also encourages administrations to participate in open CEF Calls. The Slovak government also created a number of operational programmes, which focus on the implementation of digitalisation of public administrations, thus helping to build a Digital Single Market.

> **Interesting Fact:**
> **In Slovakia, the UPVII is the main institution, which is responsible for the digitalisation of public services.**

Data management is inevitable for building eGovernment services. Since eGovernment services require cleaned, structured and categorized data, several Slovak initiatives have been started with the aim of cleaning data and connecting public bodies. According to the Slovak Digital Leader Martina Slabejová, there are approximately 100 projects that are crucial for the mission of the UPVII, which is to offer better e-services. In 2017, the "Detailed Action Plan on Digitisation of Public Administration" was published in Slovakia. The goal of this action plan is the development of an eGovernment system, which serves the needs of Slovak citizens, public administrations, businesses and academia.[105]

Stakeholders:

The ELRC Anchor Points for Slovakia represent two important stakeholders, i.e. the Ministry of Culture and the udovít Štúr Institute of Linguistics of the Slovak Academy of Sciences. The Ministry of Justice and the Ministry of Culture of the Slovak Republic contributed more than 1 million tokens of raw mono- and bilingual texts in different language combinations, mainly English-Slovak, covering a number of fields, e.g. laws, reports, letters, brochures, invitations, etc. This contribution lead to the creation of two parallel corpora in English and Slovak plus two monolingual datasets that were delivered to the ELRC-SHARE repository. In addition, the Ministry of Economy contributed a significant amount of English-Slovak parallel data after the second ELRC Workshop. Overall, more than 40 organisations participated in ELRC events such as local workshops or conferences.

[105] European Commission: *Digital Economy and Society Index (DESI) 2018, Country Report Slovakia*, 2018.

Annex
Country Profile Slovakia

Main challenges for sustainable data sharing:

- There are few multilingual digital services available in Slovakia.
- Public services are often not aware of the value of Open Data.
- In Slovakia, there is a lack of accessible multilingual data. This is also partly due to technical issues that prevent institutions from making Open Data accessible. As for the monolingual (Slovak) textual resources, there are some large and valuable resources available, e. g. the Slovak portal of judicial decisions (https://otvorenesudy.sk), collecting respectable volumes of data (texts in pdf files), as well as metadata (information on courts, judges, procedures etc.).
- Furthermore, language technology support still needs to be improved to be able to better serve the public administrations' requirements.
- Slovak public administrations are generally not aware of the range of technical solutions that could support them in their daily operations.
- There is a lack of awareness concerning the possibilities of the European Commission's funding mechanisms amongst Slovak public bodies.

Action plan:

Based on the current situation in Slovakia and the identified challenges, five objectives could be defined. Ranked by their priority, these include:

- **To increase interest in MT in public services:**
  As the use of MT is currently not common in public administrations and ministries, it is necessary to further promote the benefits of using machine translation. This can be achieved by creating synergies with national projects and initiatives on the one hand, but also by securing the support of Slovak decision makers on the other. In addition, it is important to provide more information on how MT systems work and to communicate how much data will be required to improve an MT system.

- **To raise awareness of language data as Open Data:**
  Since language data are currently not included in eGovernment initiatives, it is important to raise awareness of language data as Open Data. A first and important step towards this goal would be to identify and contact an Open Data officer.

- **To identify and gain access to outsourced translations:**
  As already mentioned, there is a general lack of openly accessible language data in Slovakia. At the same time, a high number of translations needs to be outsourced to LSPs or freelance translators. Consequently, one way to increase the number of accessible data would be to identify and gain access to outsourced translations. This could be achieved by making it a common practice to receive any by-products of the outsourced translations back.

- **To tackle legal concerns:**
  Since MT systems need to be improved to serve the public administrations' needs, as much language data as possible should be made available. However, legal issues often prevent potential contributors from sharing their data. Easy-to-apply guidelines for Intellectual Property Rights (IPR) and privacy issues could help overcome these issues and provide data holders with the necessary expertise.

- **To establish good data management practices in public services:**
  Last but not least, it would be useful to identify a data manager, supporting the creation and development of data management practices in public services.

European Language Resource Coordination
Connecting Europe Facility

References and further reading list:

Crime Map: https://mapazlocinu.sk/

European Commission: *Digital Economy and Society Index (DESI) 2018, Country Report Slovakia*, 2018, http://ec.europa.eu/information_society/newsroom/image/document/2018-20/sk-desi_2018-country-profile_eng_B4415E7E-9154-E26E-7B403212919F3F7C_52238.pdf.

European Commission: Free Flow of Non-Personal Data, https://ec.europa.eu/digital-single-market/en/free-flow-non-personal-data.

Freedom of Information Act: https://www.ujd.gov.sk/files/zakon211/Act_211.pdf
G-Cloud: https://www.sk.cloud/index-en.html

NEC-TM Report: *Slovakian National Contracts Report: Process and Findings*, 2019, https://www.nec-tm.eu/wp-content/uploads/2019/05/Slovakia-Report.pdf.

Šimková et. al: *The Slovak Language in the Digital Age*, in: META-NET White Paper Series, 2012, http://www.meta-net.eu/whitepapers/e-book/slovak.pdf.

Slovak portal of judicial decisions: https://otvorenesudy.sk.
The "Open Courts" - Slovak Portal of Judicial Decisions: https://otvorenesudy.sk.

Zumrík, Levická: *ELRC Workshop Report for Slovakia*, 2016, http://www.lr-coordination.eu/sites/default/files/Slovakia/ELRC-Workshop-Report_SLOVAKIA-public.pdf.

Zumrík, Miroslav: *ELRC Workshop Report for Slovakia*, 2018, http://www.lr-coordination.eu/sites/default/files/Slovakia/2018/ELRC%2BWorkshop%20Slovakia%20Public%20Report_FINAL.PDF.

Annex
Country Profile Slovenia

Simon Krek, Andraz Repar, Lilli Smal

State of Play:

Translation practices in ministries and public administrations in Slovenia:

In Slovenia, the majority of the translation demands in public services are handled centrally through the Translation and Interpretation Division (TID) at the Secretariat-General of the Government who manages translation and interpretation demands for most ministries. The TID handles 26% of the translations in-house, whereas the other translations are outsourced to language service providers. The Translation and Interpretation Division uses CAT tools and requires LSPs to do the same when they generate their translations. TMs are not shared, but the LSPs are required to return bilingual files for English, German and French translations.

A few ministries handle their translation needs independent from the TID, among them are the Ministry of Foreign Affairs, the Ministry of Defence, the Ministry of the Interior and the National Bank. These public bodies use CAT tools but they do not exchange data or know-how with other ministries or the TID. However, they also outsource part of their translations to the TID.

Data sharing infrastructures and Open Data in Slovenia:

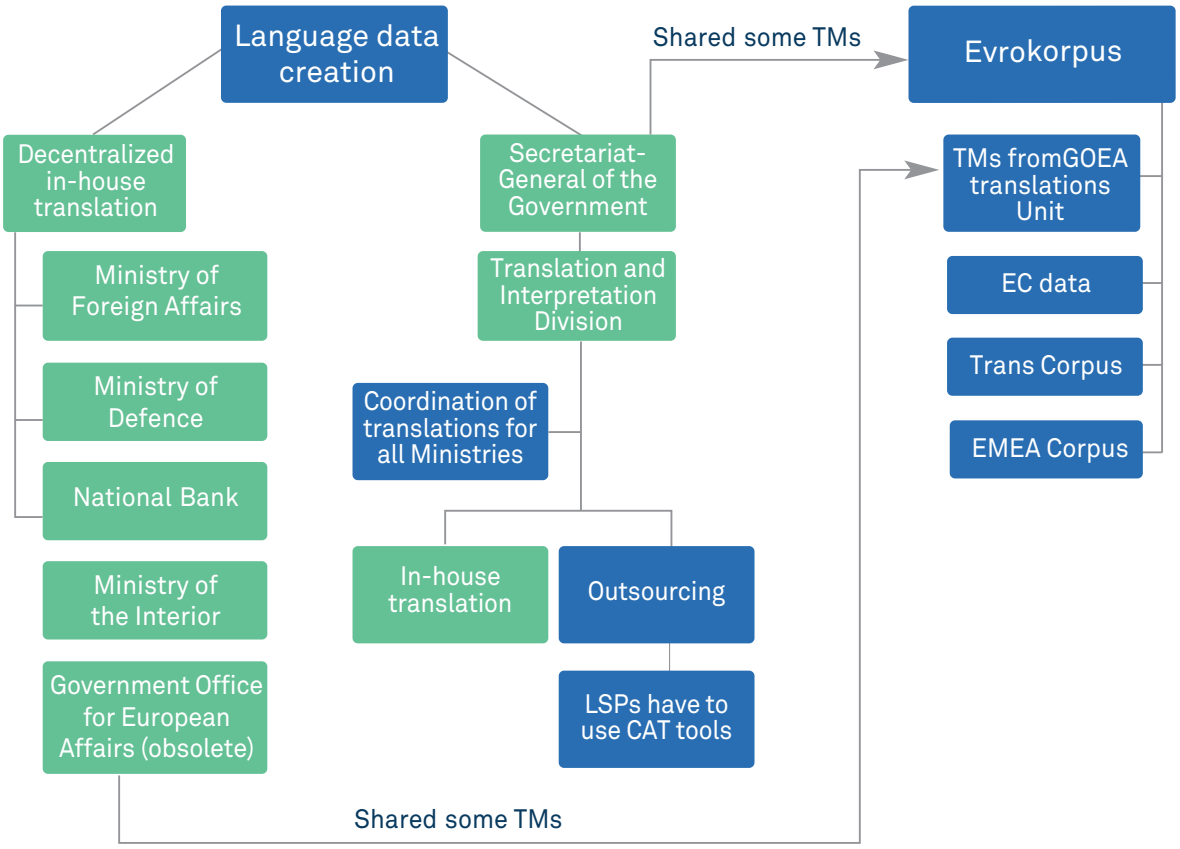
The Evrokorpus is a dedicated portal for parallel language resources for Slovene <> English, German, French, Italian, Spanish. It contains data from the (former) translation unit at the Government Office of European Affairs, data from the European commission, the Trans Corpus and the EMEA corpus, and is maintained and updated by the TID. Due to anonymization issues, however, only parts of the corpus could be shared with ELRC so far.

Evrokorpus has a companion terminology database called Evroterm which contains English, Slovene, German, Italian and Spanish terminology. The Evroterm database has been released on the Slovene Open Data portal under the CC BY-NC-ND license and is available in the ELRC-SHARE repository.

The full Slovene legislation, however, with more than 100 million tokens was made available in the Slovene Open Data portal as an open access database in JSON format which is considered an important achievement of the ELRC data collection task in Slovenia

> **Interesting fact:**
> **Full Slovene legislation with more than 100 million tokens was made available in the Slovene Open Data portal in JSON format.**

The Slovene Open Data Portal (https://podatki.gov.si/) has a dual function. The first one is to provide a central catalogue of all the records and databases of Slovenian public bodies. In this catalogue the metadata about all the Open Data from state authorities, municipalities and other public sector bodies is made available. The second function of the portal is to be the single access point for data in a machine-readable format and with an Open Data license. This includes Open Data collections which had already been published on different websites, such as Evroterm.



Language policy in Slovenia:

According to the Constitution, Slovene is the only official and state language of the Republic of Slovenia.[106] However, in municipalities, where the Italian and Hungarian speaking population resides, these languages have official status as well.[107] Romany languages have minority status in Slovenia.[108] In 2004, the Act on Public Usage of Slovenian Language came into effect monitored by the Ministry of Culture. The act determines the use of the Slovenian Language in public communication and in specific areas and resulted in the first resolution on the National Programme for Language Policy (NPLP) for a period of 4 years (2007-2011).[109] With the resolution, a budget of 12 million EUR was allocated to language policy and language planning for the first time. However, only about 300,000 EUR were actually spent by the Ministry of Culture in the framework of the resolution. The Ministry of Education on the other hand, spent over 3.2 million EUR from structural funds for the development of modern language technologies and resources for the Slovene language between 2008-2013.

It took three years after the initial resolution to pass the next Resolution on the National Programme for Language Policy (2014-2018) as the funds in the first resolution were not adequately spent. However, after the resolution passed, an action plan for Language Infrastructures was initiated in 2015 budgeted with

[106] Nećak Lük, Albina: "Slovene Language Status Planning", in: *Revista de Llengua I Dret, Journal of Language and Law,* p. 57, 2017.
[107] Nećak Lük, Albina: "Slovene Language Status Planning", in: *Revista de Llengua I Dret, Journal of Language and Law*, p. 62, 2017.
[108] Nećak Lük, Albina: "Slovene Language Status Planning", in: *Revista de Llengua I Dret, Journal of Language and Law*, p. 60, 2017.
[109] Cf. Nećak Lük, Albina: "Slovene Language Status Planning", in: *Revista de Llengua I Dret, Journal of Language and Law,*, p. 62 ff., 2017.

Annex
Country Profile Slovenia

11 million EUR. The financing bodies are the Ministry of Education, Science and Sports, the Ministry of Culture, the Slovenian Research Agency (subordinated to the Ministry of Education) and the publicly funded Slovenian Academy of Sciences and Art. Another result of the language act was the foundation of the Council for Continuous Monitoring of the Development of Language Resources and Technologies for Slovene representing several ministries, government offices and agencies who produce a yearly progress report.

According to this report, less than 250,000 EUR were spent in 2015 for the realisation of the resolution but it was announced in 2018 that the Ministry of Education had allocated up to 2 million EUR on the "Promotion of flexible and innovative learning techniques with the development of language resources and technologies".[110] The funding for the CLARIN infrastructure was increased from 42,000 EUR to 100,000 EUR per year, the funding for the Centre for Language Resources and Technologies at the University of Ljubljana were raised to 55,000 EUR per year compared to 11,000 EUR before, and a new research group for "Language Resources and Technologies for Slovene" at the University of Ljubljana received resources to fund 2,5 full time equivalent personnel per year from 2019-2024.

A third resolution for the timeframe 2019-2024 passed in 2018 followed by a public consultation on "Development of Slovene in digital environment – language resources and technologies" conducted by the Ministry of Culture, for which structural funding in the amount of four million EUR was available. The resulting call was to be published before the summer in 2019, however, as of October 2019, the call is not open yet.

Main challenges for sustainable data sharing:

The main challenges for sustainable language data sharing in Slovenia are the following:

- One of the central issues is the fact that the implementation of the Resolution on the National Programme for Language Policy is often dependent on high-ranking individuals and their disposition regarding language technology and language resources, which makes continuous and sustainable progress difficult.
- Another issue is the lack of efficient cooperation between stakeholders, although a lot of expertise is available.
- The unawareness of the value of language resources and Open Data results in reluctance to share language data as the benefits and incentives are not evident.
- In addition, concerns about personal or confidential data are holding many potential data donors back from sharing their language data.

Stakeholders:

The Council for Continuous Monitoring of the Development of Language Resources and Technologies and all its members play a crucial part in all activities related to language data and language technologies in Slovenia. As one of the main language data creators of parallel data for Slovene public administrations, the Translation and Interpretation Division subordinated to the Secretariat General of the Government is also a key stakeholder to create sustainable language data sharing infrastructures in Slovene public administrations. Many activities related to language resource collections are supported by the Josef Stefan Institute, represented by the Technology NAP, and the Centre for Language Resources and Technologies. So far, more than 30 institutions from the public sector, academia and industry have attended ELRC events and the Secretariat-General of the Government is one of the data contributors that shared language resources with ELRC.

Action plan:

Several objectives should be targeted to address the identified challenges. The first two objectives are suggested recommendations, whereas the last two objectives are partly addressed in the language resolution but their implementation should be reinforced as they are considered very important.

[110] Call for proposals: *Innovative and flexible forms of teaching and learning*, 2019.

- **Tackle legal concerns:**
  To address the concern of confidential or personal data potentially included in language resources, legal experts are needed that can advise each public administration if their data needs any kind of pre-processing before it can be shared. This also includes copyright and IPR-related issues. One venue worth exploring could be the practice of differentiating between non-personal open government data and texts that contain personal or confidential data which could help establish good data management practices in public services. In addition, appropriate guidelines for the creators of language data are needed that can be followed during or after the translation process to make data sharing easier.

- **Raising awareness of language data as Open Data and a valuable asset:**
  This objective includes activities such as integrating language data in the national Open Data policy and establishing practical guidelines for language resources as Open Data.

- **Identify and gain access to outsourced translations:**
  By establishing the value of language resources, changes in the procurement process can be initialised. As the procurement of language service is not centralized, each public administration that has a demand for translation would have to change its procurement process and ensure the provision of translation memories as well as any other by-product of translation including the transfer of copyright mandatory for outsourced translations.

- **Increasing interest in MT/LT in public services as part of the national digital policy:**
  To increase the public interest in machine translation and language technologies, it is important to create synergies between different initiatives and address the needs of the public sector. This includes e.g. the dissemination of use cases and best practices.

References and further reading list:

Call for proposals: Innovative and flexible forms of teaching and learning, 2019, https://www.gov.si/zbirke/javne-objave/inovativne-in-prozne-oblike-poucevanja-in-ucenja-2/.

Centre for Language Resources and Technologies: www.cjvt.si.

Evrokorpus: http://www.evroterm.gov.si/evrokorpus/index.php?jezik=angl

Evroterm: https://evroterm.vlada.si/index.php?jezik=angl

Nećak Lük, Albina: “Slovene Language Status Planning”, in: *Revista de Llengua I Dret, Journal of Language and Law*, 2017, https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=2ahUKEwi-UmtqzyKrlAhUDyqQKHVY1BHkQFjABegQIARAC&url=http%3A%2F%2Frevistes.eapc.gencat.cat%2Findex.php%2Frld%2Farticle%2Fdownload%2F10.2436-rld.i67.2017.2918%2Fn67-necak-en.pdf&usg=AOvVaw3v19rbPUtJ6TSRRRhEtgLL.

Slovene Open Data portal: https://podatki.gov.si/.

Annex
Country Profile Spain

Maite Melero, Núria Bel, David Perez, Lilli Smal

State of Play:
Translation practices on the national and regional level:

In Spain, there is a significant difference in meeting translation needs depending on the administrative level. In contrast to many other countries, translations carried out for or by public services on the regional level are stored in translation temories, translators make use of machine translation and the processes are mostly centralized and coordinated. The reason for this is that Spain is a multilingual country and has three co-official languages in addition to the official language Spanish. The co-official languages, Catalan, Basque and Galician, are official in the respective regions, leading to an enormous demand for translation, especially on the regional level (see Language Policy in Spain). This has led to a more principled approach to translation at this administrative level.

On the national/state level, however, there is less pressing need to translate, therefore most public administrations outsource their translations to language service providers. The translation and procurement processes are highly decentralized, sometimes even within one institution. Only very few ministries have in-house translation services, namely: the Ministry of Interior, the Ministry of Defence, and the Language Interpretation Office (OIL) at the Ministry of Foreign Affairs. According to a "White Paper on Institutional Translation and Interpretation" only one translator at the OIL used computer-aided translation (CAT) tools in 2011. More figures on the number of staff translators and the use of CAT tools provided in the "White Paper on Institutional Translation and Interpretation"[111] published in 2011 are presented below:

- OIL (Ministry of Foreign Affairs: 17 translators, 1 used CAT as of 2011)
- Ministry of Interior (230 translator no CAT as of 2011)
- Ministry of Defense (30 translators,. no use of CAT as of 2011)
- Ministry of Presidency (11 translators, no use of CAT as of 2011)

In the last few years, most translators at OIL are using a licensed CAT tool. As for the other ministries, they outsource some of their translations to language service providers but in most cases they do not request back the resulting translation memories (TMs) for internal purposes. The main reason is lack of awareness for the intrinsic value of TMs, plus the fact that they are not currently using CAT tools and therefore do not see a direct benefit from requesting back the TMs. Other public services, such as the State Administration Agency, the Tourism Agency (Segittur) or the Spanish National Centre for Legal Administration also outsource most of their translations but are starting to claim their translation memories.

In the course of the NEC TM project[112], an assessment of the translation costs was carried out. Not considering translation contracts below 10,000 EUR, the translation costs for outsourced translations amounted to approximately 44 million EUR in the period of five years (2013-2018). According to this study, requesting the translation memories from language service providers could reduce translation costs by up to 10%.

In the course of a study commanded by the State Secretariat for Telecommunications and the Information Society (SETSI) in 2016, called "Inventory of linguistic resources of the Public Administration for automatic translation"[113]) several sectors were identified that have a particularly large demand for translations. All the related public bodies outsource translations. The sectors are:

- Police (state and autonomous communities)
- Administration of justice (both state and regional level)

[111] Cf. Ministerio de Asuntos Exteriores y de Cooperación: *Libro Blanco de la traducción y la interpretación institucional*, 2012.(White Paper on Institutional Translation and Interpretation)

[112] NEC TM Country Report: *Report on Spanish National Translation Contracts*, 2018

[113] Aguado de Cea et. al: *Inventario de Recursos Lingüísticos de la Administración Pública para Traducción Automática*, 2016.

- Tourism
- Social security
- Tax agency

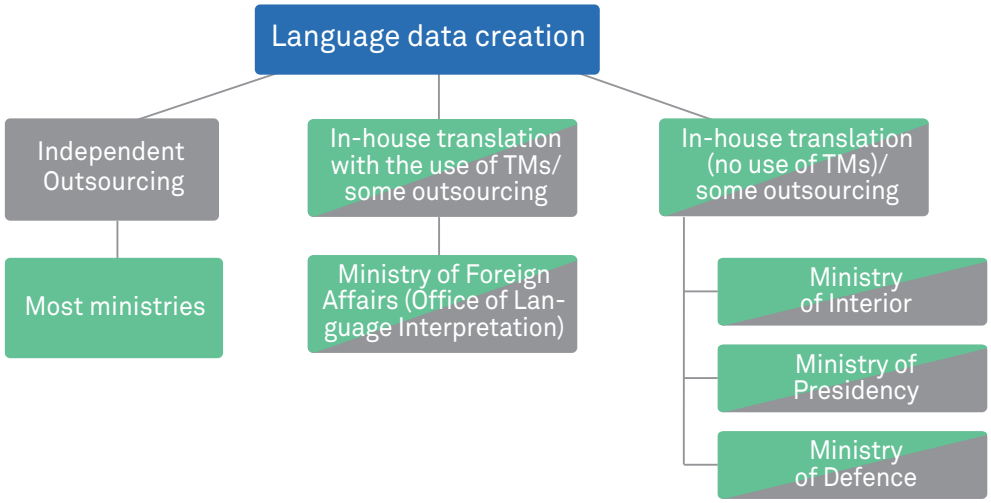
To support the need for translation, the Spanish government runs their own Machine Translation Platform called PLATA that is used as an API to translate Spanish Government web pages, but is also serving other customers such as MUFACE, and the Spanish Agency for Data Protection, to translate to and from Spanish and the co-official languages and English. To support more language pairs and provide a better service, it has been recently connected to eTranslation.

Interesting fact:
Due to the need to translate between co-official languages and Spanish, translation procedures at the regional level have a longer tradition of using CAT tools and machine translation than those at national level.

Overall, it can be said that most public administrations on the state level rely heavily on outsourcing their translations, without a standard procedure to request back translation memories. Generally speaking, there is no systematic use of translation memories.

Moreover, there is little coordination and management of language data in terms of protocols for consistent archiving, using metadata standards, differentiating between confidential documents and documents that contain personal information, or keeping original documents aligned with their translations.

The current language data creation infrastructure on the national level in Spain looks as follows:



Language and digital policy in Spain:

In Spain, the implementation of the digital agenda is coordinated by the State Secretariat for the Digital Advancement (SEAD). SEAD (previously SETSI) is strongly supporting and fostering the use of language technologies in the public and private sector and has published a Plan for the Advancement of Language Technology underlining the importance of collecting and sharing language data as a means to "foster the natural language processing and machine translation sectors through this targeted plan"[114].

This plan is focusing on three main points:

- "Increasing the amount, quality and availability of linguistic infrastructure in Spanish and in Spain's co-official languages.

[114] Plan for the Advancement of Language Technology, p.8, 2015.

Annex
Country Profile Spain

- Fostering the language industry by promoting knowledge transfer from the research field to industry. Bolstering internationalisation of companies and institutions in the sector. Improving the reach of current projects.
- Improving the quality and capacity of public services, by integrating natural language processing and machine translation technologies, while simultaneously driving market demand. Supporting creation, standardisation and distribution of language resources, created by the management activities performed by the public administrations." [115]

The governance bodies of the Spanish language are the Royal Spanish Academy and the Association of Spanish Language Academies in Ibero-America.[116]

Open Data Portal in Spain:

The web portal http://datos.gob.es federates the Open Data from different public administrations: government, municipalities and regional autonomous governments. The richest language data comes from the Basque Administration in form of translation memories. Other text collections can be found, most of them in PDF format.

Main challenges for sustainable data sharing:

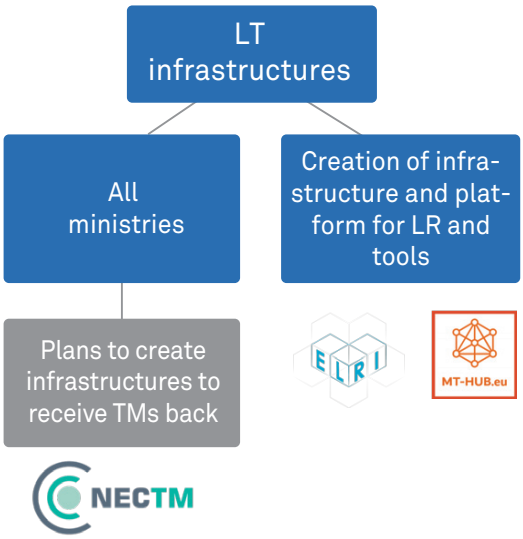
- Undervalued perception of textual data leads to a number of issues:
  - Translations are not filed consistently which makes it difficult to match the original text with the translation
  - Text and translations produced for single occasions are not stored in TMs as it seems unlikely that they will be reused
  - Most documents are stored in PDF format, which is an unsuitable format for building translation memory files, and the source format is lost
  - General lack of data plans and protocols and resistance to modify internal document management practices[117]
- Only some translators in translation services use CAT tools
- Translation needs are met very decentrally, even within ministries
- Legal uncertainty (unclear authorization chain to decide what can be shared)

Stakeholders:

About 90 unique stakeholders could be identified representing public institutions on the regional, state and municipal level. Some of them are starting to contribute with language data resources, specifically after the dissemination of the ELRI initiative.

Action plan - ongoing projects and future plans:

Following the strategy laid out by the national **Plan for the Advancement of Language Technology**, an appropriate infrastructure to facilitate language data reuse and sharing is being put in place thanks to the synergies with several CEF-funded projects participated by SEAD (see figure). The following repositories and platforms are currently being deployed:



[115] Plan for the Advancement of Language Technology, p.7, 2015.
[116] Cf. Plan for the Advancement of Language Technology, p.6, 2015.
[117] Cf. Bel, Núria, Melero, Maite: *ELRC Workshop Report for Spain*, p. 11, 2018.

- The national Relay Station resulting from the ELRI project (https://elri.plantl.gob.es/es-es/) is already being leveraged as a data storage and processing solution, with complex sharing capabilities. It is being used as the main data portal to gather language data from public institutions.

- The MT-hub platform, an outcome from the CEF project iADAATPA, is also being deployed as a translation portal for the public administration, and an access point for eTranslation. MT-HUB is compatible with multiple content management tools, and helps the user to select the best domain-adapted MT engine for their document in need of translation.

- A central translation memory server will soon be deployed, as a result of the NEC TM project. The server will be secure place to store translation memories coming from outsourced translation contracts, directly by the LS providers. It will allow sharing of TMs between administrations themselves, and administrations and providers, and will be compatible with commercial CAT tools. An open CAT tool integrated to it will also be provided.

References and further reading list:

Aguado de Cea et. al: *Inventario de Recursos Lingüísticos de la Administración Pública para Traducción Automática*, 2016, https://www.plantl.gob.es/tecnologias-lenguaje/actividades/ Estudios%20tcnicos%20y%20de%20gobernanza/Inventario%20de%20recursos%20para%20traducci%C3%B3n%20autom%C3%A1tica/inventario-recursos-traduccion-Retele.pdf.

Bel, Núria, Melero, Maite: *ELRC Workshop Report for Spain*, 2018, http://lr-coordination.eu/sites/default/files/Spain/2018/ELRC-Workshop-Report_Spain-2018-public-final-EC_.pdf.

Melero, Badia, Moreno: *"The Spanish Language in the Digital Age"*, in: *META-NET White Paper Series*, 2012, http://www.meta-net.eu/whitepapers/e-book/spanish.pdf.

Ministerio de Asuntos Exteriores y de Cooperación: *Libro Blanco de la traducción y la interpretación institucional*, 2012, http://www.ritap.es/wp-content/uploads/2012/11/libro_blanco_traduccion_vfinal_es.pdf.

NEC TM Country Report: *Report on Spanish National Translation Contracts,* 2018, https://www.nec-tm.eu/country-reports-nec-tm/country-report-spain/.

Website of the Plan of Impulse of Language Technologies: https://www.plantl.gob.es/Paginas/index.aspx. Plan for the Advancement of Language Technology, 2015, https://www.plantl.gob.es/tecnologias-lenguaje/PTL/Bibliotecaimpulsotecnologiaslenguaje/Detalle%20del%20Plan/Plan-Advancement-Language-Technology.pdf.

Annex
Country Profile Sweden

Rickard Domeij, Arne Jönsson, Lilli Smal

State of Play:
Translation practices and data sharing in ministries and public administrations in Sweden:

In Sweden, most public administrations outsource translations through public procurement. The process itself is coordinated by each public administration or agency independently but they are obliged to order translation services through Kammarkollegiet, the Legal, Financial and Administrative Services Agency that is in charge of the framework agreements for translation services in Sweden. This framework agreement includes, inter alia, that the supplier must be able to use translation memories provided by the public administration and that they have to deliver the produced translation memories upon request of the public administration without extra charge.[118] Public agencies that outsource translations under the legal framework hold the property right to the translation and the translation memories by default and can also include more explicit formulations about their specific needs in the framework agreement. If a public administration decides not to procure translation services through the framework agreement, they are obliged to notify the National Procurement Services of the reason.

Public procurement in Sweden in general follows five fundamental principles. These are: the principle of non-discrimination, the principle of equal treatment, the principle of transparency, the principle of proportionality and the principal of mutual recognition.[119] Most of the regulations for public procurement are implementations of the EU directive 2014/24.[120]

> **Interesting fact:**
> **According to the framework agreement for procuring translation services, the property rights belong to the contracting authority by default and TMs must be transferred upon request without extra charge.**

Although most translation services are outsourced, some institutions have in-house translation services such as the Swedish Police who exclusively translates in-house and the Ministry for Foreign Affairs who outsources part of their translations. The main reason for the in-house translation services is the nature of the texts that need to be translated, i.e. texts that contain confidential information, which is the reason why these translation memories are not shared with e.g. the National Language Bank. Generally, it is very common to use computer-aided translation (CAT) tools in the translation process, this applies to both in-house translation services as well as most freelance translators and language service providers. Some large commercial language service providers also have integrated machine translation systems.

Although the National Food Agency does not have an in-house translation service, they have access to eTranslation, i.e. machine translation on the institutional level to skim texts and documents in other languages or draft texts in languages other than Swedish.

Language data sharing infrastructures in Sweden:

The Swedish Language Council, Språkrådet, subordinated to the Swedish Institute of Language and Folklore (Ministry of Culture), plays a central role in collecting language data in Sweden in order to promote the development of language technology and terminology, as stated by the instruction in the regulation for the Institute of Language and Folklore.[121] For that purpose, texts and terminologies are regularly fed into Nationella Språkbanken, the National Language Bank of Sweden, who then share the data with the ELRC-Share repository. In its function as national coordinator of terminology, the Language Council manages the national termbank and also supports other public agencies in their terminology management. The Language Council envisions that this pipeline will be employed for any language resources, including translation memories.

---

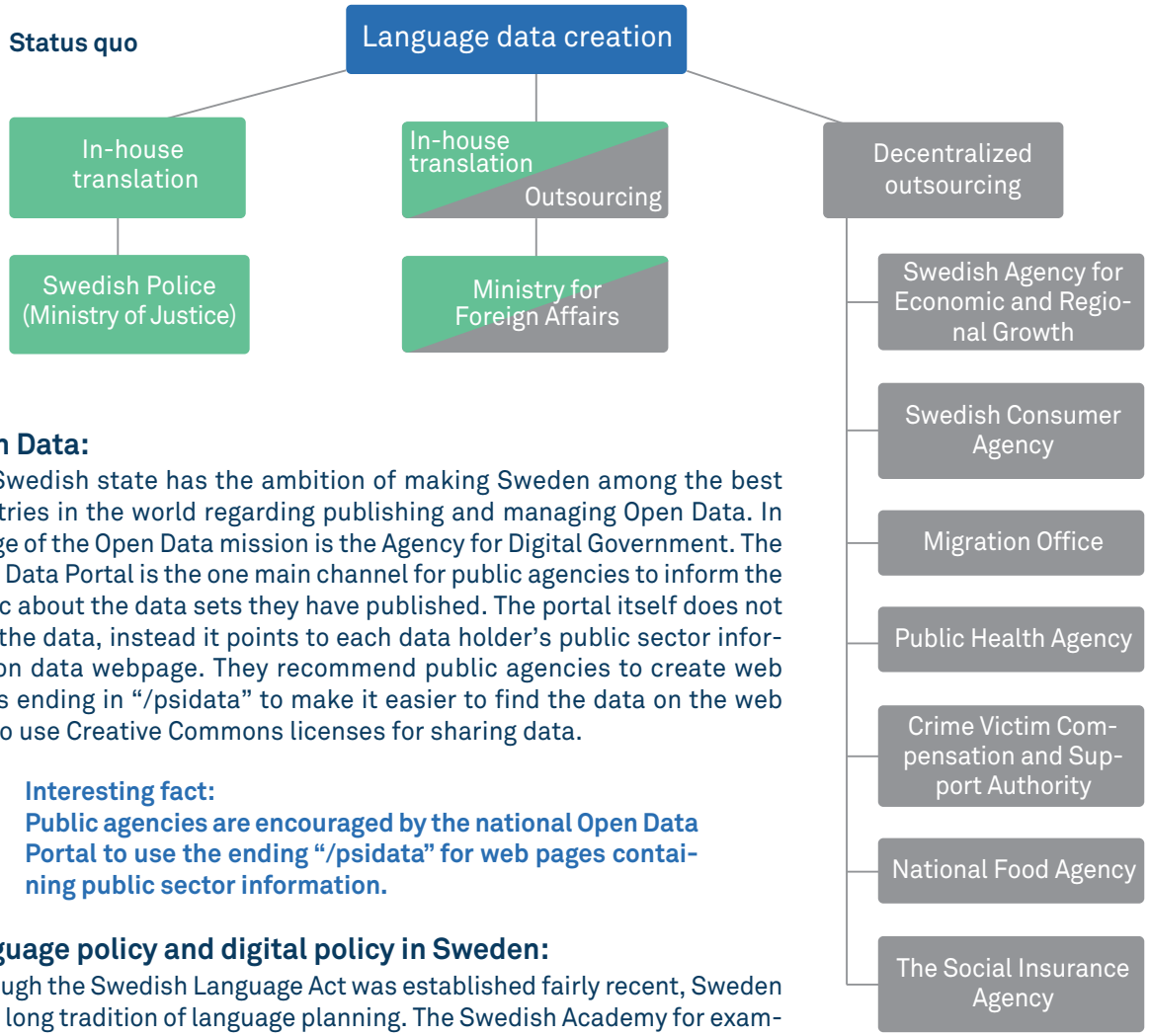
[118] Kammarkollegiet: *Procurement Framework for Translation Services*, 2005; Sverigs Rigsdak: *Kulturutskottets betänkande 2005/06*:KrU4, p.3.

[119] The National Agency for Public Procurement: Sustainable Public Procurement, 2014.

[120] DIRECTIVE 2014/24/EU OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 26 February 2014 on public procurement and repealing Directive 2004/18/EC.

Recently, a network consisting of representatives from a few different public agencies has been established to actively put this vision into practice and to discuss measures that will make language data sharing in the future easier.

The current language data sharing infrastructure in Swedish public administrations looks as follows:



Open Data:

The Swedish state has the ambition of making Sweden among the best countries in the world regarding publishing and managing Open Data. In charge of the Open Data mission is the Agency for Digital Government. The Open Data Portal is the one main channel for public agencies to inform the public about the data sets they have published. The portal itself does not host the data, instead it points to each data holder's public sector information data webpage. They recommend public agencies to create web pages ending in "/psidata" to make it easier to find the data on the web and to use Creative Commons licenses for sharing data.

Interesting fact:
Public agencies are encouraged by the national Open Data Portal to use the ending "/psidata" for web pages containing public sector information.

Language policy and digital policy in Sweden:

Although the Swedish Language Act was established fairly recent, Sweden has a long tradition of language planning. The Swedish Academy for example was founded in 1786 to "advance the Swedish language and Swedish literature"[122] and the precursor to the current Language Council has been working with language planning and cultivation for Swedish since 1944. In 2005, the Swedish Parliament adopted a bill addressing that the four main objectives of its concerted language policy are:

- Swedish is to be the main language in Sweden.
- Swedish is to be a complete language, serving and uniting society.
- Public Swedish is to be cultivated, simple and comprehensible.
- Everyone is to have a right to language: to develop and learn Swedish, to develop and use their own mother tongue and national minority language, and to have the opportunity to learn foreign languages.[123]

[121] Sverigs Riksdag: *Instruction 2007:1181*.
[122] Cf. The Swedish Academy Website.
[123] Lindberg, Inger: *Multilingual Education: a Swedish Perspective*, p.74, 2007.

Annex
Country Profile Sweden

The importance of language technology and language data collection for the Swedish language is also acknowledged in the government bill:

Central to promoting good development in the language technology area is to systematically build up large text and speech databases and to develop software. Text and speech databases store very large amounts of authentic spoken and written language in a way that makes it accessible for computerized, linguistic analysis. Such an analysis, in turn, is a prerequisite for developing programs for automatic translation, for transmitting text to speech (and vice versa), for computerized speech recognition, etc. The construction of text and speech databases is costly and labour-intensive and requires long-term planning and is about creating basic language technology resources to develop well-functioning language technology. We therefore believe that a function for coordination of language technology should exist with the new language care organization so that resources can be better coordinated and the conditions for participating in major collaboration programs in the Nordic countries and the EU is improving.[124]

It was recognized that to meet these objectives a coordinated effort was needed and as a consequence the Swedish Language Council, the national language planning authority, was founded in 2006.[125] In 2009, the Swedish Language Act entered into force, establishing Swedish as "the principle language in Sweden" that must be usable and therefore have specialist terminology in all different areas of society.[126] Five other languages have been granted minority status, these are: "Finnish, Yiddish, Meänkieli (Tornedal Finnish), Romany Chib and Sami." [127]

The Language Act has 15 sections addressing the Swedish language and its status, national minority languages, Swedish sign language, the use of language in the public sector, Swedish in international context and Individuals' access to language.[128] The sections addressing the public sector impose responsibility on the public sector to use, develop and cultivate Swedish while focusing on simple and comprehensible language. The status of Swedish as an official EU language is also addressed in the language act and is deemed important to be safeguarded.[129] At the same time, the public sector is also obliged to "protect and promote the national minority languages"[130] and to ensure that "the individual is given access to language" including official minority languages as well as other first languages spoken by residents in Sweden.[131] The Language Act is also monitored by the Swedish Language Council.

In 2008, the National Language Bank became a national research infrastructure (2017-00626) funded by the Swedish Research Council by about 1.5 million EUR per year until 2025. The overall budget including co-financing is about 3 million per year. The Language Bank is divided into three divisions: Text, Speech and Sam (for Society) and supports all fields of research related to language data including e.g. language technology, digital humanities or artificial intelligence. The SWE-CLARIN consortium including 10 organizations is part of the research infrastructure. In 2011, the Swedish government stated in the Digital agenda for Sweden (2011)[132] that a National Language Bank is an important infrastructure for the development of language technology.

Stakeholders:

Among the main stakeholders for language data collection and sharing are the Agency for Digital Government, subordinated to the Ministry of Infrastructure who is responsible for Open Data collection in Sweden and

[124] Sverigs Rigsdak: *Kulturutskottets betaȁkande 2005/06:KrU4*, p. 30.
[125] Cf. Lindberg, Inger: *Multilingual Education: a Swedish Perspective*, p.74, 2007.
[126] Ministry of Culture: *Swedish Language Act*, p. 1.
[127] Ministry of Culture: *Swedish Language Act*, p. 2.
[128] Cf. Ministry of Culture: *Swedish Language Act,* p. 1 ff.
[129] Cf. Ministry of Culture: *Swedish Language Act,* p. 3.
[130] Ministry of Culture: *Swedish Language Act*, p. 2 f.
[131] Ministry of Culture: *Swedish Language Act*, p. 3.
[132] Ministry of Enterprise, Energy and Communications: *ICT for Everyone - A Digital Agenda for Sweden*, 2011.

the Institute of Language and Folklore, subordinated to the Ministry of Culture, who is in charge of monitoring the language policy and who is the national terminology coordinator. As such they closely cooperate with the National Language Bank as well. Over 40 institutions have participated in ELRC events and several public administrations that continuously create language resources through outsourcing translations have already contributed language data to ELRC. Some of the data donors are the Swedish Agency for Economic and Regional Growth, the Public Health Agency, the Migration Office, the Consumer Agency, the National Audit Office, and the Crime Victim and Support Authority. It should be noted that not all data shared with ELRC is Open Data.

Main challenges for sustainable data sharing:

In the past years, several projects and initiatives aimed at collecting and sharing data in Sweden have increased the awareness of the value of language data for promoting the languages of Sweden and improving the efficiency of public services. Yet, there are still some challenges that need to be overcome:

- Although a considerable number of public administrations have already shared their language resources, language data are still undervalued.
- The benefits of sharing data should be more tangible and address the needs and efforts of public agencies, these however have to be identified first.
- Since most translations are outsourced, guidelines and expertise are necessary on how to request translations with maximum mutual benefit for both the contracting authority and the contractor.
- Although significant progress has been made in reaching out to public administrations and convincing them to share their language resources, the processes of continuous data sharing from public administrations to a central data bank is not yet defined. This includes the licensing of data sets. Currently, different authorities use different licenses without consensus on which ones to use.
- Current translation practices do not allow for language data sharing as translations that contain personal or confidential data are not separated from translations that fall under the public sector information directive and can be safely shared.

Action plan:

To make data sharing easier, the following objectives should be addressed:

- **Promote Language Technologies for the languages in Sweden including minority languages according to the national language policy:**
  This is the main objective for Sweden and all other objections, actions and goals are subordinated to this main objective.

- **Raising awareness of language data as Open Data:**
  In order to encourage more public administrations to share their language data, it is important to establish practical guidelines that indicate clearly what language data can be made Open Data and how.

- **Increasing interest in MT in public services:**
  By identifying specific needs that can be addressed through machine translation or language technologies and creating synergies between different actors and initiatives, the potential and benefits can be showcased.

- **Tackle legal concerns:**
  Public administrations are uncertain about how to handle legal concerns relating to language data. In practice, it is not clear what license to use. This must be clarified in the national guidelines. The discussions between the Swedish Language Council with the Agency for Digital Government about this issue are on-going.

Annex
Country Profile Sweden

- **Identify and gain access to outsourced translations:**
  The first step is to further clarify the nature of the translation contracts and to collect best practices. On that basis, changes can be discussed and introduced. Employees of contracting authorities that outsource translations also need to be advised on how to procure translation services.

- **Establish good data management practices in public services:**
  The current data management practices need to be further investigated. These activities have already started. There have also been discussions in the network of representatives from public administrations about how to introduce separation between confidential and private data from public sector information.

References and further reading list:

Avropa: Framework agreements, https://www.avropa.se/topplankar/In-English/.

Borin et. al: "The Swedish Language in the Digital Age", in: *The META-NET White Paper Series*, 2012, http://www.meta-net.eu/whitepapers/volumes/swedish.

DIRECTIVE 2014/24/EU OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 26 February 2014 on public procurement and repealing Directive 2004/18/EC: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32014L0024.

Domeij et. al: "Enhancing Information Accessibility and Digital Literacy for Minorities Using Language Technology - the Example of Sámi and Other National Minority Languages in Sweden", in: *Perspectives on Indigenous writing and literacies*, 2018, https://brill.com/view/title/31954.

Kammarkollegiet: Procurement framework for translation services, 2005, https://www.avropa.se/contentassets/175b3a49c5b74d39a3a6a1c104812554/2.-allmanna-villkor---160411.pdf.

Lindberg, Inger: *Multilingual Education: a Swedish Perspective*, 2007, http://srii.org/content/upload/documents/66f9f6e9-c2d7-47ab-8b37-c37c6fb1471a.pdf.

Ministry of Culture: *Swedish Language Act*, https://www.regeringen.se/contentassets/9e56b0c78cb5447b968a29dd14a68358/spraklag-pa-engelska.

Ministry of Enterprise, Energy and Communications: *ICT for Everyone - A Digital Agenda for Sweden*, 2011, https://de.scribd.com/document/89103979/A-Digital-Agenda-for-Sweden.

National Agency for Public Procurement: https://www.upphandlingsmyndigheten.se/en.

National Agency for Public Procurement: *Sustainable Public Procurement,* 2014, https://www.upphandlingsmyndigheten.se/en/publicprocurement/about-the-public-procurement-rules/.

National Language Bank of Sweden: http://www.sprakbanken.se/eng.

SWE-CLARIN: https://sweclarin.se/eng/about.

Swedish Open Data Portal: oppnadata.se.

Swedish Language Act: https://www.regeringen.se/contentassets/9e56b0c78cb5447b968a29dd14a68358/spraklag-pa-engelska.

Sveriges Riksdag: *Instruction 2007:1181,* https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/forordning-20071181-med-instruktion-for_sfs-2007-1181.

Sverigs Rigsdak: *Kulturutskottets betänkande 2005/06:KrU4*, https://www.riksdagen.se/sv/dokument-lagar/arende/betankande/basta-spraket---en-samlad-svensk-sprakpolitik_GT01KrU4.

The Swedish Academy: https://www.svenskaakademien.se/en.

The Swedish Institute: *The Swedish Language*, 2016, https://sharingsweden.se/app/uploads/2016/10/The-Swedish-language_high-res.pdf.

Annex
Country Profile The Netherlands

Jan Odijk, Piet van den Berg, Carole Tiberius, Andrea Lösch

State of Play:
Translation practices in ministries and public administrations in The Netherlands:

In Dutch public administrations, the creation of translations is predominantly decentralized and often, the translation process is not centralized even within one ministry. Overall, eight ministries and executing bodies have their own in-house translation department, namely: The Ministry of Foreign Affairs, the Ministry of Justice and Security, the Ministry of Defence, the National Police, the General information and Security Service (AIVD), the Military Information and Security Service (MIVD), the Social Security Bank (SVB), and the Employee Insurance Agency (UWV). The translation department of the Ministry of Foreign Affairs sometimes translates documents for other ministries, but only if the documents are considered confidential at the moment of translation (they may become public in a later stage).

Most translations produced within the Dutch public administrations are outsourced to commercial translation agencies. Public procurement data is available through PIANOo, the Dutch Public Procurement Expertise Centre. All calls for tenders are announced through TenderNed, which is fully connected with the European Tender Electronic Daily (TED). The Ministry of Foreign Affairs translates approx. 7 million words per year in-house, which corresponds to less than 5% of the total of translated texts in the government.

> **Interesting fact:**
> **The Ministry of Justice and Security (Foreigners Chain) created a standard contract for outsourcing translations that contains a clause to transfer Translation Memories (TMs) by default.**

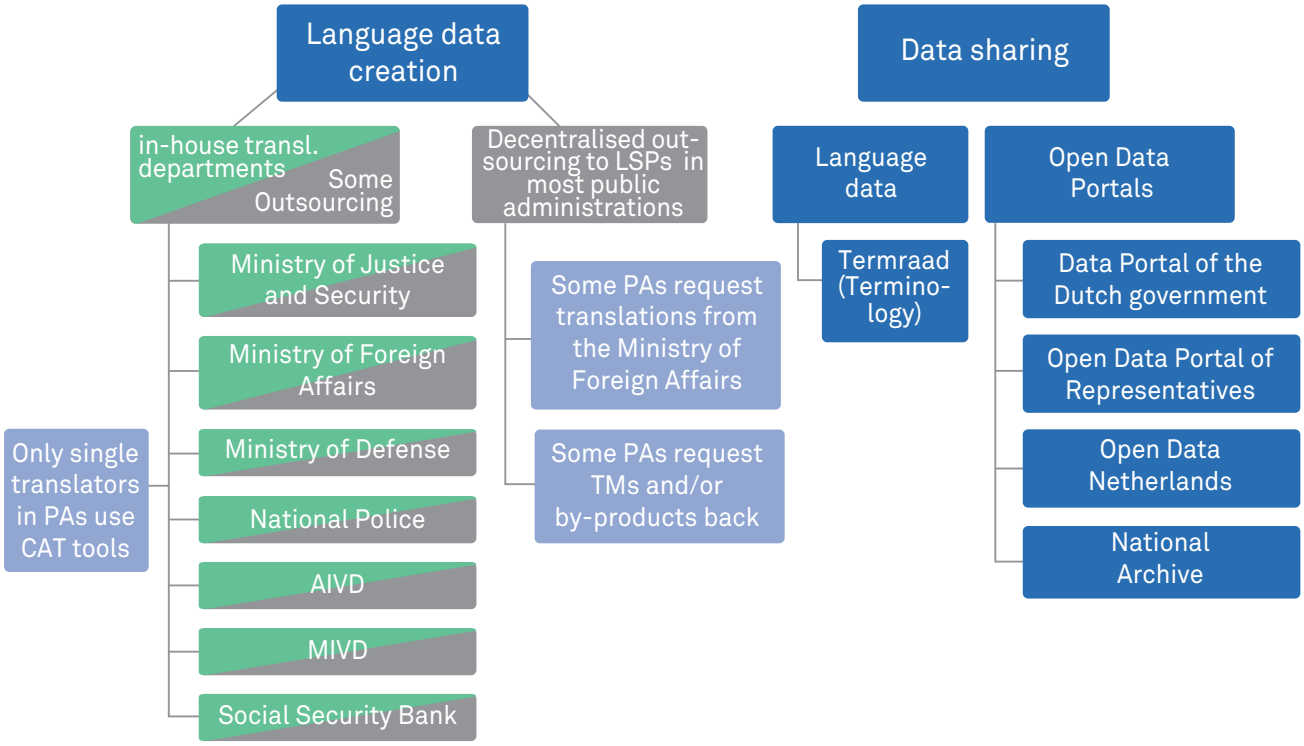
Overall, the outsourcing models for translations in The Netherlands have been – and still are – slowly changing. Originally, there was a pool of parties a ministry could source from. Now, a cascade model is considered, where one party takes all translations they can handle and the excess work goes to the second party, etc.

For translating, a computer-assisted translation software suite is often used in the Dutch Ministries. With regard to machine translation (MT), CEF eTranslation and other freely available online translation tools are used by the Ministry of Foreign Affairs.

Most documents that need to be translated are speeches, speaking notes, memos, diplomatic cables, as well as some proposals for tender (especially for embassies). Language pairs which are frequently requested include NL<>EN (45%), NL<>FR (30%) and NL-Other EU languages (ES, IT, FR - 20%). Emails are also often translated by individuals using common online translators or other freely available MT services. Until now, there is no organised or centralised exchange of language data on the national level. However, within the inter-institutional Termraad, attempts for direct collaboration on the terminological harmonisation have been made between the translation services of the various European institutions and the Dutch, Belgian and Flemish authorities.

Annex
Country Profile The Netherlands

The current language data sharing infrastructure in Dutch public bodies looks as follows:



Open Data in The Netherlands:

Open Data in The Netherlands is available through the national Open Data Portal (https://data.overheid.nl), but most of it is not natural language text; law texts are available through https://wetten.overheid.nl. So far, no systematic attention was given to multilinguality of Open Data or the availability of language data as Open Data in the Netherlands. There are a few Open Data sets with multilingual content (e.g. openstreetmap), but many of these are restricted to terms or names in metadata that consist of a single word or a few words – they do not involve running natural language text.

In addition to Open Data initiatives, The Netherlands place great emphasis on the digitization of public administrations and services. In addition to PIANOo, the Routing Institute for National and International Information Streams (RINIS) is the hub for fully-automated electronic data exchange in the public domain and tries to harmonize the exchange of information and corresponding infrastructure on the national level.

Language policy in The Netherlands:

The language policy of the Netherlands has been outsourced to the Dutch Language Union (the same is true for Flanders and Surinam).[133] The Dutch Language Union decides on the official spelling of words, for example.

Stakeholders:

So far, more than 120 potential stakeholders have been identified for The Netherlands, most of them being holders and creators of language resources. Around 30 of them participated in the last ELRC Workshop. So far, 37 language resources have been contributed to the ELRC-SHARE, mainly from the Dutch Language Institute. Main potential beneficiaries include: UWV, SVB, Ministry of Foreign Affairs, Ministry of Justice and Security.

[133] Taalunieversum: *Veelgestelde vragen over ons taalbeleid (Frequently Asked Questions about our Language Policy).*

Main challenges for sustainable data sharing:

- **Legal concerns/lack of explicit mission to share language resources:**
  There is a hesitation in ministries to share translations for various reasons including translator's rights to the texts.

- **Unavailability of translated texts:**
  - When creating a translation within the public administrations, a lot of the work is not so much to actually translate texts, but rather to "transcreate", i.e. to create a new text in a different language with contents similar to a particular source but without direct translation of these sources. Moreover, the sources are not even text in all cases. Last but not least, this process of "transcreating" a translation generally involves a good portion of localisation.
  - Last but not least, the vast majority of translations are being outsourced without transfer of respective translation memories.

Action plan:

Taking into account the main challenges in The Netherlands, corresponding actions to enable / improve the sharing of language resources focus on:

- Establishing good data management practices in public services
- Tackling legal concerns that may prevent the sharing of language resources and
- Identifying and gaining access to outsourced translations.

Especially with regard to the latter, the procurement policy needs to be changed and TMs need to be transferred to the contracting authority. As regards the tackling of legal concerns, a corresponding EU-wide initiative may help.

- Another important action is to increase interest in MT/LT in public services as part of the national digital policy. This includes on the one hand establishing synergies with related national projects and initiatives (which is part of national implementation of Regulation (EU) 2018/1724). On the other hand, it involves securing the support of decision makers to change/adapt national policy. As of 2020 a corresponding national office is foreseen. This department will examine the possibilities and conditions for arriving at national policies. National policy will encourage the use of European standards, including eTranslation. In addition, it is important to diffuse best practices. A corresponding investigation and investment in a national Proof of Concept where the functionality of the eTranslation building block for a department of the national government will be tested seems advisable. Moreover, national initiatives for the promotion of eTranslation need to be installed.

- It is of utmost importance to raise awareness of language data as Open Data and a valuable asset. This includes, above all, the integration of language data in the national Open Data policy/digital agenda, with accompanying relevant metadata (e.g., language of the text, explicit relation between source and translated texts). A first step could be to involve the national platform ECP with the former Minister of Education as the national Digital Champion.

References and further reading list:

Data Portal of the Dutch Government ('Open Data van de Overheid'): https://data.overheid.nl/.

Dutch Public Procurement Expertise Centre: https://www.pianoo.nl/en/public-procurement-netherlands.

Employee Insurance Agency (UWV): https://www.uwv.nl/overuwv/wat-is-uwv/index.aspx.

General information and Security Service (AIVD): https://www.aivd.nl/.

Annex
Country Profile The Netherlands

Military Information and Security Service (MIVD): https://www.defensie.nl/organisatie/bestuursstaf/eenheden/mivd.

National Archives ('Nationaal Archief'): https://www.nationaalarchief.nl/en.

National Open Data Portal: https://data.overheid.nl.

Odijk, Tiberius: *ELRC Workshop Report for the Netherlands*, 2018, http://www.lr-coordination.eu/sites/default/files/Netherlands/2018/ELRC%2B%20Workshop%20Report%20Dutch_Public_final.pdf.

Open Data Netherlands ('Open Data Nederland'): https://opendatanederland.org/.

Open Data Portal House of Representatives ('Open Data Portaal van de Tweede Kamer'): https://opendata.tweedekamer.nl/.

Routing Institute for National and International Information Streams (RINIS): https://www.rinis.nl/en/.

Social Security Bank (SVB): https://www.svb.nl/int/nl/index.jsp.

Taalunieversum: *Veelgestelde vragen over ons taalbeleid (Frequently Asked Questions about our Language Policy)*, http://taalunieversum.org/inhoud/veelgestelde-vragen-over-ons-taalbeleid#t560n4260.

TenderNed: https://www.tenderned.nl/tenderned-tap/aankondigingen.

Verhagen, Michel: *Language services at the Ministry of Foreign Affairs - Hurdles for sharing data*, 2019, http://lr-coordination.eu/sites/default/files/LRB%20Nice/2019/8th%20LRB%20meeting_Language%20services%20at%20MFA_Verhagen.pdf.

Wettenbank Overheid: https://wetten.overheid.nl.

References

Aguado de Cea et. al: *Inventario de Recursos Lingüísticos de la Administración Pública para Traducción Automática*, 2016, https://www.plantl.gob.es/tecnologias-lenguaje/actividades/Estudios%20tcnicos%20y%20de%20gobernanza/Inventario%20de%20recursos%20para%20traducci%C3%B3n%20autom%C3%A1tica/inventario-recursos-traduccion-Retele.pdf.

Azzano, Dino: *Placeable and localizable elements in translation memory systems, A comparative study*, 2011, https://edoc.ub.uni-muenchen.de/13841/2/Azzano_Dino.pdf .

Benoit, Kenneth: "Data, Textual", in: *International Encyclopedia of Political Science*, 2011, https://sk.sagepub.com/reference/intlpoliticalscience/n127.xml.

Directorate-General for Communications Networks, Content and Technology (European Commission):: *Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem*, 2019, https://op.europa.eu/en/publication-detail/-/publication/8494e56d-ef0b-11e9-a32c-01aa75ed71a1/language-en/format-PDF/source-106906783.

Elements of AI: https://www.elementsofai.com.

European Commission: *Building a European Data Economy*, https://ec.europa.eu/digital-single-market/en/policies/building-european-data-economy#usefullinks.

European Commission: *eTranslation Definitions*,https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/CEF+Definitions#CEFDefinitions-eTranslationDefinitions.

European Data Portal: *Analytical Report 9: The Economic Benefits of Open Data,* 2017, https://www.europeandataportal.eu/sites/default/files/analytical_report_n9_economic_benefits_of_open_data.pdf.

European Language Resource Coordination: *New Report on Language Technology for Danish*, 2019, http://lr-coordination.eu/News/New-Report-on-Language-Technology-for-Danish.

European Parliament: *Report on Language Equality in the Digital Age (2018/2018(INI))*, 2018, http://www.europarl.europa.eu/doceo/document/A-8-2018-0228_EN.html.

European Parliament and European Council: *Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on Open Data and the Re-use of Public Sector Information*, https://eur-lex.europa.eu/eli/dir/2019/1024/oj.

French Strategy for Artificial Intelligence, AI for Humanity: https://www.aiforhumanity.fr/en/.

Konrad Adenauer Stiftung: *Open Data. The Benefits, Das volkswirtschaftliche Potential für Deutschland,* 2016, https://www.kas.de/c/document_library/get_file?uuid=3fbb9ec5-096c-076e-1cc4-473cd84784df&groupId=252038.

References

Ministry of Enterprise, Energy and Communications: *ICT for Everyone - A Digital Agenda for Sweden*, 2011, https://de.scribd.com/document/89103979/A-Digital-Agenda-for-Sweden.

NEC TM Data Project: https://www.nec-tm.eu.

Palmer, Alexis: *Computational Linguistics for Low-Resource Languages*, 2011, http://www.coli.uni-saarland.de/courses/CL4LRL/slides/cl4lrl_intro.pdf.

Parliamentary Secretariat for Financial Services, Digital Economy and Innovation, Office of the Prime Minister: *Malta: The Ultimate AI Launchpad, A Strategy and Vision for Artificial Intelligence in Malta 2030*, 2019, https://malta.ai/wp-content/uploads/2019/10/Malta_The_Ultimate_AI_Launchpad_vFinal.pdf.

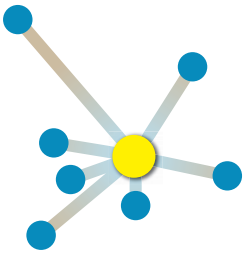
Plan for the Advancement of Language Technology, 2015, https://www.plantl.gob.es/tecnologias-lenguaje/PTL/Bibliotecaimpulsotecnologiaslenguaje/Detalle%20del%20Plan/Plan-Advancement-Language-Technology.pdf.

Slator: *Slator 2019 Language Industry Market Report*, 2019, https://slator.com/data-research/slator-2019-language-industry-market-report/.

Sverigs Rigsdak: *Kulturutskottets betaȁkande 2005/06:KrU4,* 2005, https://www.riksdagen.se/sv/dokument-lagar/arende/betankande/basta-spraket---en-samlad-svensk-sprakpolitik_GT01KrU4.

The Danish Government: *National Strategy for Artificial Intelligence,* 2019, https://eng.em.dk/media/13081/305755-gb-version_4k.pdf.

Uszkoreit, Hans: *What is LT?*, 2010, http://hans.uszkoreit.net/What_is_LT.html.



European Language
Resource Coordination

Connecting Europe Facility

LANGUAGE DATA MATTERS

ISBN-13: 978-3-943853-05-6

9 783943 853056

European Language Resource Coordination