Managing Public Sector Data for Multilingual Applications Development

Stelios Piperidis, Penny Labropoulou, Miltos Deligiannis, Maria Giagkou
Athena RC/Institute for Language and Speech Processing
Epidavrou & Artemidos, Maroussi, Athens, Greece
spip@ilsp.gr, penny@ilsp.gr, mdel@ilsp.gr, mgiagkou@ilsp.gr

Abstract

The current paper outlines the ELRC-SHARE repository, an infrastructure designed and developed in the framework of the European Language Resource Coordination action with the objective to host, document, manage and appropriately distribute language resources pertinent to machine translation, and specifically tailored to the needs of the eTranslation service of the European Commission. Due to the scope of the eTranslation service which seeks to facilitate multilingual communication across public administrations in 30 European countries and to enable Europe-wide multilingual digital services, ELRC-SHARE demonstrates a number of characteristics in terms of its technical and functional parameters, as well as in terms of its data management and documentation layers. The paper elaborates on the repository technical characteristics, the underlying metadata schema, the different ways in which data and metadata can be provided, the user roles and their respective permissions on data management, and, finally, the extensions currently being implemented.

Keywords: public sector language resources, repository, documentation, machine translation

1. Introduction

Machine Translation (MT) technology can be applied to any language pair and adapted to specific domains and text types provided that sufficient amount of Language Resources (LRs) can be made available for training, adapting and evaluating the respective translation engines. The eTranslation digital service infrastructure[1] developed by the European Commission (EC), for example, needs to be adapted to meet the quality requirements of other Digital Service Infrastructures (DSI) like eJustice or Online Dispute Resolution, besides being extended to new language pairs. To meet the objectives of eTranslation adaptation, extension and overall improvement, the EC has launched the European Language Resources Coordination (ELRC) action (Lösch et al., 2018) with the mandate to collect and manage the appropriate LRs for all languages of the EU plus Norwegian and Icelandic, and for a number of domains that are relevant to other DSIs that the EC is building under the umbrella of the Connecting Europe Facility (CEF) programme[2]. Properly collecting, documenting and managing such data and making them available for developing MT engines calls for appropriate LRs management supported by a dedicated repository with the appropriate functionalities catering for the whole LRs lifecycle.

In this paper we focus on the repository infrastructure developed for this initiative, i.e. the ELRC-SHARE repository. Section 2 briefly introduces the ELRC action and provides an overview of the ELRC-SHARE technical characteristics, section 3 describes the underlying metadata schema used for LRs documentation, section 4 explains the different ways in which data, i.e. LRs and their metadata can be provided and managed, section 5 elaborates on the user management layer, while section 6 describes some of the extensions currently being built.

2. ELRC and ELRC-SHARE

2.1. About the ELRC action

Within this framework, the ELRC Network aims to raise awareness, promote uptake and foster the acquisition, identification and collection of LRs, targeting mainly at contributions of open language data from public administrations across the CEF countries. The resources identified, collected and made available as a result of the ELRC initiatives need to be:

- documented with the appropriate information describing the resource (aka metadata)
- easily uploaded and stored in a repository accessible by all relevant actors
- updated as necessary (both metadata & data)
- indexed and, as a result, accessed and downloaded (as necessary) according to the terms and conditions of their use.

The ELRC Network makes use of different channels to offer the above functionalities, the core of which is the ELRC-SHARE repository.

2.2. About the ELRC-SHARE repository

The ELRC-SHARE repository[3] is intended for managing LRs that are considered useful for feeding the CEF eTranslation digital service. It was originally designed to store, document and render accessible LRs, but its role has evolved in accordance with the requirements set for the management of LRs in the ELRC context. It, thus, currently aims to cover the whole lifecycle of LRs: uploading, documentation, uploading of accompanying documents, monitoring and reporting, updating, browsing, delivery and downloading.

The ELRC action is looking for public sector relevant open language data that can be made available for re-use

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

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

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

based on the EU Public Sector Information (PSI) directive[4], but also for potentially restricted datasets (e.g. commercially available datasets, restricted resources pending negotiations/agreements/processing required for privacy protection, etc.). The LRs that ELRC targets can be any of the following:

- collections/sets of textual data (”corpora”) in one or more languages, including for instance:
  - monolingual sets of public administration official documents (e.g. ministerial decisions, legal acts, Board decisions etc.), as well as sets of relevant journal, newspaper, bulletin, blog articles, etc.,
  - bi/multilingual parallel corpora, i.e. sets of original documents with their translations, or ready-made translation memories, i.e. translated documents aligned with their originals.
- lexical/conceptual resources, such as:
  - terminological lexica, glossaries etc., including lists of terms, with or without any other information (e.g. definition, examples, translation equivalents, linguistic information etc.)
  - lists of words, such as person names, names of places, names of products etc.
  - lexica and dictionaries containing words with linguistic information (e.g. part of speech, inflectional information, syntactic frames etc.)
- language descriptions that comprise, for instance, computational formal grammars (i.e. sets of rules that formalize a language), as well as language and translation models (i.e. resources which contain statistical information that assigns a probability to a piece of unseen text, based on some training data).

From a user perspective, ELRC-SHARE offers:

For data contributors: basic functionalities for contributing LRs through a very simple web form

For metadata editors (ELRC Network members): a user-friendly documentation environment for the description of resources (with the ELRC-SHARE metadata schema)

For the general public: a simple and faceted search and browsing of the resources inventory.

At the backend, ELRC-SHARE employs:

- a module for storing LRs, together with their respective metadata records and accompanying documentation (e.g. deposition and licensing documents, validation report);
- a user management module assigning specific access rights to the resources and the repository operations, depending on the user’s role and the publication status of a resource;
- notification and reporting mechanisms for the efficient monitoring of updates of the hosted LRs.

ELRC-SHARE is based on a META-SHARE software instance, its latest version building on META-SHARE v3.1.1. The software has been adapted to the operational needs of ELRC and it has been evolving to respond to specific requirements of its stakeholders. The ELRC-SHARE repository can be replicated on additional servers. However, unlike the META-SHARE distributed network structure (Piperidis, 2012), ELRC-SHARE is deployed as a single repository (i.e. one central, managing, node in META-SHARE terminology) centrally managed by the ELRC consortium. Furthermore, following ELRC requirements, new user roles have been added (namely technical and legal reviewing roles with their own class of access rights), the metadata schema is updated to reflect, for instance, evolution in the open data licensing policies of countries, and new functionalities for reporting, data exporting and packaging have been implemented.

3. ELRC-SHARE metadata schema

ELRC LRs are (formally) documented using the ELRC-SHARE schema. In essence, the ELRC-SHARE metadata schema is an application profile of the META-SHARE schema appropriately modified to meet the requirements of ELRC.

META-SHARE (Gavrilidou et al., 2012) is a generic schema designed for the description of LRs in the wider area of Language Technology. It covers the description of *data* (textual, multimodal/multimedia and lexical data, grammars, language models, etc.) and *technologies (tools/services)* used for their processing. Its main features (that are also inherited to ELRC-SHARE) are:

- flexibility, empowered with a two-level approach, where the initial level (*minimal schema*) consists of a set of basic elements required for at least identifying and accessing a LR, and a second level (*maximal schema*) with a higher degree of granularity of information
- modularity, implemented in the form of *“components”*, i.e. groups of semantically coherent elements, following the Component MetaData Infrastructure (CMDI) recommendations (Broeder et al., 2008)
- standardisation: where possible, controlled vocabularies are preferred over free text for the value space of elements, especially when these can be associated with internationally acknowledged standards, best practices or widespread vocabularies (e.g. ISO 3166 for country codes, RFC 5646 for languages, IANA mimetypes etc.)
- interoperability with other related metadata schemas: links are provided to the conceptually same or similar elements mainly with Dublin Core (DC)[5] and the Data Catalog Vocabulary (DCAT)[6].

[4] https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information

[5] http://www.dublincore.org

[6] https://www.w3.org/TR/vocab-dcat/

The elements are carefully selected to capture **properties** of the resources and **relations** with other resources that are deemed important for their description and relate to any of the stages of the **resource lifecycle**, from production to consumption. In particular, relations are encoded in the metadata records of all resources which have gone through some processing, conversion or transformation indicating the type of relation between the two, essentially different, versions of a resource; e.g. a raw parallel corpus that has undergone alignment gives rise to a new version of the resource described with a new metadata record; the two metadata records (of the original and the aligned version) are linked together with a relation of type "isAlignedVersionOf".

The ELRC-SHARE schema is at the same time a restriction and an extension of META-SHARE: the adaptations include the removal of a substantial set of elements/values as well as the addition of new elements/values. For instance, all elements appropriate for non textual resources have been pruned. On the other hand, the requirements of ELRC have given rise to new elements/values mainly in the categories of licensing and classification. For instance, the legal component was adapted to include

- values for the "licence" element that specify licences recommended for public sector open data
- a value that indicates that a resource falls under the PSI directive
- elements that declare whether a resource includes personal and sensitive data and, as a result, needs special handling and further processing.

The schema includes the following mandatory metadata categories:

- **Administrative information**: features important for the *identification* of a LR (e.g. resource name, a short description of its contents, etc.), its *distribution* (e.g. the access form, i.e. whether it can be downloaded or accessed through an interface and the licensing terms under which it can be used), *contact information* (data and email of a contact person), information on the *metadata record* (e.g. data of the metadata editor, creation and update dates etc.)
- **Technical information**: features on the *language* (number of languages, language identifier(s) and name(s)), *size* of the resource and the *format(s)* in which the resource is available (e.g. plain text, PDF, XML, TMX etc.), and *subtype*, the values of which are specific to each resource type (e.g. terminological glossary, dictionary etc. for lexical/conceptual resources vs. grammar or model for language descriptions).

In addition, the following metadata categories are optional:

- information on the creation of the resource (e.g. the resource creator, whether it involved an automatic process and, if yes, the tool(s) that were used, creation date etc.), links or bibliographic data of documents that describe the resource.
- domain and text type classification, character encoding, description of the types of linguistic or extra-linguistic information (e.g. lemmas, grammatical categories, translation equivalents etc.) contained in lexical/conceptual resources and language descriptions.

It should be noted here that, although optional, **domain**, and in many cases, text type **classification** are of utmost importance to the ELRC objectives. The Eurovoc thesaurus[7] and, more specifically, the two upper levels of the hierarchy have been selected to ensure uniform domain classification across the resources and alignment with the domain classification already employed by the Directorate-General for Translation of the EC and by other European institutions. Due to the special focus on the reuse of the eTranslation service by other CEF sector-specific DSIs, an additional metadata element has been introduced, that of relevance of the LR to a DSI, thus facilitating retrieval of resources that can be used for training machine translation engines tailored to the domain(s) of a specific DSI. Finally, for the text type classification, we have devised our own controlled vocabulary based on the types of texts that are typical of the Public Sector (e.g. administrative texts, bulletins, legal documents etc.) and the ones we expect to obtain from our sources.

4. Contributing, documenting and managing resources

Contributing resources through the ELRC-SHARE repository is a process that is kept as simple as possible given that the intended contributors are primarily Public Sector employees who are not necessarily familiar with the concepts of, and around, language resources.

Providing resources and the required documentation is possible only after users have registered to the system. They can then use a simple web form to describe and upload LRs. For each LR, at least a short description and a title in English must be provided; the resource itself can be uploaded in zipped format (currently up to 100MB) or, alternatively, users may supply a link to a URL where the resource can be downloaded from.

Following this submission, the contributed data and basic metadata are received by the repository administrators and are imported into the repository. For each new contribution, proper notification email alerts are sent to designated ELRC consortium members (metadata editors) who have the responsibility to enrich the documentation, check the resource and publish it. The ELRC member may contact the contributor to ask for further information about the resource and its licensing conditions. The documentation and checking are supported with an editor form that implements the full ELRC-SHARE schema. During the documentation process, the resource descriptions are considered "internal" and can be viewed only by authorised editors of the ELRC consortium. When the documentation and the licensing conditions under which the resource is provided are finalised, the resource description (metadata record) will be made public through the inventory and the resource itself will be downloadable by users, in

[7] http://eurovoc.europa.eu/

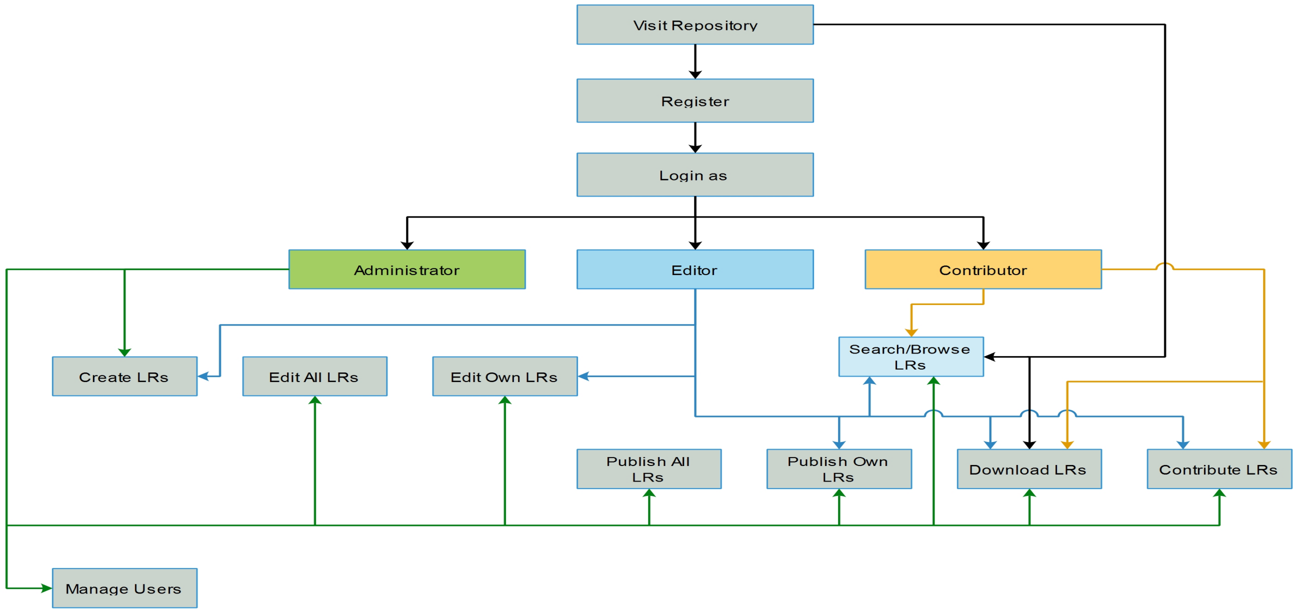

Figure 1: Available user rights per user role

accordance to its licensing conditions.

Contributions may also come from ELRC members using the editor interface, in which case the LRs are deposited straight into the repository, again as a zipped file. To upload the resource, the minimal information must have already been encoded for the resource. The description can still continue to be updated at any time, before and after uploading.
In addition, ELRC resources may be generated by using an automatic language resource discovery (Papavassiliou et al., 2018) and compilation system (ILSP-FC). ILSP-FC ((Papavassiliou et al., 2016); (Papavassiliou et al., 2013)) is essentially a pipeline of tools that, given a set of seed URLs, and optionally a domain profile in the form of a list of domain specific terms, crawl the web, fetch web pages, check their domainness, pair the translated web pages and align them at sentence level, thus generating a TMX file. The ILSP-FC system automatically extracts the required metadata and renders them in the ELRC-SHARE profile form, ready to be automatically uploaded to the repository. All LRs residing in the repository undergo a systematic validation procedure (Lösch et al., 2018), the results of which are coded in the respective validation reports stored together with any other ancillary documentation accompanying the resource.
The result of the documentation process is compiled into the ELRC-SHARE inventory of resources comprising all the descriptions (metadata catalogue).[8]

5. Managing access and editing rights

Access to ELRC-SHARE is regulated by user roles and permissions. Unregistered users can only browse and search the repository and view the resource descriptions. Registered users can be assigned one of the following roles, with respective rights and permissions: a) *contributor* (default), b) *editor*, c) *administrator*. Figure 1 illustrates the available user rights per role within the repository.
The resources hosted in the repository are assigned a Publication Status which determines the visibility of their descriptions to different user types. The Publication Statuses are:

- **internal**: used as the initial status for all metadata records;
- **ingested**: LRs with enriched and initially validated metadata records;
- **published**: LRs with finalised metadata records. Published records are available for searching and browsing on the public inventory of resources.

The combination of user access rights and resource status yields an enhanced user management module that gives resource owners extra leverage and control over the visibility and accessibility of the repository resources. Currently, the ELRC-SHARE inventory lists 225 published language resources, with 200 additional and/or derivative resources in ingested mode, i.e. waiting to be published. The evolution of the inventory is reported in the form of simple spreadsheets, while specific APIs for querying the metadata inventory are under implementation.

6. Extensions currently under way

In order to help public administrations across borders share electronic data and documents in a secure, reliable and trusted way, ELRC-SHARE is being endowed with an eDelivery[9] Access Point (AP), as an alternative to direct contribution and uploading. ELRC-SHARE will thus ensure that

[8] The inventory with the LR descriptions is licensed under the Creative Commons Attribution (CC-BY) version 4.0 or higher.

[9] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eDelivery

any sensitive or confidential data transferred from public administrations into the repository are encrypted and secure.

ELRC-SHARE is deploying an eDelivery AP using an appropriately configured Domibus EC implementation, which is an ebMS3 AP based on the e-SENS AS4 Profile. Access Points can have a message "sender" or/and message "receiver" role. Since the ELRC-SHARE Access Point will be receiving data from public administrations, it will act as a receiver in the message exchange process, accepting messages and data from public administrations that already have an available sender Access Point.

Finally, ELRC-SHARE is currently being extended to allow for the cataloging and documentation of language processing tools and services, e.g. aligners, annotators, named entities recognizers and other tools pertinent to machine translation. In the future, ELRC-SHARE will support the direct processing of resources with available tools and services. LRs will be associated with adequately documented available tools/services (e.g. text normalization for encoding detection, UTF-8 conversion and metadata extraction, language identification, text classification, deduplication, translation pair detection, text alignment, etc.), which produce new resources that will be stored in the ELRC-SHARE repository as new versions or derivatives of the original resource.

7. Acknowledgements

The European Language Resource Coordination (ELRC) is a service contract operating under the EC's Connecting Europe Facility SMART 2014/1074 programme from April 2015 to April 2017 and continues until December 2019 under SMART 2015/1091 LOT 2 "Language Resource coordination and collection with related legal and technical work" and SMART 2015/1091 LOT 3 "Acquisition of additional Language Resources and related refinement/processing services and their provision of the Language Resource Repository of CEF Automated Translation Platform". We would like to thank all ELRC consortium members, DFKI, TILDE and ELDA, and the numerous contributors from the 30 CEF-affiliated countries for contributing their language resources.

8. Bibliographical References

Broeder, D., Declerck, T., Hinrichs, E., Piperidis, S., Romary, L., Calzolari, N., and Wittenburg, P. (2008). Foundation of a component-based flexible registry for language resources and technology. In Nicoletta Calzolari, et al., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S., Papageorgiou, H., Monachini, M., Frontini, F., Declerck, T., Francopoulo, G., Arranz, V., and Mapelli, V. (2012). The META-SHARE Metadata Schema for the Description of Language Resources. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC'12)*, Paris, France, 5. European Language Resources Association (ELRA), ELRA.

Lösch, A., Mapelli, V., Piperidis, S., Vasiļjevs, A., Smal, L., Declerck, T., Schnur, E., Choukri, K., and van Genabith, J. (2018). European Language Resource Coordination: Collecting Language Resources for Public Sector Information Management. In *Proceedings of the 11th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA).

Papavassiliou, V., Prokopidis, P., and Thurmair, G. (2013). A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 43–51, Sofia, Bulgaria. Association for Computational Linguistics.

Papavassiliou, V., Prokopidis, P., and Piperidis, S. (2016). The ILSP/ARC submission to the WMT 2016 Bilingual Document Alignment Shared Task. In *Proceedings of the First Conference on Machine Translation*, pages 733–739, Berlin, Germany. Association for Computational Linguistics.

Papavassiliou, V., Prokopidis, P., and Piperidis, S. (2018). Discovering parallel language resources for training MT engines. In *Proceedings of the 11th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA).

Piperidis, S. (2012). The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).