Proceedings of the TDLE Workshop @LREC2022, pages 27–35
Marseille, 20 June 2022

Collaborative Metadata Aggregation and Curation in Support of Digital Language Equality Monitoring

Maria Giagkou, Penny Labropoulou, Stelios Piperidis, Miltos Deligiannis, Athanasia Kolovou, Leon Voukoutis
Institute for Language and Speech Processing, Athena Research Centre
{mgiagkou, penny, spip, mdel, akolovou, leon.voukoutis}@athenarc.gr

Abstract

The European Language Equality (ELE) project develops a strategic research, innovation and implementation agenda (SRIA) and a roadmap for achieving full digital language equality in Europe by 2030. Key component of the SRIA development is an accurate estimation of the current standing of languages with respect to their technological readiness. In this paper we present the empirical basis on which such estimation is grounded, its starting point and in particular the automatic and collaborative methods used for extending it. We focus on the collaborative expert activities, the challenges posed, and the solutions adopted. We also briefly present the dashboard application developed for querying and visualising the empirical data as well as monitoring and comparing the evolution of technological support within and across languages.

Keywords: language resources, language technologies, metadata aggregation, digital language equality

1. Introduction

With a large and all-encompassing consortium consisting of 52 partners covering all European countries, research and industry and all major pan-European initiatives, the European Language Equality (ELE)[1] project develops a strategic research, innovation and implementation agenda (SRIA) as well as a roadmap for achieving full digital language equality in Europe by 2030. Key component of the SRIA development process is an as accurate as possible estimation of the current standing of languages spoken in Europe with respect to their technological readiness. In turn, such estimation presupposes the existence of the necessary data, resources and services that underlie and reflect onto technological readiness.

The META-NET White Papers series (Rehm and Uszkoreit, 2012) reported, back in 2012, that more than 21 European languages were in danger of digital extinction. Despite the vast improvements in language technology (LT) performance in the last couple of years, technology support for Europe's languages is still characterised by a stark imbalance. While many resources and technologies exist for English and some of the most widely spoken European languages, the majority of other languages still suffer from lack of technology support, as attested in the Language Reports series initiated by the ELE[2] (Giagkou et al., 2022). Digital Language Equality (DLE), as conceived in the ELE project, is defined as "the state of affairs in which all languages have the technological support and situational context necessary for them to continue to exist and to prosper as living languages in the digital age" (Gaspari et al., 2021). The Digital Language Equality (DLE) Metric (Gaspari et al., 2021, Gaspari et al., 2022) is a measure that reflects the digital readiness of a language and its contribution to the state of technology-enabled multilingualism, tracking its progress towards the goal of DLE. The DLE Metric is computed for each language on the basis of various factors, grouped into technological support (technological factors, e.g., count of the available language resources, tools and technologies) and a range of situational context factors (e.g., societal, economic, educational, industrial factors)[3].

In close collaboration with its sister project, the European Language Grid (ELG)[4] (Rehm et al., 2020; Rehm et al., 2021), ELE makes use of the ELG platform functionalities and catalogue contents as the empirical base for calculating the technological factors of the DLE metric. This decision is based on the fact that the ELG catalogue is Europe's most comprehensive registry of language resources and tools/services. Despite its comprehensiveness, the ELG catalogue is not exhaustive; language resources are produced at a much higher rate than ever before due to the dominant data driven methods in language technology research and development. In addition, a number of initiatives in Europe, domain specific and general, are engaged in data and service registration activities. Therefore, the decision has been made that the ELG platform and its catalogue are further enriched by two separate procedures: (a) harvesting existing catalogues of major infrastructures and initiatives in Europe (e.g., CLARIN, ELRC, Zenodo), and (b) by an unprecedented collaborative metadata collection procedure undertaken by language experts covering over 70 languages, i.e., all the EU official languages as well as a great number of Europe's regional and minority languages and dialects[5].

All metadata resulting from these enrichment activities are not only available through the ELG catalogue, but they are also queryable through a dashboard. The ELE dashboard allows to interactively visualise the indicators of the level of LT support for the languages covered by the project, providing a detailed, empirical and dynamic map of technology support for European languages and dialects.

This paper discusses the processes used for extending the coverage of the ELG catalogue, the challenges posed, and the solutions adopted. Section 2 briefly presents the contents of the ELG catalogue and the automatic processes

[1] https://european-language-equality.eu/
[2] The research partners have prepared updates of the META-NET White Papers (Rehm and Uszkoreit, 2012) available at https://european-language-equality.eu/deliverables/ including the results of the survey.
[3] For the full list of the factors, see Gaspari et al. (2022).
[4] https://www.european-language-grid.eu/
[5] https://european-language-equality.eu/languages/

that were put in place in order to enrich the catalogue's coverage mainly through harvesting protocols and API-based access to catalogues of major European infrastructures, platforms, and initiatives. Section 3 elaborates on the collaborative metadata collection process initiated by ELE and Section 4 briefly sketches a relaxed version of the ELG metadata schema[6] (Labropoulou et al. 2020) to accommodate input from lighter schemata. In Section 5, we briefly present the ELE dashboard, and conclude with some general observations and plans for the future.

2. ELG catalogue and automatic enrichment procedures

The European Language Grid tries to tackle the observed fragmentation in the European Language Technology landscape (Soria et al. 2012) by bringing together Language Resources and Technologies (LRTs) and to support and boost the LT sector and LT activities in Europe through multiple multilevel services. ELG already provides a scalable cloud-based platform[7] through which developers and providers of LRTs can not only deposit and upload them into the ELG, but also deploy them through the grid platform. ELG offers access already to thousands of commercial and non-commercial LTs and ancillary Language Resources (LRs) for all European languages and more; these include processing and generation services, tools, applications for written and spoken language, as well as corpora, different types of lexical resources, language models and computational grammars, etc.

For the further population of the catalogue of its platform, ELG has built bridges to existing initiatives and reaches agreements for harvesting and importing information (aka metadata) and resources from other infrastructures, platforms and repositories under mutually agreed conditions and attribution of the source.

Currently, ELG has implemented a client compliant with the Open Archives Initiative Protocol for Metadata Harvesting[8] (OAI-PMH) (Lagoze et al. 2012) that supports harvesting from other repositories which expose their metadata via an ELG-compatible OAI-PMH endpoint.

OAI-PMH is used for harvesting LINDAT/CLARIAH-CZ[9], i.e., the Czech CLARIN national node, as well as the Polish (CLARIN-PL[10]) and Slovene (CLARIN-SI[11]) CLARIN nodes, given that they use the same repository software as LINDAT. Such harvesting procedure benefits from the fact that the ELG metadata model (Labropoulou et al., 2020) builds on the META-SHARE metadata model (Gavrilidou et al., 2012), while the LINDAT DSpace software supports the export of metadata in the META-SHARE minimal schema.

The same harvesting approach is followed for the harvesting of metadata records from the ELRC-SHARE repository[12], which is used for the storage of and access to language resources collected through the European Language Resource Coordination[13] initiative (Lösch et al., 2018) and considered useful for feeding the CEF Automated Translation (CEF.AT) platform[14]. The ELRC-SHARE repository (Piperidis et al., 2018) uses a metadata schema based on the META-SHARE schema tuned to text resources for Machine Translation purposes.

A different procedure (Figure 1) has been implemented for Hugging Face[15] (Wolf et al., 2019), which includes a large collection of Machine Learning (ML) models and datasets that can be used for training models, with a focus on transformers. Hugging Face exposes two distinct APIs with JSON files for datasets and models respectively, including a subset of the metadata elements displayed on their catalogue. However, not all records have values for all of the elements. Since importing into ELG presupposes that at least the mandatory elements of the minimal version are filled in, the conversion and import of records from Hugging Face into ELG has so far been restricted to datasets with at least the description, language and licence elements filled in, as these are deemed the minimum threshold for findability and usability purposes in ELG. A conversion process has been set up based on the mapping of the elements and controlled vocabularies values. Further enrichment of the resulting records has been performed for specific elements, notably the licencing information, while, where required, default values have been used for mandatory elements whose values could not be inferred from the original metadata records (e.g., all datasets have been assigned the "text" value for "media type"). Records for which the above processes did not render the mandatory elements were discarded.

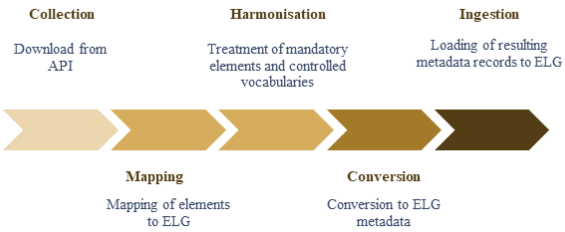

Figure 1: Workflow for the import of Hugging Face metadata records into ELG

General repositories like Zenodo[16] pose different challenges, the main one being as precise as possible filtering of the candidate records. Zenodo exposes metadata records in two channels: through a REST API[17], which outputs records as JSON files, and an OAI-PMH API[18] in a set of standard metadata formats, namely DC[19] (International Organization for Standardization 2017), DataCite[20] (DataCite Metadata Working Group 2021),

[6] https://gitlab.com/european-language-grid/platform/ELG-SHARE-schema
[7] https://live.european-language-grid.eu/
[8] https://www.openarchives.org/pmh/
[9] https://lindat.mff.cuni.cz/
[10] https://clarin-pl.eu/dspace/
[11] https://www.clarin.si/repository/xmlui/?locale-attribute=en
[12] https://www.elrc-share.eu/
[13] https://lr-coordination.eu/
[14] https://language-tools.ec.europa.eu/
[15] https://huggingface.co/
[16] https://zenodo.org/
[17] https://developers.zenodo.org/#rest-api
[18] https://developers.zenodo.org/#oai-pmh
[19] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[20] https://schema.datacite.org/

MARC21[21] (Library of Congress 1999) and DCAT[22] (Albertoni et al. 2020). With regard to import, the preferred solution is the OAI-PMH protocol, which is rate limited, hence not appropriate for big amounts of metadata records. We have, therefore, resorted to a combined solution: we have downloaded the automatically generated full dump of 2,060,674 metadata records included in Zenodo until 31/08/2021. For records added to Zenodo after this date, we are incrementally harvesting from the OAI-PMH endpoint, adding 147,621 records during a four-month period. From the resulting 2,208,295 metadata records available until 31/12/2021, 592,509 entries of type "dataset" and "software" were filtered; we are experimenting with high-precision filtering methods on these to identify records of interest for LT purposes. The conversion of the metadata records is based on the DCAT metadata schema, the richest among the ones exposed by Zenodo, while certain relaxations of the ELG schema proved necessary to take into account the DCAT features (see Section 4). Figure 2 depicts the workflow for metadata records downloaded from the OAI-PMH server.

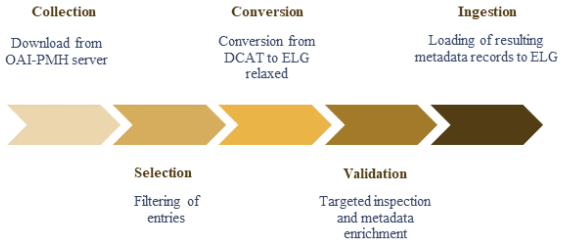

Figure 2: Workflow for the import of the Zenodo metadata records into ELG

At the time of writing, the ELG catalogue includes 977 metadata records harvested from the CLARIN nodes, and 1,299 records from ELRC-SHARE. In addition, 385 records for datasets have been imported from Hugging Face, while the conversion for models is ongoing, as is the import from Zenodo.

3. Collaborative ELE metadata collection

With the ELG Catalogue as basis and point of departure, ELE initiated a large-scale metadata collection activity in order to create an as representative as possible base on which the technological readiness of languages spoken in Europe would be estimated. At least 40 different organisations[23], ELE consortium members and other collaborators of the partners' networks, acted as language expert informants for one of the official, co-official, regional, minority and community European languages. They investigated, discovered, and appropriately documented LRTs that contribute to a language's level of technological support. These include LT tools and services, as well as language resources that can be used for the development of LT, i.e., corpora and datasets, language descriptions (language models and computational grammars), and lexical/conceptual resources. Given the availability of the respective information, the language informants additionally recorded the research or industrial providers of LRTs and the project(s) in the framework of which the LRTs have been developed.

3.1 ELE metadata collection instruments

The ELE partners were asked to only document resources that were not already included in the ELG catalogue and were thus provided with a list of its contents at the time of conducting the metadata collection.

They were given the option to describe the resources they discovered using the metadata editor which is available in the ELG platform, and/or an online form[24], and/or a spreadsheet which was automatically populated by the responses to the online form and accessible for direct manual editing and bulk import of records.

The online form (and linked spreadsheet) was appropriately configured to render a very simplified version of the ELG metadata schema. By adhering to and utilising the ELG schema, interoperability with the ELG platform was guaranteed, thus allowing for the aggregation and ingestion of the LRTs documented by the ELE partners into ELG in an as automated as possible manner. On the other hand, having set as a priority the documentation of as many LRTs as possible over a detailed documentation for each of them, and in order to respond to the variety of sources from which the ELE informants would discover relevant information, only a subset of the ELG metadata categories have been included in the ELE online form. These were carefully selected to elicit sufficient information for the ELE purposes.

The online form contained the following metadata categories (elements marked with an asterisk were mandatory):

- identification: *resource type**, *resource name**, *resource short name*, *landing page**, *description**, *publication year*, *resource provider (organisation name)*
- contact data: *name* & *homepage of source*, *contact email*
- classification: *keyword*, *domain*
- funding information: *funding project* & *funding type*
- usage information: *licence*, *access rights*
- technical information for data resources: *subclass**, *language** and, where applicable, *geographical variety*, *multilinguality type*, *media type**, *size*; in addition, for annotated corpora, *annotation type*, and, for lexical/conceptual resources, *encoding level*
- technical information for tools/services: *function**, *Technological Readiness Level (TRL)*, whether they are *language independent**, and if not, the *language* and, where applicable, *geographical variety* of the input resource, *media type* of the input resource, and, optionally, *language*, *geographical variety* and *media type* of the output resource.

Recommended controlled vocabularies, in the form of lists of values from which users could select a value, were used where possible (e.g., for language), yet informants could also add free text values. Depending on the element, adding multiple values was possible (e.g., for domains,

[21] https://www.loc.gov/marc/bibliographic/
[22] https://www.w3.org/TR/vocab-dcat-2/
[23] https://european-language-equality.eu/languages/
[24] The online form template is available at https://forms.gle/WjJZ1CZqXDPQjPHA8

languages, keywords, etc.) Mandatory elements were marked as such with validation imposed.

3.2 Curation process

This systematic collection resulted **in 6,790 new metadata records created by the ELE language experts**. Before being imported to the ELG database, these records were curated (Figure 3). The curation process concerned (semi-) automatic and manual processing of the records with the aim to ensure that they adhere to the "relaxed" version of the ELG metadata schema (Section 4) and that they can be imported in the catalogue, as well as to harmonise values and thus enhance their discoverability and contribute to more reliable statistics.

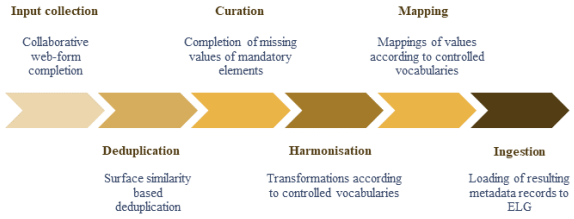

Figure 3: Workflow for the import of the ELE survey results into ELG

3.2.1 Deduplication

Duplicate records were identified first by checking the resource name, and then by inspecting those that had the same resource short name and landing page. We thus identified duplicates that had different names (e.g. "Corpus Web Salud Español" - "Spanish Biomedical Crawled Corpus", "Comprehensive Estonian-French Dictionary" - "Grand dictionnaire estonien-français"). Some records were identified as duplicates of existing ELG records while others were duplicates of other ELE informants' contributions. When the duplicate records contained different or contradicting values, e.g., different functions, licences, etc. the source was consulted, and the record was manually corrected.

3.2.2 Completion of mandatory metadata

The ELG metadata schema includes a set of mandatory elements deemed important for the documentation of LRTs, e.g., resource name, description, language and media type. Missing values in mandatory metadata result in invalid records and failure of import to the database. To minimise loss of data due to missing mandatory values, we resorted to a combination of solutions:

- We used heuristics to add the missing values, where possible. For instance, using the size unit values, keywords and/or hints in the description or resource name we automatically inferred and assigned the media type value; e.g., "text" was selected for records with size unit values such as articles, translation units, texts, etc., or containing the terms "web corpus", "Wikipedia", etc. in the description or in the resource name.
- If no value could be automatically assigned, we consulted the source and manually filled in the missing values, where possible.
- For remaining records, when the data type of the element permitted this, we used the value "unspecified".

3.2.3 Harmonisation and mapping of metadata values

The ELG schema adopts controlled vocabularies for the value space of specific metadata elements (e.g., media type, language, service function, annotation type, size unit, etc.). For some of them (e.g., service function), free text values added by users are also allowed.

During the curation, where possible and appropriate, the free text values added by ELE informants were semi-automatically mapped to values of the controlled vocabularies or aggregated under the same value. For instance, values such as "speech synthesis", "speech synthesizer", "text to speech", "TtS" for the service function element were all mapped to "Speech synthesis". For certain elements (e.g., for "domain"), broader terms were also added, to improve findability. For instance, records with the domain values "travel", "transport" "geography", "hotel" were assigned also the value "Geography, Travel & Tourism".

In addition, in the case of closed controlled vocabularies, i.e., vocabularies that do not allow the use of free text values, unmappable values left as is would result in invalid metadata records. Therefore, for specific elements deemed important for the adequate representation of resources, we manually inspected the description of the records and/or source in order to select the appropriate value. This is the case, for instance, of the element "subclass" used to distinguish models from computational grammars, as well as of "language", which is discussed in Section 3.2.4.

Moreover, in a first attempt to narrow down the wide range of size units used, for text corpora that specified size in sentences, an additional size in words was computed based on the calculation of average sentence length in words, per language, in the Universal Dependency Treebank.

Finally, despite the harmonization of service function values, the list was deemed too long for eliciting meaningful statistical observations. For ELE purposes, a set of six higher-order concepts were put forward: "Text Processing", "Speech Processing", "Translation Technologies", "Image/Video Processing", "Human Computer Interaction", "Natural Language Generation", "Information Extraction and Information Retrieval", "Support operation" and "Other". Given the fact that the values of the metadata element "service function" are from the OMTD-SHARE ontology[25] (Labropoulou et al., 2018), and most specifically the "Function" class, the grouping of the values has been made at the ontology side, and thus used for all tools and services included in the ELG catalogue. Some of the group values were already included in the ontology, but the classification of the functions could not serve ELE purposes as is. We thus decided to represent the groups as SKOS Collections and not interfere with the existing hierarchy.

3.2.4 Treatment of language values

Language occupies a central place among the documentation elements for language resources and tools. Its standardisation is therefore important while its value space must cater for the representation of language

[25] http://w3id.org/meta-share/omtd-share/

varieties, regional variants, idiolects, time delimited language forms, etc. The ELG schema has adopted the RFC recommendation[26] (Phillips and Davis, 2009), which combines the ISO 639 vocabulary[27] (International Organization for Standardization, 2007) with additional subtags for region, script and variants. Yet there are still language varieties not covered by even the ISO 639-3 part[28] (the most extensive part of the ISO 639 standard). For this reason, we use two additional elements, namely the element "glottolog code" which takes values from the Glottolog vocabulary[29] (Hammarström et al., 2021) and the "language variety" element, which takes free text values. These elements are used alongside the language subtag in the following way:

- When there's an equivalence link between the ISO value and the Glottolog code, both are added at the respective elements and the language name displayed on the ELG catalogue is that of the official name from the ISO list; this has the benefit that we can exploit alternative names from the linked data contained in Glottolog for the enhancement of search functionalities.
- If a language variety (e.g., "Abenaki") is not included in the ISO list, the value "mis" (uncoded languages) is used for the ISO value element, the Glottolog code is added and the language name displayed on the ELG catalogue is derived from Glottolog, thus serving as an additional standardization measure.
- If a language variety is not included in either of the two (e.g., "Valbonnais dialect"), the respective language name is added to the "language variety" element.

For the ELE web form, we decided to ask informants to document only the language(s) and optionally the country/region subtags of the LRTs, where appropriate and necessary. For instance, if they needed to document a resource containing Austrian German, they could indicate "German" as the language value and "Austria" as the geographical variety value. For the "language" element we added as valid values the set of names of the European languages targeted by the project, and also allowed for user added values, so that they could add languages from other countries and language varieties.

The output records included many free text values, even for cases included in the pre-filled values (e.g., alternative values such as "Greek", "Modern Greek", "el", "ell", values from different parts of the ISO 639 vocabulary, typos, etc.). Unique language values were extracted from the list and mapped to the controlled vocabularies according to the policy described above. To do so, we went through a series of repeated rounds of automatic checks, based on exact and similar match to the language identifiers and names from the ISO 639 and Glottolog vocabularies, and manual inspection and corrections[30].

The "language geographical variety" values were also harmonized and mapped to the ISO 3166 country codes[31] when possible. For regions without an ISO code (e.g., "Lower Saxony"), a value was filled in at the "language variety" element (e.g., "Variety in Lower Saxony").

3.2.5 Treatment of licensing information

Licensing information is critical for the (re-)usability of any resource and thus required in the ELG schema. For the ELE survey, anticipating that the licence value might be difficult to fill in, the web form included the "licence" element with a list of the most popular standard licences and an option for adding free text, as well as the "access rights" element with a choice between three values, namely "Licensed without a fee for all uses", "Licensed without a fee for specific uses" and "Licensed with a fee". Informants were asked to fill in at least the "access rights", which is required in the "relaxed" version of the schema (see Section 4); for standard licences, the mapping to the "access rights" would be provided by the ELG/ELE core team.

However, both elements were filled in with diverging values that needed to be harmonised and mapped in the follow up curation process. Specifically:

- Use of alternative values for the same licence (e.g., "CC-BY-4.0", "Creative commons attribution 4", etc.)
- Reference to a licence with multiple versions, without any indication of the specific version (e.g., "Creative commons attribution").
- Reference to a non-standard licence by name and no further information on the licensing terms or a hyperlink to the licence text
- Use of a free text value for licence and/or access rights besides the ELE recommended ones, such as "free for academic use", "available for research", "Copyright 2012", "not currently accessible to the public", etc.
- Total absence of a value for both elements.

Overall more than 300 values in these two elements could not be matched to known licences. Through semi-automatic and manual checks, often through searches for the specific licences, we have curated both elements, keeping the "licence" element as originally conceived in ELG (i.e., with a name and URL) and extending the notion of "access rights" to allow for any free text value. Thus, "licence" was used only when a URL with the licensing terms was found and alternative names were all mapped to a single value; if available, the name as it appears in the SPDX list of licences[32] was selected. Licences with an unspecified version were harmonized (e.g., "Creative Commons Attribution") and added as "access rights" values. Records with no licence and no access rights were added with the value "unspecified" for the access rights.

An additional element, namely "condition of use", is used for the representation of licensing information. This element takes values from a subset of popular conditions of use associated with licences (e.g., no derivatives, non-commercial use, etc.) and is deemed important for findability purposes. It was additionally deemed necessary for the calculation of the technological part of the DLE

[26] https://datatracker.ietf.org/doc/html/rfc5646
[27] https://www.iso.org/iso-639-language-codes.html
[28] https://iso639-3.sil.org/
[29] https://glottolog.org/
[30] From the initial 1,147 unique language values contained in the spreadsheet, only 937 were matched with languages in the ISO 639 set at the first step. For all remaining values, a semi-automatic curation was required, resulting in 1,263 unique values.
[31] https://www.iso.org/iso-3166-country-codes.html
[32] https://spdx.org/licenses/

metric, as it provided a higher-level representation and approximation of the "openness" scale of language resources. The appropriate "conditions of use" values are assigned to standard licences by the ELG legal team, or, in the case of non-standard licences, by the metadata creators when they describe a resource. The "access rights" values added through the ELE metadata collection have also been mapped to the same values supporting queries about resource accessibility in the DLE dashboard (Section 5).

3.3 Metadata conversion and ingestion

During metadata curation and processing, approximately 400 records of the initial 6,790 records have been discarded, mainly because of duplicates or incomplete mandatory metadata that could not be recovered.

The remaining records were automatically converted into ELG-compliant metadata records. As a result, **6,362 records** have been imported into ELG, consisting of 2,215 metadata records describing LT tools/services and 4,147 records describing data resources, i.e., corpora, lexical/conceptual resources and language descriptions (grammars or language models). They cover all the languages addressed by the ELE language reports series (Giagkou et al., 2022), i.e., the 24 official EU languages plus some other (co)official languages at the national or regional level (Norwegian, Icelandic, Serbian, Bosnian, Basque, Catalan and Galician), as well as the additional languages and dialects targeted by the ELE project.

All the metadata records are marked in the ELG catalogue as "for information", indicating that they include only a limited set of metadata elements, and they can be "claimed" for further enrichment by their owners, following the respective ELG policies and operations. Dissemination activities have been undertaken to inform persons designated as contact points for these resources as well as the broader community members about the ELE metadata collection results and their import into ELG.

3.4 Organisations and projects

Although the ELE survey focused on LRTs, the information collected was also used for the enrichment of the ELG inventory of organizations and projects, which are then automatically linked with their related LRTs in the ELG Catalogue.

More specifically, the element "resource provider" contains companies, academic institutions, public institutions, etc. that are active in the LT domain. After a round of cleanup (e.g., person and project names were included among the values) and harmonization (e.g., for alternative names and typos), these were imported and they are published in the catalogue.

A similar process of curation is ongoing for the publication of the funding projects. This process seeks to add missing mandatory values and assign the mixture of values that were filled in for "project name" to the appropriate metadata elements; indicatively, this was filled in with project names in various languages, identifiers, grant award numbers, funder names, funding programmes, etc.

4. Metadata schema adaptations

Achieving metadata interoperability across repositories is a challenging task due to the diversity and granularity of schemas used by different communities, intended purposes, types of resources described, etc. and various methods are utilized to address it (Alemu et al. 2012, Chan & Zeng 2006, Broeder et al. 2019, McCrae et al. 2015, Zeng & Chan 2006). The approach presented in this paper is based on the mapping of the source schemas into the target (ELG) schema, as well as on the enrichment of the source records with information required when this is possible without misconceptions and inconsistencies. Yet, this does not suffice for automatically aggregating records from the sources presented above.

More specifically, to be imported into the ELG platform, metadata records must comply with the minimal version of the ELG schema, i.e., the values must respect the designated data type of the elements and at least some mandatory metadata elements must be filled in. However, for metadata records automatically imported from other catalogues and repositories, as well as in sizable collaborative initiatives, such as the metadata collection undertaken by the ELE experts, the demand for filling in even the minimal version was considered challenging. The modifications required to accommodate such a collaborative population scenario resulted in the "relaxed" version, which can only be used in such cases.

The "relaxed" version of the ELG metadata schema aims to accommodate "mismatches" between the ELG schema and schemas with lighter information requirements. The main features characterising this version are the introduction of alternative elements for mandatory metadata elements that may be missing from the source records or elements that have different data types.

The first case refers to two elements that are deemed important for ELG purposes: "media type" and "licence".

- The "media type part" element is crucial for ELG purposes, as it is used for attaching important metadata properties, such as language, format, size, etc. Therefore, even in cases where these elements are included in the source records, they cannot be imported into ELG if the "media type part" value is missing. For these cases, the value "unspecified media part" can be used.
- Licence is crucial for re-usability purposes; for a licence, both a name and a URL hyperlink to the legal document with the terms and conditions are required. However, in many cases, such as legacy resources, or records in catalogues allowing free text as licence value, these two elements cannot be determined. Therefore, the "access rights" element that takes a free text value may be filled in as an alternative to "licence", specifying the rights of access and use at a higher level of abstraction.

The second case refers to metadata properties, such as size, which in the ELG schema are represented as a combination of two elements – "amount" and "sizeUnit"– while in other schemas and catalogues a single free text element is used. In this case, a new element that takes free text as a value (e.g., "sizeText") has been added in the schema as an alternative to the combination.

5. ELE dashboard

To provide a mechanism for exposing and monitoring the technological (TFs) and contextual factors (CFs) that contribute to the DLE metric (Gaspari et al., 2022), we designed and implemented an interactive dashboard as part of the ELG platform. The dashboard exposes the TFs

(based on the contents of the ELG catalogue) and the CFs as interactive visuals dynamically created by user queries.
With regard to the TFs, as the ELG catalogue organically grows over time, the resulting DLE Metric scores will be updated for all European languages, thereby providing an up-to-date and consistent measurement of the level of LT support and provision that each of them enjoys, also showing where the status is less than ideal or not at the expected level. Similarly, the situational indicators that are reflected by the CFs will be updated for the relevant languages on up-to-date data, as it becomes available from the selected sources.
The user interface of the ELE dashboard, which can be accessed through the ELG platform[33], consists of three entry points (sections). The first section displays the bar graphs of the DLE metrics for CFs and TFs for the languages selected by the user (see for instance Figure 4 in the Appendix). In the other two sections users can dive into a more detailed comparison of a subset of the TFs across languages and within a language respectively. The comparison can be made on datasets vs. software resources and, by selecting one of the two, for a number of features characteristic of the corresponding resource class. For datasets, these are the resource subclass, the linguality type, the media type and the access rights. For software, the available query criteria are: service function groups, input and output media types and access rights.
Architecturally, the ELE dashboard consists of two layers. The ELG database provides the source data to be exposed, in particular the source data for the technological factors that contribute to DLE. The ELG database contents are indexed and saved in appropriate JSON structures. Each user query retrieves the respective results from JSON and exposes them to the front end. The calculated scores pre language for the contextual part of the DLE metric are stored in a separate file and exposed to the respective tab of the dashboard front end.
All results are visualised as graphs. For the front end implementation, the react-chartjs-2[34] library for charts and the chartjs-plugin-zoom[35] library for additional features like pan and zoom options on a chart have been selected.

6. Conclusions and future plans

In this paper we have presented the methods used to construct the empirical basis on which the technological readiness of languages spoken in Europe can be estimated. With the catalogue of the ELG Platform as point of departure, we have presented the automatic and collaborative language expert-based enrichment activities, so that the empirical basis is as representative as possible. We have also discussed the challenges emerging when such large-scale metadata aggregation activities are undertaken as well as the techniques used to mitigate them. While it is becoming clear that the language resources and technologies community is gradually converging to common metadata-based documentation practices, such that this work has been possible in the end, technical and semantic interoperability issues still remain and further standardisation will only make such aggregation activities more robust, efficient and cost-effective. The automatic enrichment procedures of the ELG catalogue put in place will continue at regular intervals, ensuring that the empirical basis for monitoring the level of digital readiness of languages is expanding in proportion to community activities and achievements. In parallel, the technical means made available through the ELG Platform will help keeping the empirical basis as up to date as possible through hopefully easy to use data and metadata registration functionalities.
We have also presented the ongoing work on the ELE dashboard, the availability of which helps monitor the evolution of technological support, identify gaps for each of the languages covered, and enable cross-language comparisons.

7. Acknowledgements

The work presented in this paper has been co-financed by the European Union under grant agreement LC-01641480 – 101018166 (European Language Equality). Part of the work has also been supported by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 825627 (European Language Grid).

8. Bibliographical References

Albertoni, R., D. Browning, S. Cox, A. Gonzalez-Beltran, A. Perego, and P. Winstanley (Eds.) (2020). Data Catalog Vocabulary (DCAT) - Version 2. W3C. https://www.w3.org/TR/vocab-dcat-2/.

Alemu, G., B. Stevens, and P. Ross (2012). Towards a conceptual framework for user-driven semantic metadata interoperability in digital libraries: a social constructivist approach. *New Library World*, vol. 113, no. 1/2.

Broeder, Daan, Trippel, Thorsten, Degl'Innocenti, Emiliano, Giacomi, Roberta, Sanesi, Maurizio, Kleemola, Mari, Moilanen, Katja, Ala-Lahti, Henri, Jordan, Caspar, Alfredsson, Iris, L'Hours, Hervé, & Ďurčo, Matej (2019). SSHOC D3.1 Report on SSHOC (meta)data interoperability problems. Zenodo. https://doi.org/10.5281/zenodo.3569868.

Chan, L. M. & Zeng, M. L. (2006). Metadata Interoperability and Standardization - A Study of Methodology Part I: Achieving Interoperability at the Schema Level. *D-Lib Magazine*, vol. 12, no. 6. https://doi.org/10.1045/june2006-chan.

DataCite Metadata Working Group (2021). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4. https://doi.org/10.14454/3w3z-sa82.

Gaspari, F., Way, A., Dunne, J., Rehm, G., Piperidis, S., and Giagkou, M. (2021). Deliverable D1.1 - Digital Language Equality (Preliminary Definition). European Language Equality. https://european-language-equality.eu/wp-content/uploads/2021/05/ELE_Deliverable_D1_1.pdf.

Gaspari, F., Grützner-Zahn, A., Rehm, G., Gallagher, O., Piperidis. S., and Way, A. (2022). Digital Language Equality (Full Specification). European Language Equality. https://european-language-equality.eu/wp-content/uploads/2022/04/ELE__Deliverable_D1_3.pdf

Gavrilidou, M., Labropoulou, P., Desipri, E., Giannopoulou, I., Hamon, O., and Arranz, V. (2012). The META-SHARE Metadata Schema: Principles,

[33] Direct access to the dashboard: https://live.european-language-grid.eu/catalogue/dashboard

[34] https://react-chartjs-2.js.org/

[35] https://www.chartjs.org/chartjs-plugin-zoom/

Features,Implementation and Conversion from Other Schemas. In *Proceedings of LREC 2012 - Workshop on Describing Language Resources with Metadata*, Istanbul, Turkey. European Language Resources Association.

Giagkou, M., Piperidis, S., Rehm, G. and Dunne, J. (Eds.) (2022). Language Technology Support of Europe's Languages in 2020/2021. European Language Equality Project.

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2021). Glottolog 4.5. Zenodo. https://doi.org/10.5281/zenodo.5772642

International Organization for Standardization (2007). Codes for the Representation of Names of Languages - Part 3: Alpha-3 Code for Comprehensive Coverage of Languages. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/03/95/39534.html (08.04.2022).

International Organization for Standardization (2017). ISO 15836-1:2017 Information and documentation - The Dublin Core metadata element set - Part 1: Core elements

Labropoulou, P., Gkirtzou, K., Gavriilidou, M., Deligiannis, M., Galanis, D., Piperidis, S. Rehm, G., Berger, M., Mapelli, V., Rigault, M., Arranz, V., Choukri, K., Backfried, G., Gómez Pérez, J. M. and Garcia Silva, A. (2020). Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020).* Marseille, France, 2020. European Language Resources Association (ELRA).

Labropoulou, P., Galanis, D., Lempesis, A., Greenwood, M., Knoth, P., Eckart de Castilho, R., Sachtouris, S., Georgantopoulos, B., Anastasiou, L., Martziou, S., Gkirtzou, K. Manola, N. and Piperidis, S. (2018). OpenMinTeD: A Platform Facilitating Text Mining of Scholarly Content. In *WOSP 2018 Workshop Proceedings, Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018, pp. 7–12 European Language Resources Association.

Lagoze, C., H. Van de Sompel, M. Nelson, and S. Warner (eds.) (2002). Open Archives Initiative - Protocol for Metadata Harvesting - v.2.0. Open Archives. http://www.openarchives.org/OAI/openarchivesprotocol.html.

Library of Congress (1999). MARC 21 Format for Bibliographic Data. https://www.loc.gov/marc/bibliographic/.

Lösch, A., Mapelli, V., Piperidis, S., Vasiļjevs, A., Smal, L., Declerck, T. et al. (2018). European Language Resource Coordination: Collecting Language Resources for Public Sector Multilingual Information Management. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Miyazaki, Japan, May 2018 European Language Resources Association.

McCrae, J.P., Philipp Cimiano, Victor Rodríguez Doncel, Daniel Vila-Suero, Jorge Gracia, Luca Matteis, Roberto Navigli, Andrejs Abele, Gabriela Vulcu, and Paul Buitelaar (2015). Reconciling Heterogeneous Descriptions of Language Resources. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, Beijing, China. https://doi.org/10.18653/v1/W15-4205.

Phillips, A. and Davis, M. (2009). Tags for Identifying Languages RFC 5646. Internet Engineering Task Force.

Piperidis, S., Labropoulou, P., Deligiannis, M. and Giagkou, M. (2018). Managing Public Sector Data for Multilingual Applications Development. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* Miyazaki, Japan. European Language Resources Association.

Rehm, G., and Uszkoreit, H. (Eds.) (2012). META-NET White Paper Series: Europe's Languages in the Digital Age. Springer.

Rehm, G., Berger, M., Elsholz, E., Hegele, S., Marheinecke, K., Piperidis, S., et al. (2020). European Language Grid: An Overview. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020).* Marseille, France, 2020. European Language Resources Association.

Rehm, G., Piperidis, S., Bontcheva, K., Hajic, J., Arranz, V., Vasiļjevs, A., et al. (2021). European Language Grid: A Joint Platform for the European Language Technology Community. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Online*, April 2021, pp. 221–230 Association for Computational Linguistics.

Soria, C., Núria Bel, Khalid Choukri, Joseph Mariani, Monica Monachini, Jan Odijk, Stelios Piperidis, Valeria Quochi, and Nicoletta Calzolari (2012). The FLaReNet Strategic Language Resource Agenda. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. ArXiv preprint arXiv:1910.03771.

Zeng, M.L. & Chan, L.M. (2006) Metadata Interoperability and Standardization - A Study of Methodology Part II: Achieving Interoperability at the Record and Repository Levels. D-Lib Magazine, vol. 12, no. 6. https://doi.org/10.1045/june2006-zeng.

9. Language Resource References

Labropoulou, P., Gkirtzou, K., Gavriilidou, M., Deligiannis, M., Galanis, D., Piperidis, S. et al. (2022). ELG-SHARE metadata schema, v3.0.1. https://gitlab.com/european-language-grid/platform/ELG-SHARE-schema.

Labropoulou, P., Aubin, S., Galanis, D., Giagkou, M., Gkirtzou, K., Knoth, P., Piperidis, S., Villegas, M., Eckart de Castilho, R. (2022). OMTD-SHARE ontology, v2.0.0 (pre-release). http://w3id.org/meta-share/omtd-share/.

Appendix: Additional figures

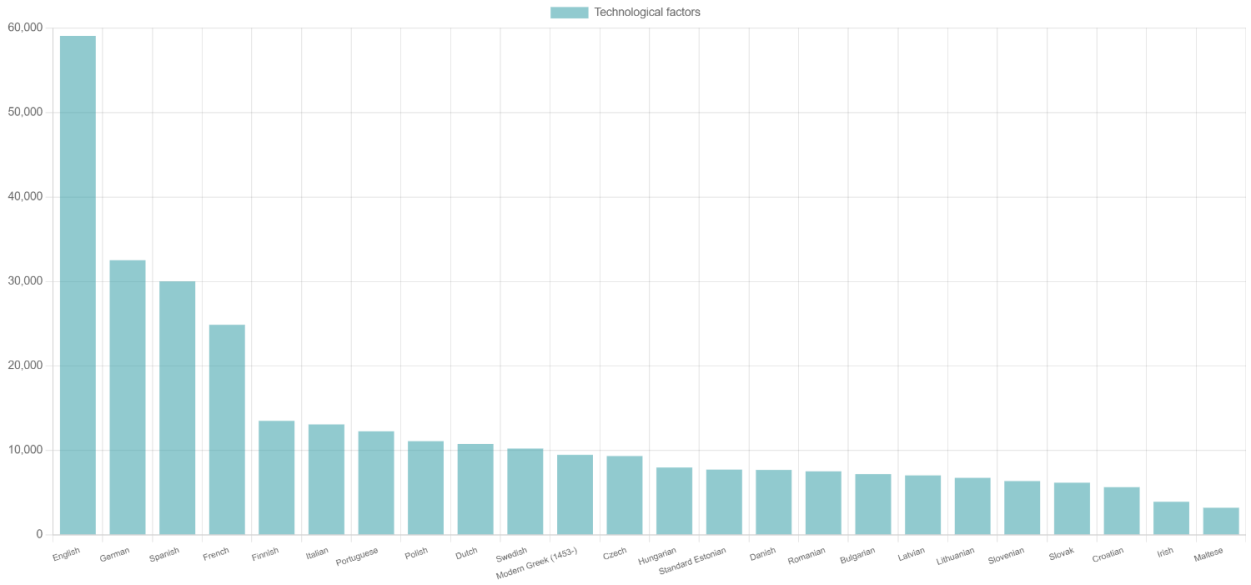

Figure 4. ELE dashboard screenshot: Technological DLE scores for the official EU languages (23 May 2022)